Genomic context sensitizes regulatory elements to genetic disruption

Raquel Ordoñez[1,9], Weimin Zhang[1,9], Gwen Ellis[1,5], Yinan Zhu[1], Hannah J. Ashe[1,6], André M. Ribeiro-dos-Santos[1], Ran Brosh[1], Emily Huang[1,7], Megan S. Hogan[1,8], Jef D. Boeke[1,2,3], Matthew T. Maurano[1,4,10,*]

[1]Institute for Systems Genetics, NYU School of Medicine, New York, NY 10016, USA
[2]Department of Biochemistry Molecular Pharmacology, NYU School of Medicine, New York, NY 10016, USA
[3]Department of Biomedical Engineering, NYU Tandon School of Engineering, Brooklyn, NY 11201, USA
[4]Department of Pathology, NYU School of Medicine, New York, NY 10016, USA
[5]Present address: Department of Biology, University of Vermont, Burlington, VT 05405, USA
[6]Present address: School of Medicine, University of Maryland, Baltimore, MD 21201, USA
[7]Present address: Highmark Health, Pittsburgh, PA 15222, USA
[8]Present address: Neochromosome Inc., Long Island City, NY 11101, USA

[9]These authors contributed equally
[10]Lead contact

[*]Correspondence: maurano@nyu.edu

Keywords: genome writing, genetic engineering, gene regulation, synthetic regulatory genomics, enhancer selectivity, genomic regulatory architecture

SUMMARY

Enhancer function is frequently investigated piecemeal using truncated reporter assays or single deletion analysis. Thus it remains unclear to what extent enhancer function at native loci relies on surrounding genomic context. Using the Big-IN technology for targeted integration of large DNAs, we analyzed the regulatory architecture of the murine *Igf2*/*H19* locus, a paradigmatic model of enhancer selectivity. We assembled payloads containing a 157-kb functional *Igf2*/*H19* locus and engineered mutations to genetically direct CTCF occupancy at the imprinting control region (ICR) that switches the target gene of the *H19* enhancer cluster. Contrasting the activity of payloads delivered to the endogenous locus or to a safe harbor locus (*Hprt*) revealed that the *Igf2*/*H19* locus includes additional, previously unknown long-range regulatory elements. Exchanging components of the *Igf2*/*H19* locus with the well-studied *Sox2* locus showed that the *H19* enhancer cluster functioned poorly out of context, and required its native surroundings to activate *Sox2* expression. Conversely, the *Sox2* locus control region (LCR) could activate both *Igf2* and *H19* outside its native context, but its activity was only partially modulated by CTCF occupancy at the ICR. Analysis of regulatory DNA actuation across different cell types revealed that, while the *H19* enhancers are tightly coordinated within their native locus, the *Sox2* LCR acts more independently. We show that these enhancer clusters typify broader classes of loci genome-wide. Our results show that unexpected dependencies may influence even the most studied functional elements, and our synthetic regulatory genomics approach permits large-scale manipulation of complete loci to investigate the relationship between locus architecture and function.

INTRODUCTION

Genomic regulatory elements are often treated as discrete peaks, but they operate within complex neighborhoods that vary in sequence content, and in the density and position of surrounding DNase I hypersensitive sites (DHSs) and genes. Cross-tissue DNA accessibility patterns frequently resemble the activity of their target genes (Thurman et al. 2012), suggesting a role for genomic context in DHS function. Since distinct expression programs often co-occupy nearby genomic space, genomic regulatory architecture must provide for a degree of selectivity. Selective expression of certain genes might be facilitated through promoter-enhancer compatibility (Bergman et al. 2022; Martinez-Ara et al. 2022), competition for regulatory elements, interposing regulatory elements or genomic distance (Peterson and Stamatoyannopoulos 1993; Zuin et al. 2022), or boundary elements that confer compartmentalization and large-scale domain organization (Symmons et al. 2014; Ringel et al. 2022). Questions of context and sufficiency have been studied in a genomic context at the level of individual DHSs or transcription factor (TF) recognition sequences (Bolt et al. 2022; Blayney et al. 2022; Brosh et al. 2023). Yet enhancer elements are typically investigated through limited genome engineering at their endogenous loci, and the lack of methods for genome editing regions upwards of 10 kb has resulted in a limited understanding of locus architecture and genomic context at larger length scales.

Studies using simpler reporter elements have shown that genomic context modulates reporter activity and acts at ranges upwards of 100 kb (Maricque et al. 2019; Hong and Cohen 2022; Ribeiro-Dos-Santos et al. 2022). Certain composite regulatory elements such as locus control regions (LCR) confer position-independent expression, and super-enhancers are argued to have properties beyond those of their constituent enhancers (Li et al. 2002; Moorthy et al. 2017; Blobel et al. 2021). Autonomy might result from inclusion of multiple regulatory elements to provide resilience, specific genomic features such as insulation or domains, or simply additional intervening distance from other surrounding regulatory elements. Thus it is possible that complex regulatory elements have evolved to permit more precise control over these features. However, the exact sequence elements required to confer robust locus activity remain unclear.

The canonical *Igf2*/*H19* imprinted locus includes multiple complex regulatory elements and is a model for enhancer selectivity (Arney 2003). The insulin-like growth factor 2 (*Igf2*) gene is expressed from the paternal allele while the non-coding *H19* gene is maternally expressed. The

"boundary model" posits that an imprinting control region (ICR) located between the two genes directs the activity of a shared enhancer cluster (**Figure 1a**) (Bell and Felsenfeld 2000; Hark et al. 2000). This ICR contains a tandem array of CTCF sites which are methylated and unoccupied by CTCF on the paternal allele, allowing the enhancers to interact with and activate the more distal *Igf2* gene. The maternal allele is unmethylated and occupied by CTCF which acts as an insulator that directs the enhancer cluster to activate the proximal *H19* instead of *Igf2* (**Figure 1b**). A transgenic YAC containing a 130-kb sequence from the *Igf2*/*H19* locus has been shown sufficient to recapitulate this pattern of imprinted gene expression when integrated randomly (Ainscough et al. 1997). Therefore, the *Igf2*/*H19* bimodal regulation offers a unique paradigm to understand how locus architecture and genomic context might influence the activity of the individual regulatory components in a specific locus and their interactions with their regulatory landscape.

Here we take advantage of our Big-IN technology (Brosh et al. 2021) for repeated delivery of large DNA sequences to develop a synthetic regulatory genomics approach to quantitatively dissect enhancer cluster context sensitivity. Using this approach to deliver payloads up to 158 kb in length, we identified a new long-range distal enhancer cluster at the *Igf2*/*H19* locus that, together with the endogenous *H19* enhancers, directs *Igf2* and *H19* expression in an ICR-dependent manner. Placement of the *H19* enhancer cluster replacing the well-described *Sox2* LCR shows that full function of the *H19* enhancer cluster relies on additional elements at its endogenous locus. In contrast, the *Sox2* LCR functions outside its native context to drive robust expression of both *Igf2* and *H19*. However, the *Sox2* LCR responds only partially to the ICR, as it is able to maintain high *H19* expression levels regardless of CTCF occupancy. We show that the independence of the *Sox2* LCR and the context dependence of the *H19* enhancers are reflected in cross-cell type patterns of regulatory DNA actuation which manifest broadly at many other loci genome-wide, providing clues to the regulatory architecture of these and other important developmentally regulated genes.

RESULTS

A synthetic *Igf2/H19* locus as a model for dissection of enhancer selectivity

To quantitatively dissect the sensitivity of enhancer elements to their broader genomic context, we developed a synthetic regulatory genomics approach to manipulate the *Igf2*/*H19* locus in

cultured cells. We employed our Big-IN technology (Brosh et al. 2021) for repeated delivery (**Figure 1c**) to mouse embryonic stem cells (mESC). To facilitate allele-specific engineering and readout, we employed a C57BL/6J (B6, maternal) × CAST/EiJ (CAST, paternal) mESC cell line derived from a *M. musculus* and *M. musculus castaneus* F1 hybrid mouse, which harbors heterozygous point variants every 140 bp on average throughout its genomic sequence (Keane et al. 2011; Eckersley-Maslin et al. 2014). We first installed a Big-IN landing pad (LP) replacing the *Igf2/H19* locus on the B6 allele (**Figure 1d**). After comprehensive quality control using PCR genotyping and targeted capture sequencing (Capture-seq), we identified clones with single-copy, on-target LP integration, named *LP-Igf2/H19* mESCs (**Figure S1**).

We then assembled in yeast a baseline payload containing a complete 157-kb *Igf2/H19* locus into a Big-IN delivery vector, starting from a B6 mouse-derived bacterial artificial chromosome (BAC). From this baseline payload, we derived a set of further engineered derivatives (**Figure S2**, **Table S1**). Payloads were sequence-verified (**Table S2**) and delivered to *LP-Igf2/H19* mESCs using Big-IN (**Figure 1d**). Multiple mESC clones were isolated for each payload and comprehensively verified for single-copy targeted payload integration using PCR genotyping and Capture-seq (**Table S3**). As both *Igf2* and *H19* are transcriptionally inactive in mESCs, we employed a mesendodermal differentiation protocol (Thomson et al. 2011), resulting in robust expression of both *Igf2* and *H19* (**Figure S3**) and increased DNA accessibility at the enhancer cluster located downstream of *H19* (**Figure 1a**). We analyzed several independent clones for each delivered payload to account for potential variability during delivery and differentiation.

While stable parent-of-origin specific methylation is established at the *Igf2*/*H19* ICR during development, the methylation pattern maintained in cultured mESCs is more labile (Swanzey et al. 2020). In addition, controlling methylation status of transfected DNA would require additional steps (Schübeler et al. 2000; Butz et al. 2022). Thus we developed a genetic approach to reliably direct CTCF occupancy at the ICR (**Figure 1d**). We engineered a methylation-insensitive ICR (MI-ICR) where CpGs were mutated in all 6 CTCF recognition sequences without reducing predicted CTCF binding energy (**Figure S4a-b**). Similarly, we engineered a ΔCTCF-ICR where all 6 CTCF recognition sequences were surgically deleted (**Figure S4a**). We profiled both parental and engineered mESCs using CUT&RUN to verify the expected CTCF occupancy. In parental cells, we detected biallelic CTCF occupancy throughout the locus, except at the ICR, where only the maternal B6 allele was occupied. Delivery of a wild-type ICR or a MI-ICR payload resulted in

specific CTCF occupancy on the engineered allele, while the un-engineered CAST allele remained unoccupied. In contrast, CTCF occupancy was absent in ΔCTCF-ICR payloads, demonstrating that our synthetic ICR sequences can genetically direct CTCF occupancy independently of parental origin of the engineered chromosome (**Figure S4c**).

To enable allele-specific expression analysis of the engineered allele, we developed an allele-specific real-time quantitative reverse transcription PCR (qRT-PCR) strategy (**Figure 1d**). The *H19* transcript included two point variants distinguishing the B6 and CAST alleles, and we designed four synonymous point variants in the *Igf2* transcript sequence (**Figure S5, Table S4**). We scaled *Igf2* and *H19* expression values between 0 (deletion of the engineered allele, Δ*Igf2/H19*) and 1 (normalizing *H19* to the "Full-length" payload with constitutive CTCF occupancy, and *Igf2* to the "Full-length" payload with abolished CTCF occupancy). This approach allowed us to rewrite the *Igf2*/*H19* locus for functional dissection and permitted a precise transcriptional readout of the engineered allele.

Allele-specific expression analysis of the "Full-length" payloads demonstrated reciprocal *Igf2* and *H19* expression dependent upon CTCF occupancy at the ICR, with higher *H19* expression in the presence of CTCF, and higher *Igf2* expression in absence of CTCF (**Figure 1e**). Thus, our experimental system recapitulates the boundary model of the *Igf2/H19* locus (Bell and Felsenfeld 2000; Hark et al. 2000) (**Figure 1b**). Surprisingly, we observed that the "ΔH19Enh" payloads, while they completely lack the *H19* enhancer cluster, still retained modest expression of both *Igf2* and *H19* (**Figure 1e**), suggesting that the canonical *H19* enhancers are not the only determinant of gene expression at this locus.

Dissection of the *Igf2/H19* regulatory architecture reveals a novel distal enhancer cluster

We thus investigated whether additional regulatory elements outside the payload boundaries play a role in *Igf2/H19* regulation (**Figure 2a**). We first engineered payloads with an inverted *Igf2*/*H19* locus and delivered them to LP-*Igf2/H19* mESCs (**Figure 2b**). These "Full-length inverted" payloads showed strongly reduced *H19* expression, while "ΔH19Enh inverted" payloads showed completely ablated *H19* expression. These results suggest that *H19* expression is highly sensitive to regulatory elements beyond the payload boundary. In contrast, inversion had little effect on *Igf2* levels in the payloads with abolished CTCF occupancy. In payloads with constitutive CTCF

occupancy, *Igf2* expression even increased slightly. As inverted payloads showed increased expression of *Igf2* and reduced expression of *H19*, our results imply the presence of regulatory elements downstream of the canonical 157-kb locus which cooperate with known regulatory elements and exhibit differing influences on *H19* and *Igf2*.

To further dissect the reliance of *Igf2/H19* expression on their broader genomic environment, we chose to relocate *Igf2/H19* payloads to a completely different genomic location. The *Hprt* locus is often used as a safe harbor for ectopic transgene expression in a permissive genomic context across many cell types (Papapetrou et al. 2011). We thus started from a LP-*Hprt* mESC line, wherein a Big-IN LP replaces the *Hprt* locus in chromosome X in a male mESC hybrid cell line (Camellato et al. 2022). We additionally engineered a CRISPR/Cas9-mediated deletion of the B6 allele of the full *Igf2/H19* locus in these mESC (LP-*Hprt*-Δ*Igf2/H19* cells) to maintain a single endogenous copy of *Igf2* and *H19* in addition to the delivered payload (**Figure S6a**). When the "Full-length" payloads were delivered ectopically to the *Hprt* locus, not only were *Igf2* and *H19* expressed, but also the ICR-regulated gene expression pattern was correctly recapitulated. However, both *Igf2* and *H19* expression was reduced by 50% as compared to the endogenous locus. Strikingly, *Igf2* and *H19* expression from the *Hprt* locus was completely ablated in "ΔH19Enh at *Hprt*" payloads lacking the *H19* enhancers (**Figure 2b**). This suggests that *Igf2/H19* expression can be deconstructed into two parts: (i) a well-known ICR-regulated contribution of the *H19* enhancers; and (ii) a newly-identified independent contribution of the surrounding genomic environment.

Based on these results, we examined several enhancer clusters downstream of *Igf2* and *H19* showing DNase-seq activity in mesendodermal cells. Three of these regions have been shown to form an allelic sub-TAD structure with *Igf2/H19* genes by 4C-seq and DNA FISH and were named HIDAD (*H19*/*Igf2* Distal Anchor Domain), *Syt8*, and *Lsp1* (**Figure 2a**) (Llères et al. 2019). We engineered three derivative LP mESC lines, each with a CRISPR/Cas9-mediated deletion of one of these regions. We then delivered *Igf2/H19* payloads lacking the *H19* enhancer cluster to each of these LP cell lines. "ΔH19Enh" payloads delivered to ΔHIDAD and ΔSyt8 LP mESC lines ("ΔHIDAD & ΔH19Enh" and "ΔSyt8 & ΔH19Enh" respectively) showed no difference in expression relative to their delivery to LP-*Igf2/H19* mESCs, while delivery to ΔLsp1 LP mESCs ("ΔLsp1 & ΔH19Enh") showed negligible *H19* expression and reduced *Igf2* expression (**Figure 2b**). This implicates the *Lsp1* enhancer cluster in the regulation of *H19*, and to a lower

extent of *Igf2.* We observed a synergistic cooperation between the *H19* enhancers and the surrounding genomic context, with the expression of both genes being greater than the sum of the individual activities (**Figure 2c**). These results highlight the contribution of the surrounding genomic environment to the *Igf2/H19* locus regulation, and demonstrate the power of delivery at defined ectopic sites to establish sufficiency.

Enhancer sufficiency and robustness can be disrupted in different genomic contexts

We wondered whether the *H19* enhancer cluster demonstrated context sensitivity similar to that of the entire *Igf2/H19* locus. The *H19* enhancer cluster is a composite element composed of 11 DHSs covering 55 kb. We contrasted its function with the *Sox2* locus control region (LCR) which, similar to the H19 enhancers, consists of 10 DHSs across 41 kb located also approximately 100 kb away from its target gene (Zhou et al. 2014; Li et al. 2014; Brosh et al. 2023) (**Figure 3a**). In contrast to the *H19* enhancers, the *Sox2* LCR displays strong activity in mESC, which is retained into early mesendoderm differentiation. We therefore exchanged enhancer clusters among these two loci to establish the extent to which their function was tied to other elements of their surrounding locus.

