nature genetics

Article

https://doi.org/10.1038/s41588-023-01450-7

Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions

Received: 15 October 2022

Accepted: 20 June 2023

Published online: 24 July 2023

Check for updates

Renhe Luo[1,2], Jielin Yan[1,2], Jin Woo Oh[3], Wang Xi[3], Dustin Shigaki[3], Wilfred Wong[4,5], Hyein S. Cho[1], Dylan Murphy[5,6], Ronald Cutler[7], Bess P. Rosen[1,5], Julian Pulecio[1], Dapeng Yang[1], Rachel A. Glenn[1,5], Tingxu Chen[1,2], Qing V. Li[1,2], Thomas Vierbuchen[1], Simone Sidoli[7], Effie Apostolou[6], Danwei Huangfu[1] ✉ & Michael A. Beer[3] ✉

Comprehensive enhancer discovery is challenging because most enhancers, especially those contributing to complex diseases, have weak effects on gene expression. Our gene regulatory network modeling identified that nonlinear enhancer gene regulation during cell state transitions can be leveraged to improve the sensitivity of enhancer discovery. Using human embryonic stem cell definitive endoderm differentiation as a dynamic transition system, we conducted a mid-transition CRISPRi-based enhancer screen. We discovered a comprehensive set of enhancers for each of the core endoderm-specifying transcription factors. Many enhancers had strong effects mid-transition but weak effects post-transition, consistent with the nonlinear temporal responses to enhancer perturbation predicted by the modeling. Integrating three-dimensional genomic information, we were able to develop a CTCF-loop-constrained Interaction Activity model that can better predict functional enhancers compared to models that rely on Hi-C-based enhancer–promoter contact frequency. Our study provides generalizable strategies for sensitive and systematic enhancer discovery in both normal and pathological cell state transitions.

Many consortia have made important progress in mapping putative enhancers based on chromatin accessibility and protein binding in a wide range of cell types and tissues[1]. Harnessing the atlas of enhancers predicted from chromatin features, functional interrogation with large-scale CRISPR screens has successfully identified some enhancers with relatively strong impacts on gene expression in various cell lines[2–15]. However, comprehensive enhancer discovery remains challenging. In some cases, enhancer perturbation only causes temporary phenotypes[7,8], while in other cases, the effect of enhancer perturbation is mitigated by the activity of 'shadow' or redundant enhancers[15–21].

[1]Developmental Biology Program, Sloan Kettering Institute, New York City, NY, USA. [2]Louis V. Gerstner Jr. Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer Center, New York City, NY, USA. [3]Department of Biomedical Engineering and McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA. [4]Computational & Systems Biology Program, Sloan Kettering Institute, New York City, NY, USA. [5]Weill Cornell Graduate School of Medical Sciences, Weill Cornell Medicine, New York City, NY, USA. [6]Department of Medicine, Weill Cornell Medicine, New York City, NY, USA. [7]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY, USA. ✉e-mail: huangfud@mskcc.org; mbeer@jhu.edu

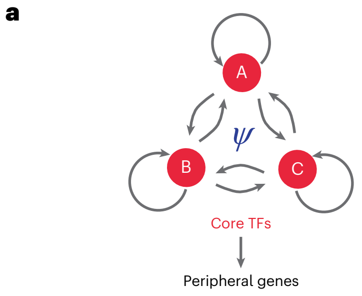
Furthermore, most genetic variants associated with common human diseases show relatively modest effects on expression in reporter assays[22,23]. A critical missing element is a quantitative model for how multiple enhancers work together at each locus to respond to physiological or experimental perturbations in a nonlinear way through altering gene regulatory network (GRN) activity. GRNs control cell states through a handful of core transcription factors (TFs), which both self-regulate and cooperatively regulate each other through core enhancers[24]. Therefore, we focused on deconstructing the core enhancers' activity in the GRN by determining the impact of their perturbation on the cell state. We developed a quantitative GRN model to simulate the dynamic process of core circuit establishment. The results predict that cell states are more susceptible to core enhancer perturbations during dynamic cell state transitions compared to post-transition when the GRN has been established.

We used guided differentiation of human embryonic stem cells (hESCs) to definitive endoderm (DE) as a dynamic system for a dCas9–KRAB-based CRISPRi screen. The CRISPR library targeted 394 putative enhancers surrounding ten core TF loci spanning 40 Mbp genomic regions. Using *SOX17* as the DE cell state readout[25], we identified multiple enhancers (4–9 per locus) for each of the core DE TFs (*EOMES*, *GATA6*, *MIXL1*, and *SOX17*). This affirms the feasibility of using a single core gene as the readout for discovering functional core enhancers during cell state transitions. The sensitive screening strategy also uncovered 12 enhancers >100 kbp away from the transcription start site (TSS) of the target gene, demonstrating the need for enhancer discovery beyond the immediate linear neighborhood. The relatively comprehensive discovery of functional enhancers allowed us to develop a CTCF-loop-constrained Interaction Activity (CIA) model that outperformed previous Hi-C contact-based enhancer prediction methods[4,26–28]. Our network-guided core enhancer mapping strategy during cell state transitions and the CIA model provide a framework for systematic enhancer discovery applicable not only to normal development but also to pathological conditions such as diabetes and cancer.

