Conserved Gsx2/Ind homeodomain monomer versus homodimer DNA binding defines regulatory outcomes in flies and mice

Joseph Salomone,[1,2] Shenyue Qin,[3] Temesgen D. Fufa,[4] Brittany Cain,[5] Edward Farrow,[2] Bin Guan,[4] Robert B. Hufnagel,[4] Masato Nakafuku,[3,6] Hee-Woong Lim,[6,7,8] Kenneth Campbell,[3,6] and Brian Gebelein[3,6]

[1]Graduate Program in Molecular and Developmental Biology, Cincinnati Children's Hospital Research Foundation, Cincinnati, Ohio 45229, USA; [2]Medical-Scientist Training Program, University of Cincinnati College of Medicine, Cincinnati, Ohio 45229, USA; [3]Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, Ohio 45229, USA; [4]Ophthalmic Genetics and Visual Function Branch, National Eye Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; [5]Department of Biomedical Engineering, University of Cincinnati, Cincinnati, Ohio 45219, USA; [6]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio 45229, USA; [7]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio 45229, USA; [8]Department of Biomedical Informatics, University of Cincinnati College of Medicine, Cincinnati, Ohio 45229, USA

How homeodomain proteins gain sufficient specificity to control different cell fates has been a long-standing problem in developmental biology. The conserved Gsx homeodomain proteins regulate specific aspects of neural development in animals from flies to mammals, and yet they belong to a large transcription factor family that bind nearly identical DNA sequences in vitro. Here, we show that the mouse and fly Gsx factors unexpectedly gain DNA binding specificity by forming cooperative homodimers on precisely spaced and oriented DNA sites. High-resolution genomic binding assays revealed that Gsx2 binds both monomer and homodimer sites in the developing mouse ventral telencephalon. Importantly, reporter assays showed that Gsx2 mediates opposing outcomes in a DNA binding site-dependent manner: Monomer Gsx2 binding represses transcription, whereas homodimer binding stimulates gene expression. In *Drosophila*, the Gsx homolog, Ind, similarly represses or stimulates transcription in a site-dependent manner via an autoregulatory enhancer containing a combination of monomer and homodimer sites. Integrating these findings, we test a model showing how the homodimer to monomer site ratio and the Gsx protein levels defines gene up-regulation versus down-regulation. Altogether, these data serve as a new paradigm for how cooperative homeodomain transcription factor binding can increase target specificity and alter regulatory outcomes.

[*Keywords*: CUT&RUN; Gsx2; Ind; lateral ganglionic eminence (LGE); transcriptional activation versus repression]

Supplemental material is available for this article.

Received July 28, 2020; revised version accepted November 19, 2020.

Homeodomain (HD) proteins constitute a large transcription factor (TF) family that regulates many developmental processes from embryonic patterning to inducing specific cell fates in virtually every organ system of metazoans (Bürglin and Affolter 2016; Zandvakili and Gebelein 2016). Like all TFs, HD proteins induce specific developmental outcomes by binding to enhancer elements and regulating target gene expression. However, biochemical studies have revealed that the vast majority of HD TFs bind highly similar AT-rich DNA sequences (Berger et al. 2008; Noyes et al. 2008; Jolma et al. 2013). For example, the conserved *Gsx* genes, which regulate specific cell fates within the nervous systems of both invertebrates and vertebrates (Hsieh-Li et al. 1995; Valerius et al. 1995; Weiss et al. 1998; Illes et al. 2009), encode HD proteins that largely bind the same DNA sequences as the Hox TFs that specify a wide variety of cell fates along the developing anterior-posterior axis (Bürglin and Affolter 2016). These findings raise a fundamental paradox: How do HD

Corresponding authors: brian.gebelein@cchmc.org, kenneth.campbell@cchmc.org

Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.343053.120.

TFs that bind the same DNA sequences in vitro regulate distinct target genes and ultimately different cell fates in vivo?

Gsx factors, which consist of the vertebrate Gsx1 and Gsx2 TFs and the *Drosophila* Ind TF, perform two key functions during nervous system development in animals from flies to mammals. First, Gsx factors regulate dorsal-ventral (D-V) patterning of the nervous system. In *Drosophila*, Ind is required for D-V patterning of the neuroectoderm within the intermediate column of the ventral nerve cord by repressing dorsal column identity (Weiss et al. 1998). Similar to fly Ind, Gsx2 is essential for D-V patterning of neural progenitors within the mouse lateral ganglionic eminence (LGE) of the ventral telencephalon and the establishment of the pallio-subpallial boundary by repressing dorsal TFs, such as Pax6 (Corbin et al. 2000; Toresson et al. 2000; Yun et al. 2001). Second, Gsx factors are required for the generation of neurogenic progenitors and the specification of distinct neuronal cell fates. For example, fly Ind is necessary for the production of intermediate column neuroblasts and the specification of their neuronal progeny (Weiss et al. 1998). In the mouse LGE, Gsx2 is necessary for the generation of secondary proliferative (i.e., subventricular zone) progenitors that are restricted to the neuronal lineage and specified to generate a variety of neuronal subtypes, including striatal projection neurons and olfactory bulb interneurons (Toresson and Campbell 2001; Waclaw et al. 2009; Pei et al. 2011; Roychoudhury et al. 2020). Accordingly, *Gsx2*-null mouse mutants have significantly diminished basal ganglia (i.e., striatum) and olfactory bulb structures (Corbin et al. 2000; Toresson et al. 2000; Toresson and Campbell 2001; Yun et al. 2001, 2003). Importantly, a human genetic study found that a homozygous loss-of-function *GSX2* variant also results in severe basal ganglia and olfactory bulb agenesis, indicating a conserved role for this HD TF between mouse and human (De Mori et al. 2019).

Despite extensive studies showing that Gsx factors are essential for neural development, little is known about the molecular mechanisms by which these TFs function. Notably, Gsx factors are one of >40 members of the antennapedia class of HD TFs, all of which bind overlapping AT-rich binding sequences (such as TAATTA) (Berger et al. 2008; Noyes et al. 2008; Jolma et al. 2013; Bürglin and Affolter 2016). Hence, it remains largely unclear how Gsx factors specifically recognize and bind the appropriate target genes required for proper nervous system development. Moreover, the mechanisms underlying whether Gsx binding to *cis*-regulatory modules conveys gene activation versus repression have not been well elucidated. Experiments in flies and frogs have shown that Gsx TFs can repress transcription, in part, via an engrailed homology (eh1) domain that recruits Groucho/Tle factors (Von Ohlen et al. 2007a; Von Ohlen et al. 2009; Von Ohlen and Moses 2009; Winterbottom et al. 2010, 2011). However, Ind can also function as an activator in *Drosophila* by positively regulating its own expression via unknown mechanisms (Von Ohlen et al. 2007b; Von Ohlen and Moses 2009). Whether the mammalian Gsx factors similarly activate gene expression is unclear, as are the parameters and cofactors that dictate whether Gsx/Ind mediates transcriptional activation versus repression once bound to DNA.

In this study, we investigated how Gsx HD factors recognize target gene sequences and regulate transcriptional output in the developing mouse and fly nervous systems. Using high-resolution in vitro and in vivo DNA binding assays, we found that Gsx factors bind DNA as independent monomers or as cooperative homodimer complexes. Importantly, we show that Gsx2 binding to homodimer (D) sites stimulates gene expression, while binding to monomer (M) sites results in transcriptional repression. In *Drosophila* neuroblasts, we similarly found that Ind mediates both positive and negative gene expression via an autoregulatory enhancer containing M and D sites. Last, we demonstrate how the ratio of M-to-D sites within an enhancer and the levels of Gsx protein can dictate whether a target gene is up-regulated or down-regulated. Collectively, our data supports a model wherein Gsx factors gain DNA binding specificity by binding as monomers or by forming cooperative homodimers on precisely spaced and oriented binding sites. Moreover, through this increased DNA binding specificity, Gsx factors gain regulatory specificity by mediating opposing transcriptional outcomes based on the ratio of repressive M sites to stimulatory D sites.

Results

Negative autoregulation of Gsx2 *expression in the developing mouse telencephalon*

Previous studies revealed that *Drosophila* Ind positively regulates its own expression in neuroblasts (Von Ohlen et al. 2007b; Von Ohlen and Moses 2009). To determine whether mouse Gsx2 similarly regulates itself during telencephalic development, we monitored *Gsx2* gene expression in embryonic (E) 12.5 forebrain tissue sections using an enhanced green fluorescent protein (EGFP) knock-in allele that interrupts the *Gsx2* locus (i.e., $Gsx2^{EGFP}$) with an *IRES-EGFP-pA* and thereby generates a null allele (Wang et al. 2009). In contrast to *Ind*, which is required for its own activation, EGFP levels in the E12.5 LGE were noticeably higher in *Gsx2*-null embryos that have both a *EGFP* knock-in null allele and a recombined $Gsx2^{RA}$-null allele (i.e., $Gsx2^{EGFP/RA}$) as compared with embryos with a wild-type *Gsx2* allele (i.e., $Gsx2^{EGFP/+}$) (Fig. 1A,B). To better quantify this effect, we conducted RNA sequencing (RNA-seq) on LGEs dissected from E12.5 embryos with either wild-type $Gsx2^{+/+}$ or the two *Gsx2*-null alleles ($Gsx2^{EGFP/RA}$). Importantly, while both the $Gsx2^{EGFP}$ and the $Gsx2^{RA}$ alleles abolish Gsx2 protein expression, each maintains the *Gsx2* transcription start site and 5′ end of exon 1, and thereby allows for comparative quantitative analysis of the 5′ end of the *Gsx2* transcript between wild-type and *Gsx2* mutant LGEs (Fig. 1C). Analysis of the RNA-seq data revealed a significant increase in the 5′ end of the *Gsx2* transcript in the absence of Gsx2 protein (1.24 log fold change; $P\text{-value} = 1.68 \times 10^{-40}$) (Fig. 1C). Thus, these observations show that, opposite to

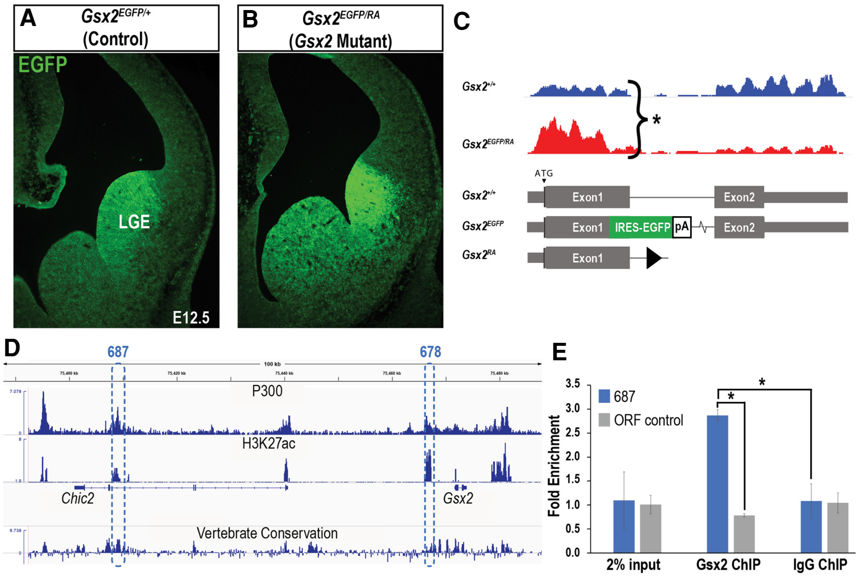

Figure 1. Gsx2 binds and negatively regulates its own expression in the mouse telencephalon. (*A*,*B*) Increased EGFP expression from the *Gsx2* locus in the LGE of E12.5 *Gsx2*$^{EGFP/RA}$ (i.e., null) embryos compared with a *Gsx2*$^{EGFP/+}$ sibling. (*C*) RNA-seq analysis from wild-type and *Gsx2*$^{EGFP/RA}$ LGE shows significant up-regulation of the *Gsx2* first exon (bracket). The experiment was performed using biological quadruplicates. (*) Log$_2$ fold change = 1.24; FDR = 7.7 × 10^{-33} by EdgeR exact test. The black triangle represents a *loxP* site. (*D*) ChIP-seq for P300 and H3K27ac reveals potential regulatory elements around the *Gsx2* locus. The locations of the *687* and *678* enhancers are highlighted and vertebrate conservation is noted at the *bottom*. (*E*) ChIP-PCR data showing Gsx2 binds to *687* in E12.5 LGEs relative to both input chromatin and control IgG samples. Blue bars denote fold enrichment using *687*-specific primers, whereas gray bars denote fold enrichment for the control *Actb* open reading frame (ORF). (*) *P*-value < 0.05 using an unpaired two-tailed Student's *t*-test.

Drosophila Ind, mouse Gsx2 negatively regulates its own expression.

To better understand the mechanism of Gsx2 negative autoregulation, we first analyzed the *Gsx2* locus for potential regulatory elements. Two *Gsx2* enhancers, namely *687* and *678*, were previously identified within 62 kb of the *Gsx2* locus on the VISTA browser (https://enhancer.lbl.gov), and each shows a high degree of species conservation and enrichment for active chromatin marks (p300 and H3K27ac) (Fig. 1D; Zhou et al. 2017). Of these two enhancers, only *687* was sufficient to drive reporter expression in a pattern similar to endogenous *Gsx2* in the mouse LGE (Visel et al. 2013; Qin et al. 2016). A motif scan of the highly conserved 1.2 kb *687* enhancer revealed numerous potential Gsx2-binding sites (see Fig. 2A; Supplemental Fig. S1A). To assess for direct in vivo Gsx2 binding to *687*, we performed chromatin immunoprecipitation-quantitative PCR (ChIP-qPCR) on dissected E12.5 LGE tissue using either a rabbit anti-Gsx2 antibody or a control IgG. qPCR showed a clear enrichment of Gsx2 binding at the *687* locus compared with a control locus (Fig. 1E). These data show that Gsx2 directly binds to the *687* enhancer in vivo, suggesting that this element contributes to Gsx2-mediated negative autoregulation within the mouse LGE.