We employed LP-*Sox2* and LP-LCR mESCs harboring Big-IN landing pads replacing the full 143-kb *Sox2* locus or just the *Sox2* LCR, respectively (Brosh et al. 2023). We quantified the expression of the engineered *Sox2* allele on a scale of 0 (ΔSox2) to 1 ("*Sox2* & Sox2LCR") using qRT-PCR (**Figure 3b**). We assembled payloads comprising just the *H19* enhancers in both orientations (**Figure 3b**). When the *H19* enhancers were delivered to LP-LCR mESCs replacing the *Sox2* LCR, *Sox2* expression remained very low even upon differentiation to mesendodermal cells (**Figure 3b**, left panel, "*Sox2* & H19Enh" and "*Sox2* & H19Enh (inv)"). To test whether the *H19* enhancers relied upon other regulatory elements within the 157-kb payload boundaries, we edited the "Full-length" *Igf2/H19* payload in yeast to replace the *Igf2* gene and its promoter with the *Sox2* gene and promoter (**Figure S7**), and delivered it to LP-*Sox2* mESCs replacing the whole *Sox2* locus (**Figure 3b**, "H19Enh & *H19* & ICR (CTCF) & *Sox2*" and "H19Enh & *H19* & ICR (no CTCF) & *Sox2*"). These too showed very low *Sox2* expression (**Figure 3b**, left panel), suggesting that the *H19* enhancers were either incompatible with the *Sox2* promoter or lacked key elements from outside the *Igf2*/*H19* "Full-length" payloads.

To investigate whether the *H19* enhancers relied on genomic elements outside the boundaries of our 157-kb payload, we delivered the same payloads at the *Igf2/H19* locus (**Figure 3b**, right panel). Starting from LP-*Igf2/H19* mESCs, we further engineered a CRISPR/Cas9-mediated deletion of the endogenous B6 *Sox2* allele (LP-*Igf2/H19*-Δ*Sox2* cells, **Figure S6a**) to ensure only two copies of *Sox2* were present after payload delivery. A payload with the full *Sox2* locus including its native LCR ("*Sox2* & Sox2LCR") resulted in a 1.5-fold increase in *Sox2* expression when delivered to the *Igf2/H19* locus, relative to its endogenous location. In comparison, both payloads wherein the *H19* enhancers replaced the *Sox2* LCR ("*Sox2* & H19Enh" and "*Sox2* & H19Enh (inv)", right panel) showed a 2.5-fold increase in *Sox2* expression when delivered to the *Igf2/H19* locus compared to at the *Sox2* locus. Notably, the intervening sequence between the *H19* enhancers and *Igf2* was not required to activate *Sox2*, as the *H19* enhancers drove similar *Sox2* expression in payloads derived from the *Sox2* locus ("*Sox2* & H19Enh" and "*Sox2* & H19Enh (inv)") and from the *Igf2/H19* locus (H19Enh & *H19* & ICR (CTCF/no CTCF) & *Sox2*"). Finally, the *H19* enhancers remained sensitive to CTCF occupancy at the intervening ICR, activating the *Sox2* promoter only in absence of CTCF (**Figure 3b**).

We next performed the cognate swap, replacing the *H19* enhancers with the 41-kb *Sox2* LCR in both orientations, yielding "Sox2LCR & *H19* & ICR & *Igf2*" and "Sox2LCR (inv) & *H19* & ICR & *Igf2*" payloads respectively. We profiled *Igf2* and *H19* expression in mESCs and scaled expression to that of the "Full-length" payloads in mesendodermal cells for comparison. This showed that the *Sox2* LCR robustly boosted the expression of both *Igf2* and *H19* in undifferentiated cells relative to their native state (**Figure 3c, Figure S3**). The convergent orientation of the *Sox2* LCR (i.e. with CTCF sites oriented towards the target genes) yielded approximately 10-fold increased expression of both *Igf2* and *H19* genes relative to the *Sox2* LCR in the divergent orientation (**Figure 3c**), consistent with observations at the endogenous *Sox2* locus (de Wit et al. 2015; Brosh et al. 2023). In either orientation, the *Sox2* LCR activated *Igf2* almost exclusively in the absence of CTCF binding at the ICR, matching the behavior of the endogenous *H19* enhancer cluster (**Figure 3c**). But, in contrast to the *H19* enhancers, the *Sox2* LCR upregulated *H19* expression even in the absence of CTCF occupancy. Thus, the *Sox2* LCR did not fully recapitulate the activity of the endogenous regulation of the *H19* enhancer cluster. Understanding this promiscuity in more detail may provide important clues to the function of the endogenous ICR and the sequence determinants of enhancer selectivity.

Characteristic patterns of coordinate DHS actuation across cell and tissue types

Regulatory elements operate within complex neighborhoods, including extended regions of coordinated transcriptional programs as well as examples of genes with distinct expression patterns relative to their neighbors (Pachano et al. 2022). The contrasting function of the *Sox2* LCR and *H19* enhancer cluster in ectopic contexts suggests they may have different relationships with their native genomic context. We analyzed cell-type selective accessibility at the *Hprt*, *Sox2*, and *Igf2*/*H19* loci using DNase-seq data from 45 representative cell and tissue types from the mouse ENCODE project (Vierstra et al. 2014). *Hprt* showed accessibility in all cell and tissue types, befitting its status as a housekeeping gene (**Figure 4a**). In contrast, the *Sox2* LCR was mESC-specific (Zhou et al. 2014), while the *Sox2*-proximal DHSs and those at *Igf2*/*H19* showed broader tissue-selective accessibility profiles. We further observed stereotypical differences in the pattern of coordinated DHS activity across tissue types surrounding each of the delivery sites (**Figure 4b)**. The *Sox2* LCR was active specifically in mESC, while the rest of the locus harbored a series of more broadly active enhancers. This is consistent with the observation that the LCR is required for *Sox2* expression in mESC but becomes silent and dispensable upon differentiation to neuroectoderm (Zhou et al. 2014). A broader pattern of coordinated activity extended across the *Igf2* and *H19* genes and partially into *Lsp1*. In contrast, the *Hprt* locus was sparsely populated by mostly constitutive DHSs, most prominently those for *Hprt* itself and the nearby gene *Phf6*, which has a similar expression pattern. The loci also differed in terms of gene and DHS density: *Sox2* lies within a gene desert, *Igf2* and *H19* lie in a region rich with other genes and DHSs, and *Hprt* is surrounded by an intermediate number of genes (**Table 1**).

Regulatory interaction can be reflected in correlated actuation of regulatory DNA across cell types (Thurman et al. 2012). We thus computed correlations in accessibility across 45 cell and tissue types between the *Hprt*, *Sox2,* and *Igf2*/*H19* loci and their surrounding 1 Mb (**Figure 4c, Figure S8**). While the *Hprt* transcription start site (TSS) showed no correlation with surrounding DHSs, the *Sox2* LCR DHSs showed correlation only amongst themselves. Finally, the *Sox2* TSS as well as the *Igf2* TSS, *H19* TSS, and *H19* enhancers showed tight correlation with additional surrounding regions. We defined these three patterns as uncorrelated, independent, and coordinated activity, respectively, and analyzed their representation more broadly across the genome (**Figure 4d-e**). A random sample of genomic DHSs exhibited uncorrelated activity (**Figure 4f**). Focusing on gene promoters (±1 kb from TSS) showed stronger correlation with

surrounding regions. Finally, mESC super-enhancers (Whyte et al. 2013) exhibited the highest rate of independent and coordinated activity, with only 27% being uncorrelated. These cross-cell type actuation patterns may provide a key feature of genomic context at tissue-selective genes and may distinguish “independent” from “context-dependent" enhancer clusters.

DISCUSSION

Our work shows that composite regulatory elements such as enhancer clusters recapitulate their characteristic function with degraded fidelity relative to their full native context. The *Sox2* LCR at first glance manifests the autonomy and independence implied by the classical definition of an enhancer (Banerji et al. 1981), in that it drives robust expression at the *Igf2*/*H19* locus. Yet its influence on *H19* expression differs from the endogenous enhancers in that the *Sox2* LCR partially overrides the ICR, upregulating *H19* even in the absence of CTCF occupancy. The *H19* enhancers and ICR both harbor multiple CTCF sites convergent on *H19*, and the *Sox2* LCR harbors CTCF sites pointing towards *Sox2*. Inverting the ICR has been shown to lower the range of expression between the paternal and maternal alleles (Matsuzaki et al. 2021), and we show that inversion of the *Sox2* LCR in the context of the ICR yields reduced expression, especially for *H19* (**Figure 3c**). This suggests a CTCF-mediated link between the ICR and the enhancer cluster it modulates, although we note that the link between LCR orientation and expression at *Sox2* is not due solely to CTCF (de Wit et al. 2015; Brosh et al. 2023). It is possible that CTCF sites underpin the robustness of the *Sox2* LCR across genomic positions (Ribeiro-Dos-Santos et al. 2022), but this does not explain why the CTCF sites in the *H19* enhancers lack similar function. More work is needed to further dissect whether there are specific sequence determinants within the *Sox2* LCR (i.e. full DHSs or TF recognition sequences) responsible for such distinctive activity, or if such discrepancy might be related to enhancer strength.

In contrast, the *H19* enhancers are highly subordinated to their enclosing genomic context, and show reduced activity upon any perturbation in the surrounding architecture, such as *Lsp1* deletion, delivery to an ectopic locus, or linkage to a heterologous promoter. While proper function of the *H19* enhancers is tied to their native locus, there are necessary element(s) for this synergy outside the 157-kb *Igf2*/*H19* locus. It is possible that the *Lsp1* enhancers themselves underpin the synergy with the *H19* enhancers, or that other elements outside the 143-kb *Sox2* locus suppress the activity of ectopic enhancer clusters.

Our analysis of cross-cell type correlation in regulatory DNA actuation suggests a broad influence of genomic context in potentiating enhancer cluster function. While there is little evidence for strict genomic regulons in complex organisms, domains of coordinated expression have been observed in yeast (Cohen et al. 2000) and fly (Rubin and Green 2013). In mammalian genomes, housekeeping and disease-associated genes are unevenly distributed genome-wide (Lercher et al. 2002; Muro et al. 2019), and topological associated domains (TADs) correlate with tissue-selective expression patterns (Symmons et al. 2014). But the strength, generalizability, and sequence determinants of these architectural context effects have been difficult to clarify across mammalian genomes. The role of context we uncover suggests that the tissue-specific accessibility and expression patterns observed at key developmental loci (Bolt and Duboule 2020) might be required to reinforce the intrinsic propensities of regulatory elements therein. This is consistent with the β-globin genes, whose promoters add developmental specificity to the activation provided by the LCR (Bender et al. 2000). The large gene deserts surrounding tissue-selective genes (e.g. *Sox2* and *Hbb*) may be required to insulate unrelated surrounding genes from this context effect, and might derive mechanistically from a reliance on cohesin for longer-range contacts for tissue-specific genes (Kane et al. 2022; Rinzema et al. 2022) or models invoking diffusion of regulatory factors (Karr et al. 2022). We note that while enhancers showing correlated cross-cell type accessibility patterns have appeared as redundant in deletion analyses (Hong et al. 2008; Frankel et al. 2010; Osterwalder et al. 2018; Hörnblad et al. 2021), our sufficiency analysis suggests that these broader activity patterns can be required for proper function.

In this study, we showcase the potential of synthetic regulatory genomics to characterize and interrelate multiple regulatory elements, to interpret complex loci, and to gain new insights into transcriptional regulation. This approach can be easily applied to characterize many other genomic regions whose size, complexity, and/or allele-specific regulation challenge the limited scope of genome editing tools currently in wide use. We expect that dissecting specific sequence elements responsible for novel activities of synthetic combinations of regulatory elements will lead to generalizable models of locus architecture and gene regulation.

ACKNOWLEDGEMENTS

We thank Brendan Camellato for the gift of pDONOR~Neo~EGFP, and Raven D. Luther for technical assistance with sequencing. This work was partially funded by National Institutes of Health (NIH) grants RM1HG009491 (to J.D.B.) and R35GM119703 (to M.T.M.).

AUTHOR CONTRIBUTIONS

R.O., M.S.H., and M.T.M. designed experiments. R.O., G.E., and R.B. performed experiments. W.Z. and Y.Z. assembled DNA payloads. G.E., H.J.A., and E.H. performed capture and sequencing. R.O. and A.M.R.-d.-S. performed computational analyses. R.O. and M.T.M. wrote the manuscript. All authors edited the manuscript. M.T.M. and J.D.B. supervised work.

DECLARATION OF INTERESTS

R.B., J.D.B. and M.T.M are listed as inventors on a patent application describing Big-IN. J.D.B. is a founder and Director of CDI Labs, Inc., a founder of and consultant to Neochromosome, Inc, a founder, SAB member of, and consultant to ReOpen Diagnostics, LLC, and serves or served on the Scientific Advisory Board of the following: Logomix, Inc., Modern Meadow, Inc., Rome Therapeutics, Inc., Sample6, Inc., Sangamo, Inc., Tessera Therapeutics, Inc. and the Wyss Institute.