Results

GRN model predicts temporal sensitivity to enhancer perturbation

We sought to develop a dynamic GRN model to study the temporal and threshold-dependent requirements for enhancers during cell state transitions. Our previous studies of cell state transitions[24,25,29,30] and sequence-based modeling[24,31,32] of a broad range of ENCODE epigenomic profiling data have identified key features of this model. Machine learning applied to chromatin-accessible peaks identifies a small set of 5–10 lineage-determining core TFs in each cell type whose binding sites can predict chromatin-accessible peaks to a high degree of accuracy[24,31,32]. Each chromatin-accessible peak contains combinations of multiple binding sites for these core TFs. Thus, the lineage-determining core TFs cooperatively auto-regulate each other through multiple enhancers flanking each core TF gene and they coregulate downstream peripheral genes (Fig. 1a and Extended Data Fig. 1a), resulting in highly nonlinear regulation of gene expression. Here we describe the activity of the network with a time-dependent state variable, $\psi(t)$, whose amplitude reflects the activity of the core TFs that determine the network state, which we use interchangeably with the term 'cell state' (Fig. 1b). Each gene is expressed at an activated level (e_1) with probability $p_{on} \sim cf(\psi)$ and at the basal level (e_0) with $p_{off} \sim b$. $f(\psi)$ is a nonlinear function of the core TF activity, which reflects co-operativity at the enhancers, and $f(\psi)$ can be modulated through the parameter c (for example, via CRISPRi). To simulate dynamic cell state transitions, we allow a time-dependent differentiation stimulus $\delta(t)$ that acts at the enhancers through a separate mechanism (for example, differentiation signaling), so $p_{on} \sim cf(\psi) + \delta(t)$. Finally, we add degradation or export of the TFs ($-r\psi$) and a stochastic noise term, $\xi(t)$.



b

$$\frac{d\psi}{dt} = -r\psi + \frac{e_1(cf(\psi)+\delta(t))}{b+cf(\psi)+\delta(t)} + \frac{e_0 b}{b+cf(\psi)+\delta(t)} + \xi(t)\psi$$

Key features:

ψ: describes time evolution of network activity
$f(\psi) \sim \psi^n$: autoregulatory enhancer activation
c: modulation of enhancer strength
δ: transition inducing stimulus
e_1: high enhancer activated transcription
e_0: low basal transcription

Fig. 1 | Dynamic gene regulatory network model. a, The schematic representation of the core circuit in the GRN. The core TFs cooperatively auto-regulate each other and coregulate downstream peripheral genes. **b**, The equation of the dynamic GRN model.

We further simplified this into a one-dimensional network state equation (Fig. 1b), where $\psi(t)$ is a scalar representing the degree to which the network is activated (Methods). In the strongly cooperative limit, the enhancers together have an activity $f(\psi) \approx \psi^n$ where n is a typical number of TFs binding at each enhancer. For concreteness and simplicity, we will take $n = 3$, but our conclusions are robust for $n \geq 3$. Stochastic simulation of this model (Methods) produces distributions of cells with either high or low network activity (Fig. 2a). This can be understood from stability analysis of this model, by plotting $d\psi/dt$ versus ψ (Fig. 2b). Over a wide parameter range, the network has three fixed points where $d\psi/dt = 0$. Two of these are stable states: the OFF state, with low activity where basal activation balances degradation, and the ON state, with high activity where enhancer-driven transcriptional activation balances degradation. There is also an intermediate unstable fixed point (Fig. 2b), above which the network activity will transition to the ON state, otherwise, it will fall to the OFF state. The simulation results are in good agreement with single-cell RNA sequencing (scRNA-seq) experiments sampling every 12 h during hESC-DE differentiation (Fig. 2a,c,d). To generalize the definition used in the equation, we used the projection of the expression of all TFs along principal component analysis (PCA) component 1 to measure the network state in each cell (Fig. 2c and Extended Data Fig. 1b,c). The scRNA-seq results show a bistable distribution of cell states during differentiation as follows: a pretransition steady state, from 0 h to 24 h, and a post-transition steady state from 48 h to 72 h, when transcriptome profiles are relatively similar over time, and a transitional period, from 24 h to 48 h, when transcriptome profiles are changing more rapidly (Fig. 2c,d and Extended Data Fig. 1d). This transition behavior is also predicted in the simulation results (Fig. 2a,b), suggesting that the simple dynamic network model is capturing key features of the establishment of the core circuit in a GRN.

To quantitatively model enhancer perturbation during the cell state transition, we decreased the total enhancer strength (c) to varying degrees (with constant stimulus strength $\delta(t)$). Mildly decreasing the total enhancer strength, which mimics perturbing one of the multiple core enhancers flanking a core TF, has a weak effect on both the ON and OFF steady states, but has a much stronger effect on the network activity required to transition between states (the unstable fixed point where $d\psi/dt = 0$ in Fig. 2b), indicating mild enhancer perturbation