Gsx2 differentially regulates transcription via independent monomer and cooperative dimer-binding sites

To identify specific Gsx2-binding sites within the *687* enhancer, we used purified mouse Gsx2 protein and DNA probes containing predicted Gsx2 sites in electrophoretic mobility shift assays (EMSAs) (Fig. 2A; Supplemental Fig. S1A). Interestingly, we observed two modes of binding depending on the probe sequence. Consistent with prior studies showing that Gsx factors bind individual AT-rich DNA sequences, the majority of probes were bound in a manner consistent with monomeric Gsx2 binding (Fig. 2B). Moreover, even if a probe contained more than one site, these probes were bound by Gsx2 in an additive manner (Supplemental Fig. S1B–F). In contrast, one probe that contained two predicted sites was unexpectedly preferentially bound by two Gsx2 proteins in a manner consistent with cooperative Gsx2 homodimer formation (Fig. 2C; Supplemental Fig. S1G). Thus, the in vitro DNA-binding assays revealed that Gsx2 interacts with DNA as either a monomer (on M sites) or as a homodimer (on D sites). Considering these distinct binding modalities, the *687* enhancer contains multiple monomer sites (M1–M9) and one dimer site (D1) (Fig. 2A).

Previous studies in *Xenopus* and *Drosophila* revealed that Gsx factors function as transcriptional repressors (Von Ohlen et al. 2007a, 2009; Von Ohlen and Moses 2009; Winterbottom et al. 2010, 2011). Thus, we developed a reporter assay to investigate the regulatory potential of mouse Gsx2 on M and D sites by creating luciferase vectors with a minimal promoter, 5xUAS sites, and either six copies of the M1 site or three copies of the D1 site from the *687* enhancer (Fig. 2D–E). In this assay, transfection with Gal4-VP16 activated luciferase expression in a mouse kidney cell line (mK4 cells) (green bar in Fig. 2D), whereas cotransfection with Gsx2 resulted in dose-dependent repression of Gal4-mediated activation on the 6xM1 reporter (orange bars in Fig. 2D). Surprisingly, the D1 site reporter behaved differently in this assay. Instead of repression, we observed enhanced Gal4-VP16 mediated luciferase activity in response to Gsx2 (Fig. 2E). However, if the D1 site was mutated so that it binds Gsx2 in a noncooperative, monomer-like manner

(Fig. 2F, the D1-to-M mutation), Gsx2 mediated dose-dependent repression (Fig. 2G). These data suggest that Gsx2 differentially mediates transcription when bound as a monomer versus a homodimer. Interestingly, Gsx2 did not activate gene expression via D sites in the absence of Gal4-VP16 (Fig. 2E), suggesting that it is insufficient to activate transcription by itself, at least in mK4 cells. Moreover, the greatest stimulation of Gal4-VP16 mediated activation of the D1 reporter occurs at low Gsx2 levels, whereas increasing Gsx2 concentrations reduced luciferase activity (Fig. 2E). Analysis of the minimal promoter (minP) sequence, which is present in each luciferase vector, revealed a predicted M site (vertical red line in Fig. 2D,E,G) that may contribute to decreased reporter activity at higher Gsx2 levels (note, the impact of combining D and M sites on transcriptional output is tested later in this study). Collectively, these data support a novel model of HD-mediated gene regulation, whereby mouse Gsx2 represses transcription when bound to M sites and stimulates transcription via D sites.

Cooperative Gsx2 dimer binding to DNA requires amino acids flanking the homeodomain and precisely spaced and oriented binding sites

As Gsx factors had not previously been shown to cooperatively bind DNA, we next defined the critical domains required for cooperative Gsx2 DNA binding using deletion constructs and EMSAs to calculate their Hill binding coefficient as a measure of cooperativity (Hill coefficient of 2 = perfect cooperativity; Hill of 1 = no cooperativity) (Weiss 1997). Strikingly, we found that the mouse Gsx2 HD alone does not preferentially form dimers on a DNA probe encoding a D site, and instead binds this probe in a noncooperative monomeric manner (Fig. 3A, right). In contrast, this same DNA probe preferentially binds two Gsx2 proteins that contain short 40-amino-acid regions N- and C-terminal of the HD, consistent with the formation of cooperative Gsx2 homodimers on DNA (Fig. 3A,

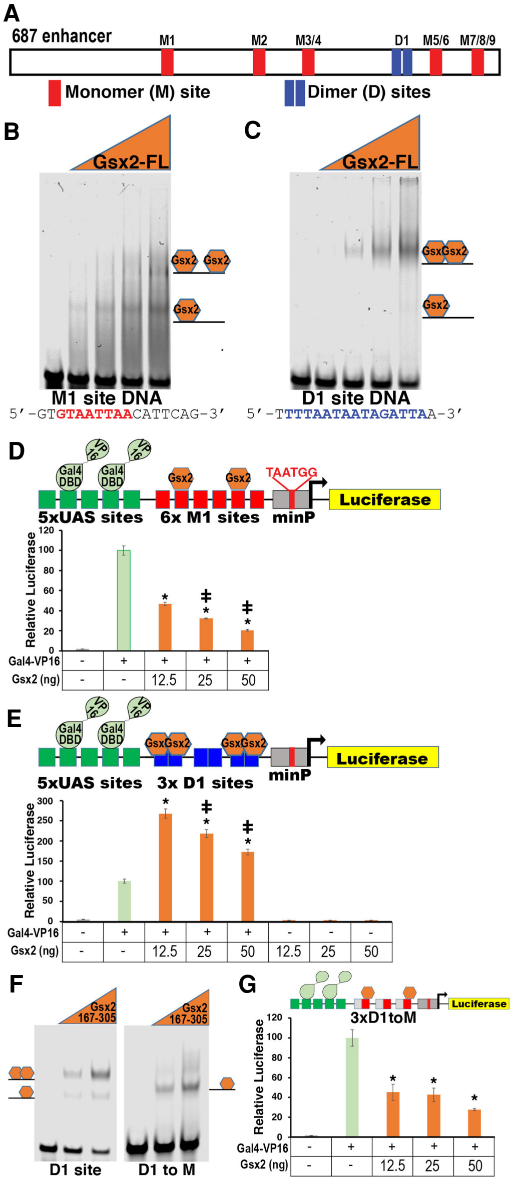

Figure 2. Gsx2 differentially regulates gene expression via two types of binding sites. (*A*) Schematic of the *687* enhancer with the M sites (red, M1–M9) and D site (blue, D1) noted. Sequence of *687* is reported in Supplemental Document 1. (*B*,*C*) Comparative EMSAs using an M1 and D1 probe with equal amounts of full-length Gsx2 reveals monomer and dimer binding, which are highlighted using schematics at right of each gel. The sequences of each site (M1 and D1) are noted below each EMSA. (*D*) Schematic of Luciferase reporter containing five UAS sites, six M1 sites, and a minimal promoter that encodes a predicted M site (red bar). Luciferase assay from mK4 cells transfected with 25 ng of *UAS-6xM1-Luciferase*, 5 ng of Gal4VP16 alone (green bar), and the indicated amounts of Gsx2 (orange bars) revealed that Gsx2 represses Gal4VP16-mediated activation. An ANOVA with Tukey post-hoc was used to determine significance. (*) $P < 0.01$ compared with Gal4VP16 alone, (‡) $P < 0.01$ compared with 12.5 ng of Gsx2. (*E*) Schematic of Luciferase reporter containing five UAS sites, three D1 sites, and a minimal promoter that encodes a predicted M site (red bar). Luciferase assays from mK4 cells transfected with 25 ng of *UAS-3xD1-Luciferase*, 5 ng of Gal4VP16 alone (green bar), and the indicated amounts of Gsx2 (orange bars) revealed enhanced Gal4VP16 activation in presence of Gsx2. Note, Gsx2 does not induce gene expression in the absence of Gal4-VP16, suggesting it is insufficient to activate transcription. An ANOVA with Tukey post-hoc was used to determine significance. (*) $P < 0.01$ compared with Gal4VP16 alone, (‡) $P < 0.01$ compared with 12.5 ng of Gsx2. (*F*) EMSAs using purified Gsx2 (167–305) protein reveals cooperative dimer binding to the D1 site but noncooperative monomer binding to the D1-to-M probe. Each EMSA binding reaction had a final concentration of 34 nM labeled DNA probe with either no protein added (first lane) or with 140 or 280 nM purified Gsx2 (167–305) protein. (*G*) *UAS-3xD1toM-luciferase* activity is repressed, and not up-regulated by Gsx2. Twenty-five nanograms of luciferase reporter, 5 ng of Gal4-VP16 if indicated by a plus sign, and the noted amount of Gsx2 were transfected. An ANOVA with Tukey post-hoc was used to determine significance. (*) $P < 0.01$ compared with Gal4VP16 alone.

left). We used this data to generate a Hill plot and found that the larger Gsx2 fragment has a Hill coefficient of ∼1.8, indicating significant cooperative binding, whereas the HD-only fragment has a Hill coefficient much closer to 1 (1.17) (Fig. 3B). These findings reveal that while the Gsx2 HD is sufficient to independently bind the sites that comprise the D site, it is not sufficient to mediate cooperative binding to the D site.

We next sought to define the DNA sequence constraints that facilitate cooperative binding to D sites. Using the 687 D1 site for this analysis, we first tested a series of mutations across the D1 site (mut1 – mut10) in EMSAs and categorized each as either cooperative (C) or noncooperative (N) (Fig. 3C; Supplemental Fig. S2B–M). When sequences were aligned, the importance of 5 key nucleotides, a TA and an ATT, separated by 7 bp was evident (Fig. 3C). To test whether this spacing is required, we designed probes that increased spacing in single-base-pair increments and discovered that adding just 1 bp disrupts cooperative binding (Fig. 3D; Supplemental Fig. S2N–R). Interestingly, adding 3 bp between sites restored cooperative binding; however, this insertion shifted a different TA into a 7-bp spacing with the ATT sequence from the original site and thereby created a new cooperative D site (Fig. 3D; Supplemental Fig. S2Q). Determining the orientation of the two sites that make up a D site was complicated by the fact that the previously defined optimal Gsx2 motif is palindromic (TAATTA) (Jolma et al. 2013). To circumvent this issue, we analyzed existing protein binding microarray (PBM) data for Gsx2 and selected the highest scoring nonpalindromic 8-mer sequence (Berger et al. 2008). We then designed and tested a series of probes containing this 8-mer in different orientations and found that only one spacing and orientation combination, the forward-forward probe with 7-bp spacing, was bound cooperatively by Gsx2 (Fig. 3E; Supplemental Fig. S2S–BB). In total, by coupling EMSA data with Gsx2's known monomer binding preferences, we predict that an optimal Gsx2 D site contains two ATTA sequences separated by 7-bp (Fig. 3F).

To determine the dimer binding preferences of Gsx2 to DNA in an unbiased manner, we analyzed published HT-SELEX data that was generated using random 20-mers (Jolma et al. 2013). Importantly, while the original study only identified an enriched M motif, the SELEX assay was performed using a human GSX2 protein containing

