Dissection of a complex enhancer cluster at the *Sox2* locus

Ran Brosh[1], Camila Coelho[1,7], André M. Ribeiro-dos-Santos[1,7], Megan S. Hogan[1,5], Hannah J. Ashe[1], Gwen Ellis[1], Nicolette Somogyi[1], Raquel Ordoñez[1], Raven Luther[1,6], Emily Huang[1], Jef D. Boeke[1,2,3], Matthew T. Maurano[1,4,*]

[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
[4] Department of Pathology, NYU School of Medicine, New York, NY 10016, USA.

[5] Present address: Neochromosome Inc., Long Island City, NY, 11101, USA.
[6] Present address: BlueRock Therapeutics, New York, NY 10016, USA.

[7] These authors contributed equally to this work.

[*] Corresponding author: maurano@nyu.edu

ABSTRACT

A complex landscape of genomic regulatory elements underpins patterns of metazoan gene expression, yet it has been technically difficult to disentangle composite regulatory elements within their endogenous genomic context. Expression of the *Sox2* transcription factor (TF) in mouse embryonic stem cells (mESCs) depends on a distal regulatory cluster of DNase I hypersensitive sites (DHSs), but the contributions of individual DHSs and their degree of independence remain a mystery. Here, we comprehensively analyze the regulatory architecture of the *Sox2* locus in mESCs using Big-IN to scarlessly deliver payloads ranging up to 143 kb, permitting deletions, rearrangements and inversions of single or multiple DHSs, and surgical alterations to individual TF recognition sequences. Multiple independent mESC clones were derived for each payload, extensively sequence-verified, and profiled for expression of *Sox2* specifically from the engineered allele. We found that a single core DHS comprising a handful of key TF recognition sequences is sufficient to sustain significant expression in mESCs, though its contribution is modulated by additional DHSs. Moreover, their overall activity was influenced by specific DHS order and/or orientation effects. We built a highly predictive model for locus regulation which includes nonlinear components indicating both synergy and redundancy among key DHSs. Our results suggest that composite regulatory elements and their influence on gene expression can be resolved to a tractable set of sequence features using synthetic regulatory genomics.

INTRODUCTION

Many developmentally important genes lie near clusters of highly cell-type specific DNase I hypersensitive sites (DHSs), typified by the ß-globin locus control region (LCR) (Grosveld et al. 1987; Evans et al. 1990). More recently, 'super enhancers' have been proposed to exhibit exceptionally cooperative binding, acting as developmental switches that regulate key TFs (Whyte et al. 2013; Pott and Lieb 2014; Blobel et al. 2021). General exploration of the broad assumptions involved have been impeded by the challenges of large-scale genomic engineering, leaving fundamental questions unanswered. Individual constituent DHSs demonstrate enhancer activity in transient expression assays (Thurman et al. 2012; Whyte et al. 2013), but it is less clear whether they act independently or synergistically in composite elements (Arnosti and Kulkarni 2005; Shin et al. 2016; Hay et al. 2016; Dukler et al. 2016; Moorthy et al. 2017). Furthermore, nearby enhancers may provide redundancy rather than augmenting gene expression per se (Hong et al.

2008; Bothma et al. 2015; Osterwalder et al. 2018; Scholes et al. 2019). Therefore, there is a need for new approaches to investigate the architecture of complex regulatory elements.

Sox2 encodes a key TF regulating mouse embryonic stem cell (mESC) self-renewal and pluripotency (Avilion et al. 2003). The *Sox2* gene is surrounded by a proximal enhancer cluster, but its expression in mESCs relies on a LCR located over 100 kb downstream (Zhou et al. 2014; Li et al. 2014). The effect of the LCR on *Sox2* expression is influenced by distance and/or intervening CTCF sites (Huang et al. 2021; Zuin et al. 2022; Chakraborty et al. 2022). Surgical inversion of a single CTCF recognition sequence within the *Sox2* LCR affects chromatin architecture but not *Sox2* expression in mESCs (de Wit et al. 2015), suggesting that proximity and expression might be separate functions (Alexander et al. 2019; Taylor et al. 2022). The *Sox2* locus therefore provides a natural model for dissecting how the composition and configuration of complex regulatory elements determine transcriptional activity.

We have recently developed the Big-IN genome engineering technology which employs recombinase-mediated cassette exchange (RMCE) for targeted delivery of DNA payloads of more than 140 kb (Brosh et al. 2021). Here, we use Big-IN to rewrite the mouse *Sox2* locus with designer payloads including deletions, inversions, translocations, and mutations of single and multiple DHSs. We show that the LCR is a complex element whose constituent DHSs are only partially interchangeable and exhibit unexpected dependencies not previously predicted by reporter assays or deletion analyses. We show that the contribution of individual TF activity depends critically on their local context. Future application of this synthetic regulatory genomics approach to assess regulatory architecture promises to cast new light on the function of key loci genome-wide.

RESULTS

Synthetic regulatory genomics of the *Sox2* locus

To enable interrogation of regulatory element function at the *Sox2* locus, we integrated Big-IN landing pads (LPs) into the BL6 allele of C57BL6/6J × CAST/EiJ (BL6xCAST) F1 hybrid mESCs, replacing either the 143 kb *Sox2* locus (LP-Sox2) or the 41 kb *Sox2* LCR (LP-LCR) (**Fig. 1a)**. Each LP engineering step was accompanied by comprehensive quality control (QC) steps to ensure single-copy on-target integration (**Fig. 1b**, **Fig. S1,** and **Methods**). Using allele-specific primers, we confirmed biallelic *Sox2* expression in parental cells, ablation of BL6 *Sox2* expression

in LP-Sox2 mESCs, and near total loss of BL6 *Sox2* expression in LP-LCR mESCs, recapitulating previous reports (Zhou et al. 2014; Li et al. 2014) (**Fig. 1c**).

To dissect the architecture of the *Sox2* locus, we delineated 28 distinct regulatory regions (**Fig. 2a**) through inspection of genomic annotation including DNase-seq and ChIP-seq data in mESC. Based on a BL6 BAC (bacterial artificial chromosome) harboring the murine *Sox2* locus, we used yeast assembly to engineer a series of Big-IN payloads including deletions of different regulatory regions (**Table S1**). We verified the resulting payloads through sequencing, delivered them to LP-Sox2 mESCs (**Fig. 1a**), isolated multiple mESC clones for each payload, and comprehensively verified single-copy targeted payload integration (**Table S2, Table S3**). We measured *Sox2* expression from the engineered payload relative to the untouched CAST allele as an internal control. Delivery of the wild-type (WT) locus fully rescued *Sox2* expression, whereas payloads lacking the LCR (ΔLCR) showed minimal expression (**Fig. 2b**, **Table S4**). Absence of DHSs 10-16 downstream of *Sox2* led to 22% reduction in expression, while a larger deletion of the entire 81-kb intervening region between *Sox2* and its LCR (including DHSs 10-18) showed expression comparable to WT. Given that LCR activation of *Sox2* is distance-dependent (Zuin et al. 2022), it is possible that reduced distance partially compensates for loss of proximal regulatory elements. Deletion of the proximal region upstream of *Sox2* (DHSs 1-8) also reduced expression. Combined deletion of both the upstream and downstream proximal DHSs led to even lower *Sox2* expression than either single deletion. A more limited deletion of DHSs 1-8 with just DHSs 10 & 15 showed similar activity, thus we conclude that the effect of the downstream deletion is largely attributable to DHSs 10 & 15. In contrast to these results showing a role for the proximal regulatory regions, single deletion of DHS7 or DHS15 or the combined deletion of DHSs 10-11 have been reported to show no expression difference (Zhou et al. 2014), suggesting that some of these DHSs are redundant and show no effect when deleted alone.

Dissection of the *Sox2* LCR

We next focused on a 41-kb region surrounding the *Sox2* LCR, which contains 10 DHSs and CTCF sites active in mESCs (**Fig. 3a**). To interrogate the necessity of each DHS, we delivered a series of LCR payloads to LP-LCR mESCs (**Fig. 1a**) and analyzed their ability to restore the expression of the BL6 *Sox2* allele (**Fig. 3b**). Delivery of the WT LCR completely rescued *Sox2* expression. Analysis of single and multiple DHS deletions delineated a core LCR region comprising DHSs 23-26. Deletion of DHS24 alone critically ablated *Sox2* expression to just 42% of WT, while deletion of DHS23 or DHS26 showed a lesser but significant impact. Within this core

LCR, only deletion of CTCF25 was tolerated with little effect. Deletions outside of the core LCR showed little or no effect.

Examining the occupancy of key TFs involved in pluripotency using ChIP-seq and ChIP-nexus data showed that each LCR DHS exhibits a different pattern (**Fig. S3**), suggesting that the individual DHSs may be functionally specialized. To investigate whether individual DHSs within the LCR function in isolation or rely on their surrounding DHSs, we delivered minimal payloads containing subsets of LCR DHSs to LP-LCR mESCs and measured their activities (**Fig. 4a** and **Fig. S3**). First, we observed that a minimal payload containing the 4 core LCR DHSs recapitulated 88% of the activity of the full WT LCR. Activities of single DHSs varied substantially: DHS24 and DHS26 were able to restore 30% and 14% of *Sox2* expression, respectively, whereas DHS23 was completely inactive. But when DHS23 was linked to either DHS24 or DHS26, it augmented their overall activity nearly twofold. Linking 2 copies of DHS24 (*2x24*) also nearly doubled expression relative to a single copy, demonstrating that increased dosage of DHS24 without addition of any new TF activity can manifest in increased expression. Strikingly, linking DHS19 and DHS20 with CTCF25 and DHS26 yielded significant expression (54% of WT) despite lacking both DHS23 and DHS24. This is particularly surprising given that deletion of DHS19 and DHS20 had no effect provided the rest of the LCR was left intact (**Fig. 3b**), DHS19 by itself had no activity (**Fig. 4a**), and both DHS19 and DHS20 were previously reported as inactive in a reporter assay (Zhou et al. 2014). We speculate that DHS19 and DHS20 can augment enhancer activity either due to the increased proximity among each other or to CTCF25 and DHS26. We conclude that the DHSs comprising the LCR are only partially redundant and many exhibit unexpected function when placed in novel contexts.

Context dependence might further manifest as differences in expression for different configurations of the same set of DHSs. Since certain genomic features have an intrinsic polarity, such as transcription or the interaction between cohesin and CTCF, we investigated the effect of DHS orientation on LCR activity. While inversion of the entire LCR resulted in a 24% decrease in its activity, inversions of smaller LCR regions had a lesser effect (**Fig. 4b**). Individual inversions of DHSs 23, 24, 25, and 26 had almost no effect on expression. This is consistent with a prior report showing that surgical inversion of a CTCF recognition sequence within CTCF25 had no effect on *Sox2* expression (de Wit et al. 2015). To explore whether the effect of larger inversions might be mediated by positional differences rather than the orientation of specific TF or CTCF recognition sites, we profiled multiple permutations of the core DHS order (**Fig. 4c**). Each

permutation of DHSs 23, 24, and 25 order reduced overall activity, whereas relocating CTCF25 downstream of DHS26 resulted in increased activity. These results suggest both the orientation and relative position of DHSs within the LCR may affect expression.

Modeling *Sox2* expression

To coherently summarize the architecture of the *Sox2* locus, we investigated linear regression models that predict *Sox2* expression from surrounding DHS composition and configuration (**Fig. 5a**). We established a baseline model considering only key proximal regions along with the four DHSs in the core LCR. Performance improved substantially upon incorporation of interactions among core LCR DHSs, in particular the contribution of DHS23 in conjunction with DHS24 or DHS26, and CTCF25 in the presence of DHS26 (**Fig. 5b**). Performance was higher when restricting to these three top interaction indicators than when considering all pairwise interactions. Given the potential influence of DHS configuration on function, we extended our top interaction regression model to include the orientation of each core LCR DHS and relative order encoded by indicators for DHS24 preceding DHS23 or CTCF25 preceding DHS24. Inclusion of either orientation or order separately improved model performance, but a model including order alone provided a better fit than orientation alone or a combination of both (**Fig. 5b**). The coefficients of the best fit model showed strong weights for the presence of proximal regulatory regions and core LCR DHSs, as well as key interactions and DHS order (**Fig. 5c**).

Given the contribution of interaction terms to model performance, we aimed to explore in more detail the extent to which activities of core LCR DHSs depend on their surrounding context. We identified pairs of payloads differing solely by the presence of each core LCR DHS. We calculated the contribution of the focus DHS in that context as the difference in *Sox2* expression between mESC clones for each payload pair (**Fig. 5d**). The synergism between DHS23 and DHS24 was the most evident contextual effect, followed by the higher activity shown by near-complete LCR payloads containing individual DHS deletions vs. minimal elements lacking their surrounding native context. Thus, the *Sox2* LCR consists of 4 DHSs whose function depends strongly on their context.

TF-scale dissection of the core LCR

Given the complex interdependencies among core LCR DHSs when considered as units, we used synthetic DNA to investigate function at the level of individual TF recognition sequences. We

designed base payloads covering the core LCR DHSs that avoided or shortened a handful of repetitive regions to facilitate chemical synthesis (**Fig. 6a**). We then designed a series of derivative payloads including surgical deletions, mutations or inversions to assess the individual and collective functions of putative TF or CTCF binding sites identified through analysis of TF motif matches and in vivo DNA accessibility and TF occupancy data (**Fig. 6b**).

We started with DHS24, which is the strongest single DHS in the LCR and is capable of reproducing 30% of WT LCR activity (**Fig. 4a**). We identified 4 distinct TF recognition sequence sites, numbered 24.1-24.4 (**Fig. 6b**). These sites were predicted to be occupied by key mESC regulators, including the ESRR family, nuclear receptors, SOX, and forkhead proteins. Deletion or mutation of all four sites was sufficient to fully ablate *Sox2* expression (**Fig. 6c**). All single deletions showed at least some effect, ranging from 42% (24.2) to 73% (24.4) of base activity (**Fig. 6c**). Double or triple deletions showed a further negative effect, essentially reaching null when 24.2 and 24.3 were both deleted. A minimal payload containing the most essential site based on the deletion analysis (24.2) with its flanking regions was unable to confer any *Sox2* expression. We analyzed the contribution of each TF site relative to its context, comparing pairs of payloads differing by the presence of a focus TF site (**Fig. 6d**). This analysis identified a clear context dependence for each of the 4 TF sites, and the contribution of each TF site increased with the number of other sites present.