FIGURES

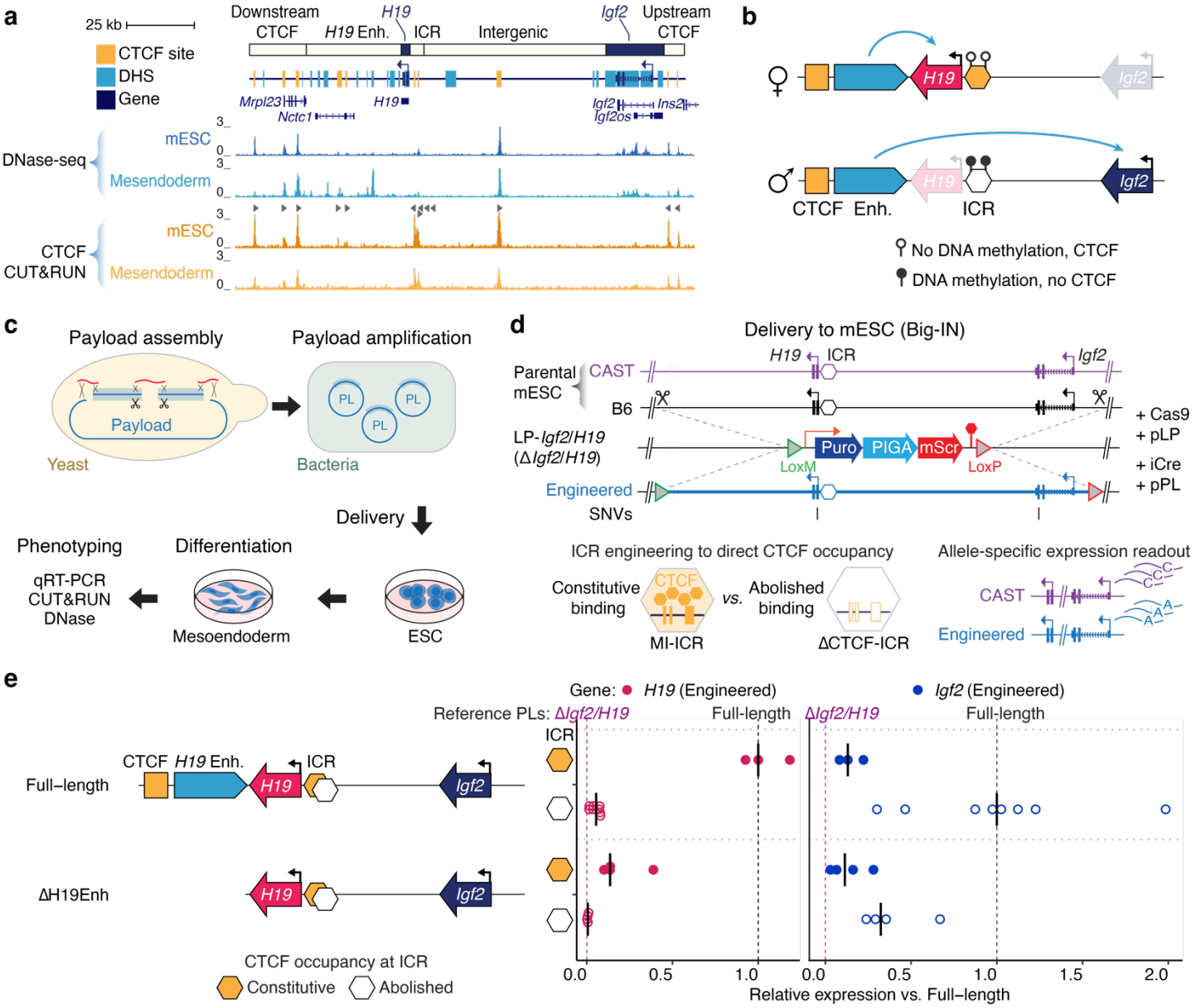
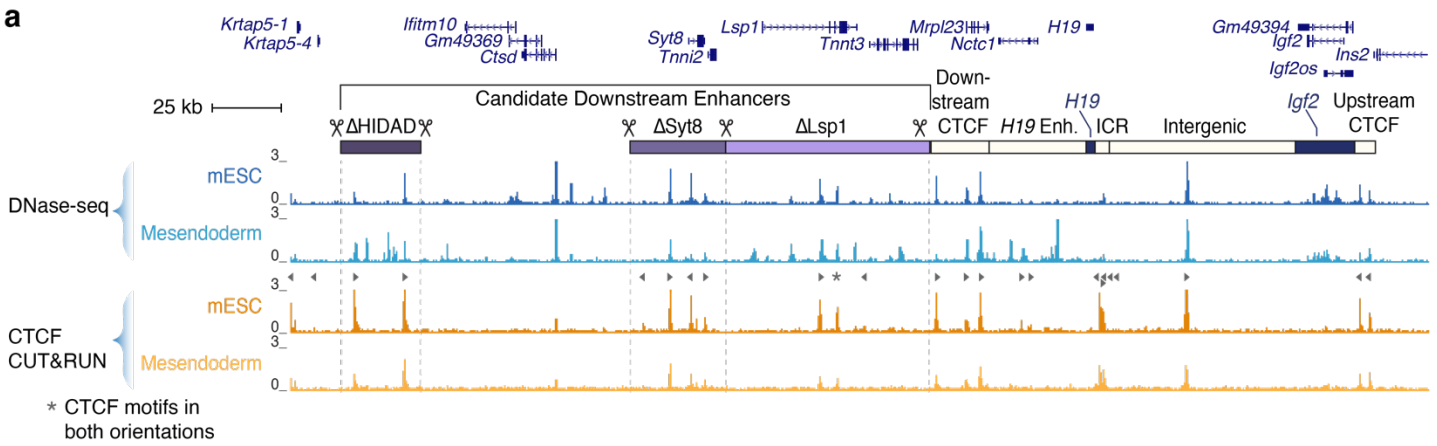
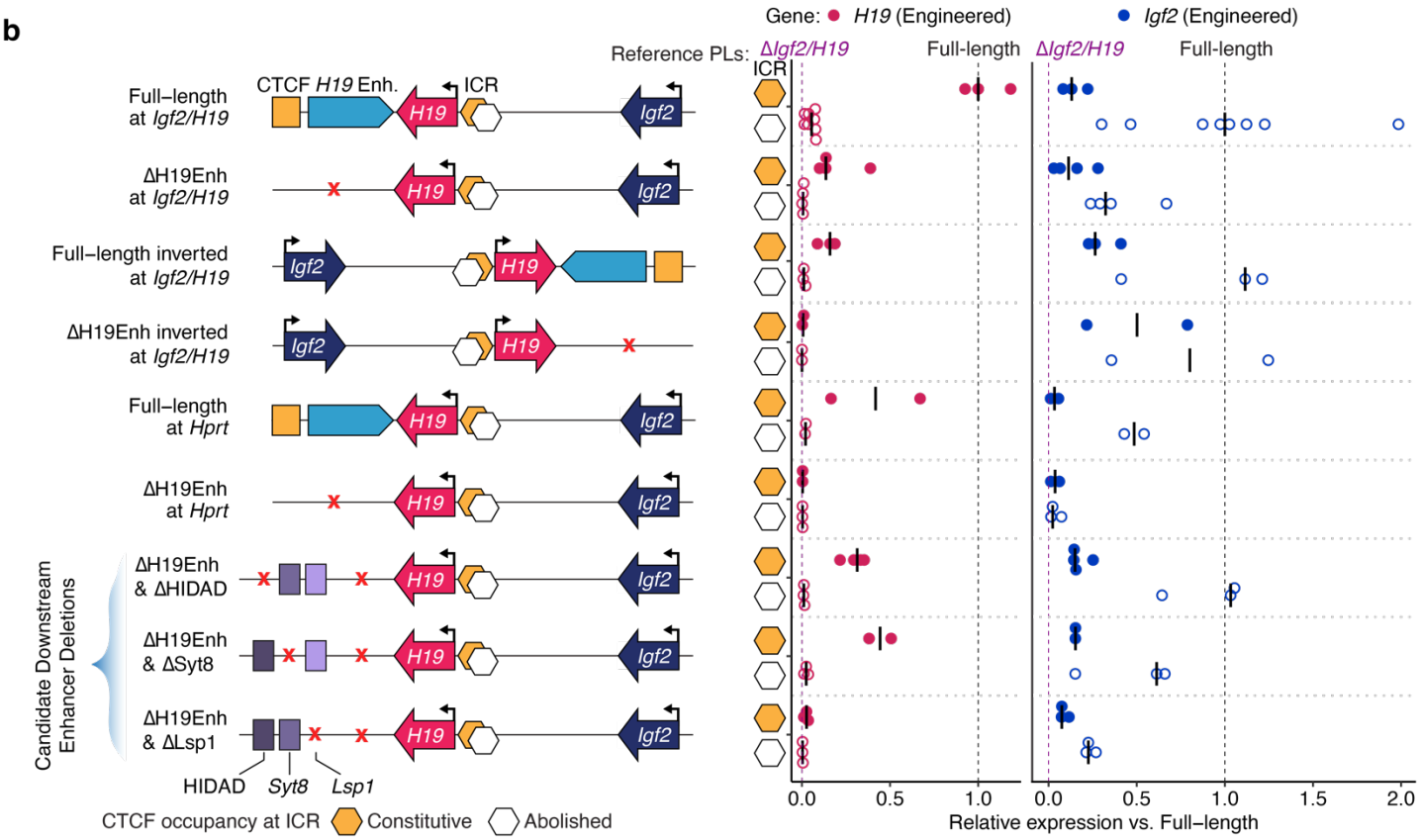


Figure 1. Engineering the *Igf2/H19* locus

a. *Igf2/H19* locus in mESC and differentiated mesendodermal cells. DNase-seq tracks show activation of the enhancer cluster upon differentiation. Triangles indicate CTCF recognition sequence orientation. **b**. Barrier model where CTCF occupancy at the imprinting control region (ICR) directs activity of the endogenous enhancer cluster towards *H19* or *Igf2* (Bell and Felsenfeld 2000; Hark et al. 2000). **c**. Schematic of engineering and delivery strategy. Payloads were assembled in yeast, amplified in bacteria, and delivered to mESCs using Big-IN. Multiple independent clones were systematically verified for correct genome engineering and differentiated into mesendodermal cells for further phenotyping. **d**. Schematic of *Igf2/H19* locus engineering using Big-IN. CpG mutations (methylation-insensitive ICR [MI-ICR]) or CTCF deletions (ΔCTCF-ICR) permit genetic control of CTCF occupancy. B6 or engineered variants in *H19* and *Igf2* genes permit allele-specific expression analysis. **e**. Expression analysis of the engineered *Igf2/H19* locus. Schemes on the left represent each delivered payload. At right, each point represents the expression of the engineered *H19* or *Igf2* gene in an independent mESC clone, measured using allele-specific qRT-PCR. Bars indicate median. Expression was normalized to *Gapdh* and then scaled between 0 (Δ*Igf2/H19* locus) and 1 ("Full-length" rescue payloads), normalizing *H19* to the "Full-length" payload with constitutive CTCF occupancy, and *Igf2* to the "Full-length" payload with abolished CTCF occupancy (dashed lines). Payloads with both constitutive CTCF occupancy (orange hexagon) or abolished CTCF occupancy (empty hexagon) at the ICR were assessed.


a
Krtap5-1
Krtap5-4
Ifitm10
Gm49369
Ctsd
Syt8
Tnni2
Lsp1
Tnnt3
Mrpl23
Nctc1
H19
Gm49394
Igf2
Ins2
Igf2os
25 kb
Candidate Downstream Enhancers
ΔHIDAD
ΔSyt8
ΔLsp1
Down-stream CTCF
H19 Enh.
H19 ICR
Intergenic
Igf2
Upstream CTCF
DNase-seq
mESC
Mesendoderm
CTCF CUT&RUN
mESC
Mesendoderm
* CTCF motifs in both orientations


b
Gene: H19 (Engineered)
Igf2 (Engineered)
Reference PLs: ΔIgf2/H19
Full-length
ΔIgf2/H19
Full-length
CTCF H19 Enh.
ICR
Full–length at Igf2/H19
ΔH19Enh at Igf2/H19
Full–length inverted at Igf2/H19
ΔH19Enh inverted at Igf2/H19
Full–length at Hprt
ΔH19Enh at Hprt
ΔH19Enh & ΔHIDAD
ΔH19Enh & ΔSyt8
ΔH19Enh & ΔLsp1
Candidate Downstream Enhancer Deletions
HIDAD
Syt8
Lsp1
CTCF occupancy at ICR
Constitutive
Abolished
0.0
0.5
1.0
1.5
2.0
Relative expression vs. Full–length

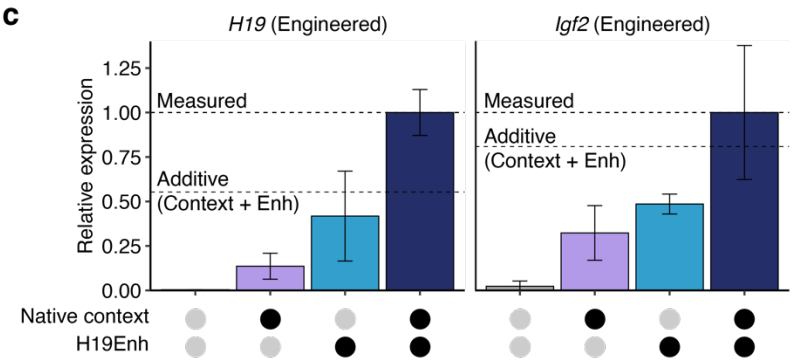

c
H19 (Engineered)
Igf2 (Engineered)
Relative expression
1.25
1.00
0.75
0.50
0.25
0.00
Measured
Additive
(Context + Enh)
Native context
H19Enh

Figure 2. Identification of long-range regulatory elements at the *Igf2/H19* locus

a. The *Igf2/H19* locus and extended downstream region showing DNase-seq and CTCF CUT&RUN in mESC and differentiated mesendodermal cells. Triangles indicate CTCF recognition sequence orientation. "Full-length" and "ΔH19Enh at *Igf2/H19*" are repeated from **Fig. 1e** for reference. **b**. Expression analysis of *Igf2/H19* payloads when delivered to the *Igf2/H19* locus, to a genomic safe harbor locus (*Hprt*), or to *Igf2/H19* locus in conjunction with distal candidate enhancers deletions. Each point represents the expression of the engineered *H19* or *Igf2* in an independent mESC clone using allele-specific qRT-PCR with bars indicating the median. Expression was normalized to *Gapdh* and then scaled between 0 (Δ*Igf2/H19* locus) and 1 ("Full-length" rescue payloads), normalizing *H19* to the "Full-length" payload with constitutive CTCF occupancy, and *Igf2* to the "Full-length" payload with abolished CTCF occupancy, which are also represented as dashed lines. Payloads with both constitutive CTCF occupancy (orange hexagon) or abolished CTCF occupancy (empty hexagon) at the ICR were assessed. **c**. Contribution of the *H19* enhancers and the native context in the expression levels of *H19* and *Igf2* genes. The presence (black dots) and absence (gray dots) of each element is indicated. Bars indicate median and error bars represent the interquartile range. Dashed horizontal lines indicate the expected transcriptional output of adding both the enhancers and the surrounding context (bottom), versus the measured transcriptional output (top). Source data are from (**b**).

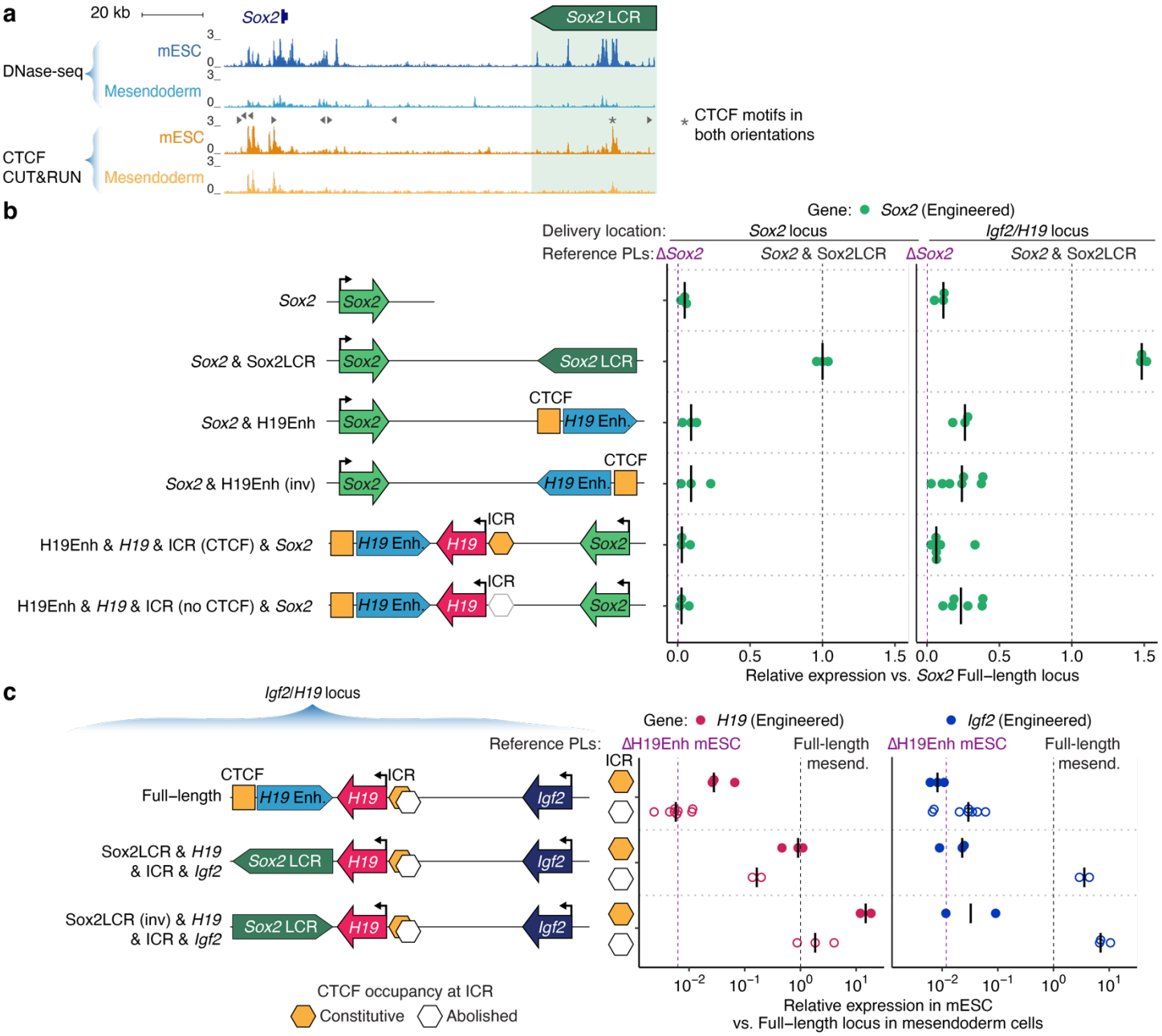


Figure 3. Enhancer interchangeability at the *Igf2/H19* and *Sox2* loci

a. *Sox2* and locus control region (LCR) showing DNase-seq and CTCF CUT&RUN in mESC and mesendodermal cells. Triangles indicate CTCF recognition sequence orientation. **b**. Expression analysis in differentiated mesendodermal cells harboring payloads including the wild-type *Sox2* locus, the *Sox2* locus with the LCR replaced by the *H19* enhancer cluster, and the *Igf2/H19* locus with the *Igf2* gene and promoter replaced by the *Sox2* gene and promoter. Each payload was delivered both to the *Sox2* locus (left) and *Igf2/H19* locus (right). Expression of the engineered *Sox2* allele was normalized to the wild-type CAST allele and then scaled between 0 (Δ*Sox2* locus) to 1 ("*Sox2* & Sox2 LCR"), which are also represented as dashed lines. Inv, inverted. **c**. *Igf2* and *H19* expression analysis in mESC engineered with a "Full-length" *Igf2/H19* rescue payload, or payloads wherein the *H19* enhancer was replaced by the *Sox2* LCR in either orientation. Expression was normalized to *Gapdh* and then scaled between 0 (Δ*Igf2/H19* locus) and 1 ("Full-length"), normalizing *H19* to the "Full-length" payload with constitutive CTCF occupancy, and *Igf2* to the "Full-length" payload with abolished CTCF occupancy. Payloads with both constitutive CTCF occupancy (orange hexagon) or abolished CTCF occupancy (empty hexagon) at the ICR were assessed. The dashed lines indicate basal expression of "ΔH19Enh" in mESCs and expression of "Full-length" *Igf2/H19* rescue payload in differentiated mesendodermal cells. Points represent independently engineered mESC clones and bars indicate the median.