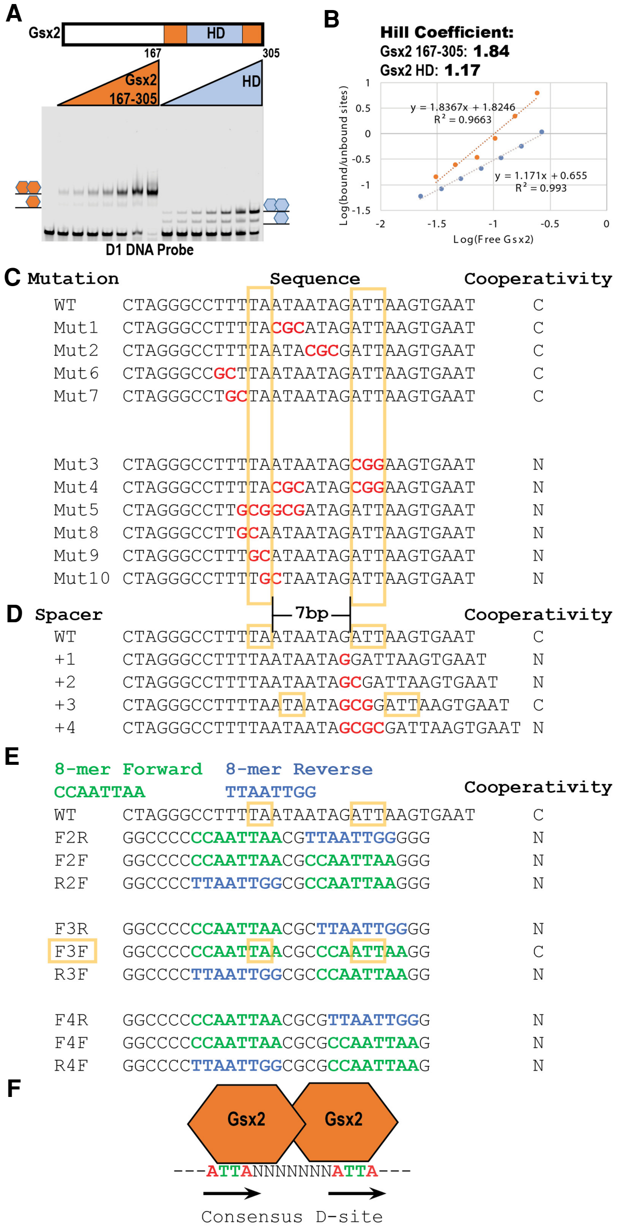

Figure 3. Gsx2 requires domains flanking the homeodomain and precisely spaced and oriented DNA sites to cooperatively bind D sites. (*A*) EMSAs using the 687-D1 probe and purified Gsx2 proteins containing the flanking regions plus homeodomain (167–305, *left*) or only the HD (light blue, *right*). Each EMSA binding reaction had 34 nM labeled DNA probe with the following concentrations of either the Gsx2 (167–305) or Gsx2 HD protein: 0 nM (i.e., probe alone), and 2.34, 4.69, 9.38, 18.75, 37.5, 75, and 150 nM. Note, the purity of the Gsx2 (167–305) and Gsx2-HD proteins are shown by SDS-PAGE analysis and gel staining in Supplemental Figure S2A. Schematics highlight the fact that the Gsx2 + flanking region protein more readily forms dimer complexes than the HD only protein. (*B*) Hill coefficient calculations from EMSAs reveal that the Gsx2+flanking protein binds D sites cooperatively (Hill coefficient = 1.84), whereas the HD alone does not (Hill coefficient = 1.17). (*C*) Summary of scanning mutagenesis and Gsx2 EMSA data highlighting the nucleotides in the 687 D1 probe required for cooperative (C) versus noncooperative (N) binding. EMSAs are shown in Supplemental Figure S2B–M. (*D*) Summary of Gsx2 EMSA data using probes with insertion of different numbers of nucleotides reveals that a 7-bp spacer is required for cooperative binding to the D1 probe. Note, the +3-bp insertion generated a new "D site" with the required 7-bp spacing. EMSAs are shown in Supplemental Figure S2N–R. (*E*) Summary of Gsx2 EMSA data using probes engineered with a nonpalindromic D site in different orientations and spacing. Note, only the F3F probe contains the required orientation and spacing to mediate cooperative binding. EMSAs are shown in Supplemental Figure S2S–BB. (*F*) Optimal Gsx2 D site with a 7-bp spacer as defined by EMSAs.

similar flanking N- and C-terminal domains that we found were required for cooperative binding by the mouse Gsx2 protein (see Fig. 3A,B). Hence, we reanalyzed the HT-SELEX data for occurrences of the optimal D site motif (Fig. 3F) after each of 4 selection cycles and observed clear enrichment of D site sequences (Fig. 4A). To control for simply enriching sequences with two M sites, we calculated the percentage of sequences with two sites in relation to spacer length from 3 to 11 bp and found a striking preference for sites separated by 7 bp (Fig. 4B), which is identical to our empirical studies (Fig. 3D; Supplemental Fig. S2). Moreover, we performed a de novo motif search for longer motifs using MEME and found that the top enriched motif contained two HD sites with a 7-bp spacer (Fig. 4C). Thus, human GSX2 enriches for D sites with the same spacing and orientation that were cooperatively bound by the mouse Gsx2 protein.

To test whether the binding site spacing and orientation rules could predict sites that would be both cooperatively bound and stimulated by Gsx2, we analyzed another recently characterized *Gsx2* enhancer, which we have called *DSG* (*downstream from Gsx2*) (Fig. 4D; Desmaris et al. 2018; Konno et al. 2019). We scanned the ∼1200-bp region for putative D sites and identified one that perfectly matches the optimal D motif as well as eight predicted high-affinity M sites (Fig. 4D). We subsequently used comparative EMSAs to show that mouse Gsx2 cooperatively bound the *DSG* D site probe (Fig. 4E), and luciferase assays further revealed that Gsx2 strongly enhanced Gal4-VP16-mediated gene activation via the *DSG*-D site (Fig. 4F). Thus, these data show that two different cooperative D sites (the *687* D1 and *DSG*-D sites) both stimulate Gal4-VP16 mediated gene expression in the luciferase assay, whereas two different M sites (the *687* M1 and the D1-to-M sites) both mediate gene repression (Fig. 2D,G).

Genome-wide analysis in the mouse LGE reveals Gsx2 binding to monomer and dimer sites

To identify the in vivo genomic targets for Gsx2 in the developing LGE, we performed ChIP-seq from chromatin isolated from mouse E12.5 LGEs using a Gsx2 antibody. While this antibody worked for ChIP-qPCR (see Fig. 1E), ChIP library sequencing revealed a low signal to noise ratio compared with IgG controls. Thus, we used genome editing to insert an N-terminal 2x-FLAG tag in-frame with the *Gsx2* endogenous locus (Fig. 5A). Importantly, homozygous *Gsx2*^{2xFLAG} mice are viable, breed normally, and are grossly normal compared with littermate controls. Furthermore, FLAG expression overlaps Gsx2 staining in the expected pattern (Fig. 5B), and embryonic ventral forebrain morphology is normal with no noticeable ventral expansion of dorsal telencephalic markers such as Pax6

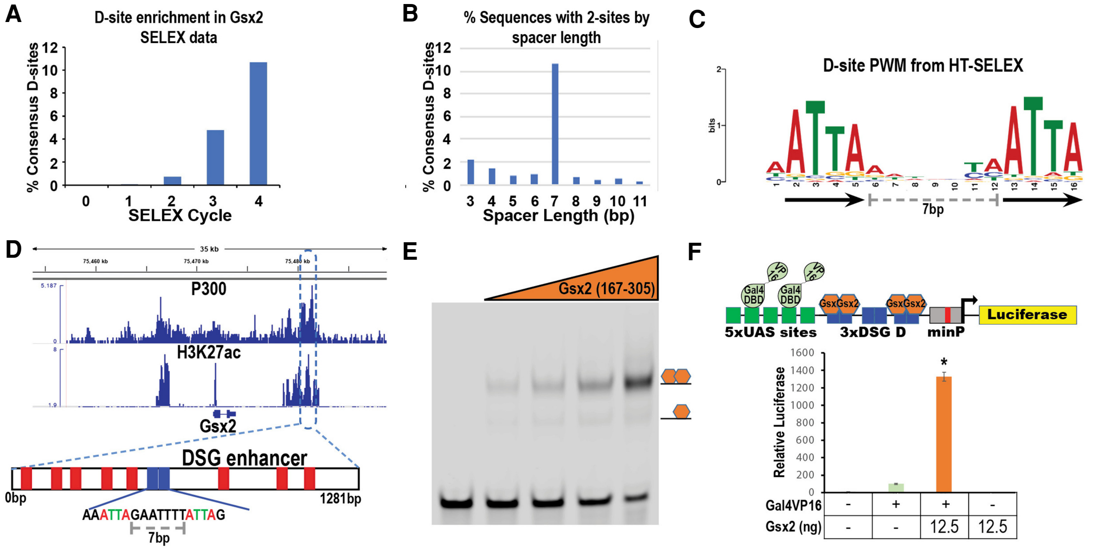

Figure 4. Analysis of HT-SELEX data reveals that the human GSX2 protein enriches for dimer DNA binding sites. (*A*) The percentage of sequences that contain an optimal D site motif as a function of SELEX cycle. Note, the "0" cycle is the starting library and enrichment for D sites is observed in successive SELEX cycles using the human GSX2 protein. (*B*) The percentage of sequences with two sites by spacer length. Note, the highest frequency occurs with the 7-bp spacer. (*C*) The D site PWM motif from MEME de novo motif search on the human GSX2 HT-SELEX data after the fourth round of selection. (*D*) ChIP-seq for P300 and H3K27ac in E12.5 forebrain tissue revealed strong signals at the characterized *Gsx2 DSG* enhancer. The location of *DSG* is boxed and the sequence of an optimal D site (blue) as well as predicted M sites (red) within the *DSG* are noted. Sequence of the *DSG* is reported in Supplemental Document 1. (*E*) EMSA using Gsx2 (167-305) reveals cooperative dimer binding to the *DSG* D site. Each EMSA binding reaction had 34 nM of labeled DNA probe and the following concentrations of the Gsx2 (167–305) protein: 0 nM (i.e., probe alone), or 46.5, 93, 186, or 372 nM. Note, a larger image of this exact same gel is shown in Supplemental Figure S2CC. (*F*) *UAS-3xDSGD-Luciferase* activity revealed enhanced Gal4VP16 activation in the presence of Gsx2. The amounts of transfected plasmid are noted. (*) $P < 0.05$ using an unpaired two-tailed Student's *t*-test compared with Gal4VP16 alone.

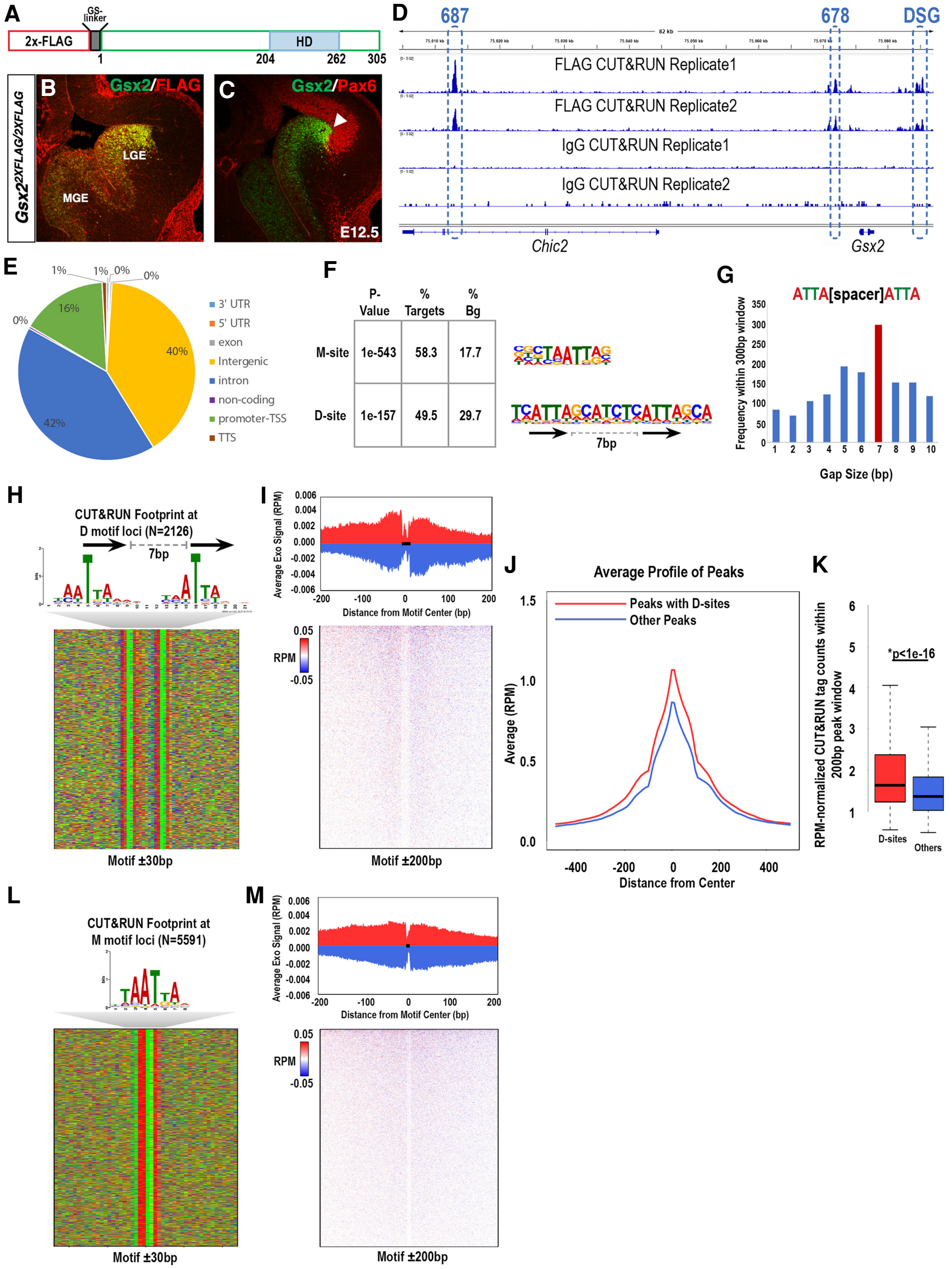

Figure 5. Genomic analysis of Gsx2 binding in the mouse LGE reveals enrichment of monomer and dimer sites. (*A*) Schematic of the 2XFLAG-Gsx2 with the homeodomain (HD) highlighted in light blue. (*B*) Immunostaining of an E12.5 *Gsx2*$^{2XFLAG/2XFLAG}$ mouse telencephalon reveals extensive colocalization of Gsx2 (green) and FLAG (red) in the expected LGE expression pattern. (*C*) Gsx2/Pax6 double staining shows that Dorsal-Ventral patterning in the E12.5 *Gsx2*$^{2XFLAG/2XFLAG}$ telencephalon appears normal. (*D*) Replicate CUT&RUN analysis of FLAG-Gsx2 genomic binding to the *Gsx2* locus in comparison with IgG controls. Note, significant FLAG-Gsx2 binding to both the *687* and *DSG* enhancers. (*E*) Genomic annotation of Gsx2 peaks using HOMER. Note, most peaks are found in intergenic and intronic regions. (*F*) Top motifs identified by HOMER reveal significant M and D site enrichment. (*G*) The number of occurrences of two ATTA sites by spacer length in Gsx2 CUT&RUN peaks. Note, like with the SELEX assay (Fig. 4B), the strongest peak occurs with the 7-bp spacer. (*H*) Alignment of the top 2126 sites containing Gsx2 D sites identified by MEME and secondary filtering for footprint contrast. Sequences are color-coded and outside of the ATTA motifs, there is limited sequence similarity between sites. (*I*) MNase digestion footprint of the genomic Gsx2 D sites. All sequences were aligned centered on the D motif with the most highly protected sites overlapping the D motif. (*J*) CUT&RUN signal at Gsx2 peaks either containing at least one D site (1591 peaks, Red) or lacking a D site (1441 peaks, blue). (*K*) Normalized tag counts (RPM) within a 300-bp window around Gsx2 peaks containing at least one D site versus all other peaks. (*) Wilcoxon test. (*L*) Alignment of the top 5591 sites containing Gsx2 M sites that do not overlap with a D site. (*M*) MNase digestion footprint of the genomic Gsx2 M sites. All sequences were aligned centered on the M motif.