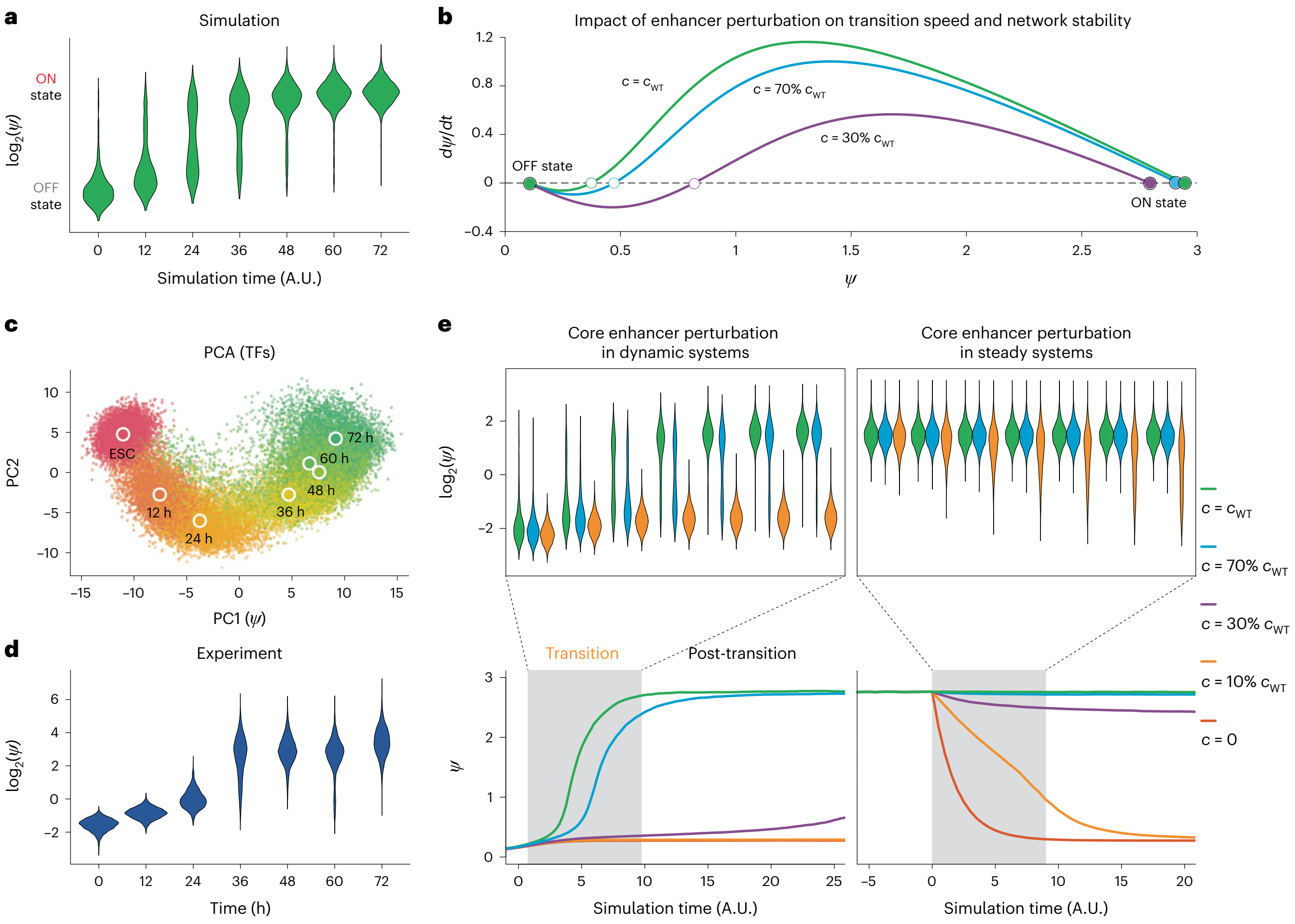


Fig. 2 | Dynamic gene regulatory network model predicts temporal sensitivity to enhancer perturbation during cell state transition. **a**, The violin plots of core circuit establishment during cell state transition by simulation. **b**, Plot of $d\psi/dt$ versus ψ showing cell state transition with different enhancer strengths. The green line represents the total enhancer strength without perturbation. The cyan line represents the total enhancer strength reduced to 70% of full strength by perturbation. The purple line represents the total enhancer strength reduced to 30% of full strength by perturbation. **c**, PCA of all TFs from scRNA-seq data collected every 12 h during hESC-DE transition. Each dot represents a cell. The large circles represent the average of all cells from the same time point (filled with the same color). **d**, The violin plots of core circuit establishment during hESC-DE transition by scRNA-seq experiments. **e**, The comparison between the same enhancer perturbation strength during cell state transition or at steady state. The line plots represent the median of simulation results. The violin plots correspond to a zoomed-in time interval denoted by the gray box on the line plots. The green, cyan, purple, yellow and orange lines represent 100%, 70%, 30%, 10% and 0% of the original total enhancer strength, respectively.

could delay the cell state transition without dramatically changing the final network state (Fig. 2b,e, left). We also simulated enhancer perturbation in a steady state by decreasing the total enhancer strength (c) after cells have transitioned. Compared with the simulation during the forward cell state transition, the same enhancer perturbations show much weaker effects in steady state (Fig. 2e). These results indicate that a time window exists during cell state transitions where enhancer perturbation screens will be more sensitive than screens conducted at a steady state. We further validated the results of the network model (Fig. 1b) with stochastic Gillespie simulations[33,34] under the same nonlinear autoregulatory assumptions. These simulations reproduce the main findings of the sensitivity to enhancer perturbation during the transition, temporal delay in the transition and relative insensitivity to enhancer perturbation after the GRN is fully activated in the ON state (Extended Data Fig. 1e; Methods).

Systematic identification of core TFs in hESC-DE transition

To discover core enhancers in cell state transitions, we used hESC-DE differentiation as a test case, for which core enhancers remain incompletely defined[35]. Optimization of our existing differentiation protocol[25] allowed us to reproducibly generate >95% SOX17$^+$/CXCR4$^+$ DE cells 72 h after the initiation of DE differentiation (DE-72 h; Extended Data Fig. 2a; Methods). To identify core enhancers, we first took a systematic approach to define core TFs (Fig. 3a). Because *OCT4* (*POU5F1*), *NANOG* and *SOX2* are well-known TFs essential for the ESC identity[36], we focused on identifying core TFs for the acquisition of the DE identity. After analyzing our previous genome-scale CRISPR–Cas9 screening data for genes regulating hESC-DE transition[25], we selected four core DE TFs (*EOMES*, *MIXL1*, *GATA6* and *SOX17*) and three signaling TFs (*SMAD2*, *SMAD4* and *JUN*; Fig. 3b and Extended Data Fig. 2b). The DE and ESC TFs showed opposing changes in gene expression during hESC-DE transition with corresponding changes in their regulatory activities predicted by gkm-SVM[24,31] trained on the ATAC-seq data (Fig. 3c–e and Extended Data Fig. 2c,d). Analysis of gene expression from scRNA-seq data (collected every 12 h during hESC-DE transition) using the Pearson correlation as the distance metric for UMAP visualization further demonstrated that the expression patterns of DE and ESC TFs clustered separately, and they each correlated among themselves