As DHS23 alone is incapable of activating *Sox2* expression (**Fig. 4a**), its core DHS peak was analyzed linked to DHS24 (*syn23-24*, **Fig. 6b**). The activity of this payload was slightly higher than the non-synthetic payload including DHSs 23-24 (**Fig. 6e**), possibly due to the decreased distance between the core DHS peaks of DHS23 and DHS24 in the synthetic configuration. Deletion of 7 recognition sequence sites within DHS23 reduced activity by only 30%, and a payload containing a deletion extending halfway through DHS23 reduced activity by 53%. These results suggest that we have identified some, but not all, TF sites responsible for the function of DHS23. This synthetic approach further permitted assessment of two conserved DHSs from the human *SOX2* locus, orthologous to mouse DHS23 and DHS24. While mouse DHS23 and DHS24 are sufficient to confer nearly half of WT *Sox2* expression, the two orthologous human DHSs surprisingly showed no activity whatsoever. While the core of the human DHSs showed detectable sequence conservation with mouse sequence, analysis of TF recognition sequences showed a marked divergence relative to the orthologous mouse sites. This suggests that, despite the sequence orthology, the human LCR has diverged functionally.

Finally, we investigated CTCF25 and DHS26, which delineate a region covered by continuous DNA accessibility that comprises two broad DHS peaks positioned closely together (**Fig. 6a-b**). We identified 9 CTCF recognition sequences in CTCF25 and DHS26, 3 of which lie in convergent orientation relative to *Sox2* (**Fig. 6b**), including the previously reported CTCF site 25.2 (de Wit et al. 2015). While many of these sites lay outside the ChIP-seq peak, they do lie within DHSs and demonstrate a trace level of CTCF occupancy (**Fig. 6b)**, so we reasoned that they might be partially redundant with 25.2. We thus investigated whether inversion or deletion of multiple sites might result in a stronger phenotype than the reported mutation of 25.2 alone. Inversion of the 41-bp CTCF footprint at the 3 sites in convergent orientation to *Sox2* in a baseline payload incorporating DHS23 through DHS27 (*23-27 (Divergent CTCF*), **Fig. 6b**) led to a 20% reduction in expression (**Fig. 6f**). Deletion of 8 CTCF recognition sequences (*23-27 (ΔCTCF)*, **Fig. 6b**) led to an even greater reduction of 36% (**Fig. 6f**), bringing activity down nearly to the baseline of the remaining DHS23 and DHS24, and showing that these CTCF sites are collectively required for full function of the CTCF25-DHS26 segment. As DHS26 can function as an enhancer replacing 13% of LCR activity (**Fig. 4a**), we investigated the function of TF sites not overlapping with CTCF recognition sequences. A synthetic DHS26 payload demonstrated slightly higher activity than the non-synthetic DHS26 (**Fig. 6f**), likely because it extended slightly into CTCF25 to include a predicted HNF4A/RARB site (25.4) in addition to one CTCF site (25.3) (**Fig. 6b**). Deletion of 5 TF recognition sequence sites in this context completely abrogated activity. Notably, a payload lacking all 5 of these TF sites but including one of the CTCF sites that was inverted/deleted (26.1, also occupied by Pbx, and Zic3) showed nearly full expression, suggesting that, like DHS24, TF binding at CTCF25-DHS26 is highly context dependent and synergistic. Our work shows the power of a synthetic regulatory genomics approach to dissect locus architecture from 100-kb scale down to individual bp-level TF binding events, and suggests the utility of further combinatorial analysis of TF activity across multiple DHSs.

DISCUSSION

Dissection of the *Sox2* LCR shows it to be a highly complex and species-specific element whose function depends on the specific conformation of its constituent DHSs. These context dependencies often are not recapitulated in reporter assays or more limited engineering of the endogenous locus, underscoring the importance of a synthetic regulatory genomics approach for comprehensive analysis of regulatory architecture in context. A key example is our investigation of CTCF25, which previously had been shown to exhibit little or no effect on expression when

deleted or inverted (de Wit et al. 2015). In our study, removal or inversion of additional nearby CTCF recognition sequences yielded a significant difference in *Sox2* expression, yet we found inversion of the full DHS to be less important than its position within the LCR. Notably, redundant CTCF binding is a feature of other insulator elements (Bell and Felsenfeld 2000; Dickson et al. 2010; Huang et al. 2021). Furthermore, while we did not identify all the specific TFs occupying the deleted recognition sequences, the overlapping motif matches suggests that many of these CTCF and TF sites may exhibit occupancy by multiple distinct TFs.

Our results suggest that TF activity at the *Sox2* LCR demonstrates a high degree of synergy. Deletions of TF recognition sites had larger effects in more complete contexts, which contrasts with prior results showing that the effect of point variants is buffered by stronger DHSs (Maurano et al. 2012). We speculate that this discrepancy is related to the size of the perturbation, as TF binding can tolerate point changes without falling below the threshold needed for activity (Singh et al. 2021). Nevertheless, it is possible that the LCR DHSs might demonstrate redundancy in other contexts. For example, while *Sox2* expression for a given payload was highly consistent across replicate clones, a single clone lacking DHSs 21-22 showed a nearly complete absence of activity in contrast to the other 8 clones which showed near-WT activity (**Fig. 3b**). This outlier clone passed extensive genotyping and Capture-seq to demonstrate that the locus is intact, suggesting that seemingly redundant sites may in fact influence robustness in certain circumstances.

While our approach investigates function at the endogenous locus, preserving its long-distance architecture and surrounding genomic context, the payloads themselves are epigenetically naive, in that they do not pass through development and their chromatin state is established after delivery based on the surrounding context. Characteristic chromatin and expression patterns are readily re-established on DNA that passes through the germline (Peterson and Stamatoyannopoulos 1993; Wilson et al. 2008), but it is possible that DNA transfected to cycling cells exhibits only partial functionality (Li et al. 2005; Lienert et al. 2011). For example, DHS23 may be an enhancer unable to establish accessibility on its own. Thus, it remains for future work to apply this approach to investigate function across additional contexts and loci.

METHODS

Cloning landing pad and CRISPR/Cas9 plasmids for genome integration

pLP-PIGA (pLP140) (Addgene #168461) was described previously (Brosh et al. 2021) and harbors a pEF1α-mScarlet-P2A-CreERT2-P2A-PuroR-P2A-hmPIGA-EIF1pA cassette flanked by loxM and loxP sites.

pLP-PIGA2 (pLP300) (Addgene #168462) (**Fig. 1a)** 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 counterselectable marker cassette.

To derive LP-PIGA3 (pLP305) (**Fig. 1b**), an intermediate pLP303a was cloned by removing mScarlet-P2A-CreERT2-P2A from pLP-PIGA using NcoI and SalI digest followed by a fill-in reaction with Klenow DNA Polymerase and self-ligation to yield a minimal LP plasmid consisting of pEF1α-PuroR-P2A-hmPIGA-EIF1pA cassette flanked by loxM and loxP sites (**Table S5**). Sleeping Beauty inverted terminal repeats (ITRs) were amplified from pMH005 (Ribeiro-Dos-Santos et al. 2022) using primers oRB_277 + oRB_278 for ITR(L) and oRB_279 + oRB_287 for ITR(R). LP-PIGA3 was cloned using a *Bsa*I Golden Gate reaction in which ITRs were cloned outside the LP region of pLP303a.

To target LPs to specific genomic loci, Cas9 gRNAs (**Table S8)** were cloned into pSpCas9(BB)-2A-Puro (pCas9-Puro, Addgene #62988) or pSpCas9(BB)-2A-GFP (pCas9-GFP, Addgene #48138) plasmids using *BbsI* Golden Gate reactions as described (Ran et al. 2013). LP homology arms (HAs) corresponding to the genomic sequence flanking the Cas9 cut sites were amplified from a BAC (**Table S6)**. HAs were cloned flanking the LoxM and LoxP sites in pLP-PIGA2 using a *Bsa*I Golden Gate reaction or flanking the ITRs in pLP-PIGA3 using a *BsmB*I Golden Gate reaction (**Table S7**). DNA suitable for transfection was prepped using the ZymoPURE maxiprep kit (Zymo Research D4203) according to the manufacturer's protocol.

Payload assembly

All payloads were assembled into pLM1110 (Brosh et al. 2021) (Addgene #168460) or pRB051 (a derivative of pLM1110). Both are multifunctional YAC/BAC (yeast and bacterial artificial chromosomes) vectors supporting low-copy DNA propagation, selection (LEU2), and efficient homology-dependent recombination in yeast; low-copy propagation, selection (KanR), and copy

number induction in TransforMax EPI300 E. coli (Wild et al. 2002), and transient visualization and selection (eGFP-BSD) in mammalian cells (Brosh et al. 2021).

Four different strategies were used to assemble payloads into these YAC/BACs (**Table S1**), using different primers (**Table S5)**, gRNAs (**Table S8),** and synthetic DNA fragments (**Table S9**):

1) *Yeast CRISPR/Cas9 Assembly*

Desired segments of *Sox2* BAC RP23-144O8 were released by an *in vitro* CRISPR/Cas9 digestion using a pair of synthetic gRNAs and recombinant Cas9 and assembled into *Bsa*I-digested pLM1110 as previously described (Brosh et al. 2021).

2) *Yeast Assembly*

PCR amplicons or synthetic DNA (IDT) tiling the desired payload with >75 bp overlap were assembled into pLM1110. Yeast cells were transformed with 20-50 ng I-SceI-digested pLM1110, 100 ng of each DNA fragment, and 50 ng terminal linker fragments (~400 bp gBlocks, IDT) to enable homologous recombination-dependent assembly followed by selection for LEU^+ phenotype.

3) *Yeast CRISPR/Cas9 Editing*

Existing payloads were modified in yeast using CRISPR/Cas9 and synthetic linker fragment(s) (IDT) to mediate up to two deletions, insertions, or inversions simultaneously (DiCarlo et al. 2013):

i. A yeast strain carrying a parental payload YAC/BAC was pre-transformed with 100 µg SpCas9 expression vector pNA0519 (carrying a ScHIS3 marker) or pCTC019 (carrying a SpHIS5 marker), and selected for a LEU^+/HIS^+ phenotype. Yeast were then transformed again with 100 µg single/dual gRNA expression plasmid pNA0304/pNA0308 (Xie et al. 2018) carrying a URA3 marker and with 100 µg synthetic linker fragment(s), and selected for a $LEU^+/HIS^+/URA^+$ phenotype. gRNAs (**Table S8)** were cloned into the single/dual gRNA yeast expression plasmids pNA304/pNA0308, respectively, using *Hind*III and/or *Not*I Gibson Assembly reactions as described (Xie et al. 2018). Yeast Cas9 expression vector pNA0519 pRS413-TEF1p-Cas9 was derived from pNA0306 (Agmon et al. 2020) by subcloning the TEF1p-Cas9-CYC1t cassette into a pRS413 (ATCC 87518) in order to replace the LEU2 marker with HIS3 (Pinglay et al. 2021).

ii. Alternatively, yeast cells were transformed with pYTK-Cas9 plasmids (carrying HIS3 marker) that co-express SpCas9 and a gRNA and with synthetic linker fragment(s), and

selected for LEU^+/HIS^+ phenotype. gRNAs (**Table S8)** were cloned into pYTK-Cas9 using a *BsmB*I Golden Gate reaction.

4) Golden Gate Assembly (GGA)

DNA fragments were assembled into pLM1110 using *Bsa*I or into pRB051 using *Esp3I*. pRB051 was cloned by PCR-amplifying the RFP transcriptional unit from pLM1110 using primers oRB_564 + oRB_565, digesting the product with *Esp3*I, and performing a *Bsa*I Golden Gate Assembly into pLM1110. For payload assembly, 100 ng of pLM1110 or pRB051 was mixed with 20 ng of each DNA fragment containing terminal *Bsa*I or *Esp3*I sites, respectively, designed to mediate assembly with neighboring fragments (DNA fragments were sourced either from synthetic fragments (IDT) or from existing payloads by PCR-amplification), and with 0.4 µL *Bsa*I-HF V2 (NEB R3733S) or *Esp3*I (NEB R0734S), 1.5 µL 1 mg/mL BSA, 1 µL T4 Quick Ligase (NEB M2200S) and 1.5 µL T4 Ligase Buffer (NEB) in a total volume of 15 µL. Reactions were cycled 25 times between 37 °C (3 min) and 16 °C (4 min), heat inactivated at 50 °C (5 min) and 80 °C (5 min) before transforming into TransforMax EPI300 cells.

Yeast transformation

Yeast transformations were performed using the Lithium acetate method (Gietz and Schiestl 2007) with *Saccharomyces cerevisiae* strain BY4741. 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). YAC/BACs were isolated from candidate yeast clones using the Zymo yeast miniprep I protocol (Zymo Research D2001) and transformed into TransforMax EPI300 bacteria cells by electroporation using the manufacturer's protocol (Lucigen EC300150).

Payload DNA screening and preparation

For initial verification, TransforMax EPI300 bacterial 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. YAC/BACs smaller than 20 kb were isolated using the Zyppy Plasmid Miniprep Kit (Zymo Research D4020).

For larger YAC/BACs, crude DNA extraction was performed as follows: 2-3 mL of induced culture was spun down, resuspended with 300 µL RNase A-supplemented Buffer P1 (QIAGEN 19051), and topped with 300 µL of Buffer P2 (QIAGEN 19052). The suspensions were inverted 10 times, incubated at room temperature for 2 min, topped with 300 µL Buffer P3 (QIAGEN 19053), inverted 10 times and spun down at 12000 RPM for 5 min. Supernatants were transferred to new 2 mL tubes, topped with 900 µL isopropanol, mixed by inverting, and spun down at 12000 RPM for 5 min. Supernatants were discarded, pellets were washed with 500 µl of 70% EtOH, and spun down at 12,000 RPM for 1 min. Supernatants were discarded and the DNA pellets were air-dried and resuspended in 30 µL TE buffer. Tubes were spun down at 12000 RPM for 1 min and clear supernatants transferred to new 1.5 mL tubes.

Payload assembly verification was performed by diagnostic digestion with restriction enzymes, PCR verification of new junctions, and Sanger sequencing, depending on the payload. All YAC/BACs were subsequently sequenced to high coverage depth.