(Fig. 5C), as would be expected in *Gsx2*-null mutant embryos (Corbin et al. 2000; Toresson et al. 2000; Yun et al. 2001).

We next performed CUT&RUN assays (Skene and Henikoff 2017; Skene et al. 2018) on LGE tissue from homozygous E12.5 *Gsx2*^{2xFLAG} embryos using the anti-FLAG-M2 antibody and an IgG control. Analysis of FLAG CUT&RUN biological replicates revealed a highly reproducible peak pattern ($N = 3032$) against matching IgG CUT&RUN control samples (Supplemental Fig. S3A). Notably, specific peaks in the *687* and *DSG* enhancers were observed, confirming that Gsx2 directly binds these enhancer regions (Fig. 5D). Gsx2 peaks also overlapped with the *678* enhancer (Fig. 5D), which is insufficient to activate LGE gene expression on its own, but results in robust LGE expression when coupled with the *687* enhancer in a transgenic reporter assay (Qin et al. 2016). Thus, Gsx2 binds multiple regulatory elements near the *Gsx2* locus, consistent with its direct role in autoregulation (see Fig. 1A–C).

Genomic annotation of FLAG CUT&RUN peaks (Gsx2-binding sites) revealed that the majority fall within intergenic or intronic regions with only a small subset associated with promoters (Fig. 5E). Consistent with many Gsx2 binding events being associated with regulatory elements, we found a high association with E12.5 forebrain marks of active enhancers such as p300 binding (Zhou et al. 2017), H3K27ac chromatin marks (ENCODE project ENCSR966AIB), and transposase accessible regions (i.e., open chromatin; ENCODE project ENCSR559FAJ) (Supplemental Fig. S4). Moreover, analysis with either HOMER's or MEME's de novo motif enrichment tools revealed significant enrichment of both M and D motifs (Fig. 5F; Supplemental Fig. S3B). Further analysis of called Gsx2 peaks for ATTA sequences with different spacer lengths revealed that similar to the HT-SELEX data (Fig. 4E), the Gsx2 CUT&RUN data showed a strong preference for dimer sites with a 7-bp spacer (Fig. 5G).

Based on the idea that bound TF sites are protected from MNase digestion during CUT&RUN analysis (Skene and Henikoff 2017), we performed footprint analysis to identify Gsx2 binding sites in high resolution. First, we scanned 300-bp windows of the 3032 called peaks for D and M sites using FIMO to obtain candidate binding sites (Grant et al. 2011). We then measured MNase digestion signals across each loci and defined loci displaying lower signal than flanking windows as true Gsx2-binding sites (see Materials and Methods for details). Through this analysis, we found that approximately half of the called Gsx2 peaks contained one or more D sites (1591 peaks) for a total of 2126 D sites (Fig. 5H,I). Moreover, we found that the CUT&RUN signal at the 1591 peaks with at least one D motif was significantly higher compared with the 1441 peaks lacking a D motif (Fig. 5J,K), consistent with D sites promoting robust Gsx2 DNA binding. Importantly, high-resolution Gsx2 binding analysis also identified 5591 Gsx2 binding events to M sites that did not overlap with the predicted D motifs, and as expected these binding events revealed a smaller protective footprint over a single TAATTA M motif (Fig. 5L,M). Altogether, these data demonstrate that mouse Gsx2 binds to both D and M sites in LGE progenitors.

To determine how Gsx2 binding correlates with changes in gene expression, we performed differential RNA-seq analysis on wild-type and *Gsx2*-null (*Gsx2*$^{EGFP/RA}$) E12.5 mouse LGEs and found 719 up-regulated and 289 down-regulated genes in *Gsx2* mutants (fold change >1.5; FDR <0.05) (Fig. 6A). As expected, gene ontology (GO) analysis revealed a significant change in the expression of genes associated with many aspects of neural development (Fig. 6B,C). For example, in addition to *Gsx2* itself (which is negatively autoregulated) (see Fig. 1A–C), a group of TFs known to regulate forebrain development were differentially expressed between WT and *Gsx2* mutant animals, including *Ascl1*, *Dlx1*, *Dlx2*, *Dlx5*, *Sp8*, *Dbx1*, *Gsx1*, and *Pax6* (Fig. 6A; Corbin et al. 2000; Toresson et al. 2000; Toresson and Campbell 2001; Yun et al. 2001, 2003; Waclaw et al. 2006, 2009). Consistent with potential direct regulation of these target genes by Gsx2, analysis of Gsx2 binding events revealed peaks near each of *Ascl1*, *Dlx1/2*, *Gsx1* and *Pax6* (Supplemental Fig. S5).

We next intersected the CUT&RUN and RNA-seq data by comparing the spatial proximity of Gsx2 peaks with distinct groups of genes that were either significantly up-regulated, down-regulated, or unchanged in *Gsx2*$^{EGFP/RA}$ LGEs. Moreover, we divided the Gsx2-binding events into four groups: Gsx2 peaks containing at least one M and D site (Fig. 6D), Gsx2 peaks containing neither an M nor a D site (Fig. 6E), Gsx2 peaks containing at least one D site (Supplemental Fig. S6A), and Gsx2 peaks containing at least one M site (note only the first two groups are mutually exclusive) (Supplemental Fig. S6B). Importantly, while Gsx2 peaks containing D and/or M sites were all strongly associated with significant gene expression changes in *Gsx2* mutant animals (Fig. 6D; Supplemental Fig. S6), those binding events that lack an M or a D site were not positively correlated with significant up-regulation or down-regulation of nearby genes in *Gsx2* mutants (Fig. 6E). These data suggest that Gsx2 binding to M and/or D sites significantly contributes to the regulation of LGE gene expression. However, there was no obvious correlation between the direction of gene expression changes (up vs. down) and Gsx2 binding to D site versus M site containing Gsx2 peaks. These findings suggest that the simple presence of an M or D site within a Gsx2 peak is not sufficient to predict the direction of in vivo transcriptional output within the LGE.

Our inability to predict gene expression changes in *Gsx2* mutant LGEs based solely on M versus D site binding is not unexpected given that *Gsx2* mutants misregulate numerous other TFs (e.g., *Ascl1*, the *Dlx* factors, etc) that also play significant roles in LGE gene expression (Corbin et al. 2000; Toresson et al. 2000; Yun et al. 2001). Among these genes, *Dlx1*, *Dlx2*, and *Dlx5* encode homeodomain TFs that bind highly similar, if not identical, DNA sequences as the Gsx2 monomer site in published HT-SELEX assays (Fig. 6F). To determine whether the Dlx factors bind the same genomic regions as Gsx2, we analyzed recently published ChIP-seq data for Dlx1, Dlx2, and Dlx5 from the E11.5 and E13.5 mouse forebrain

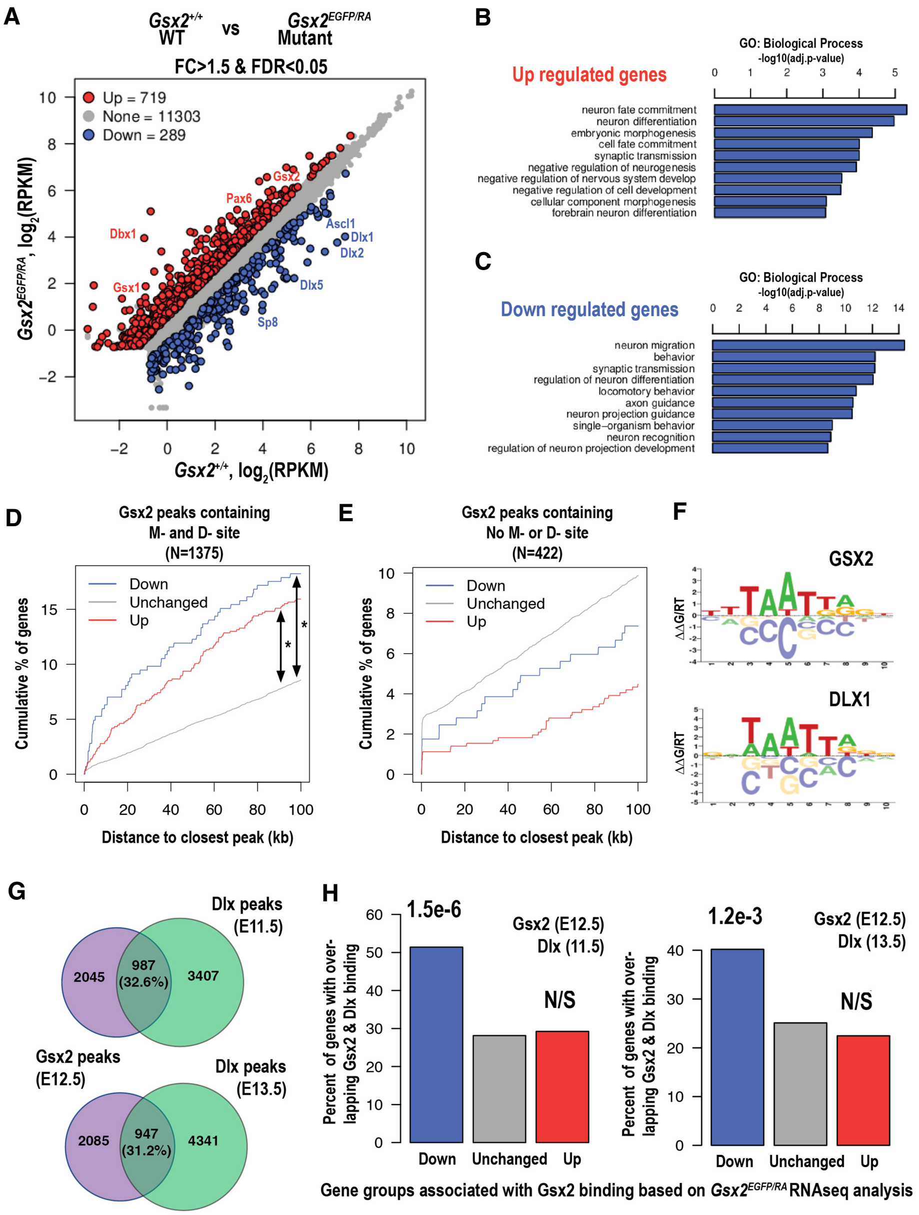

Figure 6. Peaks containing D and M sites are associated with gene expression changes in the LGE of *Gsx2*-null embryos. (*A*) RNA-seq analysis from LGE tissue of E12.5 WT (*Gsx2*$^{+/+}$) and *Gsx2*-null (*Gsx2*$^{EGFP/RA}$) embryos reveals genes that are significantly up-regulated (red) and down-regulated (blue) in the absence of Gsx2. Significantly altered expression of transcription factors important for forebrain patterning and neurogenesis are labeled. Differentially expressed genes were defined by fold change >1.5 and FDR < 0.05. (*B*,*C*) GO analysis on genes upregulated (*B*) and downregulated (*C*) in *Gsx2*-null embryos reveals Biological Process GO terms related to neural development. (*D*,*E*) Genes were divided into down-regulated (blue), up-regulated (red), or unchanged (gray) groups. The cumulative percentage of genes in each group that had a Gsx2 CUT&RUN peak with an M and D site (*D*) or with no M or D site (*E*) within a certain distance up to 100 kb from their TSS is plotted. Note, only those peaks with an M and D site are significantly associated with up-regulated and down-regulated genes in *Gsx2*-null LGEs, whereas those lacking M and D sites are not associated with gene expression changes of nearby genes. (*) *P*-value < 0.05. Similar analysis for Gsx2 CUT&RUN peaks selected only on the basis of having a D site (*F*) or a M site (*G*) is shown in Supplemental Figure S6. (*F*) Comparative PWM logos generated from previously published HT-SELEX assays (Jolma et al. 2013) reveals nearly identical GSX2 monomer (*top*) and DLX1 (*bottom*) DNA-binding sites, consistent with these TFs binding largely the same DNA sites. (*G*) Comparative genomic binding analysis of the Gsx2 CUT&RUN data from E12.5 mouse LGEs and the previously published Dlx (Lindtner et al. 2019) ChIP-seq data from E11.5 (*top*) and E13.5 (*bottom*) mouse forebrains reveals significant overlap in genomic binding between Gsx2 and the Dlx factors. (*H*) Analysis of Gsx2 and Dlx binding to the same genomic regions associated within a 100 kb window around gene TSSs that are either down-regulated (blue bars), unchanged (gray bars), or up-regulated (red bars) in *Gsx2* mutant LGEs. The percentage of Gsx2 genomic binding events that were also bound by at least two Dlx factors (see the Materials and Methods) for each group of genes is calculated using the published E11.5 (*left*) and E13.5 (*right*) ChIP-seq data for Dlx1, Dlx2, and Dlx5 (Lindtner et al. 2019). Note, those genes that are down-regulated in *Gsx2* mutant LGEs are significantly enriched for nearby genomic regions that bind both Dlx and Gsx2 factors compared with the unchanged group (*P*-value by Fisher's exact test). In contrast, the up-regulated gene group is not significantly different from the unchanged gene group. Thus, a substantial portion of genes down-regulated in *Gsx2* mutant animals is likely due to the indirect loss of Dlx transcription factor expression.