https://doi.org/10.1038/s41588-023-01450-7

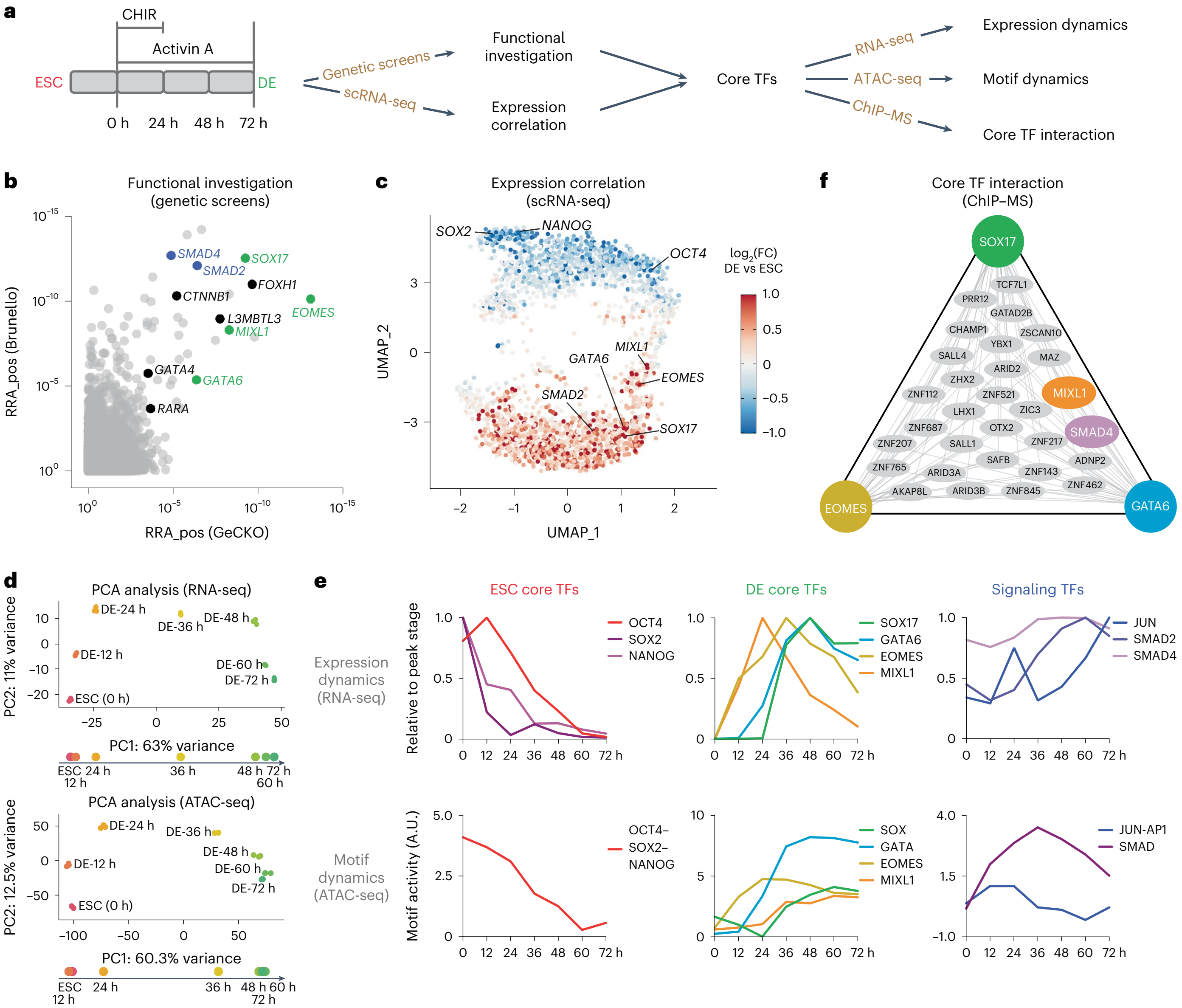


Fig. 3 | Systematic identification of core TFs during ESC-DE cell state transition. a, The schematic representation of using a systematic approach to identify the core TFs during hESC-DE transition. **b**, MAGeCK robust ranking aggregation (RRA) scores for positive hits in two genome-scale DE screens from ref. 25. We overlapped the top 100 hits from each screen, and there were 11 TFs (labeled). DE core TFs and signaling TFs selected for the core enhancer perturbation screen are shown in green and blue, respectively. **c**, The expression correlation analysis from scRNA-seq of hESC-DE transition. Each dot represents a significantly expressed gene during the hESC-DE transition (Methods). The distance between each pair of genes is defined by Pearson correlation. The color scale is defined from the expression fold change from ESC to DE-72 h. **d**, PCA analysis of bulk RNA-seq (top) and ATAC-seq (bottom) during hESC-DE transition. Datapoints are projected to PC1 to determine the time window of cell state transition. **e**, The expression dynamics (top) and motif dynamics (predicted by gkm-SVM trained on the ATAC-seq data, bottom) of core TFs during hESC-DE transition. **f**, The common protein interactive partners between core TFs identified by ChIP–MS using EOMES, GATA6 and SOX17 as baits during hESC-DE transition.

(Fig. 3c and Extended Data Fig. 2d; Methods). Compared to the ESC and DE TFs, the signaling TFs showed less dynamic changes in transcriptional and regulatory activities during differentiation (Fig. 3d,e and Extended Data Fig. 2c,d). We further conducted chromatin immunoprecipitation followed by mass spectrometry (ChIP–MS) for EOMES, GATA6 and SOX17. The proteomics data highlighted that all three TFs interacted with each other, and they also interacted with many common partners including the DE TF MIXL1 and the signaling TF SMAD4 (Fig. 3f, Extended Data Fig. 2e and Supplementary Table 1). In summary, we identified ten core TFs for our study by using functional genomics data and corroborating our findings with gene expression, chromatin accessibility and proteomics data (Fig. 4a).