Sequence-verified YAC/BAC 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 bacteria using the ZymoPURE maxiprep kit (Zymo Research D4203) for YAC/BACs <20 kb or using the Nucleobond XtraBAC kit (Takara 740436) for YAC/BACs >20 kb. YAC/BAC preps were stored at 4 °C.

mESC culture

C57BL6/6J × CAST/EiJ (BL6xCAST) mESCs were cultured as described (Brosh et al. 2021). Specifically, 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 on average twice per week. MK6 C57BL/6J mESCs were provided by the NYU Rodent Genetic Engineering Laboratory.

mESC genome engineering

Landing pad integrations were performed (**Table S3)** using the Neon Transfection System as previously described into BL6xCAST Δ*Piga* mESCs, in which the endogenous *Piga* gene was deleted (Brosh et al. 2021). LP-PIGA A1 mESCs, in which LP-PIGA replaces the BL6 *Sox2* allele were previously described (Brosh et al. 2021).

LP-PIGA2 integration at *Sox2* (replacing a 143-kb genomic region) was performed using 5 µg of pLP-PIGA2 and 2.5 µg of each pCas9-GFP plasmid expressing BL6-specific gRNAs mSox2-5p-1 and mSox2-3p-5, 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.

LP-PIGA3 integration at the *Sox2* LCR (replacing a 41-kb genomic region) was performed similarly with pCas9-Puro plasmids expressing the non-allele specific gRNA mSox2-DHS18-19 and the BL6-specific gRNA mSox2-3p-5, followed by 1 µg/mL puromycin selection. PCR genotyping and Capture-seq validation were used to identify clone LP-LCR H1 (**Fig. S1b** and **Fig. S1d**), to which a few LCR payloads were delivered (**Table S3**). DELLY analysis (**Methods**) later identified a 7.7 kb deletion (**Fig. S1d**) in the CAST allele including the mSox2-DHS18-19 gRNA site in a small number of payload clones derived from LP305_H01. We therefore isolated LP-LCR C1, and confirmed it did not harbor the CAST allele deletion (**Fig. S1 c-d**), for use in subsequent LCR deliveries.

Payload deliveries were performed as previously described (Brosh et al. 2021) with 1-5 × 10^6 mESCs, 1-10 µg YAC/BAC payload DNA and 2-5 µg pCAG-iCre plasmid (Addgene #89573) per transfection, depending on the payload size (larger payloads required more cells, PL DNA and pCAG-iCre plasmid to obtain sufficient correct mESC clones). 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 6 or 7 post-transfection.

For both LP integration and payload delivery, approximately 9 days post-transfection, individual mESC clones were manually picked into gelatinized 96-well plates prefilled with 100 µL 80/20 media. Two days post-picking, clones were replicated into two gelatinized 96-well plates at 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 (Brosh et al. 2021) 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. 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 YAC/BAC backbone, and for large (>100 kb) payloads, allele-specific assays to detect delivered regions of the *Sox2* locus. Genotyping primers are listed in **Table S10 and Table S11**.

High-throughput sequencing verification

Illumina sequencing libraries were prepared from purified DNA using:

1. A previously described method (Brosh et al. 2021)
2. One-Pot dsDNA
3. NEBNext Ultra FS II kit (NEB E7805L)

Hybridization capture (Capture-seq) for targeted resequencing of engineered regions was performed as previously described (Brosh et al. 2021). Biotinylated bait was generated using nick translation from BACs covering the *Sox2* locus (RP23-144O8 or RP23-274P9; see **Table S6**), landing pad plasmid (LP-PIGA2 or LP-PIGA3), payload plasmid backbone, pCAG-iCre and pSpCas9 plasmids. Bait sets and sequencing statistics are listed in **Table S12**.

Sequencing libraries were sequenced in paired-end mode on an Illumina NextSeq 500 operated at the Institute for Systems Genetics. Reads were demultiplexed with Illumina bcl2fastq v2.20 requiring perfect match to the indexing BC sequence. All whole-genome sequencing and Capture-seq data were processed using a uniform mapping pipeline. Illumina sequencing adapters were trimmed with Trimmomatic v0.39 (Bolger et al. 2014). Reads were aligned using BWA v0.7.17 (Li and Durbin 2009) to the appropriate reference genome (GRCm38/mm10 or GRCh38/hg38), including unscaffolded contigs and alternate references, as well as to independent custom references for relevant vectors. PCR duplicates were marked by samblaster v0.1.24 (Faust and Hall 2014). Per-base coverage depth tracks were generated using BEDOPS v2.4.4 (Neph et al. 2012).

Variant calling was performed using a standard pipeline based on bcftools v1.14 (Bonfield et al. 2021):

```
bcftools mpileup --excl-flags UNMAP,SECONDARY,DUP --redo-BAQ --adjust-MQ 50 --
gap-frac 0.05 --max-depth 10000 --max-idepth 200000 -a DP,AD --output-type u |
bcftools call --keep-alts --ploidy [1|2] --multiallelic-caller -f GQ --output-type u |
bcftools norm --check-ref w --output-type u |
bcftools filter -i “INFO/DP>=10 & QUAL>=10 & GQ>=99 & FORMAT/
DP>=10” --set-GTs . --output-type u |
bcftools view -i ’GT=”alt”’ --trim-alt-alleles --output-type z
```

Bcftools call --ploidy was set to 1 for custom references and 2 for autosomes in mm10.

Large deletions and structural variants were called using DELLY v0.8.7 excluding telomeric and centromeric regions (Rausch et al. 2012). Variants were required to PASS filters, to have at least 10 paired-end reads and 20% of paired-end reads supporting the variant allele. DELLY results for YAC/BAC and capture sequencing data were inspected manually and used to choose clones for successive analysis.

Data were visualized and explored using the University of California, Santa Cruz Genome Browser. The full processing pipeline is available at https://github.com/mauranolab/mapping.

Payload YAC/BAC sequence verification

To identify potential sample swaps or contamination, mean sequencing coverage was calculated for the genomic regions corresponding to the payload sequence and to the non-payload

sequence, defined as the engineered *Sox2* regions that do not overlap the payload. To minimize coverage abnormalities associated with sequence termini, 400 bp were clipped from each end of continuous genomic regions. Mean coverage was normalized by the payload YAC/BAC backbone coverage. We expect a normalized coverage of 1±0.4 for the payload region and <0.1 for the non-payload region. Human orthologous sequences and regions smaller than 200 bp after clipping were ignored. Coverage analysis results and quality control (QC) calls for each sample are reported in **Table S13**.

To detect payload DNA variants that might have been introduced during assembly or propagation in yeast or bacteria, we analyzed high confidence variant calls relative to the payload custom reference having sequencing depth above 50 and quality score above 100 (DP ≥ 50 & QUAL ≥ 100) (**Table S14)**.

mESC clone sequence verification

To verify correct genomic engineering of mESC clones, mean sequencing coverage was calculated for the genomic regions corresponding to the payload and non-payload (as defined above); and for the payload YAC/BAC backbone, landing pad (LP), LP backbone, and pCAG-iCre custom references. Terminal clipping of 400 bp was performed as described above for all genomic regions. Coverage was normalized by the mean genomic coverage of the engineered region flanks (regions captured by the *Sox2* bait, but unmodified) such that a single copy corresponds to a value of ~0.5. For LP clones we expect normalized coverage of 0.5±0.25 for the non-payload region and <0.1 for LP backbone. For payload clones we expect normalized coverage of 1±0.25 for the payload region, 0.5±0.25 for the non-payload region and <0.1 for the LP, payload YAC/BAC backbone and pCAG-iCre. Human orthologous sequences and regions smaller than 200 bp after clipping were ignored. Coverage analysis results and QC calls for each clone are reported in **Table S15**.

We further verified the presence of BL6 allele variants, which are lost in regions replaced by landing pads and restored by delivered payloads (**Table S15)** by calculating the mean proportion of reads supporting the reference (BL6) allele (propREF) for the genomic regions described above. Values of 0.5±0.2 and 0±0.2 were expected for payload and non-payload regions, respectively. Regions with less than 10 variants were ignored.

To verify genomic integration sites, we detected sequencing read pairs mapping to two different reference genomes using *bamintersect* as previously described (Brosh et al. 2021) 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 (<1 and 5 reads/10M reads sequenced for LP and payload samples, respectively) were excluded. *Bamintersect* junctions were classified hierarchically based on position (**Table S16)** and reported in **Table S17**.

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

RNA was isolated from fresh or frozen cells using the Qiagen RNeasy-mini protocol. Since *Sox2* has no introns, additional steps were taken to ensure that RNA was not contaminated with trace residual genomic DNA. DNase treatment was performed on extracted RNA with the Turbo DNA-free kit (Ambion, Fisher AM1907) using the "rigorous DNAse treatment" protocol prescribed by the manufacturer. Following DNase treatment, cDNA was synthesized from 1-2 µg total RNA with the Multiscribe High-Capacity cDNA Reverse Transcription Kit (Fisher 4368814), including a "-RT" no-reverse-transcriptase control for a subset of samples.

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). Primers were designed to detect *Sox2* in an allele-specific manner (**Table S18**). An Echo 550 liquid handler was used for loading as described above. Thermal cycling parameters were as follows: 3 min pre-incubation at 95 °C, followed by 40 amplification cycles of 3 sec at 95 °C, 20 sec at 57 °C and 20 sec at 72 °C. A subset of -RT control samples were verified for lack of amplification.

Assays were performed in duplicate and replicate wells were averaged. ΔCt values for the BL6 and CAST *Sox2* alleles were computed relative to *Gapdh*. Replicate measurements of the same clone across different plates were averaged. *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 (WT) 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 WT payload samples (full locus or LCR) (**Table S4**).

Modeling

We fitted linear regression models to predict *Sox2* expression based on DHS composition and configuration using R v3.5.2 (R Core Team 2018). Models were compared based on Bayesian Information Criterion (BIC) to evaluate the performance of predictor combinations. The composition of proximal (DHSs 1-8 and DHSs 10-15) and core LCR (DHS23, DHS24, CTCF25, and DHS26) regulatory elements was represented by their copy number in the payload. Regulatory element configuration was represented by indicator variables of their relative order (DHS24-DHS23 and CTCF25-DHS24; read as DHS24 before DHS23 and CTCF25 before DHS24) and sequence inversion (inv_DHS23, inv_DHS24, inv_CTCF25, and inv_DHS26). Payloads which did not map clearly onto these features were excluded from the analysis (*Δ26-27 & 28* and *Δ20-23.3*, and the synthetic payloads in **Fig. 6**). Interaction between two elements was represented by indicator variables of both elements' presence (DHS23and24, DHS23and26, and CTCF25andDHS26).

Transcription factor motif analysis

We used motif matches previously derived from scanning the reference genomes using FIMO v4.10.244 with TF motifs as previously described (Halow et al. 2021)

ACKNOWLEDGEMENTS

We thank Brendan Camellato, Florrie Zhu, Leslie Mitchell, Sudarshan Pinglay, Weimin Zhang, and Yu Zhao for help with yeast assembly. This work was partially funded by National Institutes of Health (NIH) grants RM1HG009491 (to J.D.B.) and R35GM119703 (to M.T.M.).

Competing interests

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: Sangamo, Inc., Modern Meadow, Inc., Rome Therapeutics, Inc., Sample6, Inc., Tessera Therapeutics, Inc. and the Wyss Institute.

REFERENCES

Agmon N, Temple J, Tang Z, Schraink T, Baron M, Chen J, Mita P, Martin JA, Tu BP, Yanai I, et al. 2020. Phylogenetic debugging of a complete human biosynthetic pathway transplanted into yeast. *Nucleic Acids Res* **48**: 486–499.

Alexander JM, Guan J, Li B, Maliskova L, Song M, Shen Y, Huang B, Lomvardas S, Weiner OD. 2019. Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. *Elife* **8**.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of cellular biochemistry* **94**: 890–898.

Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R. 2003. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes & Development* **17**: 126–140.

Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, Lachanski CV, Gillis DR, Phillips-Cremins JE. 2017. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* **27**: 1139–1152.

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

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

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A, et al. 2017. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**: 557-572.e24.

Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, Keane T, Davies RM. 2021. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10**: giab007.

Bothma JP, Garcia HG, Ng S, Perry MW, Gregor T, Levine M. 2015. Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. *eLife* **4**.

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.

Chakraborty S, Kopitchinski N, Eraso A, Awasthi P, Chari R, Rocha PP. 2022. High affinity enhancer-promoter interactions can bypass CTCF/cohesin-mediated insulation and contribute to phenotypic robustness. 2021.12.30.474562. https://www.biorxiv.org/content/10.1101/2021.12.30.474562v1 (Accessed June 12, 2022).

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.

DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. 2013. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res* **41**: 4336–4343.

Dickson J, Gowher H, Strogantsev R, Gaszner M, Hair A, Felsenfeld G, West AG. 2010. VEZF1 elements mediate protection from DNA methylation. *PLoS Genetics* **6**: e1000804.

Dukler N, Gulko B, Huang Y-F, Siepel A. 2016. Is a super-enhancer greater than the sum of its parts? *Nat Genet* **49**: 2–3.

Evans T, Felsenfeld G, Reitman M. 1990. Control of globin gene transcription. *Annual review of cell biology* **6**: 95–124.

Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**: 2503–2505.

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

Grosveld F, van Assendelft GB, Greaves DR, Kollias G. 1987. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**: 975–985.

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.

Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen LLP, Kassouf MT, Oudelaar AM, Sharpe JA, Suciu MC, et al. 2016. Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics* **48**: 895–903.

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

Huang H, Zhu Q, Jussila A, Han Y, Bintu B, Kern C, Conte M, Zhang Y, Bianco S, Chiariello AM, et al. 2021. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nat Genet* **53**: 1064–1074.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**: 430–435.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li Q, Emery DW, Han H, Sun J, Yu M, Stamatoyannopoulos G. 2005. Differences of globin transgene expression in stably transfected cell lines and transgenic mice. *Blood* **105**: 3346–3352.

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.

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schübeler D. 2011. Identification of genetic elements that autonomously determine DNA methylation states. *Nature Genetics* **43**: 1091–1097.

Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA. 2012. Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genetics* **8**: e1002599.

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.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**: 1919–1920.

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.

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.

Pinglay S, Bulajić M, Rahe DP, Huang E, Brosh R, German S, Cadley JA, Rieber L, Easo N, Mahony S, et al. 2021. Synthetic genomic reconstitution reveals principles of mammalian Hox cluster regulation. 2021.07.07.451065. https://www.biorxiv.org/content/10.1101/2021.07.07.451065v1 (Accessed April 29, 2022).