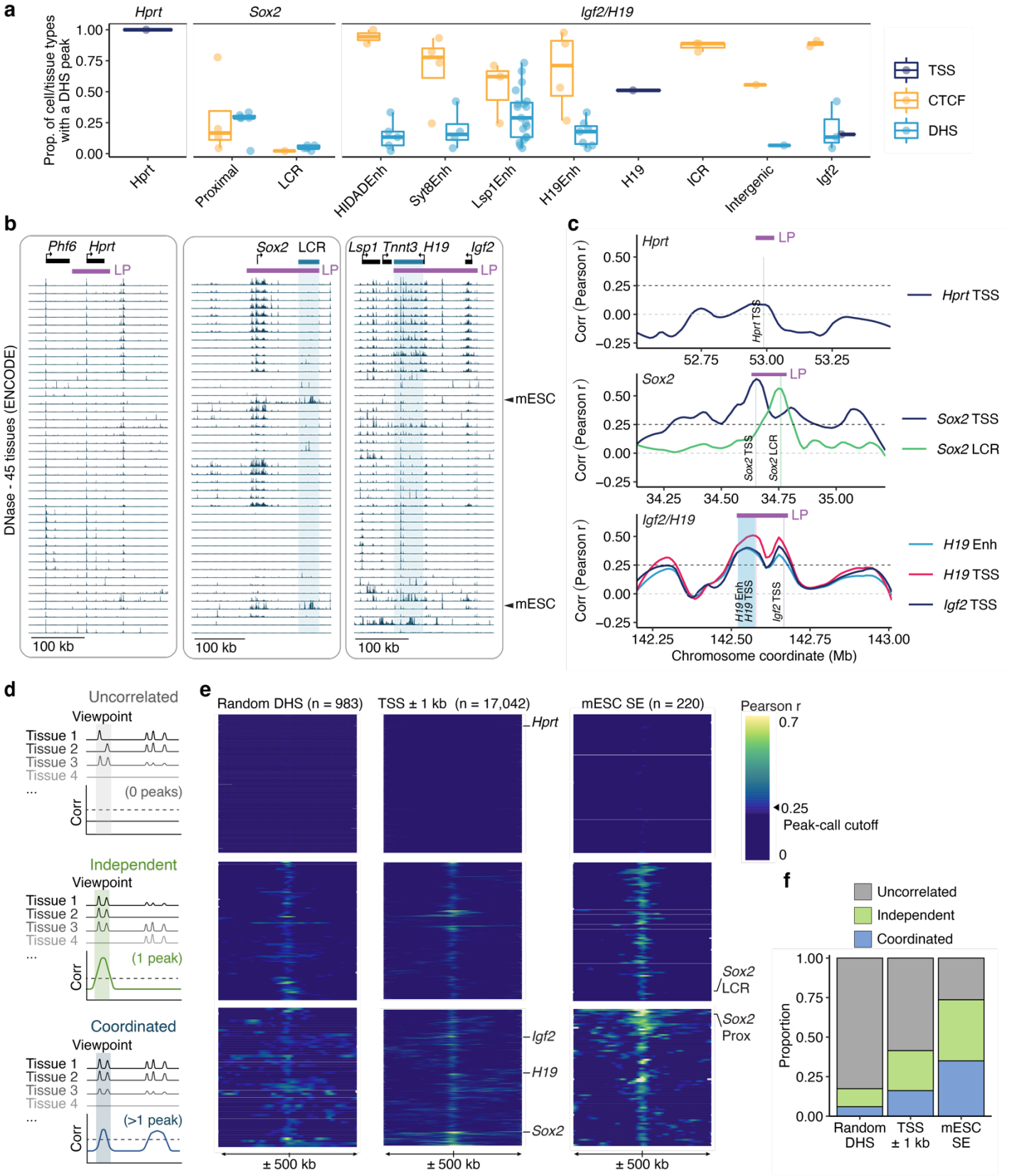

a
Hprt
Sox2
Igf2/H19
Prop. of cell/tissue types with a DHS peak
1.00
0.75
0.50
0.25
0.00
Hprt
Proximal
LCR
HIDADEnh
Syt8Enh
Lsp1Enh
H19Enh
H19
ICR
Intergenic
Igf2
TSS
CTCF
DHS
b
Phf6
Hprt
LP
Sox2
LCR
Lsp1
Tnnt3
H19
Igf2
DNase - 45 tissues (ENCODE)
mESC
100 kb
c
Corr (Pearson r)
Hprt TSS
Sox2 TSS
Sox2 LCR
H19 Enh
H19 TSS
Igf2 TSS
Chromosome coordinate (Mb)
d
Uncorrelated
Viewpoint
Tissue 1
Tissue 2
Tissue 3
Tissue 4
(0 peaks)
Corr
Independent
(1 peak)
Coordinated
(>1 peak)
e
Random DHS (n = 983)
TSS ± 1 kb (n = 17,042)
mESC SE (n = 220)
Pearson r
0.7
0.25
Peak-call cutoff
0
Hprt
Sox2 LCR
Sox2 Prox
Igf2
H19
Sox2
± 500 kb
f
Uncorrelated
Independent
Coordinated
Proportion
Random DHS
TSS ± 1 kb
mESC SE

Figure 4. Cross-cell type accessibility patterns reflect locus function

a. Cell and tissue selectivity of DHSs at the *Hprt*, *Sox2*, and *Igf2*/*H19* loci. Accessibility was assessed as the proportion of 45 samples from mouse ENCODE which had a DHS. DHSs overlapping CTCF sites and TSS were broken out separately. Bars represent the median, boxes represent the interquartile range, and whiskers the minimum and maximum values. **b**. Accessibility patterns at the *Hprt*, *Sox2* and *Igf2/H19* loci across 45 cell and tissue types, ordered by hierarchical clustering of genomic DHSs as in (Halow et al. 2021). Shading represents the *Sox2* LCR and *H19* enhancers. **c**. Correlation of DHS accessibility across *Hprt*, *Sox2* and *Igf2/H19* loci. Each line represents the Pearson correlation coefficient (r) between the DHS located at the viewpoint with every other DHS in a 1-Mb window. Viewpoints are shaded in each plot. **d**. Schematic representation of uncorrelated, independent and coordinated DHS clusters. Multiple regions of high Pearson r values indicate coordinated DHS activity potentially due to synchronized activation in specific tissues. Corr: Pearson correlation coefficient. **e**. Heatmaps showing the Pearson correlation coefficient (r) between the DHS located at the viewpoint and the surrounding DHS within a 1-Mb window. Three different sets of viewpoints are represented: random DHSs, promoters of protein coding genes (± 1 kb from TSS), and mESC super-enhancers (SE) (Whyte et al. 2013); each row portrays the region surrounding a different viewpoint. Rows are not shown to scale. DHS correlation patterns were grouped into uncorrelated, independent and coordinated based on the number of peaks exceeding $r > 0.25$. **f**. Representation of DHSs correlation patterns for various genomic features.

TABLES

Locus	Gene density (Genes / Mb)	DHS density in mESC (DHS / Mb)	DHS density in Mesendoderm (DHS / Mb)
Hprt	8	42	52
Sox2	1	37	30
Igf2/H19	32	58	75

Table 1. Regulatory environment complexity.

Gene and DHS density was calculated by counting the number of genes or DHS hotspots (FDR <0.05), respectively, in a window of 1 Mb sequence centered on each locus. Genes include all validated (level 1) or manually annotated (level 2) Gencode VM23 protein coding transcripts.

REFERENCES

Ainscough JF, Koide T, Tada M, Barton S, Surani MA. 1997. Imprinting of Igf2 and H19 from a 130 kb YAC transgene. *Development* **124**: 3621–3632.

Arney KL. 2003. H19 and Igf2--enhancing the confusion? *Trends Genet* **19**: 17–23.

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, et al. 2021. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53**: 354–366.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308.

Bell AC, Felsenfeld G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**: 482–485.

Bender MA, Bulger M, Close J, Groudine M. 2000. Beta-globin gene switching and DNase I sensitivity of the endogenous beta-globin locus in mice do not require the locus control region. *Mol Cell* **5**: 387–393.

Bergman DT, Jones TR, Liu V, Ray J, Jagoda E, Siraj L, Kang HY, Nasser J, Kane M, Rios A, et al. 2022. Compatibility rules of human enhancer and promoter sequences. *Nature* **607**: 176–184.

Blayney JW, Francis H, Camellato BR, Mitchell L, Stolper R, Boeke J, Higgs DR, Kassouf M. 2022. Super-enhancers require a combination of classical enhancers and novel facilitator elements to drive high levels of gene expression. *bioRxiv* 2022.06.20.496856.

Blobel GA, Higgs DR, Mitchell JA, Notani D, Young RA. 2021. Testing the super-enhancer concept. *Nat Rev Genet* **22**: 749–755.

Bolt CC, Duboule D. 2020. The regulatory landscapes of developmental genes. *Development* **147**: dev171736.

Bolt CC, Lopez-Delisle L, Hintermann A, Mascrez B, Rauseo A, Andrey G, Duboule D. 2022. Context-dependent enhancer function revealed by targeted inter-TAD relocation. *Nat Commun* **13**: 3488.

Brosh R, Coelho C, Ribeiro-Dos-Santos AM, Ellis G, Hogan MS, Ashe HJ, Somogyi N, Ordoñez R, Luther RD, Huang E, et al. 2023. Synthetic regulatory genomics uncovers enhancer context dependence at the Sox2 locus. *Mol Cell* **83**: 1140-1152.e7.

Brosh R, Laurent JM, Ordoñez R, Huang E, Hogan MS, Hitchcock AM, Mitchell LA, Pinglay S, Cadley JA, Luther RD, et al. 2021. A versatile platform for locus-scale genome rewriting and verification. *Proc Natl Acad Sci U S A* **118**: e2023952118.

Butz S, Schmolka N, Karemaker ID, Villaseñor R, Schwarz I, Domcke S, Uijttewaal ECH, Jude J, Lienert F, Krebs AR, et al. 2022. DNA sequence and chromatin modifiers cooperate to confer epigenetic bistability at imprinting control regions. *Nat Genet* **54**: 1702–1710.

Camellato B, Brosh R, Maurano MT, Boeke JD. 2022. Genomic analysis of a synthetic reversed sequence reveals default chromatin states in yeast and mammalian cells. *bioRxiv* 2022.06.22.496726.

Chakraborty S, Kopitchinski N, Zuo Z, Eraso A, Awasthi P, Chari R, Mitra A, Tobias IC, Moorthy SD, Dale RK, et al. 2023. Enhancer-promoter interactions can bypass CTCF-mediated boundaries and contribute to phenotypic robustness. *Nat Genet* **55**: 280–290.

Cohen BA, Mitra RD, Hughes JD, Church GM. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* **26**: 183–186.

de Wit E, Vos ESM, Holwerda SJB, Valdes-Quezada C, Verstegen MJAM, Teunissen H, Splinter E, Wijchers PJ, Krijger PHL, de Laat W. 2015. CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell* **60**: 676–684.

Eckersley-Maslin MA, Thybert D, Bergmann JH, Marioni JC, Flicek P, Spector DL. 2014. Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev Cell* **28**: 351–365.

Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**: 490–493.

Georgolopoulos G, Psatha N, Iwata M, Nishida A, Som T, Yiangou M, Stamatoyannopoulos JA, Vierstra J. 2021. Discrete regulatory modules instruct hematopoietic lineage commitment and differentiation. *Nat Commun* **12**: 6790.

Gietz RD, Schiestl RH. 2007. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc* **2**: 31–34.

Halow JM, Byron R, Hogan MS, Ordoñez R, Groudine M, Bender MA, Stamatoyannopoulos JA, Maurano MT. 2021. Tissue context determines the penetrance of regulatory DNA variation. *Nat Commun* **12**: 2850.

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**: 486–489.

Hong CKY, Cohen BA. 2022. Genomic environments scale the activities of diverse core promoters. *Genome Res* **32**: 85–96.

Hong J-W, Hendrix DA, Levine MS. 2008. Shadow enhancers as a source of evolutionary novelty. *Science* **321**: 1314.

Hörnblad A, Bastide S, Langenfeld K, Langa F, Spitz F. 2021. Dissection of the Fgf8 regulatory landscape by in vivo CRISPR-editing reveals extensive intra- and inter-enhancer redundancy. *Nat Commun* **12**: 439.

John S, Sabo PJ, Canfield TK, Lee K, Vong S, Weaver M, Wang H, Vierstra J, Reynolds AP, Thurman RE, et al. 2013. Genome-scale mapping of DNase I hypersensitivity. *Current protocols in molecular biology / edited by Frederick M Ausubel . [et al]* **Chapter 27**: Unit 21.27.

Kane L, Williamson I, Flyamer IM, Kumar Y, Hill RE, Lettice LA, Bickmore WA. 2022. Cohesin is required for long-range enhancer action at the Shh locus. *Nat Struct Mol Biol* **29**: 891–897.

Karr JP, Ferrie JJ, Tjian R, Darzacq X. 2022. The transcription factor activity gradient (TAG) model: contemplating a contact-independent mechanism for enhancer-promoter communication. *Genes Dev* **36**: 7–16.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**: 289–294.

Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31**: 180–183.

Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. 2002. Locus control regions. *Blood* **100**: 3077–3086.

Li Y, Rivera CM, Ishii H, Jin F, Selvaraj S, Lee AY, Dixon JR, Ren B. 2014. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**: e114485.

Llères D, Moindrot B, Pathak R, Piras V, Matelot M, Pignard B, Marchand A, Poncelet M, Perrin A, Tellier V, et al. 2019. CTCF modulates allele-specific sub-TAD organization and imprinted gene activity at the mouse Dlk1-Dio3 and Igf2-H19 domains. *Genome Biol* **20**: 272.

Maricque BB, Chaudhari HG, Cohen BA. 2019. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat Biotechnol* **37**: 90–95.

Martinez-Ara M, Comoglio F, van Arensbergen J, van Steensel B. 2022. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. *Mol Cell* **82**: 2519-2531.e6.

Matsuzaki H, Miyajima Y, Fukamizu A, Tanimoto K. 2021. Orientation of mouse H19 ICR affects imprinted H19 gene expression through promoter methylation-dependent and -independent mechanisms. *Commun Biol* **4**: 1410.

Moorthy SD, Davidson S, Shchuka VM, Singh G, Malek-Gilani N, Langroudi L, Martchenko A, So V, Macpherson NN, Mitchell JA. 2017. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res* **27**: 246–258.

Muro EM, Ibn-Salem J, Andrade-Navarro MA. 2019. The distributions of protein coding genes within chromatin domains in relation to human disease. *Epigenetics Chromatin* **12**: 72.

Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS, et al. 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**: 239–243.

Pachano T, Haro E, Rada-Iglesias A. 2022. Enhancer-gene specificity in development and disease. *Development* **149**: dev186536.

Papapetrou EP, Lee G, Malani N, Setty M, Riviere I, Tirunagari LMS, Kadota K, Roth SL, Giardina P, Viale A, et al. 2011. Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. *Nature Biotechnology* **29**: 73–78.

Peterson KR, Stamatoyannopoulos G. 1993. Role of gene order in developmental control of human gamma- and beta-globin gene expression. *Molecular and Cellular Biology* **13**: 4836–4843.

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**: 2281–2308.

Ribeiro-Dos-Santos AM, Hogan MS, Luther RD, Brosh R, Maurano MT. 2022. Genomic context sensitivity of insulator function. *Genome Res* **32**: 425–436.

Ringel AR, Szabo Q, Chiariello AM, Chudzik K, Schöpflin R, Rothe P, Mattei AL, Zehnder T, Harnett D, Laupert V, et al. 2022. Repression and 3D-restructuring resolves regulatory conflicts in evolutionarily rearranged genomes. *Cell* **185**: 3689-3704.e21.

Rinzema NJ, Sofiadis K, Tjalsma SJD, Verstegen MJAM, Oz Y, Valdes-Quezada C, Felder A-K, Filipovska T, van der Elst S, de Andrade Dos Ramos Z, et al. 2022. Building regulatory landscapes reveals that an enhancer can recruit cohesin to create contact domains, engage CTCF sites and activate distant genes. *Nat Struct Mol Biol* **29**: 563–574.

Rubin AF, Green P. 2013. Expression-based segmentation of the Drosophila genome. *BMC Genomics* **14**: 812.

Schübeler D, Lorincz MC, Cimbora DM, Telling A, Feng YQ, Bouhassira EE, Groudine M. 2000. Genomic targeting of methylated DNA: influence of methylation on transcription, replication, chromatin structure, and histone acetylation. *Mol Cell Biol* **20**: 9103–9112.

Swanzey E, McNamara TF, Apostolou E, Tahiliani M, Stadtfeld M. 2020. A Susceptibility Locus on Chromosome 13 Profoundly Impacts the Stability of Genomic Imprinting in Mouse Pluripotent Stem Cells. *Cell Rep* **30**: 3597-3604.e3.

Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, Ettwiller L, Spitz F. 2014. Functional and topological characteristics of mammalian regulatory domains. *Genome Research* **24**: 390–400.

Thomson M, Liu SJ, Zou L-N, Smith Z, Meissner A, Ramanathan S. 2011. Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell* **145**: 875–889.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.

Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ, Byron R, Humbert R, et al. 2014. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**: 1007–1012.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**: 307–319.

Wild J, Hradecna Z, Szybalski W. 2002. Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. *Genome Res* **12**: 1434–1444.

Zhao Y, Coelho C, Lauer S, Majewski M, Laurent JM, Brosh R, Boeke JD. 2023. CREEPY: CRISPR-mediated editing of synthetic episomes in yeast. *Nucleic Acids Res* **51**: e72.