(Lindtner et al. 2019). Since Lindtner et al. found that the intersection of Dlx1, Dlx2 and/or Dlx5 genomic binding provided a high-confidence dataset, we intersected our E12.5 Gsx2 CUT&RUN data with the combined Dlx genomic binding regions at E11.5 and E13.5. Importantly, we found that over 30% of the Gsx2 bound genomic regions are also bound by two or more Dlx TFs (Fig. 6G). Next, we asked whether those genomic regions that bind both Gsx2 and Dlx TFs within a 100 kb window of a gene's transcription start site are more highly associated with genes that are either down- or up-regulated in *Gsx2* mutant LGEs. Intriguingly, this analysis revealed that down-regulated genes in *Gsx2* mutant LGEs were significantly more likely to be bound by both Gsx2/Dlx factors than genes that are up-regulated in *Gsx2* mutant LGEs (Fig. 6H). These findings are consistent with the idea that Dlx factors, which are known to function as activators (Stuhmer et al. 2002; Le et al. 2017), and Gsx2 regulate many of the same target genes through the same enhancer elements. Thus, these results reveal how the misregulation of other TFs in *Gsx2* mutants complicates our ability to predict Gsx2-dependent output in the mouse LGE based on the simple presence of M versus D sites. Nevertheless, our CUT&RUN, RNA-seq, and bioinformatics data do clearly reveal that mouse Gsx2 binds to both M and D sites in vivo and that Gsx2 binding to regulatory elements with such sites significantly influences LGE gene expression.

The contribution of monomer versus dimer sites to the regulatory specificity of Gsx2/ind *autoregulation*

Despite the many similarities between vertebrate Gsx factors and *Drosophila* Ind, a significant difference in autoregulation exists. Mouse Gsx2 negatively regulates itself (see Fig. 1), whereas *Drosophila* Ind positively regulates its own expression (Von Ohlen et al. 2007b; Von Ohlen and Moses 2009). To better define how M and D sites influence gene regulatory outcomes, we next focused on binding sites within defined enhancers implicated in *Gsx2* and *ind* autoregulation. First, we sought to determine whether fly Ind autoregulates itself via cooperative binding to D sites. Like mouse Gsx2, purified Ind protein cooperatively binds to a D site, but not an M site in EMSAs (Fig. 7A), and that an Ind protein with amino acids flanking its HD had a similar Hill cooperativity factor as mouse Gsx2 (~1.6, Supplemental Fig. S7A–C). These data support the idea that cooperative D site binding and monomeric M site binding are conserved features of the Gsx/Ind family.

Next, we characterized the 1.6 kb *ind* autoregulatory enhancer that was previously shown to drive Ind-dependent gene expression in *Drosophila* neuroblasts (Von Ohlen et al. 2007b). To simplify the analysis of this enhancer, we mapped the region responsible for neuroblast expression to a 980-bp fragment and used a combination of position weight matrices (PWMs) and EMSAs to identify four D sites and six M sites that were appropriately bound by Ind (Fig. 7B; Supplemental Fig. S7D–K). To determine whether these sites regulate *ind* enhancer activity in vivo, we generated transgenic reporters where all six M sites (i.e., Mmut) (Fig. 7C,C′; Supplemental Fig. S7D–I) or two of the four D sites were mutated (i.e., Dmut) (Fig. 7D,D′; Supplemental Fig S7J,K). A single copy of each transgene was integrated into the same locus as the WT reporter line (attP-51C) and β-gal levels were assessed in Ind-expressing cells of age-matched embryos (Fig. 7B-D′). Importantly, we found that the M mutant (Mmut) reporter expressed significantly higher β-gal than the WT reporter, whereas the D mutant (Dmut) reporter produced significantly lower levels of β-gal than wild type (Fig. 7C–E). As a control, endogenous Ind protein levels were also quantified and did not vary significantly between embryos carrying WT and mutant reporters (Fig. 7F). Thus, reducing the ratio of D to M sites from 4:6 in the wild-type element to 2:6 in the *Dmut* enhancer changes it from a positively to a negatively autoregulated enhancer. Furthermore, these in vivo *Drosophila* data support the model that the differential transcriptional regulation via M and D sites is a conserved feature of the Gsx/Ind factors.

Our DNA binding and reporter data in mammalian cells and *Drosophila* embryos support a model in which

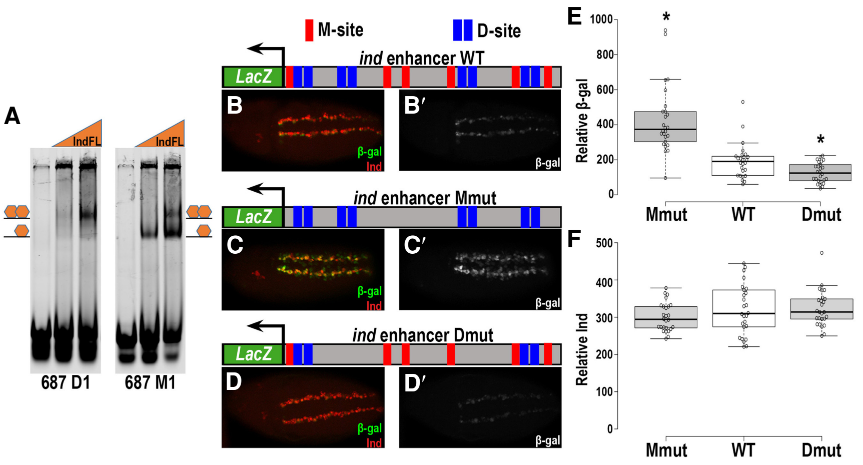

Figure 7. Relative numbers of monomer and dimer sites determine transcriptional response to Ind in *Drosophila* embryos. (*A*) Comparative EMSAs using the 687 M1 and D1 probes with equal amounts of full-length Ind reveals monomer versus dimer binding, which are schematically highlighted adjacent to each gel. (*B*–*D*) Ventral view of stage 10 *Drosophila* embryos with wild type *ind-lacZ*, *M mut ind-lacZ*, and *D mut ind-lacZ* immunostained for β-gal (green). Note, the two stripes of β-gal-positive cells are neuroblasts that express endogenous Ind (red). (*E*, *F*) Box plot of β-gal (*E*) and Ind (*F*) intensities from at least 10 embryos for each transgene. Each dot represents average β-gal or Ind intensity in an embryo, center lines show median, box limits indicate 25th and 75th percentile, and asterisks denotes significance ($P < 0.01$).

Gsx factors facilitate both positive and negative regulatory outcomes in a DNA binding site-dependent manner. Given that Gsx factors cooperatively bind D sites, we hypothesize that at low levels Gsx would preferentially bind and stimulate enhancers via the more stably bound D sites, whereas further increasing Gsx levels would result in M site binding and transcriptional repression (see Fig. 2E). In addition, we hypothesize that differing the ratios of D-to-M sites within regulatory DNA elements would alter transcriptional output; specifically, higher D-to-M ratios would increase expression levels, whereas lower D-to-M ratios would decrease expression levels. To test these hypotheses, we first designed an EMSA experiment where differentially labeled probes encoding the high-affinity *DSG* D site (magenta) or the high affinity *687* M1 site (green) compete for Gsx2 binding in the same reaction (Fig. 8A). Consistent with the D site having a higher affinity for Gsx2 than the M site, we found that the amount of free D site probe is depleted faster than the free M site probe as Gsx2 protein concentrations are increased (Fig. 8A,B).

Next, we tested how the balance between D and M sites alters transcriptional output in a context where mouse Gsx2 levels are easily manipulated using cell transfection. To do so, we constructed luciferase reporters containing five UAS sites and three copies of the D site from the *DSG* enhancer, and either two copies of an M site or two copies of a mutated M site to preserve the spacing between activating sites and the promoter (Fig. 8C). It should be noted that each reporter also contains a predicted M site in the minimal promoter (see Fig. 2). Comparative analysis between the two reporters revealed that both display a nonmonotonous pattern in which increasing Gsx2 levels stimulated higher luciferase levels until a maximum is reached, and then luciferase expression decreased in response to additional increases in Gsx2 levels (Fig. 8D). However, the Gsx2 mediated increase in expression is only about twofold over Gal4-VP16 alone for the M site reporter (Fig. 8D, red bars), while the maximum increase for the M site mutant reporter is approximately ninefold (Fig. 8D, gray bars). In addition, while the relative luciferase expression decreases from its maximum at very high Gsx2 concentrations, it never falls below the level stimulated by Gal4-VP16 alone (dashed line in Fig. 8D) for the M site mutant reporter, but there is clear repression at high Gsx2 levels on the M site-containing reporter. Taken together, this data is consistent with a model of regulation

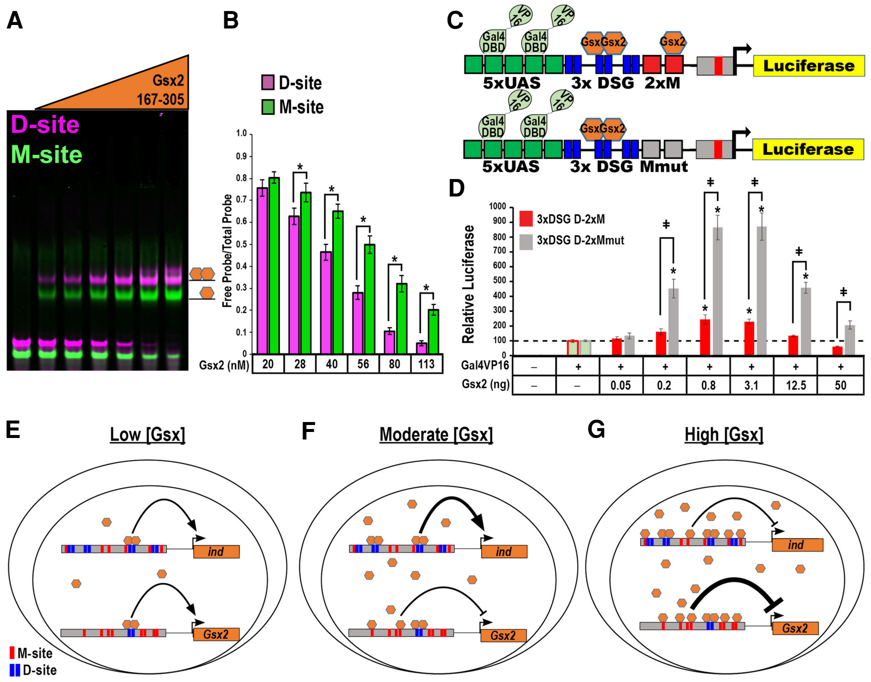

Figure 8. Relative numbers of monomer and dimer sites and concentration of Gsx2 determine transcriptional response in mouse mK4 cells. (*A*) Gsx2 EMSA competition for binding to the DSG-D site probe (magenta) and the M1 probe (green) in the same reaction. Each EMSA binding reaction had 3.4 nM each labeled DNA probe with either no protein added (first lane) or with 20, 28, 40, 56, 80, or 113 nM purified Gsx2 (167–305). Note, the free D site probe is depleted more rapidly than free M site probe as Gsx2 protein is increased. (*B*) EMSA in *A* was performed in quadruplicate and quantified. The Gsx2 concentration in each lane is indicated. The ratio of free probe to total probe is plotted as a function of Gsx2 protein concentration with error bars indicating standard deviation. (*) $P <$ 0.05 using an unpaired two-tailed Student's *t*-test comparing M and D values at each Gsx2 concentration. (*C*) Schematics of Luciferase reporters containing three DSG-D sites, and either two wild-type or mutant M1 sites. (*D*) Luciferase assays in mammalian mK4 cells using reporters containing 3xDSG D sites and either 2× wild-type M1 (red bars) or mutant M sites (gray bars). Note, Gsx2 strongly stimulates the construct containing the mutant M site but only weakly stimulates the reporter with the wild-type M sites. 25 ng of indicated luciferase reporter and 5 ng of Gal4-VP16 expression vector were transfected where indicated (+). The amount of Gsx2 plasmid transfected is noted. A two-way ANOVA with Tukey post hoc was used to determine significance. (*) $P < 0.05$ compared with Gal4VP16 alone, (*) $P < 0.05$ comparing M site reporter with M mutant reporter at indicated concentration of Gsx2. (*E*–*G*) Model of Gsx regulation of enhancers containing different ratios of D and M sites. (*E*) At low concentrations, Gsx factors bind dimer sites, and stimulate target gene transcription on each enhancer with D sites. (*F*) At moderate concentrations, Gsx factors differentially regulate target gene expression based on the ratio of D-to-M sites: Those enhancers with a relatively high D-to-M site ratio will increase in activity, whereas those with low D-to-M site ratios will recruit Gsx2 to M sites and repress gene expression. (*G*) At high Gsx concentrations, Gsx2 will bind M sites to repress enhancers with both high and low D-to-M site ratios. As examples, we are modeling two different enhancers; one that represents the fly *ind* enhancer and the other is the mouse *687* enhancer.

where low Gsx2 levels stimulate expression as D sites are preferentially filled (Fig. 8E), while high Gsx2 levels result in repression due to increased Gsx2 binding to M-sites (Fig. 8G). Notably, moderate Gsx2 levels reveal differential gene regulation depending on the ratio of D to M sites in the enhancer (Fig. 8F).