Pott S, Lieb JD. 2014. What are super-enhancers? *Nature Genetics* **47**: 8–12.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333–i339.

Ribeiro-Dos-Santos AM, Hogan MS, Luther RD, Brosh R, Maurano MT. 2022. Genomic context sensitivity of insulator function. *Genome Res* gr.276449.121.

Scholes C, Biette KM, Harden TT, DePace AH. 2019. Signal Integration by Shadow Enhancers and Enhancer Duplications Varies across the Drosophila Embryo. *Cell Rep* **26**: 2407-2418.e5.

Shin HY, Willi M, Yoo KH, Zeng X, Wang C, Metser G, Hennighausen L. 2016. Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nature Genetics*.

Singh G, Mullany S, Moorthy SD, Zhang R, Mehdi T, Tian R, Duncan AG, Moses AM, Mitchell JA. 2021. A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells. *Genome Res* **31**: 564–575.

Taylor T, Sikorska N, Shchuka VM, Chahar S, Ji C, Macpherson NN, Moorthy SD, Mitchell JA, Sexton T. 2022. Transcriptional regulation and chromatin architecture maintenance are decoupled modular functions at the Sox2 locus. 2022.02.09.479674. https://www.biorxiv.org/content/10.1101/2022.02.09.479674v1 (Accessed March 4, 2022).

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.

Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, Cowley SM, Young RA. 2012. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* **482**: 221–225.

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.

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJ, Fisher EMC, Tavaré S, Odom DT. 2008. Species-specific transcription in mice carrying human chromosome 21. *Science* **322**: 434–438.

Xie Z-X, Mitchell LA, Liu H-M, Li B-Z, Liu D, Agmon N, Wu Y, Li X, Zhou X, Li B, et al. 2018. Rapid and Efficient CRISPR/Cas9-Based Mating-Type Switching of Saccharomyces cerevisiae. *G3 (Bethesda)* **8**: 173–183.

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*.

FIGURES

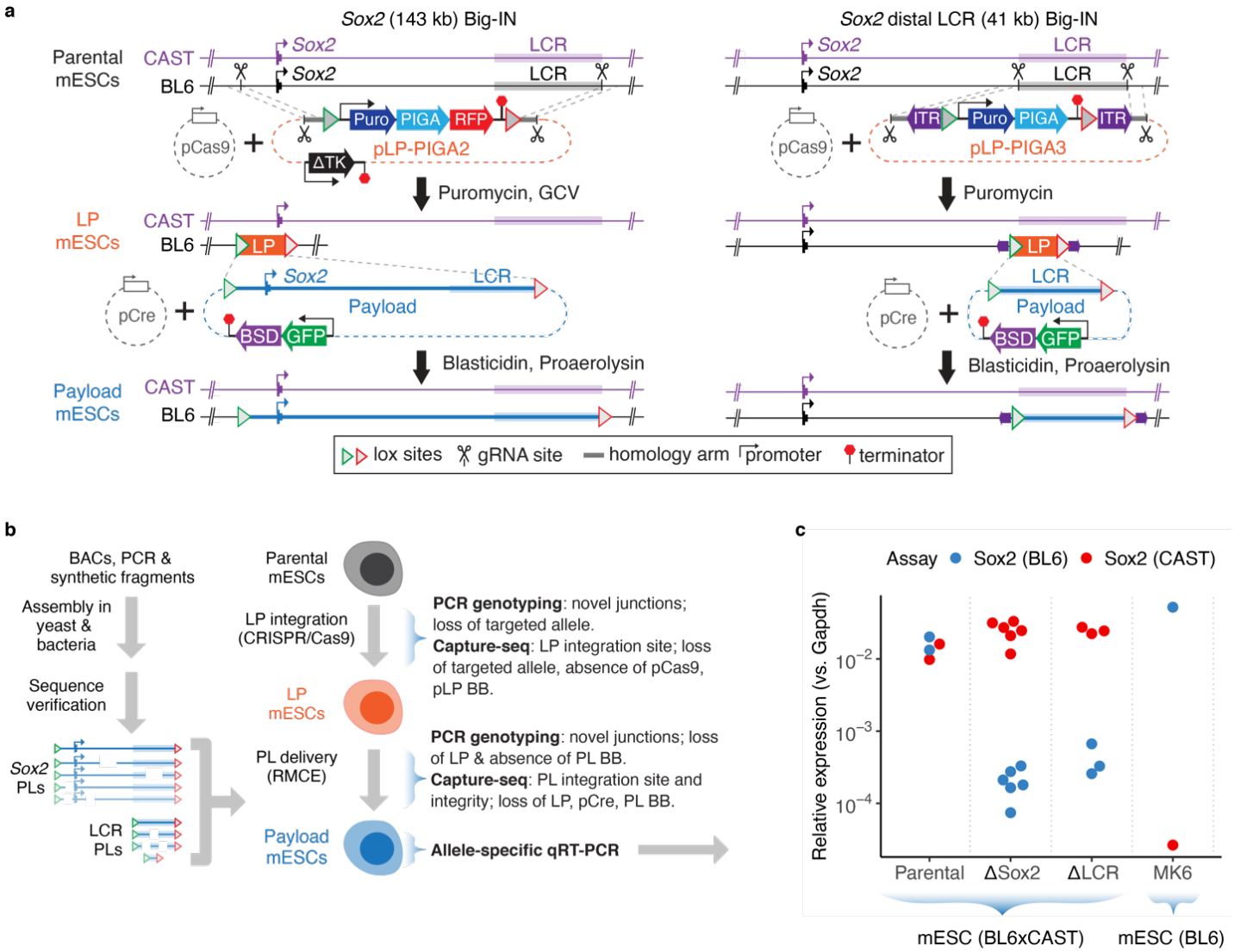

Fig. 1. Engineering the *Sox2* locus.

a. mESC engineering strategy. The BL6 allele of the 143-kb *Sox2* locus (left) or the 41-kb *Sox2* LCR (right) were replaced with LP-PIGA2 or LP-PIGA3, respectively. LP integration was aided by Cas9 and 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 pLP-PIGA2's backbone integration. Cre recombinase-mediated cassette exchange (RMCE) enabled replacement of each LP with a series of payloads. Transfected cells were transiently selected with blasticidin, followed by counterselection of LP mESCs cells with proaerolysin.

b. Schematic of DNA assembly, mESC engineering, verification and analysis pipelines.

c. Allele-specific qRT-PCR assay for *Sox2* expression. BL6 and Castaneus (CAST) expression was measured using allele-specific primers in parental BL6xCAST mESCs, in mESCs with deletion of the 143-kb Sox2 locus (ΔSox2) or 41-kb deletion of the LCR (ΔLCR), as well as in BL6 MK6 mESCs.

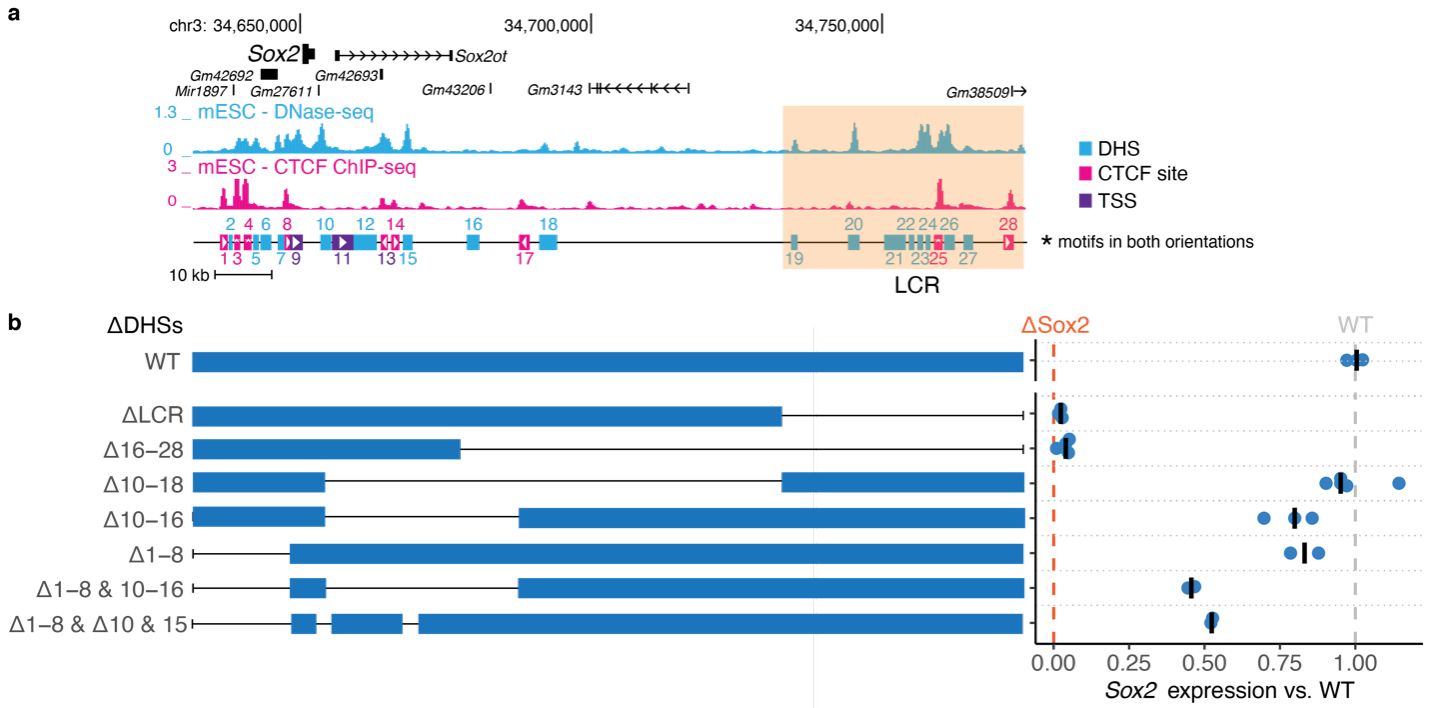

Fig. 2. Regulatory architecture of the *Sox2* locus.

a. Schematic of the *Sox2* locus. Shown are DNase-seq and CTCF ChIP-seq data in mESCs. DHSs, CTCF binding sites, and transcription start sites (TSSs) were numbered from 1 to 28. Orientation of CTCF recognition sequences is indicated where applicable.

b. *Sox2* expression analysis for payloads delivered to the *Sox2* locus in LP-Sox2 mESCs. Blue rectangles demarcate genomic regions included in each payload. Each point represents the expression of the engineered *Sox2* allele in an independent mESC clone. Bars indicate median. Expression was scaled between 0 (*ΔSox2*) to 1 (*WT LCR*).

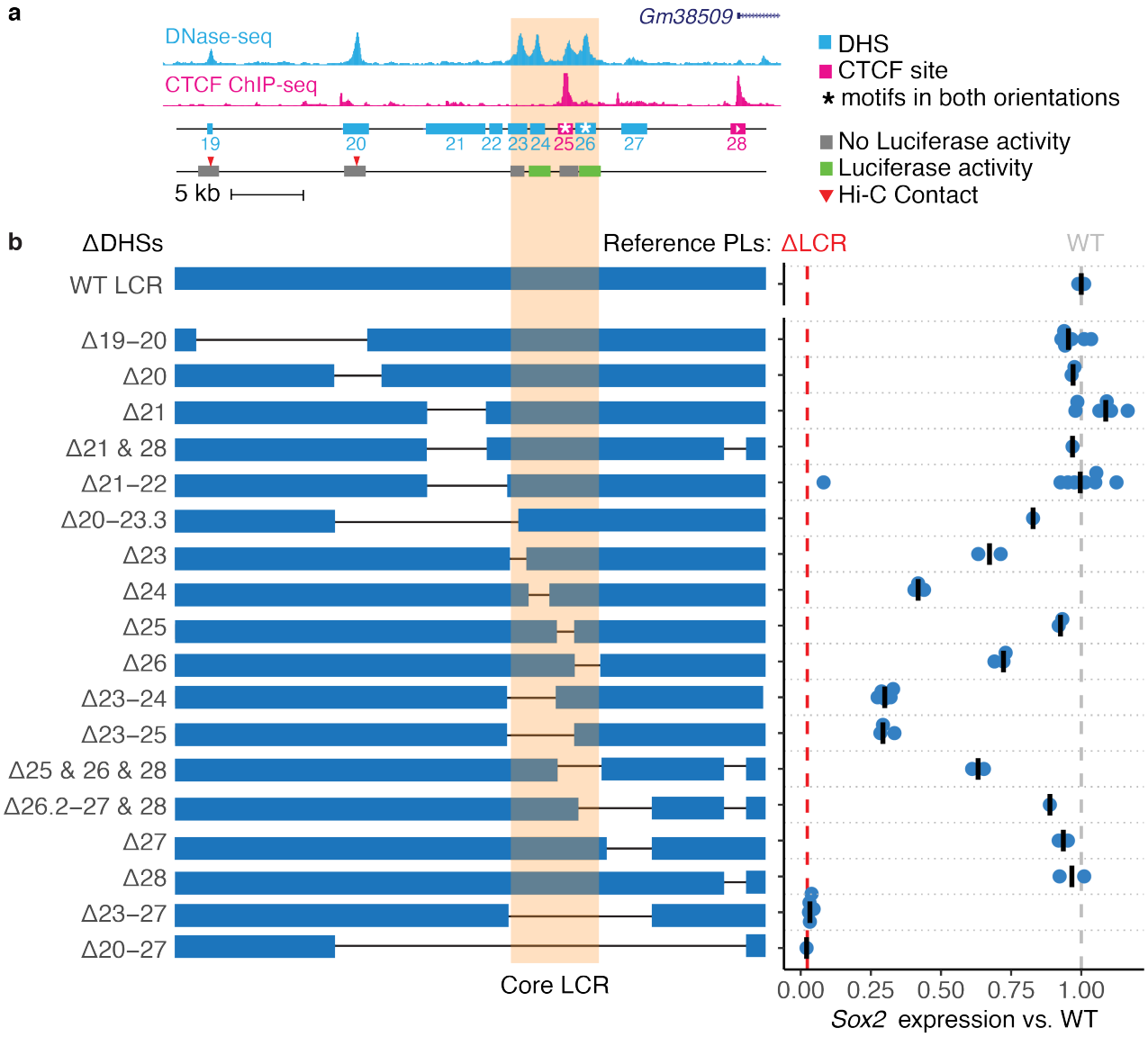

Fig. 3. Mapping essential DHSs within the *Sox2* LCR.

a. Schematic of the *Sox2* LCR showing DNase-seq and CTCF ChIP-seq data in mESCs. Regions previously tested in luciferase reporter assays (Zhou et al. 2014) and sites of Hi-C contact with the *Sox2* promoter (Bonev et al. 2017) are indicated.

b. *Sox2* expression analysis for payloads with deletions of single and multiple DHSs within the core LCR delivered to LP-LCR mESCs. Blue rectangles demarcate genomic regions included in each payload. Each point represents the expression of the engineered *Sox2* allele in an independent mESC clone. Bars indicate median. Expression was scaled between 0 (*ΔSox2*) to 1 (*WT LCR*).