Discovery of core enhancers in hESC-DE transition

To discover core enhancers, we examined 4 Mbp genomic regions surrounding each of the ten core TFs and identified 394 regions that are accessible at either ESC or DE stage (including those accessible at both stages) based on ATAC-seq (excluding promoter regions). We designed a tiling gRNA lentiviral library targeting these regions with 11,050 total gRNAs (including controls; Fig. 4a, Extended Data Fig. 3a–d and Supplementary Table 2). We generated an hESC line with a doxycycline-inducible dCas9–KRAB cassette and a DE lineage reporter *SOX17*$^{eGFP/+}$ (Extended Data Fig. 4a–d; Methods) and infected the cells with the gRNA library at a multiplicity of infection (MOI) of ~0.3 to ensure that most infected cells received a single gRNA (Fig. 4a).

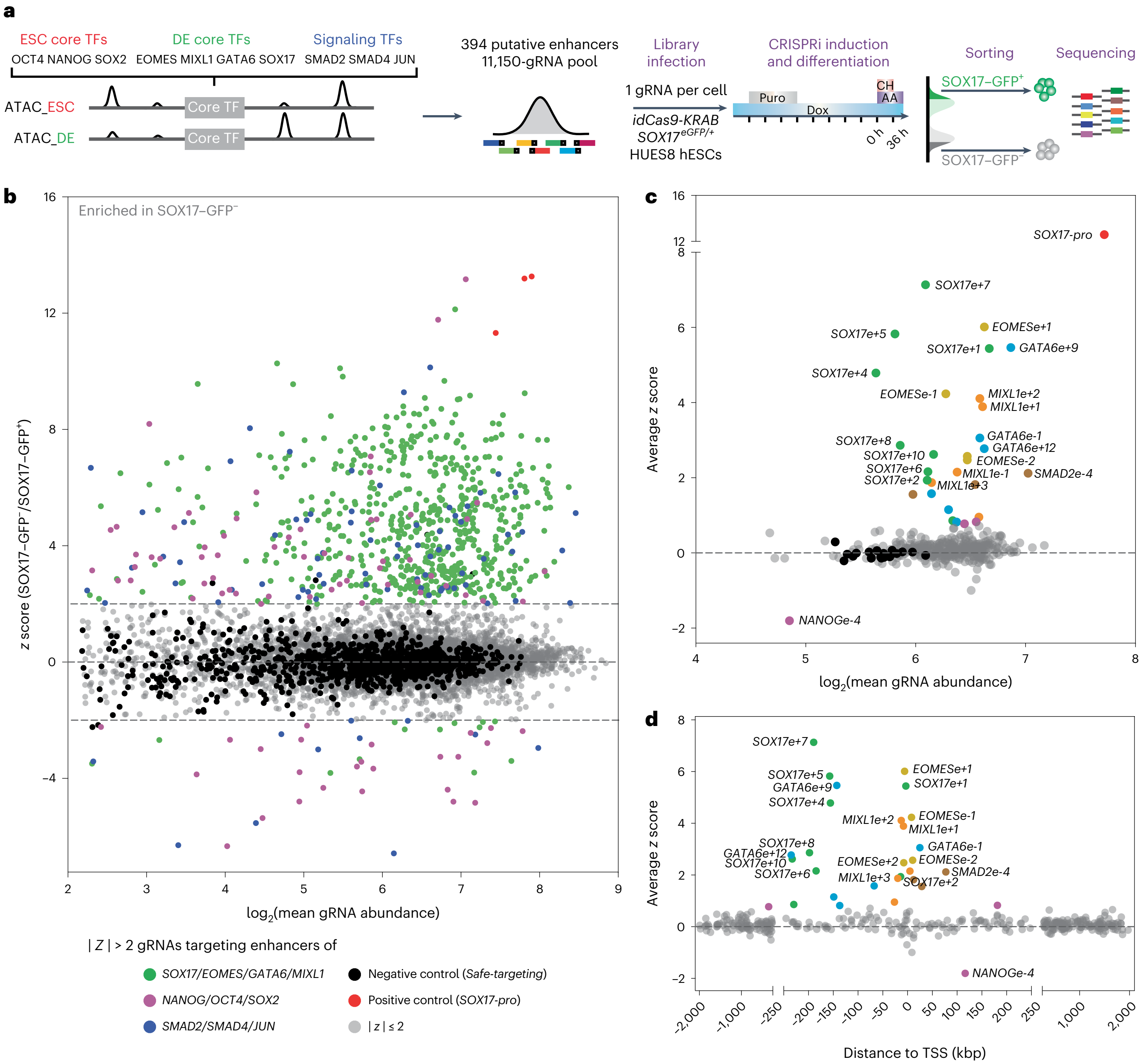


Fig. 4 | A dynamic network-guided enhancer screen identified core enhancers during hESC-DE cell state transition. a, The design of dynamic network-guided core enhancer perturbation screening. **b**, The scatter plot of gRNA *z* score distribution from the screen. Each dot represents an individual gRNA. Dashed lines indicate the threshold of |*z* score| = 2. Gray dots represent gRNAs of |*z* score| ≤ 2. Black dots represent negative control safe-targeting gRNAs. Red dots represent positive control gRNAs. Green dots represent gRNAs of |*z* score| > 2 targeting on *SOX17*, *EOMES*, *GATA6* and *MIXL1* loci. Purple dots represent gRNAs of |*z* score| > 2 targeting on *NANOG*, *OCT4* and *SOX2* loci. Blue dots represent gRNAs of |*z* score| > 2 targeting on *SMAD2*, *SMAD4* and *JUN* loci. **c**, The scatter plot of average *z* score distribution of each putative enhancer region from the screening. Each dot represents an individual region. Safe-targeting gRNAs were grouped based on chromosomes and shown as black dots. Twenty-nine enhancer hits are labeled by different colors representing the core TFs they are surrounding. **d**, Scatter plots showing the distance of 29 enhancers to the TSS of their target genes.