Discussion

In this study, we investigated how the mouse Gsx2 and *Drosophila* Ind TFs recognize target genes and regulate their expression. First, we found that like *Drosophila* Ind, mouse Gsx2 mediates autoregulation in the developing nervous system. However, unlike the positive autoregulatory behavior of Ind (Von Ohlen et al. 2007b), Gsx2 mediates negative autoregulation. Moreover, we show that Gsx2/Ind use distinct types of DNA binding sites to yield opposing transcriptional outcomes. Consistent with previous studies (Von Ohlen et al. 2009; Jolma et al. 2013), we found that mouse Gsx2 represses target gene expression by binding to AT-rich monomer (M) sites. However, we also found that the fly, mouse, and human Gsx2/Ind factors can also cooperatively bind DNA via precisely spaced and oriented homodimer (D) sites. Importantly, the D sites mediate gene stimulation rather than repression in both mammalian cell culture and transgenic *Drosophila* reporter assays. These findings led to a model wherein the ratio of M to D sites within an enhancer and the levels of the Gsx/Ind factor can result in distinct transcriptional responses (see Fig. 8E–G). Altogether, these results provide new insights into the *cis*-regulatory logic of Gsx/Ind-mediated gene expression during neural development, and thereby revealed a novel mechanism used by HD TFs to gain both DNA binding and regulatory specificity.

The cis*-regulatory logic of Gsx-mediated gene regulation*

Gsx factors regulate D-V patterning in the fly and mouse nervous system by repressing the molecular identity of cells in the adjacent developmental compartment (Weiss et al. 1998; Corbin et al. 2000; Toresson et al. 2000; Yun et al. 2001). In fact, studies in flies and frogs have led to the notion that Gsx factors function primarily as repressors (Von Ohlen et al. 2007a; Von Ohlen and Moses 2009; Winterbottom et al. 2010, 2011). However, Ind can also positively regulate its own expression in embryonic *Drosophila* neuroblasts (Von Ohlen et al. 2007b), although the mechanisms determining positive versus negative output were unknown. Our findings that Gsx factors mediate opposing transcriptional outcomes via M versus D sites reveals an unanticipated mechanism underlying how these factors control gene expression during neural development. By focusing on binding sites within enhancers near *Gsx2* in mice and *Ind* in *Drosophila*, we found that autoregulation is a common feature of Gsx factors, and their impact on autoregulation is complex. For instance, consistent with direct Ind-mediated positive autoregulation in *Drosophila* neuroblasts, we found that Ind cooperatively binds D sites from its autoregulatory enhancer and mutating these D sites dramatically decreased enhancer activity in vivo. However, we also found that Ind binds multiple M sites from the same autoregulatory enhancer and mutating these M sites results in a significant increase in enhancer activity in intermediate column neuroblast cells. These findings support the model that Ind autoregulation is both positive via cooperative D sites and negative via independent M sites.

Consistent with these results, we found that the D-to-M ratio within Gsx-regulated enhancers can have a profound impact on regulatory outcomes in mammalian cells. For instance, while both low and high Gsx2 levels repressed reporters with only M sites, reporters with enhancers containing a mixture of D + M sites revealed a more complex, nonmonotonous pattern of enhancer activity as a function of Gsx2 protein levels. At relatively low levels, Gsx2 stimulated the D + M enhancer activity, consistent with Gsx2 cooperativity resulting in preferential binding to D sites (Fig. 8E). However, as Gsx2 levels were increased, the D + M reporter activity decreased, consistent with higher Gsx2 levels resulting in repression via M sites (Fig. 8F,G). The net impact of this mechanism on autoregulation is that Gsx2/Ind transcription factors would maintain their own expression levels in a narrow window based on the ratio of stimulatory D sites to repressive M sites in each respective enhancer.

We used the CUT&RUN assay to show that Gsx2 binds to a substantial number of potential regulatory elements that contain M and D sites within the developing mouse forebrain. Consistent with these Gsx2 CUT&RUN peaks being in functional enhancers, a high percentage of the binding events mapped to genomic regions with active chromatin marks in E12.5 whole telencephalon samples. Moreover, transcriptomic studies revealed that Gsx2 binding events to genomic regions with D sites, M sites, or D + M sites were each significantly associated with nearby genes that exhibited altered expression in *Gsx2* mutants. However, there was no direct correlation between the mere presence of a D site and gene up-regulation or an M site and gene down-regulation. In hindsight, our inability to predict gene expression changes based on only D versus M sites within each individual Gsx2 peak is not unexpected for several reasons. First, most Gsx2 bound enhancers contain a mixture of D and M sites, and as we found by studying the Gsx2 and Ind autoregulatory elements, the behavior of D + M elements is complex due to the opposing activities mediated by Gsx binding to each type of site. Second, transcriptomics data revealed the expression of a number of key LGE TFs are significantly changed in *Gsx2* mutants. Hence, the germline removal of *Gsx2* is likely to have numerous indirect impacts on gene regulation in the E12.5 mouse LGE. For example, analysis of existing ChIP-seq data for the misregulated Dlx TFs, which bind highly similar if not identical DNA sequences as Gsx2, revealed significant overlap with the genomic Gsx2 binding profile (Lindtner et al. 2019). Moreover, we found that only down-regulated genes in *Gsx2* mutant LGEs are preferentially enriched for genomic regions bound by both Gsx2 and Dlx factors. Hence, genes that are significantly down-regulated in *Gsx2* mutants could be indirectly decreased due to the loss of the Dlx TFs, which are known to function as transcriptional activators (Stuhmer et al. 2002; Le et al. 2017). This result is not surprising because the LGE contains many more Dlx-expressing cells than Gsx2-expressing cells; thus, altered gene expression in the *Gsx2* mutant LGE is likely to be heavily influenced by the loss of such downstream effectors. Overall, these data highlight how the complexity of *cis*-regulatory elements is likely to make it very difficult to predict their transcriptional behaviors based solely upon Gsx2 D and M site binding information using bulk transcriptomic data from *Gsx2* germ line mutations. Thus, future studies focused on shorter temporal windows following *Gsx2* removal or addition are needed to ascertain how Gsx2 M versus D site binding directly impacts gene regulatory outcomes in the mouse LGE.

An additional unanswered question raised by our findings is how do the Gsx/Ind transcription factors mediate opposing outcomes when bound to DNA as monomers versus dimers? Prior studies found that the Gsx/Ind factors bind the Groucho corepressor protein through a conserved N-terminal domain (Von Ohlen et al. 2007a, 2009; Von Ohlen and Moses 2009; Winterbottom et al. 2010, 2011). However, Ind has been shown to have two additional regulatory activities: a second repression domain and an undefined activation domain (Von Ohlen and Moses 2009). Unfortunately, we do not know the identity of such additional corepressors/coactivators that interact with the Ind/Gsx factors, and therefore we lack an understanding of how dimer formation on DNA alters the ability of Gsx/Ind factors to recruit specific cofactors. Moreover, it should be noted that Gsx2 alone is insufficient to activate gene expression, at least in mK4 cells, but is able to stimulate gene activation in conjunction with the Gal4-VP16 protein. How the Gsx/Ind proteins synergize with other nearby transcriptional regulators to stimulate gene expression is currently unknown and an important issue to be addressed in future studies.

The impact of cooperative Gsx/Ind DNA binding on HD target specificity

Gsx TFs are members of a large HD family that regulate diverse developmental processes ranging from anterior-posterior axis specification to cell fate specification within numerous tissues and organs (Bürglin and Affolter 2016; Zandvakili and Gebelein 2016). How such factors gain in vivo target specificity, while having highly similar in vitro DNA binding properties has been a long-standing problem in developmental biology (Mann et al. 2009; Zandvakili and Gebelein 2016). While some HD proteins have been shown to bind DNA as homodimers, the vast majority of these factors contain known dimerization domains such as the LIM domain (Bürglin and Affolter 2016). In addition, other HD TFs have been shown to form functional heterodimers with each other, resulting in increased DNA binding specificity. Perhaps the best example is the Hox factors that form cooperative heterodimer complexes with members of the three-amino-acid loop (TALE) HD proteins such as Pbx and Meis to regulate anterior-posterior patterning (Gebelein et al. 2004; Li-Kroeger et al. 2008; Mann et al. 2009; Uhl et al. 2010; Zandvakili and Gebelein 2016). In contrast to these HD proteins, the Gsx factors, as well as the majority of other HD TFs, do not contain known homodimerization or heterodimerization domains. Importantly, we found that while a mouse Gsx2 protein encoding only the HD is sufficient to bind M sites, it is insufficient to form cooperative homodimers on D sites. Instead, domains flanking the HD were required to form cooperative Gsx2 homodimers on binding sites with precise DNA spacing and orientation requirements. Intriguingly, the sequences flanking the mouse Gsx2 and fly Ind HDs do not have extensive sequence conservation, and yet we provide evidence that proteins containing these regions each cooperatively binds to D sites. Thus, these data suggest that cooperative binding to D sites is a conserved feature of Gsx TFs, and future studies are needed to understand how the flanking sequences mediate cooperative DNA binding.

Overall, we found that the net effect of cooperative Gsx dimer binding is threefold: First, the increased DNA binding sequence requirements for forming cooperative complexes increases DNA binding specificity. For example, while other related HD TFs, such as the Dlx proteins, enrich for highly similar monomer DNA sites as the Gsx factors, the Dlx factors did not significantly enrich for dimer sites (Jolma et al. 2013; Lindtner et al. 2019). Hence, while Gsx2 forms cooperative complexes on AT-rich binding sites with 7 base-pair spacing and a forward-forward orientation, the Dlx factors are only predicted to bind such sites in a noncooperative monomer-like fashion. Second, the cooperative interactions between two Gsx molecules increases DNA binding affinity. Thus, low concentrations of Gsx factors are predicted to more consistently bind and regulate gene expression using D sites rather than M sites. Third, Gsx factors gain regulatory specificity when bound to D versus M sites. While the exact mechanisms underlying the ability of Gsx factors to stimulate gene expression on D sites and repress gene expression on M sites is unclear, these activities provide a unique mechanism that begins to explain how the same TF mediates opposing outcomes based on DNA binding site architecture. Intriguingly, published HT-SELEX data defining HD DNA binding preferences of the HoxL and NkxL subfamilies, which includes the Gsx factors, revealed that several other HD TFs that lack known dimerization motifs also enriched for both monomer and homodimer sites (Jolma et al. 2013). Moreover, the homodimer sites were found to often vary in length, suggesting that homodimer formation by different HD TFs may require distinct binding site spacing. Overall, these data raise the possibility that the selective formation of monomers versus homodimers will be a generalizable mechanism used by a subset of HD TFs to increase both DNA binding and regulatory specificity.

Materials and methods

Molecular cloning

Oligonucleotide sequences used to generate *luciferase, lacZ,* and expression constructs are listed in FASTA format in Supplemental Document 1. The following plasmids were used: pET14b (Novagen) for bacterial protein expression, pCDNA6 (Thermo Fischer) for mammalian cell expression, and pGL3basic (Promega) for luciferase reporter assays. When necessary, PCR was performed using Accuzyme DNA polymerase (Bioline). Ligation reactions were performed using T4 DNA Ligase (NEB). All generated plasmid constructs were verified by DNA sequencing.

Mouse lines

Genotyping of the *Gsx2*^{EGFP} and *Gsx2*^{RA} mouse lines have been described previously (Waclaw et al. 2009; Wang et al. 2009). The *Gsx2*^{2xFLAG} allele was generated via CRISPR-Cas9 using a gRNA to the 5′ UTR of *Gsx2*. A ssDNA donor oligonucleotide, which contained cross-species nucleotide mutations 5′ to the PAM sequence to prevent retargeting by Cas9, was used to insert the 2xFLAG tag and GSG-linker coding sequence in-frame with *Gsx2* (see Supplemental Document 1). To confirm 2xFLAG insertion, we used two oligonucleotide primers (5′-Primer_for_2xFLAG_genotyping and 3′-Primer_for_2xFLAG_genotyping) (see Supplemental Document 1) to identify the *Gsx2*^{2XFLAG} allele. Two founders were positive for the 2xFLAG insertion and each line was bred to homozygosity. The locus surrounding the transcription start site (TSS) and coding region of *Gsx2* was amplified and sequenced to confirm that the 2XFLAG tag was properly inserted in frame and the modified *Gsx2* reading frame was intact (see Supplemental Document 1). To confirm homozygous *Gsx2*^{2XFLAG} animals were "phenotypically normal," we collected E12.5 embryos from a homozygous cross, fixed each in 4% PFA for 6 h and processed for histology as described (Wang et al. 2009). For immunohistochemistry, the following primary antibodies were used; mouse anti-FLAG (1:500; Sigma), rabbit anti-Gsx2 (1:2000) (Toresson et al. 2000) and mouse anti-Pax6 (1: 1000; Invitrogen). Fluorescent secondary antibody detection was carried out using donkey anti-rabbit IgG conjugated to Alexa 647 (1:200; Jackson Immunoreseach) and goat anti-mouse IgG$_1$ conjugated to Alexa 568 (1:500; Invitrogen). Sections were imaged for confocal microscopy using a Nikon A1 LSM system with a GsAsP solid-state laser.