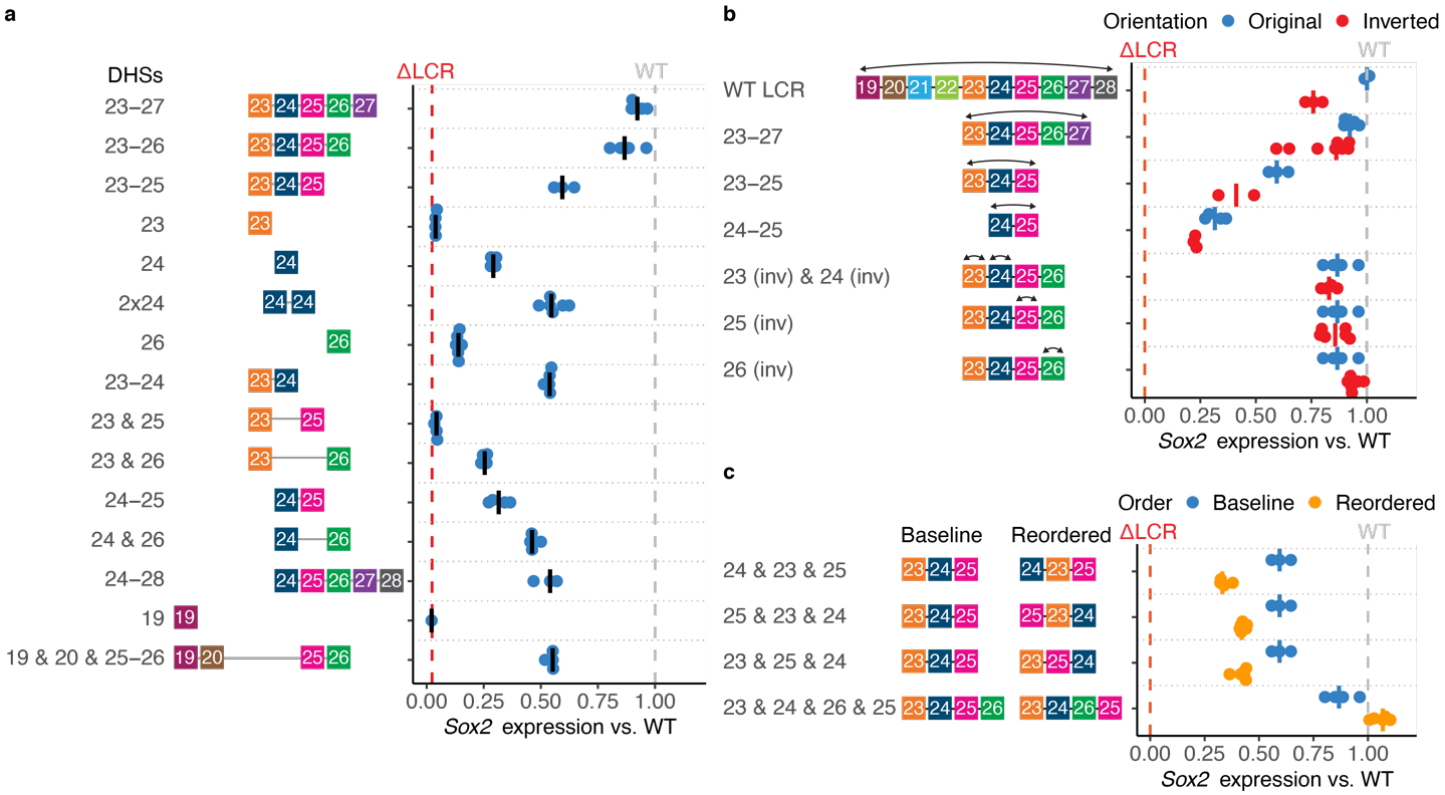

Fig. 4. Context-dependent DHS function within the core LCR.
Sox2 expression analysis for selected LCR DHSs. Each point represents the expression of the engineered *Sox2* allele in an independent mESC clone. Bars indicate median. Expression was scaled between 0 (*ΔSox2*) to 1 (*WT LCR*). Payload structures are shown in **Fig. S3**. **a**. Sufficiency of minimal payloads. **b-c**. Effect of DHS orientation (**b**) and order (**c**). Expression of corresponding baseline payloads is repeated in blue for each payload. Arrows indicate extent of inversion(s) in **b**.

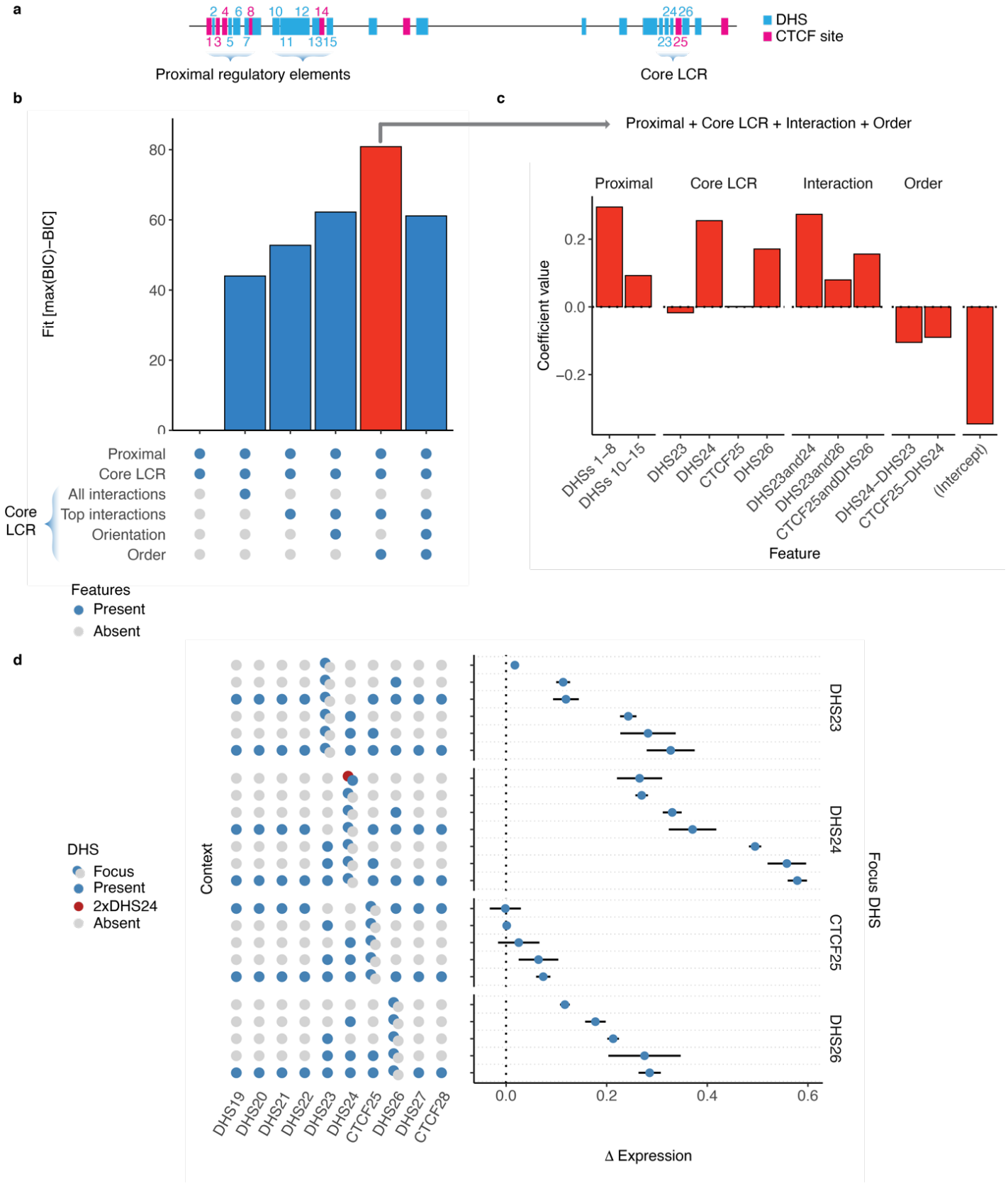
a
2 4 6 8 10 12 14
13 5 7 11 13 15
24 26
23 25
DHS
CTCF site
Proximal regulatory elements
Core LCR
b
Fit [max(BIC)−BIC]
0
20
40
60
80
Proximal
Core LCR
All interactions
Top interactions
Orientation
Order
Core LCR
Features
Present
Absent
c
Proximal + Core LCR + Interaction + Order
Proximal
Core LCR
Interaction
Order
Coefficient value
0.2
0.0
−0.2
DHSs 1–8
DHSs 10–15
DHS23
DHS24
CTCF25
DHS26
DHS23and24
DHS23and26
CTCF25andDHS26
DHS24–DHS23
CTCF25–DHS24
(Intercept)
Feature
d
DHS
Focus
Present
2xDHS24
Absent
Context
DHS19
DHS20
DHS21
DHS22
DHS23
DHS24
CTCF25
DHS26
DHS27
CTCF28
0.0
0.2
0.4
0.6
Δ Expression
DHS23
DHS24
CTCF25
DHS26
Focus DHS

Fig. 5. Modeling the regulatory architecture of the *Sox2* locus.

a-c. Analysis of linear regression models to predict *Sox2* expression. **a**. Models include features indicating the presence of selected DHSs. DHSs proximal to the gene were grouped and encoded as two features (DHSs 1-8 and DHSs 10-15). **b**. Model performance measured by Bayesian Information Criteria (BIC) and presented as Fit (difference from maximum BIC). Feature sets included in each model are indicated by blue dots. *Interaction* includes all interaction terms for core LCR DHSs or only top interactions (DHS23and24, DHS23and26, and CTCF25andDHS26). *Orientation* includes features indicating inversion of each core LCR DHS. *Order* includes features indicating that DHS24 is positioned *Sox2*-proximal to DHS23 (DHS24-DHS23), and DHS25 *Sox2*-proximal to DHS24 (CTCF25-DHS24). Red indicates the model with best fit. **c**. Coefficients for the best model. **d**. Context-sensitive activity of DHSs in the core LCR. ΔExpression was computed as the difference in expression between payload pairs differing solely by the presence of each focus DHS. The presence (blue dots) or absence (gray dots) of each DHS is indicated; two copies of DHS24 are shown in red. Points indicate mean and bars indicate SD across all pairwise combinations of clones.

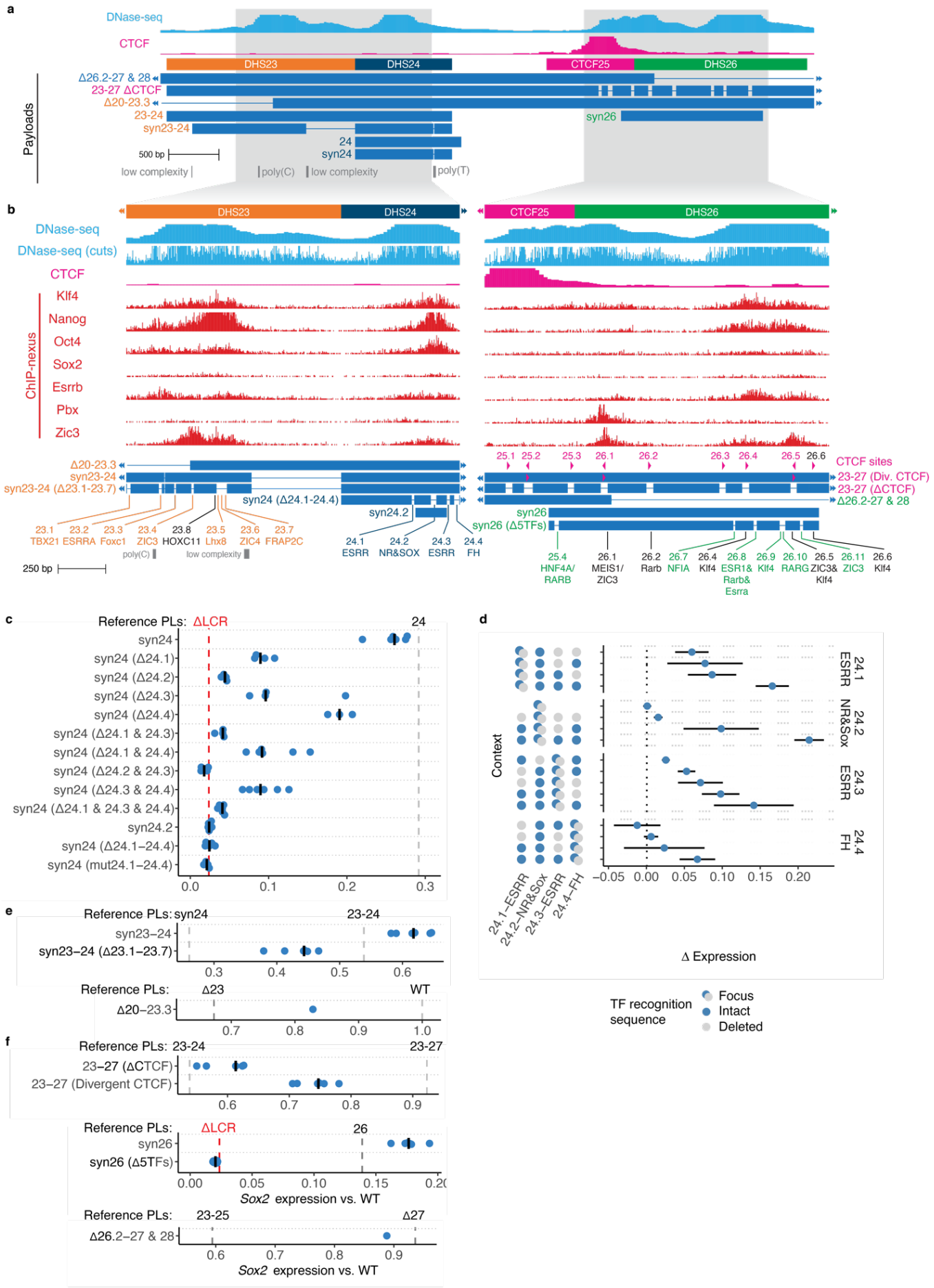
a
DNase-seq
CTCF
DHS23
DHS24
CTCF25
DHS26
Payloads
Δ26.2-27 & 28
23-27 ΔCTCF
Δ20-23.3
23-24
syn23-24
24
syn24
syn26
500 bp
low complexity
poly(C)
poly(T)
b
DNase-seq (cuts)
ChIP-nexus
Klf4
Nanog
Oct4
Sox2
Esrrb
Pbx
Zic3
syn23-24 (Δ23.1-23.7)
syn24 (Δ24.1-24.4)
syn24.2
syn26 (Δ5TFs)
CTCF sites
23-27 (Div. CTCF)
23-27 (ΔCTCF)
23.1 TBX21
23.2 ESRRA
23.3 Foxc1
23.4 ZIC3
23.8 HOXC11
23.5 Lhx8
23.6 ZIC4
23.7 FRAP2C
24.1 ESRR
24.2 NR&SOX
24.3 ESRR
24.4 FH
25.4 HNF4A/ RARB
26.1 MEIS1/ ZIC3
26.2 Rarb
26.7 NFIA
26.4 Klf4
26.8 ESR1& Rarb& Esrra
26.9 Klf4
26.10 RARG
26.5 ZIC3& Klf4
26.11 ZIC3
26.6 Klf4
250 bp
c
Reference PLs: ΔLCR
24
syn24
syn24 (Δ24.1)
syn24 (Δ24.2)
syn24 (Δ24.3)
syn24 (Δ24.4)
syn24 (Δ24.1 & 24.3)
syn24 (Δ24.1 & 24.4)
syn24 (Δ24.2 & 24.3)
syn24 (Δ24.3 & 24.4)
syn24 (Δ24.1 & 24.3 & 24.4)
syn24.2
syn24 (Δ24.1–24.4)
syn24 (mut24.1–24.4)
d
Context
24.1–ESRR
24.2–NR&Sox
24.3–ESRR
24.4–FH
Δ Expression
TF recognition sequence
Focus
Intact
Deleted
e
Reference PLs: syn24
23-24
syn23–24
syn23–24 (Δ23.1–23.7)
Reference PLs:
Δ23
WT
Δ20–23.3
f
Reference PLs: 23-24
23-27
23–27 (ΔCTCF)
23–27 (Divergent CTCF)
Reference PLs:
ΔLCR
26
syn26
syn26 (Δ5TFs)
Sox2 expression vs. WT
Reference PLs:
23-25
Δ27
Δ26.2–27 & 28
Sox2 expression vs. WT