Zhou HY, Katsman Y, Dhaliwal NK, Davidson S, Macpherson NN, Sakthidevi M, Collura F, Mitchell JA. 2014. A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes & Development* **28**: 2699–2711.

Zuin J, Roth G, Zhan Y, Cramard J, Redolfi J, Piskadlo E, Mach P, Kryzhanovska M, Tihanyi G, Kohler H, et al. 2022. Nonlinear control of transcription through enhancer-promoter interactions. *Nature* **604**: 571–577.

METHODS

Data and code availability

DNase-seq data were obtained from https://www.encodeproject.org for ES_E14 (ENCLB890BTE, DS21450). ChIP-nexus (Avsec et al. 2021) for Nanog (GSM4072778) was obtained from the GEO repository.

Landing Pads and CRISPR/Cas9 plasmids for genome integration

pLP-PIGA2 (pLP300) (Addgene 168462) (**Figure 1c** and **Figure S1a**) was described previously (Brosh et al. 2021) and harbors a pEF1α-PuroR-P2A-hmPIGA-P2A-mScarlet-EIF1pA cassette flanked by LoxM and LoxP sites, as well as a pPGK1-ΔTK-SV40pA backbone counter-selectable marker cassette. To target LPs to specific genomic loci, guide RNAs (gRNAs) (**Table S5**) were cloned into pSpCas9(BB)-2A-Puro (pCas9-Puro, Addgene 62988) using BbsI Golden Gate reactions as described (Ran et al. 2013). LP homology arms (HAs) corresponding to the sequence flanking genomic Cas9 cut sites were amplified from a BAC (**Table S6**). HAs were cloned to flank the LoxM and LoxP sites in pLP-PIGA2 using a BsaI Golden Gate reaction (**Table S5**). DNA suitable for transfection was prepped using the Zymo-PURE maxiprep kit (Zymo Research D4203) according to the manufacturer's protocol.

Payload assembly

All payloads were assembled into Yeast Assemblon Vector (YAV) pLM1110 (Brosh et al. 2021) (Addgene #168460). This is a multifunctional vector supporting (i) low-copy DNA propagation, selection (*LEU2*), and efficient homology-dependent recombination in yeast; (ii) low-copy propagation, selection (KanR), and copy number induction in TransforMax EPI300 *E. coli* (Wild et al. 2002); and (iii) transient visualization and selection (eGFP-BSD, enhanced green fluorescent protein - Blasticidin-S deaminase) in mammalian cells (Brosh et al. 2021).

Three different strategies were used to assemble payloads into pLM1110 (**Table S1**), using different primers, gRNAs, and synthetic DNA fragments (**Table S5**):

<u>(1) BAC fragment assembly in yeast</u>

Desired segments of *Igf2/H19* BAC RP23-50N22 (**Table S6**) were released by an *in vitro* CRISPR/Cas9 digestion using a pair of synthetic gRNAs and recombinant Cas9 and assembled into BsaI-digested pLM1110 as previously described (Brosh et al. 2021).

(2) CRISPR/Cas9 Engineering of EPisomes in Yeast (CREEPY)

Existing payloads were modified in yeast, as in Zhao et al. (Zhao et al. 2023) using CRISPR/Cas9 and synthetic linker fragment(s) (IDT) to mediate sequence deletions, insertions, or inversions. Yeast cells carrying a parental payload were transformed with pYTK-Cas9 plasmids (single-plasmid system co-expressing a yeast-codon optimized Cas9, the targeting gRNA and a *HIS3* marker) together with 100 ng of the synthetic linker fragment(s) (**Table S1**), and selected for Leu^+/His^+ phenotype. gRNAs (**Table S5**) were cloned into pYTK-Cas9 using a BsmBI Golden Gate reaction.

(3) Payload inversion

To deliver payloads in an inverted orientation, the flanking Lox sites were swapped in yeast. Source constructs were first digested using I-SceI, releasing the payload sequence from the backbone including the Lox sites. This digestion product was then co-transformed into yeast cells with I-SceI-digested pLM1110_URA3 (a derivative of pLM1110 carrying a *URA3* marker) and terminal linker fragments including the repositioned Lox sites (400-500 bp gBlocks, IDT) (**Table S1**, **Table S5**) to enable homologous recombination-dependent assembly. Yeast were then selected for a Ura^+ phenotype, depleting the wild-type Leu^+ recombinants.

Yeast transformation, payload validation, and transfer to *E. coli*

The following steps were performed as previously described (Brosh et al. 2023). Specifically, yeast transformations were performed using the Lithium acetate method (Gietz and Schiestl 2007). For screening of correct clones, DNA was isolated from individual yeast colonies by resuspension in 10-40 μL of 20 mM NaOH and boiling for 3 cycles of 95 °C for 3 min and 4°C for 1 min. 2 μL of yeast lysate was used as a template in a 10 μL GoTaq Green reaction (Promega M7123) with 0.25 μM of primers. Clones were screened for the intended payload structure using primers that target newly formed junctions (positive screen), and in some cases using primers that target regions that are present in the parental plasmid but absent in the intended plasmid (negative screen). Assembled payloads were isolated from candidate yeast clones using the Zymo yeast miniprep I protocol (Zymo Research D2001) and transformed into TransforMax EPI300 *E. coli* cells by electroporation using the manufacturer's protocol (Lucigen EC300150).

Payload DNA screening and preparation

Payload DNA was prepped and verified as previously described (Brosh et al. 2023). Briefly, for initial verification, TransforMax EPI300 *E. coli* colonies were picked into 3-5 mL LB-Kan supplemented with CopyControl Induction Solution (Lucigen CCIS125) and cultured overnight at 30 °C with shaking at 220 RPM. Crude DNA was extracted and payload assembly verification was performed by diagnostic digestion with restriction enzymes, PCR verification of new junctions, and Sanger sequencing, depending on the payload. All assembled payloads were subsequently sequenced to high coverage depth (**Table S7**).

Sequence-verified payload clones were grown overnight in 2.5 mL cultures of LB-Kan at 30 °C with shaking, diluted 1:100 in LB-Kan supplemented with CopyControl Induction Solution, and incubated for an additional 8-16 hours at 30 °C with shaking. For transfection, DNA was purified from induced *E. coli* using the Nucleobond XtraBAC kit (Takara 740436) and preps were stored at 4 °C.

mESC culture

mESC were cultured as previously described (Brosh et al. 2023). Specifically, C57BL/6J × CAST/EiJ (B6xCAST) mESCs were cultured on plates coated with 0.1% gelatin (EMD Millipore ES006-B) in 80/20 medium comprising 80% 2i medium and 20% mESC medium. 2i medium contained a 1:1 mixture of Advanced DMEM/F12 (ThermoFisher 12634010) and Neurobasal-A (ThermoFisher 10888022) supplemented with 1% N2 Supplement (ThermoFisher 17502048), 2% B27 Supplement (ThermoFisher 17504044), 1% GlutaMAX (ThermoFisher 35050061), 1% Pen-Strep (ThermoFisher 15140122), 0.1 mM 2-mercaptoethanol (Sigma M3148), 1,250 U/mL LIF (ESGRO ESG1107l), 3 μM CHIR99021 (R&D Systems 4423), and 1 μM PD0325901 (Sigma PZ0162). mESC medium contained knockout DMEM (ThermoFisher 10829018) supplemented with 15% FBS (BenchMark 100- 106), 0.1 mM 2-mercaptoethanol, 1% GlutaMAX, 1% MEM nonessential amino acids (ThermoFisher 11140050), 1% nucleosides (EMD Millipore ES008-D), 1% Pen-Strep, and 1,250 U/mL LIF. Cells were grown at 37 °C in a humidified atmosphere of 5% CO2 and passaged twice per week on average.

mESC genome engineering

Landing pad integrations were performed (**Table S3**) using the Neon Transfection System as previously described into B6xCAST ΔPiga mESCs, in which the endogenous Piga gene was

deleted (Brosh et al. 2021). LP-PIGA2 integration at the *Igf2/H19* locus (replacing a 157-kb genomic region) was performed using 5 μg of pLP-PIGA2 and 2.5 μg of each pCas9-GFP plasmid expressing B6-specific gRNAs mIgf2_1b and mIgf2_4b, followed by 1 μg/mL puromycin (ThermoFisher A1113803) selection for LP-harboring mESCs and 1 μM ganciclovir (GCV, Sigma PHR1593) selection against LP-PIGA2 backbone's HSV1-ΔTK gene.

The genomic deletion of the *Igf2/H19* B6 allele was performed by replacing the 157-kb genomic region with a selectable module (**Figure S6a**). The donor vector pDONOR~Neo~EGFP contained a positive selection cassette expressing eGFP, a Neomycin resistance gene and a negative selection marker expressing tagBFP and was derived from pDONOR-tagBFP-PSM-EGFP (Addgene 100603). Flanking the positive selection cassette, 500 bp homology arms corresponding to the genomic sequence flanking the Cas9 cut sites were cloned. LP cells were further transformed using 5 μg of pDONOR~Neo~EGFP vector and 2.5 μg of each pCas9-GFP plasmid expressing B6-specific gRNAs mIgf2_1b and mIgf2_4b, followed by 400 μg/mL Geneticin G418 (Thermo Fisher, 10131035) selection for 4 days.

Genomic deletions of the *Sox2* locus (**Figure S6b**) and the downstream enhancer clusters (*Lsp1*, *Syt8*, and *HIDAD*) were performed by co-transfecting cells with 2.5 µg of each pCas9-GFP and 2.5ul of 100 μM ssODN (IDT) bridging oligo with an overlap of 50 bp on each side of the deletion and modified with 5' and 3' Phosphorothioate bonds to prevent degradation. The following gRNAs were used for each deletion: mSox2_5p-1 and mSox2_3p-5 for the *Sox2* locus; mIgf2_26b and mIgf2_27b for the HIDAD DHS cluster; mIgf2_28b and mIgf2_29b for the *Syt8* DHS cluster; mIgf2_29b and mIgf2_30b for the *Lsp1* DHS cluster.

Payload deliveries were performed as previously described (Brosh et al. 2021) with 10×10^6 mESCs, 10 µg payload DNA and 10 µg pCAG-iCre plasmid (Addgene #89573) per transfection. Transfected mESCs were selected with 10 µg/mL blasticidin for 2 days starting day 1 post-transfection and with 2 nM proaerolysin for 2 days starting day 7 post-transfection.

For landing pad integration, payload delivery, and genomic deletions, individual mESC clones were manually picked approximately 10 days post-transfection into gelatinized 96-well plates prefilled with 30 μL TrypLE-Select, which was neutralized with 200 μL mESC medium. Immediately after, 25 μL were transferred into a new plate with 75 μL of 2i media, obtaining two

replica plates with 90% and 10% relative densities. Three days later, crude gDNA was extracted from the 90% plate as described (Brosh et al. 2021) and used in PCR genotyping to identify candidate clones, which were then expanded from the 10% density plate for further verification and phenotypic characterization. Genomic DNA was extracted from expanded clones using the DNeasy Blood & Tissue kit (QIAGEN 69506).

mESC genotyping

Genotyping mESC clones was performed either using PCR followed by gel electrophoresis as described or using real-time quantitative PCR (qPCR), which was performed with the KAPA SYBR FAST (Kapa Biosystems KK4610) on a LightCycler 480 Real-Time PCR System (Roche) using either 96-well or 384-well qPCR plates as previously described (Brosh et al. 2023). In most cases loading was performed using an Echo 550 liquid handler (Labcyte): a 384-well qPCR plate was prefilled with 5 μL KAPA SYBR FAST and 4 μL water per well. 100 nL of each 100 µM primer and 0.5 μL of each crude gDNA sample were transferred by the Echo. Genotyping payload clones typically included, in addition to assays designed to detect the newly-formed left and right junctions, assays to detect the loss of the landing pad, the absence of the payload vector backbone, and for large (>100 kb) payloads, allele-specific assays to detect internal delivered regions. Genotyping primers are listed in **Table S5**.

High-throughput sequencing verification

Illumina sequencing libraries were prepared from purified DNA as previously described (Brosh et al. 2023). The library construction method for each sample is listed in **Table S7**.

Hybridization capture (Capture-seq) for targeted resequencing of engineered regions was performed as previously described (Brosh et al. 2021). Briefly, biotinylated bait was generated using nick translation from BACs covering the *Igf2/H19* (RP23-50N22 or RP23-137N15), *Hprt* (RP23-412J16) or *Sox2* locus (RP23-144O8 or RP23-274P9) (see **Table S6** for BAC coordinates), landing pad plasmid (pLP-PIGA2), payload plasmid backbone, pCAG-iCre and pSpCas9 plasmids. Bait sets and sequencing statistics are listed in **Table S7**.

The sequencing parameters and the analysis pipeline are detailed in (Brosh et al. 2023). The full processing pipeline is also available at https://github.com/mauranolab/mapping.

Payload sequence verification

Payload sequence verification to identify potential sample swaps or contamination was performed as previously described (Brosh et al. 2023). Briefly, mean sequencing coverage was calculated for the genomic regions corresponding to the payload sequence and to the non-payload sequence (i.e. the portion of the original locus not included in the payload). Mean coverage was then normalized by the payload vector backbone coverage. Coverage analysis results and quality control (QC) calls for each sample are reported in **Table S8** and summarized in **Table S2**.

To detect payload DNA variants that might have been introduced during assembly or propagation in yeast or bacteria, we analyzed variant calls relative to the payload custom reference having sequencing depth above 10 and quality score above 50 (DP ≥ 10 and QUAL ≥ 50). Payload variants are reported in **Table S9** and QC calls are summarized in **Table S2**.

mESC clone sequence verification

mESC clone sequence verification was performed as previously described (Brosh et al. 2023) with slight modifications. Briefly, mean sequencing coverage depth was calculated for the genomic regions corresponding to the payload and non-payload (as defined above); and for the payload vector backbone, landing pad (LP), landing pad backbone, and pCAG-iCre custom references. Coverage depth of engineered regions was normalized by the mean genomic coverage of their flanks (i.e. the un-engineered portion of the genomic region targeted by the bait), such as that a single copy corresponds to a value of ~0.5. For landing pad clones, we expect normalized coverage depth of 0.5 ± 0.25 for the non-payload region and <0.1 for LP backbone. For payload clones, the expected coverage depth was adjusted according to the expected ploidy of the site, such as the limits are ploidy / 2 ± 0.3 for the payload region (0.2-0.8 for 1x; 0.7-1.3 for 2x; 1.2-1.8 for 3x), 0.5 ± 0.3 for the non-payload region and <0.1 for the landing pad, payload vector backbone and pCAG-iCre. Coverage depth analysis QC calls for each clone are reported in **Table S10**.

We further verified the presence of BL6 allele variants, which are lost in regions replaced by landing pads and restored by delivered payloads, by calculating allelic ratios as the mean proportion of reads supporting the reference (BL6) allele (propREF) for the genomic regions described above. Expected proportions of the reference allele were adjusted according to ploidy to: 0.3-0.7 for 2x; 0.47-0.87 for 3x for payload regions, and 0-0.2 for non-payload regions. Regions

with 10 or fewer variants were ignored. Allelic ratio QC calls for each clone are reported in **Table S10**.

To verify genomic integration sites, we detected sequencing read pairs mapping to two different reference genomes using *bamintersect* (Brosh et al. 2021) as previously described (Brosh et al. 2023) with slight modifications: same-strand reads mapping within 500 bp were clustered; a minimum width threshold of 125 bp was required for reporting; and junctions with few reads (ReadsPer10M < 20 or junctions representing less than 5% of the total ReadsPer10M) were excluded. *Bamintersect* junctions were classified hierarchically based on position (**Table S11**). Results and QC calls for each clone are reported in **Table S12**.

Mesendodermal differentiation

Differentiation to mesendodermal cells was adapted from a published protocol (Thomson et al. 2011). mESCs were seeded on plates coated with 0.1% gelatin (EMD Millipore ES006-B) in N2/B27 medium, containing a 1:1 mixture of DMEM/F12 (ThermoFisher 11330032) and Neurobasal (ThermoFisher 21103049), supplemented with 0.5% N2 Supplement (ThermoFisher 17502048), 1% B27 Supplement (ThermoFisher 17504044), 1% GlutaMAX (ThermoFisher 35050061), 1% Pen-Strep (ThermoFisher 15140122), 1% MEM nonessential amino acids (ThermoFisher 11140050), 100 µM sodium pyruvate (ThermoFisher 11360070), 0.1 mM 2-mercaptoethanol (Sigma M3148). The differentiation media was supplemented with 10 µM ROCK inhibitor Y-27632 (Cell Signaling 13264) the first 24 h after seeding the cells. For the next 3 days, 3 µM CHIR99021 (R&D Systems 4423) was added to the media, and replaced every 24 h. Four days after differentiation, plates were rinsed twice using PBS and cells were collected for further studies.