Drosophila *transgenic reporter assays*

The 1.6-kb Ind autoregulatory element (Von Ohlen et al. 2007b) and proximal 980 bp of that element relative to the *ind* TSS were PCR-amplified from an Ind-BAC using two primers (Ind_autoregulatory_element_Forward and either Ind_1.6 kb_autoregulatory_element_Reverse or Ind_980 bp_autoregulatory_element_Reverse). Fragments were cloned into *pattBLacZ* (Bischof et al. 2007) using XbaI and XhoI sites. The Mmut and Dmut constructs were synthesized by GenScript and cloned using the same enzyme sites into *pattBLacZ*. The WT 980 bp element, Mmut, and Dmut element sequences can be found in Supplemental Document 1. To create transgenic flies, each construct was injected into *Drosophila* embryos by Rainbow Transgenics and the φ-C31 integrase system was used to insert each into the 51C locus (Bischof et al. 2007). Embryos were harvested from fly lines homozygous for each *lacZ* transgene for 2 h at 25°C, aged an additional 3 h at 25°C before being collected, fixed and immunostained using standard procedures (Zandvakili et al. 2018, 2019). Chicken anti-β-gal (1:1000; Abcam) and rabbit anti-Ind (1:1000) (Von Ohlen and Moses 2009) primary antibodies and AlexaFlour fluorescent secondary antibodies were used to immunostain each sample. All transgenic embryos were imaged under identical subsaturating conditions using a Zeiss Axio Imager upright microscope with an Apotome filter for optical sectioning and an AxioCam MRm digital camera. Pixel intensities for β-gal and Ind were quantified using Fiji, and the average intensity of β-gal levels in Ind-positive regions were determined for each embryo and plotted as a single data point in the box and whisker plot in Figure 7. An unpaired two-tailed students T-test was performed to compare pixel intensities from each mutant to wild-type reporter.

Protein purification

Coding DNA for all protein constructs used in EMSAs were PCR-amplified, cloned in-frame with an N-terminal 6x-His tag into the pET-14b bacterial expression vector (Novagen), and sequence confirmed. All Gsx2 constructs were from *Mus musculus*. All Ind constructs were from *Drosophila melanogaster*. The Gsx2 HD construct (amino acids 203–264) was cloned using the Gsx2_HD_F and Gsx2_HD_R primers between NdeI and XhoI sites. The Gsx2 construct containing the HD and flanking regions (amino acids 167-305) was cloned using Gsx2_167-305_F and Gsx2_167-305_R primers into NdeI and XhoI sites. The Gsx2 full-length protein (amino acids 1-305) was cloned using the Gsx2_FL_F and Gsx2_FL_R primers between NdeI and XhoI sites. The Ind construct containing the HD and flanking regions (amino acids 175–320) was cloned using the Ind_175-320_F and Ind_175-320_R primers between NdeI and KpnI sites. Full-length Ind containing amino acids 1-320 was cloned using the Ind_FL_F and Ind_FL_R primers between NdeI and KpnI sites. Full-length Gsx2 and Ind proteins were purified via Ni-chromatography under denaturing conditions and allowed to refold while still bound to Ni-beads as previously described (Zhang et al. 2019). Gsx2 and Ind subfragments were purified via Ni-chromatography under native conditions as described previously (Uhl et al. 2010). Full-length proteins were confirmed via Western blot, whereas the purity of protein subfragments was confirmed via SDS-PAGE analysis and GelCode blue staining (Thermo Scientific). The indicated protein concentrations in the Figure legends were determined via Bradford assays.

Electrophoretic mobility shift assays (EMSA)

EMSAs were performed as described previously (Uhl et al. 2016; Kuang et al. 2020). Probe sequences are listed in FASTA format in Supplemental Document 2. Binding reactions containing the indicated DNA probes and protein concentrations listed in each Figure legend were mixed and incubated in the dark at room temperature for 10 min prior to gel electrophoresis. EMSAs were imaged via a Li-Cor Odyssey CLx scanner, and quantified using the Li-Cor image studio software as described previously (Roychoudhury et al. 2020).

Luciferase assays

Sequences for all reporters can be found in Supplemental Document 1. 5xUAS sites (highlighted in green), Gsx2-binding sites (highlighted red for M sites and blue for D sites), and a minimal promoter (gray) were synthesized by GenScript and subcloned into pGL3-basic luciferase (Promega) between KpnI and NcoI sites. The Gsx2 pCDNA6 construct has been previously described (Roychoudhury et al. 2020). The Gal4-VP16 coding DNA was cloned into pCDNA6 between KpnI and XbaI sites.

Luciferase assays were performed in triplicate using the mouse mK4 cell line as previously described (Roychoudhury et al. 2020). Each well was transfected with 5 ng of Renilla-luciferase, 25 ng of the indicated firefly luciferase reporter, 5 ng of Gal4-VP16 in pCDNA6 if indicated, and a titration of the indicated amounts of Gsx2 in pCDNA6. To ensure each well was transfected with the same amount of DNA, empty pCDNA6 vector was used to fill to 85 ng. Forty-eight hours after transfection, cells were harvested, lysed, and analyzed for luciferase activity using the Promega dual-luciferase assay kit and GloMax luminometer. All firefly luciferase values were normalized to Renilla-luciferase to control for transfection efficiency. Relative luciferase values presented in each bar graph represent the mean ± standard deviation with Gal4-VP16 alone condition set to 100.

ChIP-qPCR

LGE tissue was dissected from 35 E12.5 mouse forebrains, and ChIP assays were performed as described previously (Castro et al. 2011) with the following changes: Sonication was performed at 4°C using a Fisher Scientific Sonic Dismembrator model 100 at setting 5 for 25 cycles (10 sec on/30 sec off). Two percent of chromatin was set aside as input, and the remaining sample was split in half for immunoprecipitation with either a rabbit anti-Gsx2 antibody at a 1:100 dilution or 1 µg of control rabbit IgG. Quantitative PCR was performed using a QuantStudio 3 real-time PCR system and PowerUP SYBR Green Master Mix (Thermo Fisher). Primers for negative control regions, *Actb* and *Dll1* open reading frames (ORFs), were previously published (Castro et al. 2011). Primers for the *687* locus are 687_ChIP_PCR_F and 687_ChIP_PCR_R. qPCR using two sets of negative control primers and the *687* primers were performed in triplicate on 2% input DNA, Gsx2-ChIP sample, and IgG ChIP sample. Data are presented as fold enrichment over *Dll1*-ORF for each condition.

RNA sequencing

E12.5 embryos were collected from $Gsx2^{RA/+}$ x $Gsx2^{EGFP/+}$ crosses in ice cold PBS and the LGE and a section of embryonic tail were dissected from each embryo. LGEs were stored in RNAlater (Ambion) and tails were used for genotyping. A total of four pairs of LGEs from $Gsx2^{+/+}$ and $Gsx2^{EGFP/RA}$ genotypes were submitted for directional RNA-seq by the Genomics, Epigenomics and Sequencing Core at the University of Cincinnati. Five-hundred nanograms of total RNA was used to isolate polyA RNA using NEBNext Poly(A) mRNA Magnetic Isolation Module (New England BioLabs). Samples were further enriched using SMARTer Apollo NGS library preparation system (Takara Bio USA). A dUTP-based stranded library was generated using NEBNext Ultra II Directional RNA library preparation kit (New England BioLabs). The library was indexed and amplified for nine PCR cycles. Individually indexed libraries were proportionally pooled and sequenced at a depth of 25 million single-end 51-bp reads using a HiSeq 1000 sequencer (Illumina). The RNA-seq data are available from the GEO of the NCBI as GSE162590.

CUT&RUN assays

CUT&RUN assays were performed on dissected LGE tissue from 65 homozygous $Gsx2^{2xFLAG}$ E12.5 mouse embryos (6 litters). LGE tissue from each litter was pooled, dissociated by pipetting in 0.05% Trypsin-EDTA (Gibco), passed through a 40-µm cell strainer (Corning), step-frozen in CryoStor freezing media (STEMCELL Technologies), and stored at −80°C. Samples were thawed at 37°C and washed in DMEM with 5% FBS. Cells were pooled and washed in DPBS with 0.04% BSA and protease inhibitor cocktail. After binding to Concavalin A beads, cells were divided into separate samples of $\sim 1 \times 10^6$ cells each and duplicates were incubated with either the M2-FLAG antibody (Sigma) or an IgG control antibody. CUT&RUN was performed as previously described with the following exceptions (Skene and Henikoff 2017; Skene et al. 2018). A rabbit anti-mouse secondary antibody was used due to the low affinity of protein A for mouse IgG1 (M2-FLAG antibody). MNase digestion after the addition of $CaCl_2$ was performed for 45 min at 0°C. After stopping MNase digestion, released DNA fragments were purified using the Nucleospin DNA cleanup kit. Libraries were prepared using the NEBNext Ultra II DNA library preparation kit for Illumina sequencers, and PCR products between ∼150 and 350 bp were size selected using AMPure XP beads. Bar-coded libraries were sequenced on an Illumina NextSeq 550 sequencer at a depth of ∼20 million paired-end 75-bp reads.

Bioinformatics analysis

RNA-seq reads were aligned to mouse genome, mm10, using STAR aligner (Dobin et al. 2013). Raw read counts aligned to each gene were measured using featureCounts (Liao et al. 2014). Genes within autosomes were considered for analysis to avoid sex chromosome gene bias from mixed samples. To assess *Gsx2* gene expression in *Gsx2* mutant LGEs, we only considered the first *Gsx2* exon that is located upstream of the genetic deletion. Differential gene expression analysis was done using EdgeR (McCarthy et al. 2012). Genes were defined as "expressed" if it displayed >0.5 RPKM in at least one sample. Genes with fold change >1.5 and FDR <0.05 were selected as differentially expressed for gene ontology (GO) analysis using EnrichR (Kuleshov et al. 2016). Top significant GO terms in Biological Process categories were selected for presentation.

CUT&RUN reads were aligned to the mouse genome, mm10, using STAR after adapter-trimming. Read pairs were connected with each other to form fragments, then split into two groups by fragment size, nucleosome free reads (termed NFR, <120 bp) and nucleosomal reads (termed NUC, >150 bp). For peak calling, we considered nucleosome free reads only and performed peak calling using the IgG sample as control by "findPeaks" in Homer. Peaks falling into ENCODE blacklist regions were discarded, and robust peaks with more than one RPM were retained for downstream analysis. To capture both M and D motifs in the Homer analysis, we selected parameters that searched for both short 8-, 10-, and 12-bp motifs and long 16-, 18-, and 20-bp motifs. Using Homer's annotatePeaks.pl tool the genomic location of each peak was defined as promoter-TSS (1 kb upstream of or 100 bp downstream from a TSS), 5′-UTR, 3′-UTR, exon, intron, intergenic, noncoding, or TSS. CUT&RUN profiles were examined at Gsx2 peaks across biological duplicates of Gsx2/FLAG and IgG samples using bwtools (Pohl and Beato 2014). De novo motif search within Gsx2 peaks was performed using Homer and MEME (Whitington et al. 2011), and we obtained robust enrichment of D and M motifs using both approaches. Two types of BigWig files were made using bedtools (Quinlan and Hall 2010) and bedGraphToBigWig (Kent et al. 2010): (1) using NFR fragments to visualize peaks in an unstranded manner where fragments were resized to 100 bp to normalize for sample bias, and (2) using 5′ end 1-bp aligned reads in a strand-specific manner for footprint analysis.

For CUT&RUN footprint analysis of Gsx2 D and M site binding, we first scanned 300 bp peak windows for D and M motifs using FIMO (Grant et al. 2011). For D sites, we selected genomic loci with P-value < 0.001 as candidate Gsx2 D sites. Since simply relying on motif score is subject to a high rate of false positives, we performed filtering based on MNase digestion signal; i.e., 5′ ends of CUT&RUN sequencing reads. Assuming that true Gsx2 binding protects the underlying DNA from MNase digestion, we measured average MNase digestion frequency within the motif window (M_m) and flanking 40 bp regions on both sides (M_f) then calculated a contrast, $(M_f + \alpha)/(M_m + \alpha)$, where α is a pseudo value of 0.01 to avoid division by zero. Among candidate sites from the motif scan, ones with contrast >1 only were retained as D sites. If there is a group of D sites overlapping each other, the one with the maximum contrast was selected. Likewise, we repeated this procedure for the M motif with a relaxed P-value (<0.01) to allow for mismatches. Also, any M motif regions overlapping with previously identified D sites were discarded.

To compare spatial proximity of Gsx2 binding to the distinct group of genes from RNA-seq, we checked the distance from transcription start sites (TSS) to the closest Gsx2-binding sites for each gene and visualized the cumulative percentage of genes having at least one Gsx2 binding event up to 100 kb for each group of genes. CUT&RUN data are available from the GEO of the NCBI as GSE162589.

For integrative analysis with Dlx ChIP-seq data, we downloaded six peak calling results for Dlx1, Dlx2, and Dlx5 in E11.5 and E13.5 from GEO (GSE124936). Peak coordinates were converted to hg38 using liftOver, and redundant coordinates were condensed into one. Dlx1, Dlx2, and Dlx5 peaks were first resized to 100 bp, then were pooled and merged for each of E11.5 and E13.5. Peaks bound by at least two Dlx factors were retained for downstream analysis for each embryonic stage. Gsx2 peaks located within 100-kb window around TSS of KO down-regualted, unchanged, up-regulated genes were selected to compare colocalization with Dlx. P-value for the Dlx colocalization was calculated using Fisher's exact test against Gsx2 nearby unchanged genes.

Acknowledgments

We thank Aaron Zorn, Ron Waclaw, and members of the Gebelein and Campbell laboratories for comments on this work. We thank Dr. Matthew Weirauch (Cincinnati Children's Hospital Medical Center [CCHMC]) for bioinformatics advice. We thank Dr. Yueh-Chiang Hu and the CCHMC genome-editing core for helping with the design and creation of the *Gsx2*^{2xFLAG} mouse line, and the University of Cincinnati Sequencing Core for RNA sequencing. This work was supported by the National Institutes of Health (R01 NS044080 to K.C. and B.G., R01 NS069893 to M.N. and K.C., and R01 GM079428 to B.G. and H.-W.L.), the National Science Foundation (1715822 to B.G.) and Cincinnati Children's Research Foundation (Trustee Award to H.-W.L.).