Based on flow cytometric analysis for SOX17–GFP expression during differentiation (Extended Data Fig. 2a), as well as PCA analysis of scRNA-seq, RNA-seq and ATAC-seq data, we identified DE-36 h as an optimal mid-transition point for interrogation of enhancer perturbation effects (Fig. 2c,d, Fig. 3d,e and Extended Data Fig. 2c,d). SOX17–GFP$^+$ (top 20%) and SOX17–GFP$^-$ (bottom 20%) cells were isolated through fluorescence-activated cell sorting (FACS) for gRNA enrichment analysis (Fig. 4a; Methods). We calculated the *z* score for each gRNA based on its logarithm of fold change (log$_2$(FC)) in the SOX17–GFP$^-$ versus the SOX17–GFP$^+$ cells (Fig. 4b). Most gRNAs in the same hit regions had similar *z* scores (Extended Data Fig. 3e,f), supporting that the screen is both sensitive and robust. Through calculating the average gRNA *z* score for each region, we discovered 29 enhancer hits with *z* scores ranging from 0.75 to 7.14 (Fig. 4c and Supplementary Table 3). Relatively few enhancers were found for the ESC and signaling TFs, likely reflecting that the ESC TFs mainly exert their regulatory effects at the ESC stages and that the signaling TFs are primarily regulated post-transcriptionally (for example, through protein phosphorylation). In contrast, we discovered many enhancers for the core DE TFs, ranging from 4 to 9 for each gene (Fig. 4c, Extended Data Fig. 3f and Supplementary Table 3). These findings demonstrate the high sensitivity of the screening strategy using a dynamic transition and a single gene readout (for example, *SOX17* for the DE identity) for the discovery of core enhancers in multiple loci. Most of these enhancers were previously unknown, and their discovery expands the atlas of regulatory elements required for human development and provides a basis for understanding the complex GRNs that govern cell state transitions. Furthermore, 41% of the identified core enhancers are more than 100 kbp away from the TSS of the cognate gene (Fig. 4d), highlighting the need for examining putative enhancers in relatively broad genomic windows.

Core enhancers show temporal sensitivity to perturbation

Individual CRISPRi perturbations of all 20 top enhancer hits resulted in substantially reduced numbers of SOX17–GFP$^+$ cells at DE-36 h based on flow cytometric analysis (Fig. 5a,b). We also confirmed the downregulation of the corresponding cognate gene expression after perturbation by real-time quantitative PCR (RT–qPCR) analyses (Extended Data Fig. 5a). In contrast to the strong phenotypes observed at DE-36 h, at DE-72 h, the perturbations of most of the top 20 enhancers had little or no impact on the induction of SOX17–GFP$^+$/CXCR4$^+$ cells or their cognate gene expression (Fig. 5a–d and Extended Data Fig. 5a,b), which is reminiscent of the temporarily phenotypic enhancers previously described[7,8]. To investigate the enhancer perturbation effect on hESC-DE transition with a finer temporal resolution, we focused on two *GATA6* enhancers (*GATA6e+9* and *GATA6e+12*) and measured the differentiation efficiency based on SOX17 expression every 6 h. Consistent with our dynamic GRN model (Fig. 2e), the results show exquisite temporal sensitivity to enhancer perturbations—a significant impact was observed only in a narrow time window during the transition (Fig. 5e,f). In summary, our CRISPRi screening and individual enhancer perturbation results show that the expression of core TF genes is frequently regulated by multiple enhancers, ensuring robustness in the regulatory network. Perturbing a single enhancer can substantially decrease target gene expression and delay cell state transitions, but most perturbations do not substantially alter the post-transition state. The consistency of our experimental results with our dynamic GRN model suggests a cooperative autoregulatory mechanism underlying this effect. Together they support the utility of cell state transitions in perturbation screens as a generalizable approach for sensitized enhancer discovery.

Enhancer deletions exert stronger impacts than CRISPRi

We reasoned that the GRN is robust in part because multiple enhancers could interact additively or synergistically at a locus to regulate target gene expression. We applied CRISPR–Cas9 to generate single and double deletions of *GATA6e+9* and *GATA6e+12* (Fig. 6a and Extended Data Fig. 5c). The deletion of both *GATA6* enhancers showed a stronger impact on transition efficiencies compared to the expected additive effects of individual enhancer deletions at both DE-36 h and DE-72 h (Fig. 6b and Extended Data Fig. 5d,e). These results support synergistic interactions of these two core *GATA6* enhancers. Compared to single enhancer deletions, the double deletion produced a clearer separation of differentiated and undifferentiated cells at DE-72 h. These experimental data closely match results from the dynamic GRN modeling (Fig. 6c). We also compared the enhancer deletion results with single or double perturbations using CRISPRi and found that the latter had mild or no impact at DE-72 h (Extended Data Fig. 5d). Thus, although CRISPRi is more amenable for large-scale enhancer screens than CRISPR–Cas9-mediated deletion, it may miss bona fide enhancers especially when examining the perturbation effects in a steady state. The weaker effect of dCas9–KRAB could be due to competition with endogenous TFs and other technical differences between CRISPRi and deletion[37]. Overall, our findings demonstrate that the GRN is robust post-transition but can still be disrupted with strong perturbations as achieved here through the double deletion of *GATA6* enhancers.