DNase-seq

DNase digestion was performed as described previously (John et al. 2013; Georgolopoulos et al. 2021) and adapted to 200 μL thermocycler tubes. Briefly, nuclei were extracted from cells and incubated with limiting concentrations of the DNA endonuclease DNase I (Sigma D4527-20KU) supplemented with Ca^{2+} and Mg^{2+} for 3 min at 37 °C. Digestion was stopped by the addition of EDTA, and the samples were treated with proteinase K and RNase A. Short double-hit fragments were isolated from DNaseI digestion using magnetic bead polyethylene glycol (PEG)

fractionation. Illumina libraries were generated and sequenced on an Illumina NextSeq 500. Sequencing statistics are listed in **Table S7**.

CUT&RUN

CUT&RUN experiments were conducted using the CUT&RUN (Cleavage Under Targets & Release Using Nuclease) assay kit (Cell Signaling Technology, #86652) following the manufacturer's protocol. For each reaction, 5×10^6 cells were used. Positive (H3K4me3) and negative (Rabbit mAb IgG) controls were performed using the antibodies provided in the kit. The CTCF antibody (Cell Signaling Technology, #2899S) was used at a dilution of 1/50 for overnight incubation at 4°C. DNA fragments were purified using the DNA Clean & Concentrator-5 Kit (Zymo, D4014) according to the manufacturer's procedures, and eluted in 15 μL. Up to 11.15 μL purified DNA was taken directly into end repair for library construction using the Illumina dsDNA protocol as previously described (Brosh et al. 2023). During PCR enrichment of adaptor-ligated DNA, the extension time was reduced to 15 sec to exclude amplification of large library DNA fragments (>1,000 bp) and reduce background signal. The sequencing parameters and the analysis pipeline are detailed in (Halow et al. 2021). The full processing pipeline is also available at https://github.com/mauranolab/mapping. Sequencing statistics are listed in **Table S7**.

RNA Isolation, cDNA synthesis and mRNA expression analysis by real-time qRT-PCR

Expression assays were performed as previously described (Brosh et al. 2023). Briefly, RNA was isolated using the RNeasy-mini (QIAGEN 74106) protocol followed by DNase treatment using the Turbo DNA-free kit (ThermoFisher AM1907). cDNA was synthesized from 1-2 µg total RNA with the Multiscribe High-Capacity cDNA Reverse Transcription Kit (ThermoFisher 4368814).

Real-time quantitative reverse transcription PCR (qRT-PCR) was performed using KAPA SYBR FAST (Kapa Biosystems KK4610) on a 384-well LightCycler 480 Real-Time PCR System (Roche) and threshold cycle (Ct, also called Cp) values were calculated using Abs Quant/2nd Derivative Max analysis. Primers were designed to detect *Igf2* and *H19* transcripts in an allele-specific manner (**Figure S5** and **Table S6**). In most cases loading was performed using an Echo 550 liquid handler (Labcyte): a 384-well qPCR plate was prefilled with 5 μL KAPA SYBR FAST and 4 μL water per well. 100 nL of each 100 µM primer and 1 μL of each cDNA sample were transferred by the Echo. Thermal cycling parameters were as follows: 3 min pre-incubation at 95 °C, followed by 40 amplification cycles of 3 sec at 95 °C, 10 sec at 60 °C and 10 sec at 72 °C. *Sox2* expression

was quantified using a similar protocol but with annealing temperature lowered to 57 °C as in (Brosh et al. 2021).

Assays were performed in duplicate and replicate wells on the same plate were averaged after masking wells with no or very low amplification (Cp > 30) and low RNA quality (Gapdh Cp > 19). Raw Ct values are listed in **Data S2**. ΔCt values for the BL6 and CAST alleles were computed relative to *Gapdh* (*Hprt* in the time course differentiation experiment). Clones demonstrating non-specific differentiation under ESC culture conditions were identified as having high *Igf2* CAST allele expression (ΔCt > 7 in ESC) and discarded (n=5). Replicate ΔCt measurements of the same clone across different plates were averaged. Fold change was defined as the difference between the BL6 and CAST alleles, calculated as $2^{\Delta Ct[CAST-BL6]}$. Fold change was scaled to yield expression values ranging from 0 (Δ*Igf2/H19*) and 1 ("Full-length"), by subtracting the mean fold change calculated for the Δ*Igf2/H19* samples from all data points, then dividing by the mean fold change of the appropriate "Full-length" payload sample: constitutive CTCF occupancy for *H19,* and abolished CTCF occupancy for *Igf2* (**Table S4**).

Sox2 expression was scaled similarly, and as previously described in (Brosh et al. 2023)**.** Briefly, *Sox2* fold change was defined as the difference between the BL6 and CAST *Sox2* alleles, calculated as $2^{\Delta Ct[CAST-BL6]}$. Fold change was scaled to yield expression values ranging from 0 (ΔSox2) to 1 ("*Sox2* & Sox2LCR") by subtracting the mean fold change calculated for the ΔSox2 samples from all data points, then dividing by the mean fold change of the appropriate wild-type payload samples ("*Sox2* & Sox2LCR" or "Sox2LCR").

Cross-cell type DHS correlation analysis

Tissue-specific accessibility patterns were based on a master peak list of 45 representative samples from mouse ENCODE DNase-seq data (Halow et al. 2021). The cell-type activity spectrum (MCV, for multi-cell verified) was computed using 'bedmap --count' to consider overlaps with the center ± 25 bp of each feature. MCV was then scaled from 0 to 1 by dividing by 45.

Accessibility per DHS peak per cell/tissue type was quantified using bedmap --max from density tracks. The Pearson correlation coefficient (r) was calculated pairwise between each viewpoint and each DHS within the surrounding 1-Mb window. We then fitted a loess regression curve (span

= 0.2). The coordinates for the *Sox2* LCR correspond to the DHS23 to DHS26 (Brosh et al. 2023) (**Table S6)**. Gene TSS viewpoints were computed as TSS ± 1 kb.

DHS accessibility correlation was assessed for three viewpoint sets: (i) 1000 random DHSs selected from a peak master list of 45 representative samples from mouse ENCODE DNase-seq data (Halow et al. 2021) and located at least 10 kb away from a gene TSS; (ii) promoters of protein-coding genes (±1 kb from TSS, Gencode release M23, genes located in mitochondrial DNA were excluded); (iii) mESC super-enhancers (Whyte et al. 2013). Viewpoints with less than 80% mappable sequence in the 1 Mb surrounding window were excluded. DHS correlation patterns were grouped based on the number of instances within a 1-Mb window that the smoothed Pearson correlation exceeded 0.25: 0 times (uncorrelated), 1 time (independent), and 2+ times (coordinated).

Supplementary Materials

TABLE OF CONTENTS

SUPPLEMENTARY FIGURES **37**

Figure S1. Big-IN landing pad at *Igf2/H19*. 37

Figure S2. Schematic representation of engineered payloads. 38

Figure S3. Mesendodermal differentiation time course. 39

Figure S4. ICR engineering. 40

Figure S5. Allele-specific qRT-PCR primer details. 41

Figure S6. *Igf2/H19* and *Sox2* deletions. 42

Figure S7. *H19*, *Igf2* and *Sox2* gene and promoter regions. 43

Figure S8. DHS accessibility correlation. 44

SUPPLEMENTARY TABLES **45**

Table S1. Payloads and assembly details. 45

Table S2. Summary of payload sequencing QC. 45

Table S3. mESC clones. 45

Table S4. qPCR expression values and activity. 46

Table S5. Primers, gRNAs, and synthetic DNA fragments. 46

Table S6. Genomic coordinates. 46

Table S7. Sequencing libraries. 46

Table S8. Payload sequencing coverage QC. 47

Table S9. Payload variant calling. 47

Table S10. mESC clone capture sequencing coverage QC. 47

Table S11. *Bamintersect* junction categories. 48

Table S12. mESC clones *bamintersect* results. 49

SUPPLEMENTARY DATA **50**

Data S1. Landing pad and payload sequences. 50

Data S2. Raw qRT-PCR Ct values. 50

SUPPLEMENTARY FIGURES

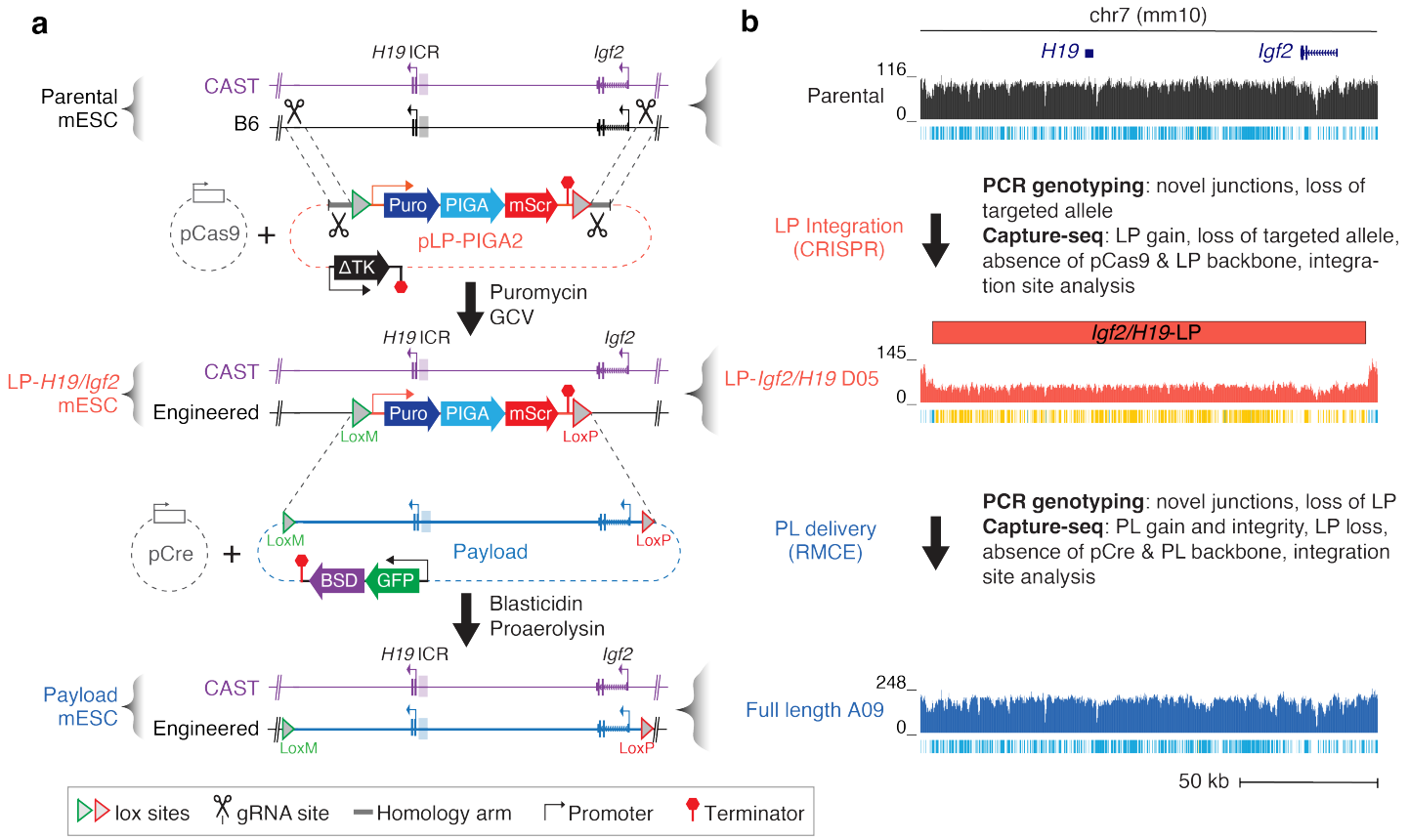


Figure S1. Big-IN landing pad at *Igf2/H19*.
a. Engineering strategy for LP-*Igf2/H19* mESCs. The B6 allele of the 157-kb *Igf2/H19* locus was replaced with a landing pad (LP-PIGA2). LP integration was driven by Cas9 using a pair of gRNAs targeting both the replaced allele and the LP plasmid and by short homology arms that facilitate homology-directed repair. LP mESCs were selected with puromycin, while ganciclovir (GCV) selects against LP backbone integration. Cre recombinase-mediated cassette exchange (RMCE) enabled replacement of the LP with a series of payloads. Payload-transfected cells were transiently selected with blasticidin, followed by counterselection of remaining LP mESCs cells with proaerolysin. **b**. Schematic of mESC clone verification pipeline. Representative examples of targeted capture sequencing validation throughout the engineering process: parental (ΔPiga) mESCs, LP-*Igf2/H19* clone D05, and full-length A09 payload clones. Ticks under each coverage track indicate heterozygous (B6xCAST, blue) or homozygous non-reference (CAST, yellow) variants.

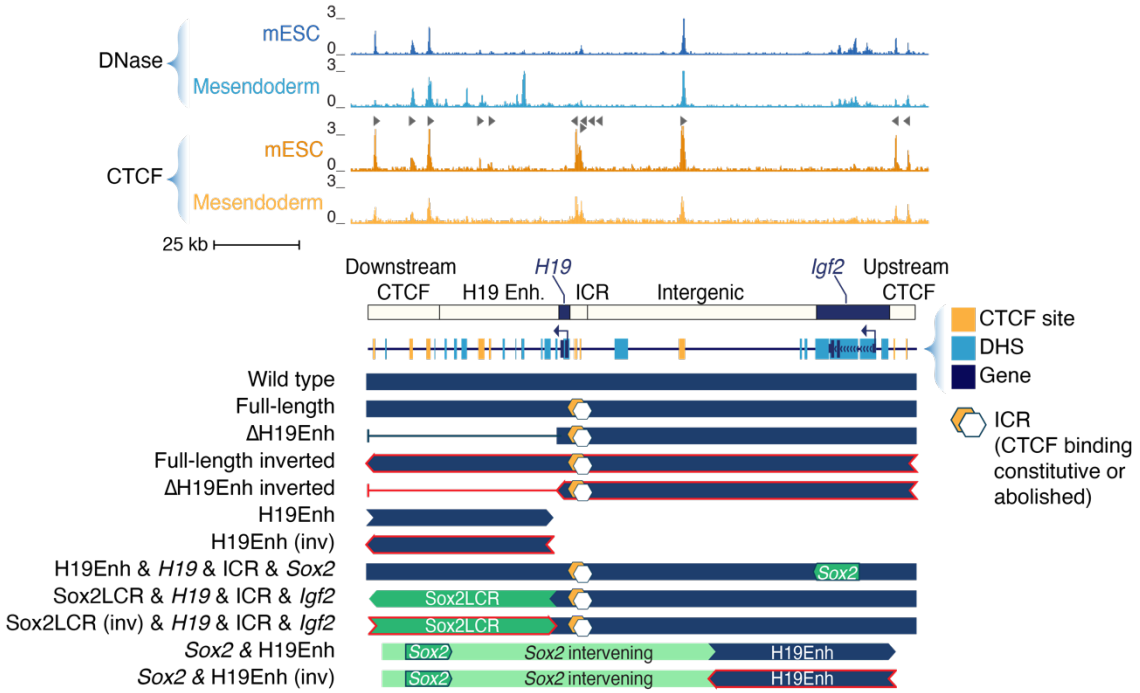


Figure S2. Schematic representation of engineered payloads.
Shown are DNase-seq and CTCF CUT&RUN tracks and the CTCF sites and DNase Hypersensitive Sites (DHS) identified throughout the *Igf2/H19* locus. Inverted sequences are colored in red. Inv, inverted. Hexagons denote payloads pairs that are identical except for their ICR structure.

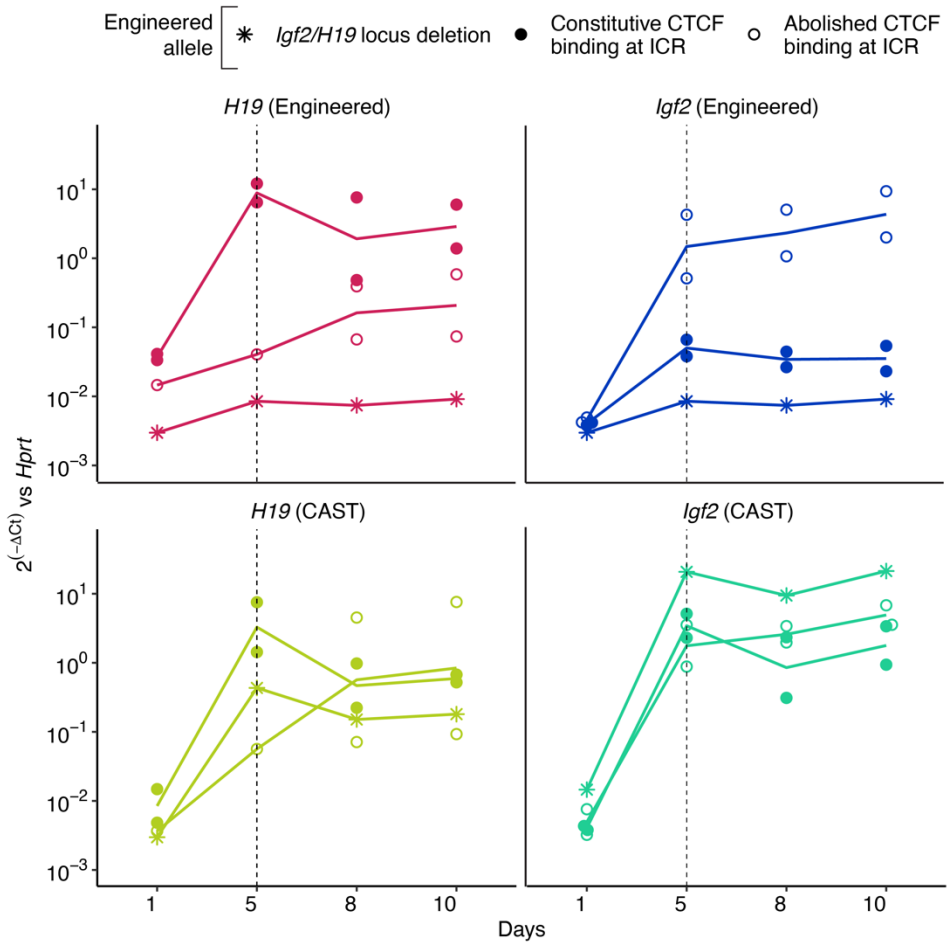


Figure S3. Mesendodermal differentiation time course.
H19 and *Igf2* allelic expression from both the engineered and the CAST alleles during a differentiation time course from mESC to mesendodermal cells, as determined by allele-specific qRT-PCR. Each point represents the expression of an independently engineered mESC clone normalized to *Hprt*. Lines represent medians. The selected time point for further differentiation experiments is represented as a vertical dashed line.