Author contributions: This scientific study was conceived and planned by J.S., K.C., M.N., and B.G. The design and making of the *Gsx2*^{2xFLAG} mouse allele were done by S.Q., J.S., and K.C. in conjunction with the CCHMC genome-editing core. The transgenic *Drosophila* constructs and experiments were designed and performed by J.S. and B.G. Protein purification and gel shift analysis were performed by J.S., E.F., and B.G. Embryonic mouse dissections and RNA-seq experiments were performed by S.Q. and K.C. CUT&RUN assays were designed and performed by T.D.F., B.G., R.B.H., and J.S. ChIP-PCR experiments were performed by J.S. Mouse immunostaining was performed by S.Q. Bioinformatics analysis was performed by B.C., J.S., and H.-W.L. The manuscript was written by J.S., K.C., and B.G.

References

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, et al. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133:** 1266–1276. doi:10.1016/j.cell.2008.05.024

Bischof J, Maeda RK, Hediger M, Karch F, Basler K. 2007. An optimized transgenesis system for *Drosophila* using germ-line-specific φC31 integrases. *Proc Natl Acad Sci* **104:** 3312–3317. doi:10.1073/pnas.0611511104

Bürglin TR, Affolter M. 2016. Homeodomain proteins: an update. *Chromosoma* **125:** 497–521. doi:10.1007/s00412-015-0543-8

Castro DS, Martynoga B, Parras C, Ramesh V, Pacary E, Johnston C, Drechsel D, Lebel-Potter M, Garcia LG, Hunt C, et al. 2011. A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. *Genes Dev* **25:** 930–945. doi:10.1101/gad.627811

Corbin JG, Gaiano N, Machold RP, Langston A, Fishell G. 2000. The Gsh2 homeodomain gene controls multiple aspects of telencephalic development. *Development* **127:** 5007–5020.

De Mori R, Severino M, Mancardi MM, Anello D, Tardivo S, Biagini T, Capra V, Casella A, Cereda C, Copeland BR, et al. 2019. Agenesis of the putamen and globus pallidus caused by recessive mutations in the homeobox gene GSX2. *Brain* **142:** 2965–2978. doi:10.1093/brain/awz247

Desmaris E, Keruzore M, Saulnier A, Ratié L, Assimacopoulos S, De Clercq S, Nan X, Roychoudhury K, Qin S, Kricha S, et al. 2018. DMRT5, DMRT3, and EMX2 cooperatively repress *Gsx2* at the pallium-subpallium boundary to maintain cortical identity in dorsal telencephalic progenitors. *J Neurosci* **38:** 9105–9121. doi:10.1523/JNEUROSCI.0375-18.2018

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Gebelein B, McKay DJ, Mann RS. 2004. Direct integration of Hox and segmentation gene inputs during Drosophila development. *Nature* **431:** 653–659. doi:10.1038/nature02946

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018. doi:10.1093/bioinformatics/btr064

Hsieh-Li HM, Witte DP, Szucsik JC, Weinstein M, Li H, Potter SS. 1995. Gsh-2, a murine homeobox gene expressed in the developing brain. *Mech Dev* **50:** 177–186. doi:10.1016/0925-4773(94)00334-J

Illes JC, Winterbottom E, Isaacs HV. 2009. Cloning and expression analysis of the anterior parahox genes, *Gsh1* and *Gsh2* from *Xenopus tropicalis*. *Dev Dyn* **238:** 194–203. doi:10.1002/dvdy.21816

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152:** 327–339. doi:10.1016/j.cell.2012.12.009

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. Bigwig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26:** 2204–2207. doi:10.1093/bioinformatics/btq351

Konno D, Kishida C, Maehara K, Ohkawa Y, Kiyonari H, Okada S, Matsuzaki F. 2019. Dmrt factors determine the positional information of cerebral cortical progenitors via differential suppression of homeobox genes. *Development* **146:** dev174243. doi:10.1242/dev.174243

Kuang Y, Golan O, Preusse K, Cain B, Christensen CJ, Salomone J, Campbell I, Okwubido-Williams FV, Hass MR, Yuan Z, et al.

2020. Enhancer architecture sensitizes cell specific responses to Notch gene dose via a bind and discard mechanism. *Elife* **9:** e53659. doi:10.7554/eLife.53659

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44:** W90–W97. doi:10.1093/nar/gkw377

Le TN, Zhou QP, Cobos I, Zhang S, Zagozewski J, Japoni S, Vriend J, Parkinson T, Du G, Rubenstein JL, et al. 2017. GABAergic interneuron differentiation in the basal forebrain Is mediated through direct regulation of glutamic acid decarboxylase isoforms by *Dlx* homeobox transcription factors. *J Neurosci* **37:** 8816–8829. doi:10.1523/JNEUROSCI.2125-16.2017

Liao Y, Smyth GK, Shi W. 2014. Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656

Li-Kroeger D, Witt LM, Grimes HL, Cook TA, Gebelein B. 2008. Hox and senseless antagonism functions as a molecular switch to regulate EGF secretion in the *Drosophila* PNS. *Dev Cell* **15:** 298–308. doi:10.1016/j.devcel.2008.06.001

Lindtner S, Catta-Preta R, Tian H, Su-Feher L, Price JD, Dickel DE, Greiner V, Silberberg SN, McKinsey GL, McManus MT, et al. 2019. Genomic resolution of DLX-orchestrated transcriptional circuits driving development of forebrain GABAergic neurons. *Cell Rep* **28:** 2048–2063.e8. doi:10.1016/j.celrep.2019.07.022

Mann RS, Lelli KM, Joshi R. 2009. Hox specificity unique roles for cofactors and collaborators. *Curr Top Dev Biol* **88:** 63–101. doi:10.1016/S0070-2153(09)88003-4

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res* **40:** 4288–4297. doi:10.1093/nar/gks042

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133:** 1277–1289. doi:10.1016/j.cell.2008.05.023

Pei Z, Wang B, Chen G, Nagao M, Nakafuku M, Campbell K. 2011. Homeobox genes Gsx1 and Gsx2 differentially regulate telencephalic progenitor maturation. *Proc Natl Acad Sci* **108:** 1675–1680. doi:10.1073/pnas.1008824108

Pohl A, Beato M. 2014. Bwtool: a tool for bigWig files. *Bioinformatics* **30:** 1618–1619. doi:10.1093/bioinformatics/btu056

Qin S, Madhavan M, Waclaw RR, Nakafuku M, Campbell K. 2016. Characterization of a new *Gsx2-cre* line in the developing mouse telencephalon. *Genesis* **54:** 542–549. doi:10.1002/dvg.22980

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Roychoudhury K, Salomone J, Qin S, Cain B, Adam M, Potter SS, Nakafuku M, Gebelein B, Campbell K. 2020. Physical interactions between Gsx2 and Ascl1 balance progenitor expansion versus neurogenesis in the mouse lateral ganglionic eminence. *Development* **147:** dev185348. doi:10.1242/dev.185348

Skene PJ, Henikoff S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6:** e21856. doi:10.7554/eLife.21856

Skene PJ, Henikoff JG, Henikoff S. 2018. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat Protoc* **13:** 1006–1019. doi:10.1038/nprot.2018.015

Stuhmer T, Anderson SA, Ekker M, Rubenstein JL. 2002. Ectopic expression of the Dlx genes induces glutamic acid decarboxylase and Dlx expression. *Development* **129:** 245–252.

Toresson H, Campbell K. 2001. A role for Gsh1 in the developing striatum and olfactory bulb of Gsh2 mutant mice. *Development* **128:** 4769–4780.

Toresson H, Potter SS, Campbell K. 2000. Genetic control of dorsal-ventral identity in the telencephalon: opposing roles for Pax6 and Gsh2. *Development* **127:** 4361–4371.

Uhl JD, Cook TA, Gebelein B. 2010. Comparing anterior and posterior Hox complex formation reveals guidelines for predicting *cis*-regulatory elements. *Dev Biol* **343:** 154–166. doi:10.1016/j.ydbio.2010.04.004

Uhl JD, Zandvakili A, Gebelein B. 2016. A Hox transcription factor collective binds a highly conserved distal-less *cis*-regulatory module to generate robust transcriptional outcomes. *PLoS Genet* **12:** e1005981. doi:10.1371/journal.pgen.1005981

Valerius MT, Li H, Stock JL, Weinstein M, Kaur S, Singh G, Potter SS. 1995. Gsh-1: a novel murine homeobox gene expressed in the central nervous system. *Dev Dyn* **203:** 337–351. doi:10.1002/aja.1002030306

Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV, McKinsey GL, Pattabiraman K, Silberberg SN, Blow MJ, et al. 2013. A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152:** 895–908. doi:10.1016/j.cell.2012.12.041

Von Ohlen TL, Moses C. 2009. Identification of Ind transcription activation and repression domains required for dorsoventral patterning of the CNS. *Mech Dev* **126:** 552–562. doi:10.1016/j.mod.2009.03.008

Von Ohlen T, Syu LJ, Mellerick DM. 2007a. Conserved properties of the *Drosophila* homeodomain protein, Ind. *Mech Dev* **124:** 925–934. doi:10.1016/j.mod.2007.08.001

Von Ohlen TL, Harvey C, Panda M. 2007b. Identification of an upstream regulatory element reveals a novel requirement for Ind activity in maintaining ind expression. *Mech Dev* **124:** 230–236. doi:10.1016/j.mod.2006.11.003

Von Ohlen T, Moses C, Poulson W. 2009. Ind represses *msh* expression in the intermediate column of the *Drosophila* neuroectoderm, through direct interaction with upstream regulatory DNA. *Dev Dyn* **238:** 2735–2744. doi:10.1002/dvdy.22096

Waclaw RR, Allen ZJ 2nd, Bell SM, Erdélyi F, Szabó G, Potter SS, Campbell K. 2006. The zinc finger transcription factor Sp8 regulates the generation and diversity of olfactory bulb interneurons. *Neuron* **49:** 503–516. doi:10.1016/j.neuron.2006.01.018

Waclaw RR, Wang B, Pei Z, Ehrman LA, Campbell K. 2009. Distinct temporal requirements for the homeobox gene Gsx2 in specifying striatal and olfactory bulb neuronal fates. *Neuron* **63:** 451–465. doi:10.1016/j.neuron.2009.07.015

Wang B, Waclaw RR, Allen ZJ 2nd, Guillemot F, Campbell K. 2009. Ascl1 is a required downstream effector of Gsx gene function in the embryonic mouse telencephalon. *Neural Dev* **4:** 5. doi:10.1186/1749-8104-4-5

Weiss JN. 1997. The Hill equation revisited: uses and misuses. *FASEB J* **11:** 835–841. doi:10.1096/fasebj.11.11.9285481

Weiss JB, Von Ohlen T, Mellerick DM, Dressler G, Doe CQ, Scott MP. 1998. Dorsoventral patterning in the *Drosophila* central nervous system: the intermediate neuroblasts defective homeobox gene specifies intermediate column identity. *Genes Dev* **12:** 3591–3602. doi:10.1101/gad.12.22.3591

Whitington T, Frith MC, Johnson J, Bailey TL. 2011. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res* **39:** e98. doi:10.1093/nar/gkr341

Winterbottom EF, Illes JC, Faas L, Isaacs HV. 2010. Conserved and novel roles for the Gsh2 transcription factor in primary neurogenesis. *Development* **137:** 2623–2631. doi:10.1242/dev.047159

Winterbottom EF, Ramsbottom SA, Isaacs HV. 2011. Gsx transcription factors repress iroquois gene expression. *Dev Dyn* **240:** 1422–1429. doi:10.1002/dvdy.22648

Yun K, Potter S, Rubenstein JL. 2001. Gsh2 and Pax6 play complementary roles in dorsoventral patterning of the mammalian telencephalon. *Development* **128:** 193–205.

Yun K, Garel S, Fischman S, Rubenstein JL. 2003. Patterning of the lateral ganglionic eminence by the Gsh1 and Gsh2 homeobox genes regulates striatal and olfactory bulb histogenesis and the growth of axons through the basal ganglia. *J Comp Neurol* **461:** 151–165. doi:10.1002/cne.10685

Zandvakili A, Gebelein B. 2016. Mechanisms of specificity for Hox factor activity. *J Dev Biol* **4:** 16. doi:10.3390/jdb4020016

Zandvakili A, Campbell I, Gutzwiller LM, Weirauch MT, Gebelein B. 2018. Degenerate Pax2 and senseless binding motifs improve detection of low-affinity sites required for enhancer specificity. *PLoS Genet* **14:** e1007289. doi:10.1371/journal.pgen.1007289

Zandvakili A, Uhl JD, Campbell I, Salomone J, Song YC, Gebelein B. 2019. The *cis*-regulatory logic underlying abdominal Hox-mediated repression versus activation of regulatory elements in Drosophila. *Dev Biol* **445:** 226–236. doi:10.1016/j.ydbio.2018.11.006

Zhang X, McGrath PS, Salomone J, Rahal M, McCauley HA, Schweitzer J, Kovall R, Gebelein B, Wells JM. 2019. A comprehensive structure-function study of Neurogenin3 disease-causing alleles during human pancreas and intestinal organoid development. *Dev Cell* **50:** 367–380.e7. doi:10.1016/j.devcel.2019.05.017

Zhou P, Gu F, Zhang L, Akerberg BN, Ma Q, Li K, He A, Lin Z, Stevens SM, Zhou B, et al. 2017. Mapping cell type-specific transcriptional enhancers using high affinity, lineage-specific Ep300 bioChIP-seq. *Elife* **6:** e22039. doi:10.7554/eLife.22039