CIA model improves enhancer prediction

Our CRISPRi screen and validation studies identified multiple enhancers around the core DE regulator genes (*GATA6*, *EOMES*, *SOX17* and *MIXL1*; Supplementary Table 4). This high-quality dataset allowed us to explore genomic features that can distinguish the enhancers that were positive in the screen (hits) from those that were not (nonhits). As expected, all enhancer hits for the core DE TFs had elevated levels of H3K27ac and increased accessibility during hESC-DE transition, accompanied by the binding of DE core TFs EOMES, GATA6 and SOX17 (Extended Data Figs. 6 and 7). Curiously, we also noticed that the hits were often located on one side of the TSS. We performed Hi-C and CTCF ChIP–seq experiments in ESC and DE and observed strong concordance among bounded domains of increased Hi-C contact frequency (often referred to as topologically associated domains or TADs), CTCF loops measured by ChIA–PET in H1 hESCs (ref. 38), CTCF loops predicted by our loop competition and extrusion model (LE model)[39] and CTCF binding (Fig. 7a). While Hi-C detects increased interactions between the promoter and distal enhancer hits (e.g. *SOX17e+10*) in DE, Hi-C contact frequency between the promoter and distal enhancers is only weakly correlated with the effect of enhancer perturbation (Fig. 7b). In addition, enhancer hits are often distributed broadly around the target gene (Fig. 4d and Extended Data Fig. 3f), while the Hi-C contact signals are primarily enriched near the promoter (Fig. 7a). In contrast, all the *SOX17* enhancer hits fall into a CTCF-loop enclosing the promoter (Fig. 7a), both as measured by CTCF ChIA–PET[38] and as predicted by the LE model[39]. To quantify, we calculated the Hi-C contact frequency (with the promoter) for each enhancer along the locus. We also computed the probability (denoted as *P*(in loop)) that each distal enhancer is enclosed within the same CTCF loop as the promoter, using the ratio of the sum of counts for all loops enclosing both enhancer and promoter to the sum of all counts for all loops enclosing the promoter (Methods). Comparing the genomic intervals with large *P*(in loop) against those intervals with large Hi-C contact frequency, the former is broader and encompasses more hits (Fig. 7a). Similar observations hold for the other loci (Extended Data Fig. 8a). This suggests that we are less likely to miss impactful enhancers based on predictions by their enclosure within a CTCF loop than by enhancer–promoter Hi-C contact frequency.

To quantitatively compare the predictive power of these chromatin conformation features individually, we plot precision–recall curves for predicting hits ($\log_2$(FC) > 0.15) from nonhits for all DE gene enhancers (Fig. 7c, Extended Data Fig. 8b and Supplementary Table 4). *P*(in loop) is more predictive (area under precision–recall curve (AUPRC) = 0.818) than Hi-C contact frequency in DE (AUPRC = 0.692), Hi-C contact frequency in ESC (AUPRC = 0.598) or 1/|distance| from the promoter (AUPRC = 0.604). Many nonhit enhancers with a high *P*(in loop) have low chromatin accessibility, TF binding and H3K27ac (Fig. 7a and Extended Data Figs. 6 and 7). Therefore, we tested all combinations of these features and the enhancer–promoter interaction information based on CTCF looping or Hi-C, using logistic regression, and assessed both AUPRC and correlation with $\log_2$(FC) (Fig. 7b,d). For the potential enhancers targeted in our study, enclosure within a promoter-containing CTCF loop is more predictive than any other single feature. Combining interaction information with ATAC, H3K27ac and TF binding improves AUPRC, and in all cases, CTCF-loop-based models are more predictive than Hi-C contact frequency-based models. The combination of CTCF loop, ATAC and H3K27ac achieves a very accurate prediction with AUPRC = 0.898. There is a further slight improvement by adding core TF binding data from EOMES, GATA6 and SOX17 ChIP–seq (AUPRC = 0.925). Adding H3K4me1 can slightly improve the predictions, but is less informative than H3K27ac when either is combined with ATAC and core TF binding