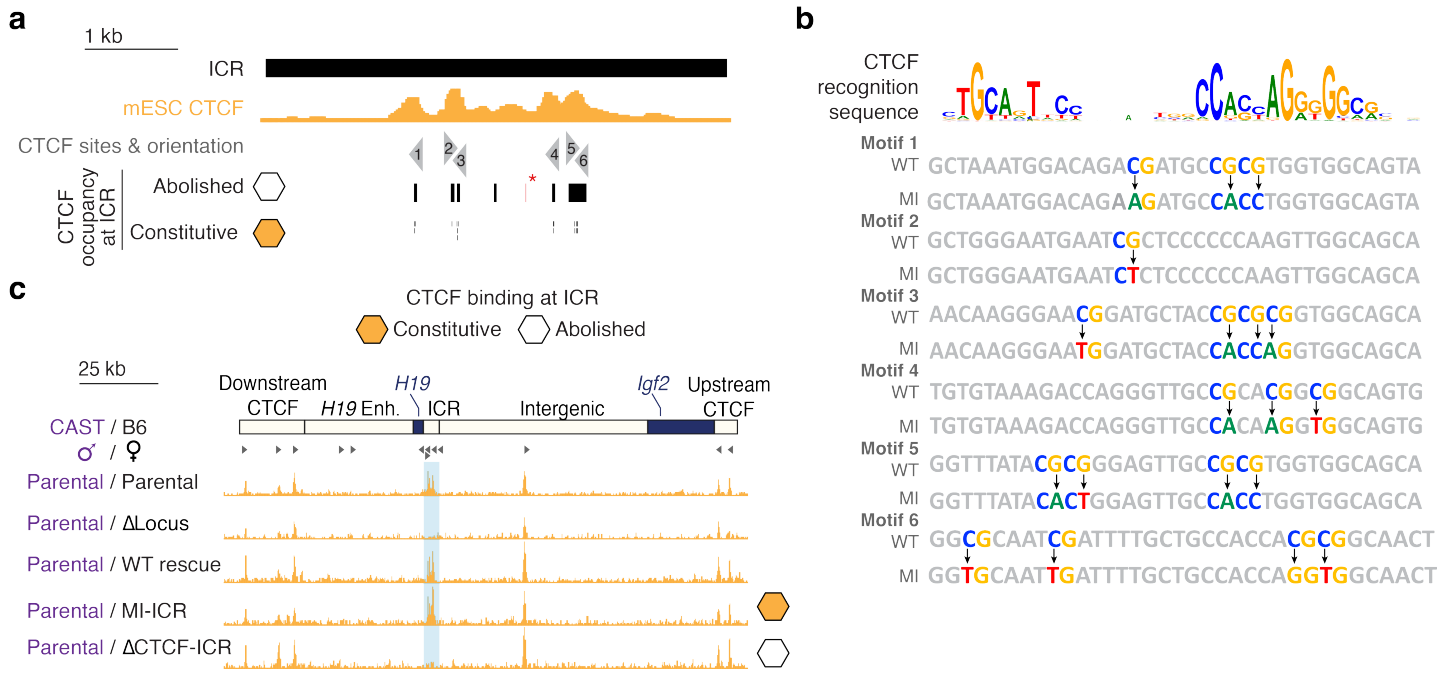
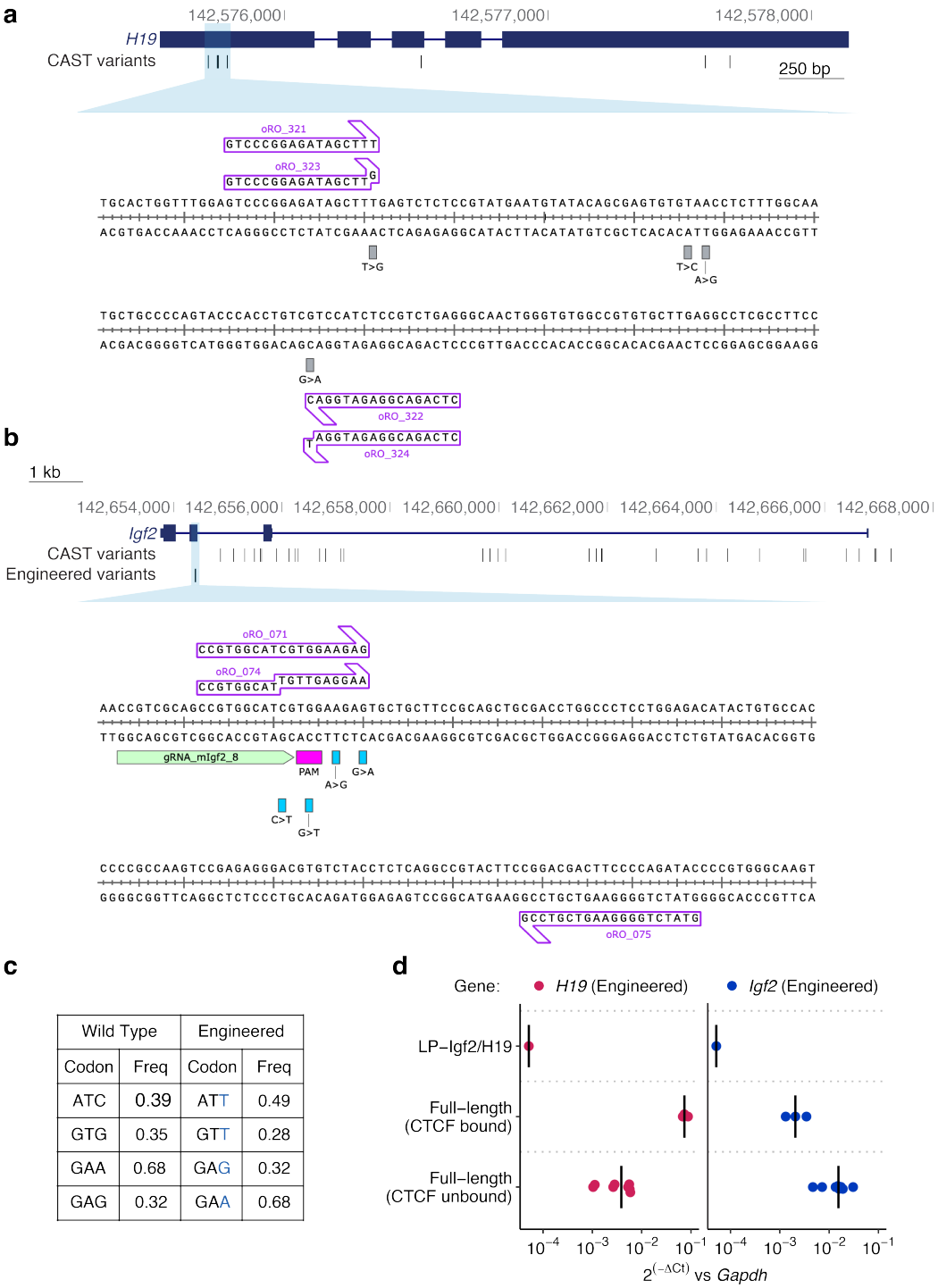


Figure S4. ICR engineering.

a. Detail of the *Igf2/H19* ICR showing CTCF occupancy in mESC. Triangles indicate CTCF recognition sequence orientation. Assembled payloads included either CTCF site deletions resulting in abolished CTCF occupancy (ΔCTCF-ICR, empty hexagon) or CpG mutations resulting in constitutive CTCF occupancy (methylation-insensitive ICR or MI-ICR, orange hexagon). *, Point deletion detected in the "Full-length" ΔCTCF-ICR payload (see BS09866A in **Table S9**). This point variant is not expected to confound the measured activities of the integrated payload. **b**. Detail on the methylation-insensitive mutations in the MI-ICR design where all CpGs in the 6 CTCF recognition sequences were mutated without significant alteration of the CTCF canonical recognition sequence. **c**. CUT&RUN CTCF signal at the *Igf2/H19* locus in differentiated mesendodermal cells. Genotypes for each track are indicated at the left. The shaded region denotes the ICR. The orientation of CTCF recognition sites is indicated on top as triangles.



Wild Type		Engineered	
Codon	Freq	Codon	Freq
ATC	0.39	ATT	0.49
GTG	0.35	GTT	0.28
GAA	0.68	GAG	0.32
GAG	0.32	GAA	0.68

Figure S5. Allele-specific qRT-PCR primer details.
a. Schematic representation of the *H19* gene, including the CAST variants present in the unengineered allele and a detailed sequence view including the allele-specific primers used in the qRT-PCR assays (as detailed in **Table S5**). **b.** Schematic representation of the *Igf2* gene. Shown below are CAST variants present in the unedited allele, and engineered synonymous variants in exon 4. The detailed sequence view includes the allele-specific primers used in the qRT-PCR assays (as detailed in **Table S5**) and the gRNA used for payload engineering. **c.** Codons overlapping the 4 point substitutions in the wild-type and engineered *Igf2* showing their genomic frequencies. **d**. Allele-specific qRT-PCR assay for *H19* and *Igf2* expression. Transcripts from engineered alleles were measured using allele-specific primers in LP-*Igf2*/*H19* mESCs as well as "Full-length" payload clones with constitutive or abolished CTCF binding to the ICR. Each point represents the expression of the engineered *H19* or *Igf2* gene in an independent mESC clone, measured using allele-specific qRT-PCR. Bars indicate median.

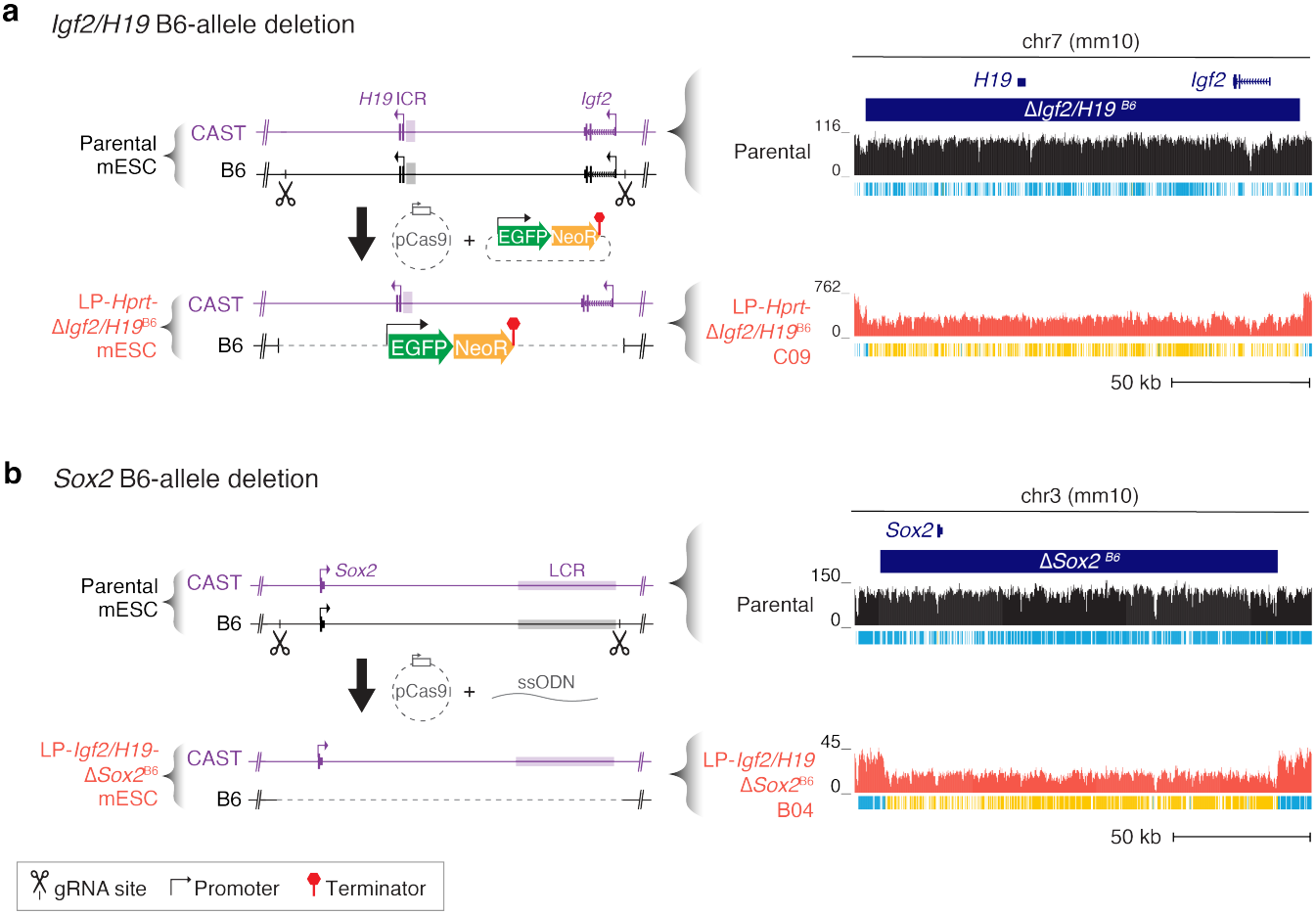


Figure S6. *Igf2/H19* and *Sox2* deletions.

a. Engineering of *Igf2/H19* B6-allele deletion. The B6 allele of the *Igf2/H19* locus in chromosome 7 was deleted through replacement by pDONOR~Neo~EGFP. Representative examples of targeted capture sequencing validation are shown: coverage tracks for parental (ΔPiga) mESCs and LP-*Hprt*-Δ*Igf2/H19* mESCs including deletion of the B6 *Igf2/H19* allele (Δ*Igf2/H19*^{B6}) clone D05. Ticks under each coverage track indicate heterozygous (B6xCAST, blue) or homozygous non-reference (CAST, yellow) variants. **b**. Engineering of *Sox2* B6-allele deletion. The B6 allele of the *Sox2* locus at chromosome 3 was deleted using Cas9 and a single-stranded oligodeoxynucleotide (ssODN) as a donor for homology-directed repair. Representative examples of targeted capture sequencing validation are shown: coverage tracks for parental (ΔPiga) mESCs and LP-*Igf2/H19*-Δ*Sox2* mESCs including deletion of the *Sox2* B6 allele clone B04. Ticks under each coverage track indicate heterozygous (B6xCAST, blue) or homozygous non-reference (CAST, yellow) variants. Puro, puromycin-resistance gene; PIGA, human mini *PIGA* gene; mScr, mScarlet; ΔTK, truncated thymidine kinase; BSD, blasticidin-resistance gene; eGFP, enhanced green fluorescent protein gene; NeoR, Neomycin-resistance gene.