Fig. 6. TF-scale dissection of DHS function.

a. Schematic of the *Sox2* LCR. Shown are DNase-seq and CTCF ChIP-seq data in mESCs. Selected payloads are illustrated. Gray regions include poly(C) and poly(T) sequences shortened in synthetic (syn) versions of DHSs 23 or 24.

b. Enlargement of engineered regions within DHS23 to DHS24 (left) and CTCF25 to DHS26 (right). DNase-seq windowed density and per-nucleotide cleavages, and ChIP-nexus data for selected TFs in mESC are shown. Payload schemes indicate surgical deletions of predicted TF recognition sites. Magenta triangles denote CTCF sites and their native orientation.

c-f. *Sox2* expression analysis for perturbations of TF recognition sequences within the core LCR. Payloads include deletions (Δ) of TF or CTCF recognition sequences shown in **b**. Each point represents the expression of the engineered *Sox2* allele in an independent mESC clone relative to baseline payloads (dashed vertical lines). Bars indicate median. Expression was scaled between 0 (*ΔSox2*) to 1 (*WT LCR*), and vertical gray lines indicate median expression of relevant baseline payloads.

c. *syn24.2* contains a minimal region surrounding DHS24.2 (see **b**). *syn24 (mut24.1-24.4)* contains point substitutions that ablate TF recognition sequences 24.1-24.4.

d. Context-sensitive activity of TF sites 24.1-24.4. ΔExpression was computed as the difference in expression between payload pairs differing solely by the presence of each focus TF recognition sequence. The presence (blue dots) or absence (gray dots) of each TF site is indicated. Points indicate mean and bars indicate SD across all pairwise combinations of clones.

f. *23-27 (ΔCTCF)* has 8 of 9 identified CTCF sites surgically deleted. *23-27 (Divergent CTCF)* has surgical inversion of 3 CTCF sites (TF sites 25.2, 26.1 and 26.5) so that all 9 CTCF sites lie in divergent orientation relative to *Sox2*. *syn26 (Δ5TFs)* has 5 TF recognition sequences surgically deleted. Deletions and inversions are detailed in panel **b**.

SUPPLEMENTARY FIGURES

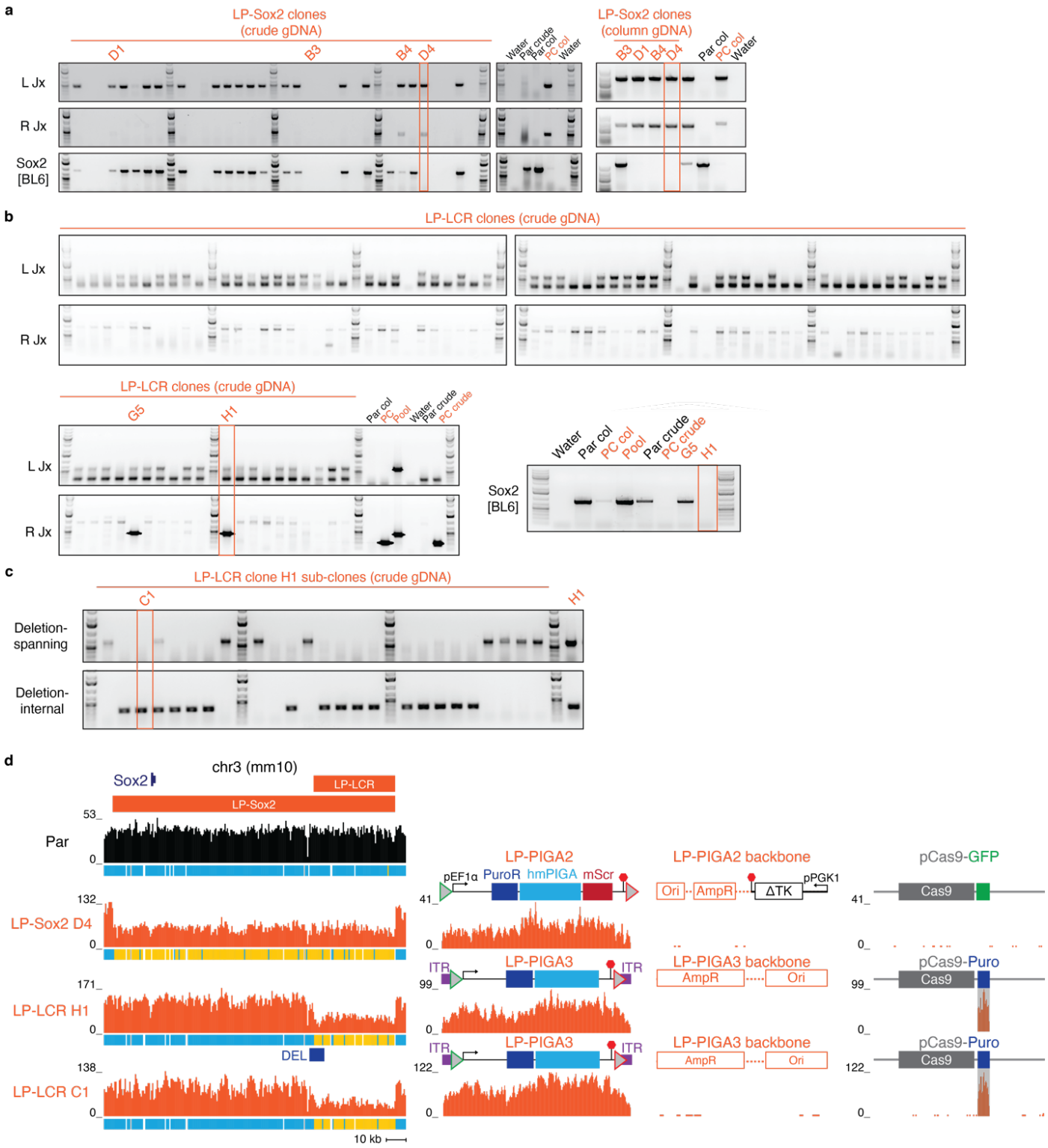

Fig. S1. Landing pad integrations.

a. Screening of BL6xCAST LP-Sox2 clones using PCR genotyping primers targeting the novel right junction (R Jx) and left junction (L Jx), as well as allele-specific primers to detect the loss of the targeted *Sox2* BL6 allele. Right, selected clones were revalidated using column gDNA. Clone D4 (orange rectangle) was selected for further verification by Capture-seq. Par, BL6xCAST Δ*Piga* mESCs; col, column gDNA; PC, BL6xCAST LP-Sox2 A1 mESCs (Brosh et al. 2021). **b**. Screening of BL6xCAST LP-LCR clones using PCR genotyping primers targeting the novel right junction (R Jx) and left junction (L Jx), as well as allele-specific primers to detect the loss of the targeted *Sox2* BL6 allele. Clone H1 (orange rectangle) was selected for further verification by Capture-seq. Par, BL6xCAST Δ*Piga* mESCs; col, column gDNA; PC, BL6xCAST LP-Sox2 A1 mESCs (which do not share the L Jx with LP-LCR clones and should produce a shorter amplicon from the R Jx). **c**. Screening of sub-clones of LP-LCR clone H1 using PCR genotyping primers spanning or internal to the CAST allele deletion. **d.** Capture-seq analysis of parental Δ*Piga* mESCs, LP-Sox2 clone D4, and LP-LCR clones H1 and C1. Reads were mapped to the references indicated above. The PuroR gene in pCas9-Puro cross-maps with the LP-PIGA3 and is shaded gray. Ticks under each coverage track indicate heterozygous (blue) or homozygous non-reference (yellow) variants. ITR, inverted terminal repeats; pEF1α, human EEF1A1 promoter; PuroR, puromycin-resistance gene; hmPIGA, human mini PIGA gene; mScr, mScarlet; pPGK1, human PGK1 promoter.

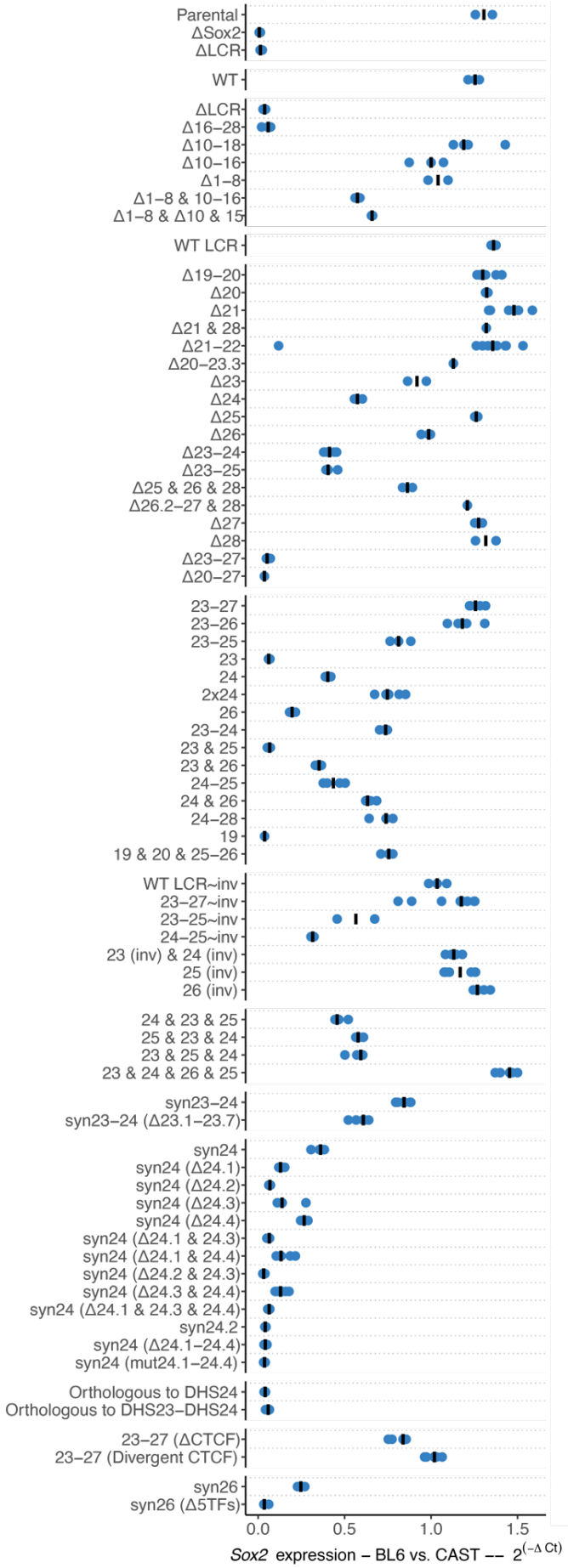

Fig. S2. qRT-PCR results.

Sox2 expression from all mESC clones described in this manuscript. Displayed is the expression of the BL6 allele of *Sox2* relative to the CAST allele, calculated as $2^{\Delta Ct[CAST-BL6]}$. Points represent individual clones and bars indicate median.