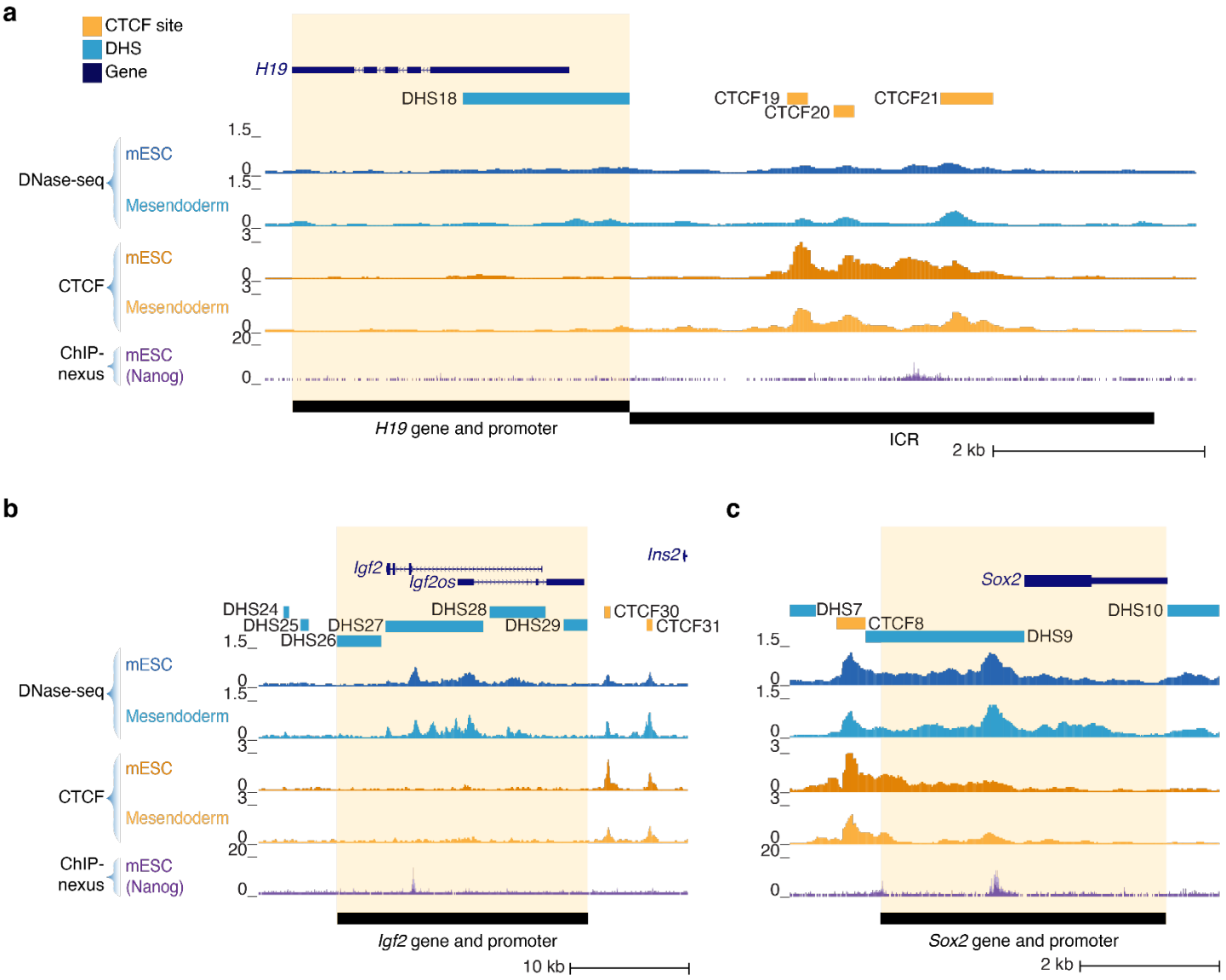


Figure S7. *H19*, *Igf2* and *Sox2* gene and promoter regions.
Genomic region surrounding *H19* gene and the ICR (**a**), *Igf2* gene (**b**), and *Sox2* gene (**c**), showing DNase-seq and CTCF CUT&RUN in mESC and differentiated mesendodermal cells, together with ChIP-nexus data for Nanog transcription factor occupancy in mESC. The region engineered in those payloads including gene and promoter swaps is represented by the black bar and highlighted in yellow (see **Table S6** for region coordinates).

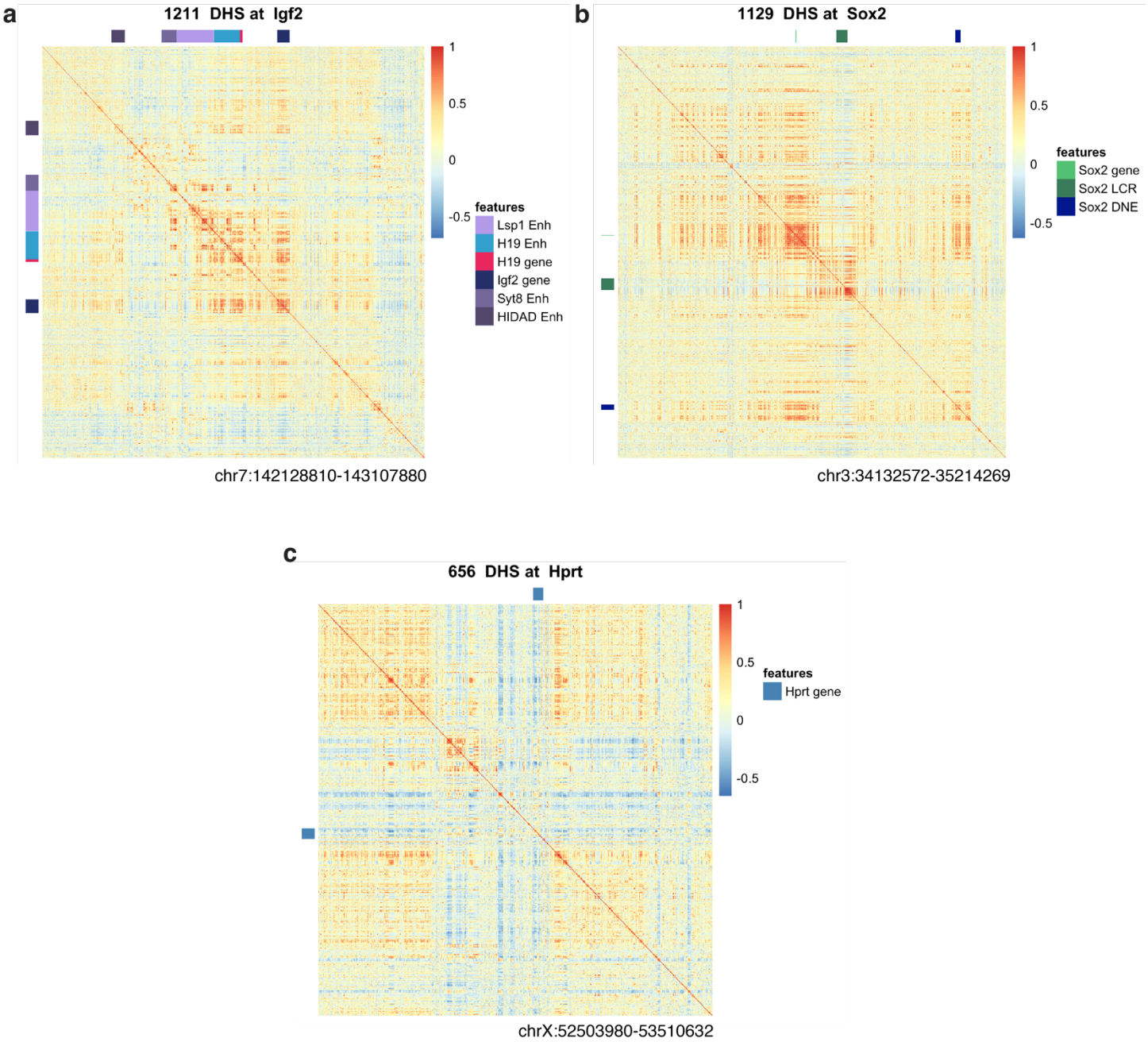


Figure S8. DHS accessibility correlation.
Heatmap of pairwise correlations (Pearson r coefficient) of DHS accessibility in the 1-Mb regions surrounding *Igf2/H19* (**a**), *Sox2* (**b**) and *Hprt* (**c**). Correlations were computed across 45 mice ENCODE DNase-seq samples. Relevant regulatory elements or genes are labeled with distinct colors. DNE, Distal Neural Enhancer (Chakraborty et al. 2023).

SUPPLEMENTARY TABLES

Table S1. Payloads and assembly details.

Strategies and reagents used for payload, as well as sequencing data. See **Methods** for details of each assembly strategy, **Table S5** for cloning primer sequences, gRNA sequences and synthetic fragment sequences. Source or intermediate payloads not delivered to mESCs were assigned "Intermediate" in the "Assembly Step" field and those not including an ICR sequence in the design are assigned n/a in the "ICR design" field. Payload ID, a unique name for each payload.

This table is provided as a separate file.

Table S2. Summary of payload sequencing QC.

	Delivered payloads (n=19)		
	Passed	**Failed**	**No call**
Coverage	18 (94.7%)	1 (5.3%)[(a)]	0
Variant calls	19 (100%)	0	0

[(a)] We detected a subclonal internal deletion by a drop in coverage in Igf2_Igf2_dCTCFinICR (BS09866A in **Table S8**). mESC clones were additionally screened for correct sequence integration after delivery.

Table S3. mESC clones.

Summary of mESC clones described in this manuscript. Genetic Modification details DNA elements integrated and their integration site in square brackets; transiently transfected plasmids (e.g. pSpCas9 or pCAG-Cre) are indicated by null. See **Table S4** for payload labels used in figures. "Bl6/C_LP300(Igf2_1_4)_D5" corresponds to LP-*Igf2/H19* mESCs, "BL6/C LP300(Igf2_1_4)_D5 dSox2_201026_B4" corresponds to LP-*Igf2/H19*-Δ*Sox2* mESCs including the deletion of the *Sox2* B6 allele, "BL6/C LP300(Hprt_3g_4g)_20210414_C04" corresponds to LP-*Hprt* mESCs, "Bl6/C_LP300(Hprt_3_4)_C04_dIgf2_C9" corresponds to LP-*Hprt*-Δ*Igf2/H19* cells including deletion of the *Igf2/H19* B6 allele, "BL6/C LP300(Sox2_1g_5g)_201007_D4" corresponds to LP-*Sox2* mESCs, and "Bl6/C LP305(Sox2_6g_7g)_20210326_C01" corresponds to LP-LCR mESCs. Sample ID, a unique sequencing library identifier; Clone ID, a unique descriptor of the mESC clone; Payload ID, LIMS name for relevant payload or LP; Source, publication in which this mESC clone was first described.

This Table is provided as a separate file.

Table S4. qPCR expression values and activity.

Igf2, H19 and *Sox2* expression in engineered clones. Payload ID, a unique name for each payload. Location, delivery location in the genome. Additional Deletions, presence of additional deletions engineered in that mESC clone. Label, display name used in Figure panels. Clone ID, a unique identifier for each engineered mESC clone. Differentiation, indicates whether expression was measured in mESC or mesendodermal cells differentiated for 5 days (Mesendoderm_5). Assay, targeted gene. deltaCt, ΔCt values relative to *Gapdh* for each gene. Fold change, $2^{-\Delta Ct}$ value. Activity, scaled fold change as detailed in **Methods**.

This Table is provided as a separate file.

Table S5. Primers, gRNAs, and synthetic DNA fragments.

Primer pairs (qRT-PCR, genotyping PCR, amplification of DNA fragments for payload assembly, homology arms cloning), gRNA sequences and synthetic DNA sequences used in this study. For the primers used to clone left and right homology arms (HAs) using Golden Gate reactions, primer sequences contain BsaI sites in lower case, gRNA binding sites (underlined), and PAMs (bold). For the gRNAs used for LP integrations and payload assembly, allele-specific gRNAs are identified as having CAST variants overlapping the PAM, the gRNA seed sequence or both.

This Table is provided as a separate file.

Table S6. Genomic coordinates.

Coordinates (mm10) for payloads, genomic locus deletions (marked with a “d”), BACs, enhancer clusters (Enh), and engineered variants. *Igf2*SV, *Igf2* synonymous variants; MI-ICR, methylation-insensitive ICR with CpGs in CTCF sites mutated; and dCTCFinICR, CTCF motif deletions in the ΔCTCF-ICR and locus features. See **Table S4** for payload labels used in figures.

This Table is provided as a separate file.

Table S7. Sequencing libraries.

Summary of payload and mESC capture sequencing libraries. Sample ID, a unique sequencing library ID; Payload ID, a descriptor of the relevant landing pad or payload (see **Table S4** for payload labels used in figures); Clone ID, a unique descriptor of mESC clones. For each sample, the number (#) of sequenced, analyzed and duplicate reads is provided. Bait Set is provided for capture samples.

This Table is provided as a separate file.

Table S8. Payload sequencing coverage QC.

Summary Table for payload DNA sequencing coverage depth analysis. Length (bp) and normalized (norm) coverage were calculated for regions corresponding to the payload sequence, including different chromosomes for those that apply, as defined in **Methods**. Quality control (QC) calls are included with comments for unexpected values. See **Table S4** for payload labels used in figures.

This Table is provided as a separate file.

Table S9. Payload variant calling.

Summary Table for payload variants relative to custom references, related to **Table S2**. Ref, reference sequence; Alt, variant sequence; Qual, quality score; DP, sequencing depth; Feature, the custom reference feature overlapping the variant; RefSeq, variant and surrounding sequence; Comments, annotation and interpretation of variant location and relevance. None of the detected variants are expected to confound the measured activities of integrated payloads. See **Table S4** for payload labels used in Figures.

This Table is provided as a separate file.

Table S10. mESC clone capture sequencing coverage QC.

Summary of QC analyses of Capture-seq samples derived from engineered mESC clones, including sequencing coverage, allelic ratios, and integration site analysis (*bamintersect*). See **Methods** for more detail. Samples captured with both *Igf2/H19* and *Sox2* baits (**Table S7**) show borderline low coverage signal at *Sox2* (~0.65x) due to lower capture efficiency of this last bait, resulting in apparent low payload coverage (marked as FAIL: low coverage [PL]). Clone ID, a unique descriptor of the mESC clone; Sample ID, a unique sequencing library identifier; Payload ID, unique identifier for relevant payload or LP. See **Table S4** for payload labels used in figures.

Table S11. *Bamintersect* junction categories.

Hierarchical criteria used to classify *bamintersect* junctions. bam1 and bam2 are the two references to which junction reads were mapped. LP, Landing Pad; PL, payload; BB, backbone; Jx, Junction; Chr, chromosome.

			bam1			bam2		
Clone Type	**Priority**	**Jx class**	**Chr**	**Position**	**Strand**	**Chr**	**Position**	**Strand**
LP	1	L Jx	HAL's chr	0-1250 bp left of HAL	+	LP	First 1250 bp	-
LP	2	R Jx	HAR's chr	0-1250 bp right of HAR	-	LP	Last 1250 bp	+
LP	3	LP-BB	bam1 or bam2 maps to LP backbone					
LP	4	off-target	None of the above					
PL	1	L Jx[a]	HAL's chr	0-1250 bp left of HAL	+	PL	First 1250 bp	-
PL	2	R Jx[a]	HAR's chr	0-1250 bp right of HAR	-	PL	Last 1250 bp	+
PL	3	L ITR-PL[b]	LP	First 1250 bp	+	PL	First 1250 bp	-
PL	4	R ITR-PL[b]	LP	Last 1250 bp	-	PL	Last 1250 bp	+
Any	5	Non-informative[c]	Within engineered region (+1250 bp left and right)			Any		
Any	6	PL-BB[d]	bam1 or bam2 maps to PL backbone					
Any	7	off-target	None of the above					

[a] For payloads with wild-type termini, CAST allele reads are indistinguishable from B6 allele reads.
[b] Only applicable to payloads delivered to LP-LCR mESCs as the LP ITRs remain flanking the integration site after PL delivery (Brosh et al. 2023).
[c] Non-informative junctions often originate from the CAST allele within the engineered locus
[d] LP-BB and PL-BB junctions are often a result of contamination of gDNA samples by LP/PL plasmids.

Table S12. mESC clones *bamintersect* results.

Summary of *bamintersect* integration site analysis for mESC clones capture sequencing samples. Each line describes a junction (Jx) detected by *bamintersect* (see **Methods**). Samples BS17489A and BS17493A missed the left junction due to a known low mappability region overlapping the *Sox2* left homology arm. Chrom (chromosome), chromStart, chromEnd, and Strand for bam1 and bam2 refer to the mapped coordinates for the first and second reference genome, respectively. NearestGene indicates the nearest UCSC gene, including distance in bp (negative value indicates an upstream gene) for mm10 or nearest feature for custom references. ReadsPer10M is the number of read pairs supporting the junction per 10 million sequencing reads. Sample ID, a unique sequencing library identifier; Payload ID, LIMS name for relevant payload or LP; Clone ID, a unique descriptor of the mESC clone; Genomes indicate the reference genomes compared. Junc_type indicates the junction classification based on criteria detailed in **Table S11**. See **Table S4** for payload labels used in figures.

This Table is provided as a separate file.

SUPPLEMENTARY DATA

Data S1. Landing pad and payload sequences.

Sequences for landing pads and payloads described in this paper, provided in fasta format. For landing pads, sequences include the integrated region (excluding the homology arms). For payloads, sequences include the region from LoxM to LoxP. See **Table S4** for payload labels used in the Figures.

This Table is provided as a separate file.

Data S2. Raw qRT-PCR Ct values.

Raw qRT-PCR Ct values for all clones and assays. Payload ID, a unique name for each payload; Label, name as in Figure panels. Location, delivery location in the genome. Additional Deletions, presence of additional deletions engineered in that mESC clone. Label, display name used in Figure panels. Clone ID, a unique identifier for each mESC engineered clone. Differentiation, indicates whether expression was measured in mESC or mesendodermal cells differentiated for 5, 8 or 10 days. Assay, targeted gene and allele (engineered or CAST). Cp, threshold cycle.

This Table is provided as a separate file.