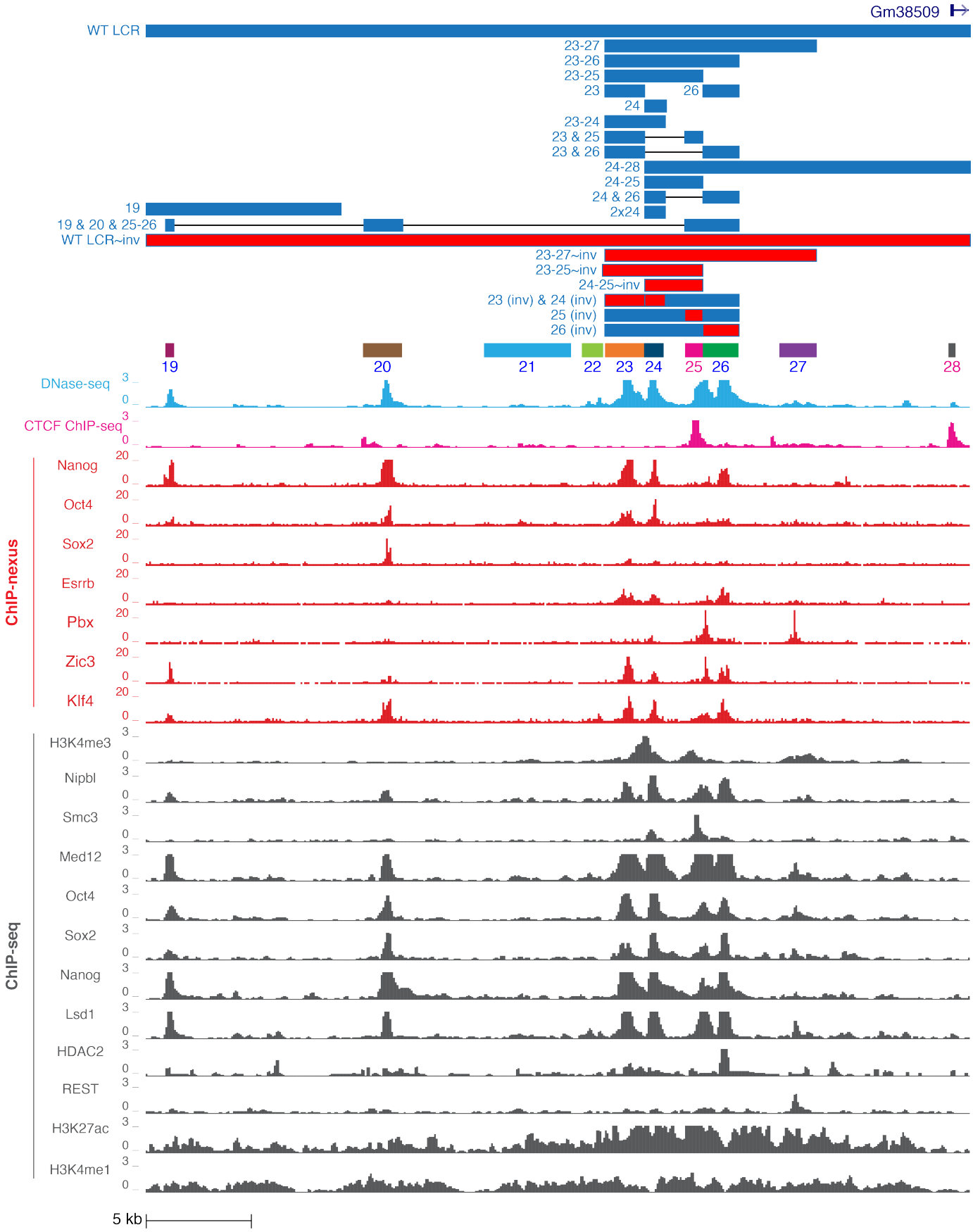

Fig. S3. *Sox2* LCR structure and payloads.

Browser shot for the *Sox2* distal LCR. DHSs and payloads included in **Fig. 3** are demarcated along with DNase-seq, ChIP-seq, and ChIP-nexus tracks.

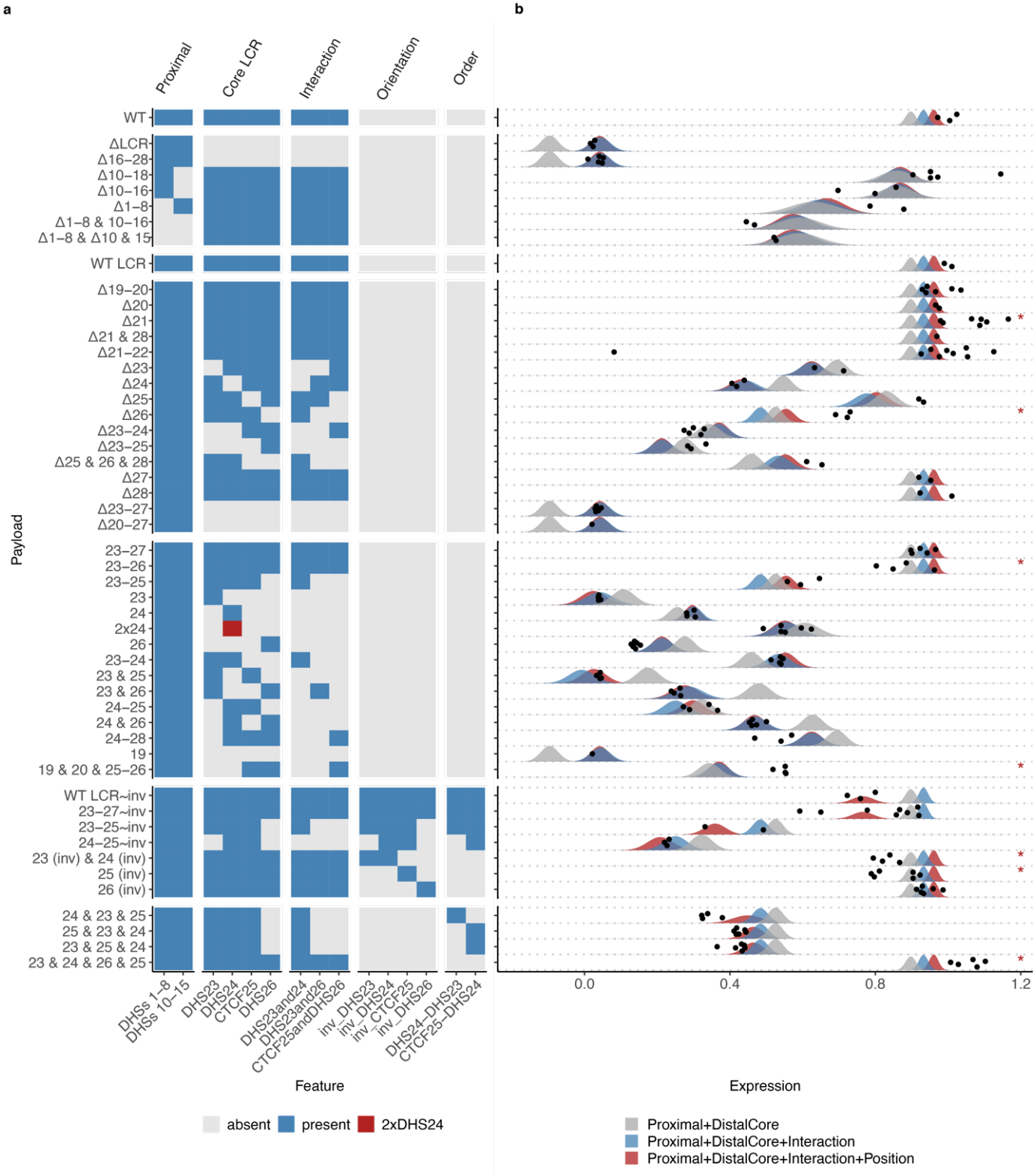

Fig. S4. Linear regression model specifics.

a. Feature composition matrix. Gray, absence; blue, presence; red, two copies.

b. Model predictions vs. actual data. Points represent the observed activity (*Sox2* BL6 allele relative expression) for individual mESC clones. Density curves indicate the expected outcome distribution for each regression calculated based on a Student t distribution using the estimated mean and standard error of each PL. Asterisks denote payloads for which measured data diverge from *Proximal+core LCR+Interaction+Position* prediction by >5 standard errors (SE).

SUPPLEMENTARY TABLES

Table S1. Payloads and assembly details.
Strategies and reagents used for payload YAC/BAC assembly, as well as sequencing data. See **Methods** for details of each assembly strategy, **Table S5** for cloning primer sequences, **Table S8** for gRNA sequences, **Table S9** for synthetic fragments sequences and **Table S4** for conversions of payload IDs to labels used in figures. Sequence-verified unintended payload structures observed in bacteria or in mESCs are indicated in the Assembly strategy field and were assigned n/a values in relevant fields. For two payloads with detected variants (see **Table S14**), two different YAC/BAC clones were delivered to mESCs.

This table is provided as a separate file.

Table S2. Summary of YAC/BAC and mESC clone QC.
Summary table for YAC/BAC and mESC clones QC analyses. See Table S13 and Table S14 for details for YAC/BACs, and **Table S15** and **Table S17** for mESC clones.

	Passed	**Failed**	**No call**
Payload YAC/BACs (n=71):			
Sequencing coverage QC	52 (73.2%)	0	19 (26.8%)[c]
Variant calling	71 (100%)	0	0
Landing pad mESC clones (n=3):			
Genotyping PCR	3 (100%)	0	0
Sequencing coverage QC	3 (100%)	0	0
Sequencing variants QC	3 (100%)	0	0
bamintersect QC	3 (100%)	0	0
Payload mESC clones (n=303)			
Genotyping PCR	303 (100%)	0	0
Sequencing coverage QC	198 (65.3%)	1 (0.3%)[a]	105 (34.7%)[c]
Sequencing variants QC	179 (59.1%)	0	124 (40.9%)[c]
bamintersect QC	299 (98.7%)	4 (1.3%)[b]	0

[a] gDNA sample with LP backbone coverage attributed to unrelated plasmid contamination.

[b] Only one of two junctions was verified. No off-target junctions were reported.

[c] Small regions, regions with small number of variants, and human orthologous sites were not analyzed (see **Table S15** and **Methods** for details).

Table S3. mESC clones.
Summary table for 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 conversions of payload IDs to labels used in figures. “LP-PIGA A1” corresponds to “Bl6/C LP140(Sox2_1g_5g)_A1”, “LP-Sox2 D4” to “BL6/C LP300(Sox2_1g_5g)_201007_D4”, “LP-LCR H1” to “Bl6/C LP305(Sox2_6g_7g)_20200915_H1”, and “LP-LCR C1” to “Bl6/C LP305(Sox2_6g_7g)_20210326_C01”.

This table is provided as a separate file.

Table S4. *Sox2* expression values and activity.
Sox2 expression in engineered mESC clones. The columns Sox2 (CAST) and Sox2 (BL6) list qRT-PCR ΔCt values relative to *Gapdh* for each *Sox2* allele. Fold change indicates the BL6/CAST $2^{\Delta Ct}$ ratio. Activity indicates scaled fold change as detailed in **Methods**.

This table is provided as a separate file.

Table S5. Cloning primers.
Primers used for cloning.

Primer name	Primer sequence (5' to 3')
oRB_277	ACGTGGTCTCACCCTCCCTCGAGACGTTATACGCACCGTCTCAAACGTACAGTTGAAGTCGGAAGTTTAC
oRB_278	ACGTGGTCTCACGTTACTTAGTGTATGTAAACTTCTGACC
oRB_279	ACGTGGTCTCGTATGATTGAGTGTATGTAAACTTCTGAC
oRB_287	ACGTGGTCTCAGGATGGATTGAGACGGTGCGTATAACGTCTCGCATATACAGTTGAAGTCGGAAGTTTAC
oRB_413	CTCATTTACATTTCCAATGCTATACCAAAAG
oRB_414	AGGACAGCCAGGGCTACACAG
oRB_419	TGCAGCCCCCTCCTCATCTG
oRB_420	GTTCTAAGCCAGTAGAGAACTATATATCAAGGTCC
oRB_465	TTGAGTACTGGGATGACAAGCAAG
oRB_466	GGGAAGGACCTTGATATATAGTTCTCTACTGGCTTAGAACTGCAGCCCCCTCCTCATCTG
oRB_467	CCCAGCTTTCCGAGCCAGATGAGGAGGGGGCTGCAGTTCTAAGCCAGTAGAGAACTATATATCAAG
oRB_468	GAAGATGGAGCAAGAGGATTACCG
oRB_468	GAAGATGGAGCAAGAGGATTACCG
oRB_470	GGGAAGGACCTTGATATATAGTTCTCTACTGGCTTAGAACGAAGATGGAGCAAGAGGATTACCG
oRB_472	TTGGATATGGTAGTGCTTGCTTGTCATCCCAGTACTCAACTAAATACCTAATAAAAAATGAAAAAAAAATGTACG
oRB_473	GATTTTGAGTTTAAACGTACATTTTTTTTTCATTTTTTATTAGGTATTTAGTTGAGTACTGGGATGACAAGCAAG
oRB_474	TATGGGGGACTTTTGGTATAGCATTGGAAATGTAAATGAGTGCAGCCCCCTCCTCATCTG
oRB_475	CCAGCTTTCCGAGCCAGATGAGGAGGGGGCTGCACTCATTTACATTTCCAATGCTATACC
oRB_476	CTAGCAAAGTTTTGTCAAAGGGCAAGCTAGCCACTGAGAGTTGTCTGAGAAGAAGATGGAGCAAGAGGATTACCG
oRB_477	CTACCTCGAACTCACGGTAATCCTCTTGCTCCATCTTCTTCTCAGACAACTCTCAGTGGC
oRB_478	AACAATCTTGGATATGGTAGTGCTTGCTTGTCATCCCAGTACTCAATCTTTTTTCTATCTTTTAAGGAAGCATCC
oRB_479	TGCGCCACTACGCCCTGCCCTAGGATGCTTCCTTAAAAGATAGAAAAAAGATTGAGTACTGGGATGACAAGCAAG
oRB_483	CTAAATACCTAATAAAAAATGAAAAAAAAAATGTACGTTTAAACTC
oRB_484	TCTTTTTTCTATCTTTTAAGGAAGCATCCTAG
oRB_486	TCAGGGTCTCACAGTTCTCATTTACATTTCCAATGCTATACCAAAAG
oRB_487	TCAGGGTCTCTACAGATATAAAATGTCTAAGAGATGGACATGGTGG
oRB_488	TCAGGGTCTCACTGTTCTCATTTACATTTCCAATGCTATACCAAAAG
oRB_489	TCAGGGTCTCTAAAAATATAAAATGTCTAAGAGATGGACATGGTGG
oRB_492	TATGGGGGACTTTTGGTATAGCATTGGAAATGTAAATGAGAAGACTTTAAAATGCATACACATATGTATTCAAGG
oRB_493	TTCAAAATGTGCACCACCATGTCCATCTCTTAGACATTTTATAGTTCTAAGCCAGTAGAGAACTATATATCAAGG
oRB_494	GACGGGGGAAGGACCTTGATATATAGTTCTCTACTGGCTTAGAACTATAAAATGTCTAAGAGATGGACATGGTGG
oRB_495	AAAAAAAATCCTTGAATACATATGTGTATGCATTTTAAAGTCTTCTCATTTACATTTCCAATGCTATACCAAAAG
oRB_496	ACCCAGCTTTCCGAGCCAGATGAGGAGGGGGCTGCAGAAGATGGAGCAAGAGGATTACCG
oRB_497	GCCTACCTCGAACTCACGGTAATCCTCTTGCTCCATCTTCTGCAGCCCCCTCCTCATCTG
oRB_498	CAAAGGGCAAGCTAGCCACTGAGAGTTGTCTGAGAATTGAGTACTGGGATGACAAGCAAG
oRB_499	TGGATATGGTAGTGCTTGCTTGTCATCCCAGTACTCAATTCTCAGACAACTCTCAGTGGC
oRB_500	GCCCTGCCCTAGGATGCTTCCTTAAAAGATAGAAAAAAGATGCAGCCCCCTCCTCATCTG
oRB_501	TTCCGAGCCAGATGAGGAGGGGGCTGCATCTTTTTTCTATCTTTTAAGGAAGCATCCTAG
oRB_502	TTCTCAGACAACTCTCAGTGGCTAG
oRB_504	TGGGATTAAAAGCATGTGCCACC
oRB_505	TTCAATTCCCAGCAACCATTTGACAGGTCACAACTTTCCAGAAGAAATGCAGCAACCAGC
oRB_506	CTTCTGGCTATGGGTGCTGGTTGCTGCATTTCTTCTGGAAAGTTGTGACCTGTCAAATGG
oRB_507	ATGGTAGTGCTTGCTTGTCATCCCAGTACTCAAGCATGCCATACCATCTCTTTTTTGTTG
oRB_508	AAGTTGGTTCAACAAAAAAGAGATGGTATGGCATGCTTGAGTACTGGGATGACAAGCAAG
oRB_560	TGTCAAAGGGCAAGCTAGCCACTGAGAGTTGTCTGAGAACTAAATACCTAATAAAAAATGAAAAAAAAAATGTACG
oRB_561	ATTTTGAGTTTAAACGTACATTTTTTTTTCATTTTTTATTAGGTATTTAGTTCTCAGACAACTCTCAGTGGCTAG
oRB_562	AAGTTTTGTCAAAGGGCAAGCTAGCCACTGAGAGTTGTCTGAGAATATAAAATGTCTAAGAGATGGACATGGTGG
oRB_563	GCTGGGATTCAAAATGTGCACCACCATGTCCATCTCTTAGACATTTTATATTCTCAGACAACTCTCAGTGGCTAG
oRB_564	TCAGCAGTAGAGACGGCCTGGCTCTAGTAGCGATCTAC
oRB_565	GCCTAAAATGAGACGCAATACGCAAACCGCCTCTCCCCGCGCG
oRB_568	TCAGCGTCTCAGGAATCTCATTTACATTTCCAATGCTATACCAAAAG
oRB_569	TCAGCGTCTCTAAAAATATAAAATGTCTAAGAGATGGACATGGTGGTG

Table S6. Genomic coordinates for engineered loci.
Coordinates for payloads, genomic deletions (marked with a “d”) and BACs. For some payloads, coordinates include mismatched regions derived from designed sequence mutation. See **Table S4** for conversions of payload IDs to labels used in figures.

This table is provided as a separate file.

Table S7. Homology arm cloning primers.
Primer pairs used to clone left and right homology arms (HAs) using Golden Gate reactions Primer sequences contain *BsaI*/*BsmB*I sites in lower case, gRNA binding sites (underlined), and PAMs (bold).

Locus	HA	Length (bp)	Forward primer sequence (5’ to 3’)	Reverse primer sequence (5’ to 3’
Sox2	Left	142	GTGTggtctcACCCTCAAGTCTGAAGTAGTTCAGG**AGG**GTTTGAGGCCAGGAAGGGAT	GTGGTGggtctcACGTTGAACTACTTCAGACTTGGGC
	Right	183	GTGGTGggtctcGTATGGTTCGGGGACGGTGTTAATATTCTTC	GGTGggtctcAGGATGAGCTGCAAAGGCTCCCGTT**AGG**AATGAATGCGGATGCCTTGC
Distal LCR	Left	100	GGTGcgtctcACCCTGTCTTAACCAACCAAAGGTC**AGG**TTATGCTGAGGGTCATCACACAAG	GGTGcgtctcACGTTCTTTGGTTGGTTAAGACGGGTCAC
	Right	138	GGTGcgtctcGTATGGTTCGGGGACGGTGTTAATATTCTTC	GGTGcgtctcAGGATGAGCTGCAAAGGCTCCCGTT**AGG**CCAAGCCTTACGGGAACGCC

Table S8. gRNAs.

gRNAs used for LP integrations and payload assembly. Allele-specific gRNAs are identified as having CAST variants overlapping the PAM or gRNA seed sequence.

gRNA Names	gRNA Sequence (5' to 3')	Allele specificity
dDHS23	AATTAGTACTCCCCCACGCA	none
invCTCF4.1	GCGTAGAGCTTTCGTCTTTG	none
invCTCF8	TAAAAGCAAGTCCACCAGCA	none
mSox2-3p-5	GAGCTGCAAAGGCTCCCGTT	BL6 (PAM)
mSox2-5p-1	CAAGTCTGAAGTAGTTCAGG	BL6 (PAM)
mSox2-DHS15-16	CATTGGCAGTGTTGTATAGG	none
mSox2-DHS18-19	GTCTTAACCAACCAAAGGTC	none
mSox2-Prom-1	GACAATGAGCGGAGCGTGGC	BL6 (seed)
mSox2_15g	CTGCACCGTCTTAAGATCTC	BL6 (seed)
mSox2_16g	CAGCTGAAAGCGAGCTGCAA	none
mSox2_17g	GTCTAGCGTTTTACAGTCTT	BL6 (seed)
oBRC0453	GTACACCAGAGATCTTAAGA	BL6 (seed)
oBRC0454	TTGGTTAAGACGGGTCACAC	none
oNKS0012	GTATGCATATGCACCACATG	none
oNKS0015	ACGCCGGAACGGTAGCTCAG	none
oNKS0018	TGACTGGCAACCGAGCCGGG	none
oNKS0022	CCTTTACAGAGGAACCCAGG	none
oNKS0023	TGCCAAGTGAAGAAAAGGCA	BL6 (seed)
oNKS0024	CTTGTGTCTATAATCCCAGA	none
oNKS0025	ATGGGGATGCCACTCACTGG	none
oNKS0026	GATCTCATGGTTTGGCAGAG	none
oNKS0028	CTTCACACTGTATCCCAGGA	BL6 (seed)
oNKS0029	GAGTTGAAAAGATGGCTCAG	BL6 (PAM)
oNKS0030	CATGCCTTTAATCCCAGCAG	none
oNKS0031	CTTTAATCTCAACACTTGGG	BL6 (seed)
oNKS0032	GTTTCTATATGTATATCCTG	none
oNKS0033	CACAAAGAAATGAAAGACTC	none
oNKS0034	CTGCTGATACCTGTTCAGCA	none
oNKS0035	GACCCGTCTTAACCAACCAA	none
oNKS0036	GACAGTTCTTGCTCACACAT	none
oNKS0037	GGAGCACTTGTTGGTCTCGG	none
oNKS0038 or dCTCF8	TAAAAGCAAGTCCACCAGCA	none
oNKS0039 or dCTCF14	AGCTAGGAGCCCCTAGGTAC	BL6 (seed)
oNKS0040 or dCTCF17	GAACGCCACTTACTGCTTAC	none
oNKS0041 or dCTCF25	CACAGAAGAAAAAGCGCGGC	none
oNKS0042 or dCTCF1	CTTGAGCAGCGCCCCCACAA	BL6 (seed)
oNKS0043	CCCATGTACTGCAGTTCCTG	none
oNKS0044	TCGTCCGATTTGTCATTAGA	none
oNKS0045	CGCCCCCTATTGGACGCAGT	none
oNKS0048	AGTTCCAGCACACGCGCAGA	BL6 (seed)

Table S9. Synthetic fragments.
Sequences of synthetic fragments used for payload assembly.

This table is provided as a separate file.

Table S10. LP and PL integration junction genotyping primers.
Primer pairs used to verify newly formed junctions in engineered cells.

This table is provided as a separate file.

Table S11. LP and PL integration additional genotyping primers.
Primer pairs used to verify or characterize engineered cells.

Assay	**Forward primer sequence (5' to 3')**	**Reverse primer sequence (5' to 3')**
Sox2[Bl6]	CTGAAGCATGAGCTTCCTCCCT	GCAGTACATGGGTCGTACACATG
CAST allele deletion spanning	TCCAGCGTCTTTCCCAGTTC	AGGAGGGTTCAAGACAGGGT
CAST allele deletion internal	CTGTTGTCCTAAGCAAGAAGGTAGGC	TTGGGACTCTCAAACCAAAGACTGTAAAAG
Sox2[BL6]_107kb_down	GCCTTGTGAGCTCAGGAGAAGTC	TACTCAAGGTCAGACGTGCG
Sox2[BL6]_7kb_down	GGCCAAGCCATACTGGGCTC	GCTCCAATCAAGCCAGTAGATGTAGAG
Sox2[BL6]_45kb_down	TTGCTCCAGGTCAGTCCAGC	ACTGTACACTCTGGTCGACCT
Sox2[BL6]_80kb_down	GACTTGAATTTGACCCATGAACCAAT	GCAGATGAGGACTTTTCGAGAGT
Sox2[BL6]_15kb_down	GGCAAGTGAAAGAGAGGTGAAAAGATG	GTCACATGAAGAGCACACAGACAC
Loss LP-Sox2	GGCCAGGAAGGGATATTTGAGA	GGGACTGTGGGCGATGTG
Loss of LP-LCR (L Jx)	CTGTTGTCCTAAGCAAGAAGGTAGGT	GGGACTGTGGGCGATGTG
Loss of LP-LCR (R Jx)	TGCGTCCTGCAGGCGCGTTC	CATCACATTCCCAGTGGGTCAG
YAC/BAC backbone (OriV)	CTGTCTATGGCCCTGCAAAC	CAGTGAGGGCCAAGTTTTCC
BSD	TCGCGACGATACAAGTCAGG	GGACCTTGTGCAGAACTCGT

Table S12. Sequencing libraries.
Summary table for DNA and DNA 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 conversions of payload IDs to 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 S13. YAC/BAC sequencing coverage QC.
Summary table for YAC/BAC payload DNA sequencing coverage. Length (bp) and normalized (norm) coverage were calculated for regions corresponding to the payload and non-payload, as defined in **Methods**. Quality control (QC) calls are included with comments for unexpected values. See **Table S4** for conversions of payload IDs to labels used in figures.

This table is provided as a separate file.

Table S14. YAC/BAC variant calling.
Summary table for YAC/BAC payload variants relative to custom references. 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 conversions of payload IDs to labels used in figures.

This table is provided as a separate file.

Table S15. mESC clone capture sequencing QC.
Summary of QC analyses of Capture-seq samples derived from engineered mESC clones. 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. Analyses were performed separately for: payload (genomic regions covered by the delivered payload), non-payload (genomic regions overlapping engineered locus and absent from payload), flanks (regions captured by the *Sox2* bait, but not engineered), PL[BB] (YAC/BAC backbone), LP (landing pad), LP[BB] (landing pad plasmid backbone), and iCre (pCAG-iCre plasmid). norm coverage, normalized mean coverage; propREF, proportion of reference to non-reference allele reads. QC calls for the coverage and variants analyses are included, as well as for the *bamintersect* analysis (see **Table S16**). R/L Jx, right/left junction. See **Methods** for more detail. See **Table S4** for conversions of payload IDs to labels used in figures.

This table is provided as a separate file.

Table S16. *bamintersect* junction categories.

Hierarchical criteria used to classify *bamintersect* junctions. bam1 and bam2 are the two references to which junctional 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	chr3	0-1000 bp upstream of HAL's 5'	+	LP	First 1000 bp	-
LP	2	R Jx	chr3	0-1000 bp downstream of HAR's 3'	-	LP	Last 1000 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]	chr3	0-1000 bp upstream of HAL 3'	+	PL	First 1000 bp	-
PL	2	R Jx[a]	chr3	0-1000 bp downstream of HAR's 5'	-	PL	Last 1000 bp	+
PL	3	L ITR-PL[b]	LP	First 1000 bp	+	PL	First 1000 bp	-
PL	4	R ITR-PL[b]	LP	Last 1000 bp	-	PL	Last 1000 bp	+
any	5	Non-informative[c]	chr3	Engineered region	any	any	any	any
any	6	PL-BB[d]	bam1 or bam2 maps to YAC/BAC PL backbone					
any	7	off-target	None of the above					

[a] For payloads with wild type termini, CAST allele reads are indistinguishable from BL6 allele reads.

[b] Only applicable to payloads delivered to LP-PIGA3 (LP305) as the LP's ITRs remain flanking the integration site after PL delivery (see **Fig. 1**).

[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 S17. mESC *bamintersect* results.
Summary of *bamintersect* analysis for mESC clones capture sequencing samples. Each line describes a junction (Jx) detected by *bamintersect* (see **Methods**). Chrom (chromosome), chromStart, chromEnd, and Strand for bam1 and bam2 refer to the mapped reads location 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 S16**. See **Table S4** for conversions of payload IDs to labels used in figures.

This table is provided as a separate file.

Table S18. qRT-PCR primers.
Primer pairs used for qRT-PCR.

Assay	**Forward primer sequence (5' to 3')**	**Reverse primer sequence (5' to 3')**
Sox2[BL6]	AACCGATGCACCGC	GAGCATTATCAGATTTTTCC
Sox2[Cast]	AGCCGATGCACCGA	TGAGCATTATCAGATTTTTCT
Gapdh	AGAACATCATCCCTGCATCC	CACATTGGGGGTAGGAACAC