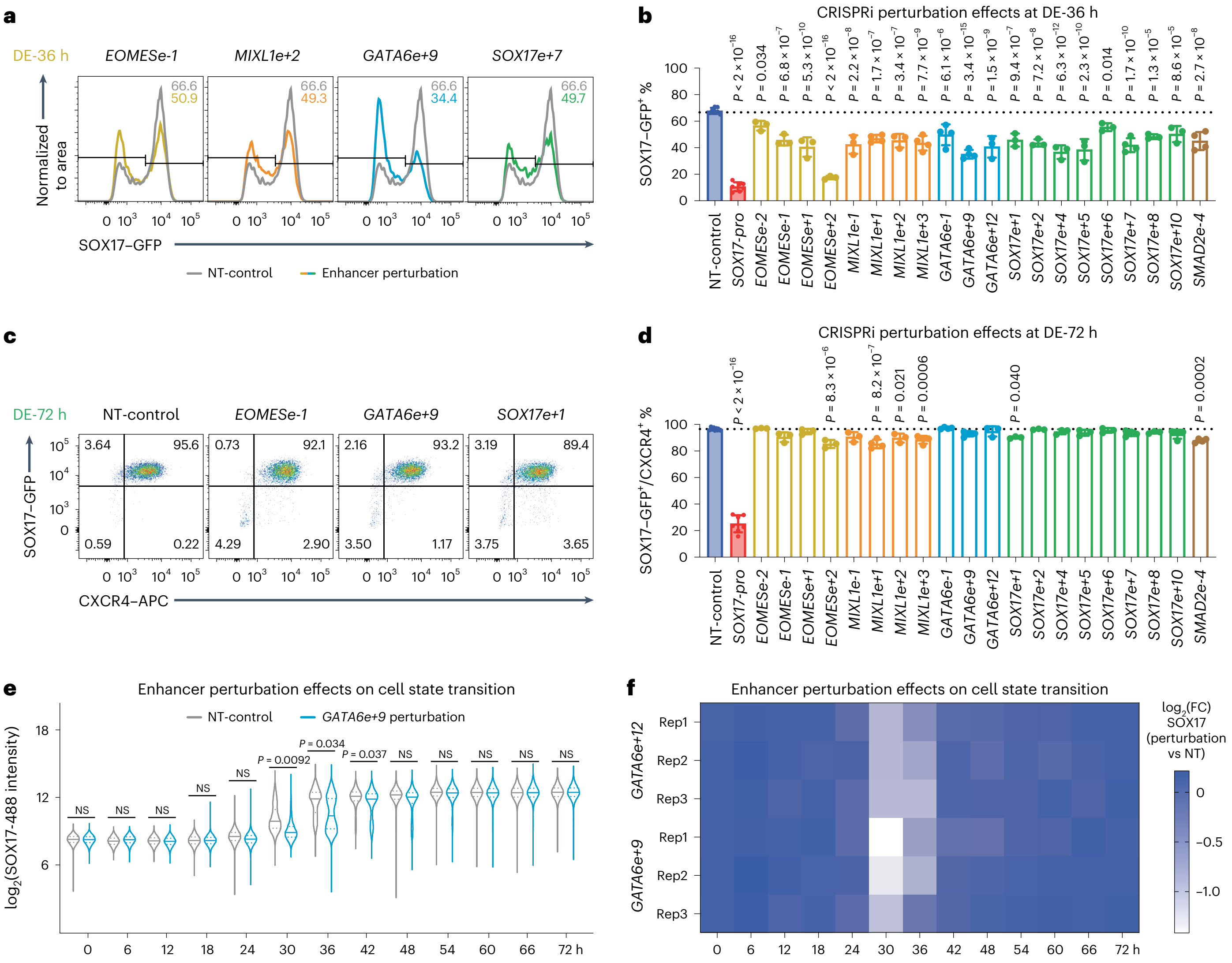


Fig. 5 | Validation of identified core enhancers using CRISPRi perturbation. **a**–**d**, Representative flow plots showing individual core enhancer perturbations decrease the hESC-DE transition efficiency measured by SOX17–GFP$^+$ at DE-36 h (**a**) and SOX17–GFP$^+$/CXCR4–APC$^+$ at DE-72 h (**c**). The bar graphs show the percentage of SOX17–GFP$^+$ cells at DE-36 h (**b**) and SOX17–GFP$^+$/CXCR4–APC$^+$ at DE-72 h (**d**). n = 3–9 biologically independent experiments. Error bars indicate mean ± s.d. Statistical analysis was performed by two-tailed unpaired multiple comparison test with Dunnett correction. **e**, Violin plots showing the effect of *GATA6e+9* perturbation on hESC-DE transition efficiency (SOX17 intensity) measured every 6 h through flow cytometry. n = 3 biologically independent experiments. Solid lines indicate median. Dashed lines indicate quartiles. Statistical analysis was performed by two-tailed paired Student t-test with mean of each replicate. NS: not significant. **f**, Heatmap showing the comparison between the mean SOX17 expression intensity of cells with nontargeting control (NT) versus *GATA6e+9* or *GATA6e+12* perturbation measured every 6 h through flow cytometry during hESC-DE transition. A significant impact was observed only in a narrow time window (around DE-36 h) during the transition.

(Extended Data Fig. 8c). The ABC model[4] combines ATAC and H3K27ac into $\text{Activity} = \sqrt{\text{ATAC} \times \text{H3K27ac}}$ and uses either Hi-C contact frequency or 1/|distance| as contact measurement. We found that CTCF-loop-constrained models outperformed both (Fig. 7b,d). Nevertheless, we found that combining $\text{Activity} = \sqrt{\text{ATAC} \times \text{H3K27ac}}$ with CTCF-loop information is simpler and performs comparably to logistic regression using ATAC and H3K27ac (Extended Data Fig. 8d). The combination of $P(\text{in loop}) > 0.5$ and Activity can classify most hits more cleanly than Activity in combination with Hi-C (Extended Data Fig. 8e). Our best classification result is with both P(in loop) above a threshold near 0.5 and $\text{Activity} = \sqrt{\text{ATAC} \times \text{H3K27ac}}$ greater than a threshold value near 1; as shown in Fig. 7e, almost all hits (green) are correctly classified, and most nonhits are correctly classified either by being outside a CTCF loop (gray) or below the Activity threshold curve (blue). Because both P(in loop) and Activity are required, we will refer to the combined rule $\text{CIAscore} = P(\text{in loop}) \times \text{Activity}$ as the CTCF loop-constrained Interaction Activity, or CIA model.

We evaluated the generalizability of this model with additional CRISPRi datasets from K562 cells[9,40]. The latter study[40] summarized screening data from multiple sources[4,6,11–15]. After mapping gRNAs to DNase hypersensitive peaks in K562 cells, we identified 36 hits and 414 nonhits around 12 genes from ref. 9 (Extended Data Fig. 9a and Supplementary Table 5) and 69 hits and 1862 nonhits from ref. 40 (Extended Data Fig. 10a and Supplementary Table 6). Again, we found that CTCF loop information was more predictive than Hi-C contact frequency (Extended Data Fig. 9b and Extended Data Fig. 10b), and CTCF loop + Activity was more predictive than the ABC model or Hi-C + Activity (Extended Data Fig. 9c,d and Extended Data Fig. 10c,d). An Activity threshold distinguished hits within loops (Extended Data Fig. 9e and Extended Data Fig. 10e). To better compare the CIA

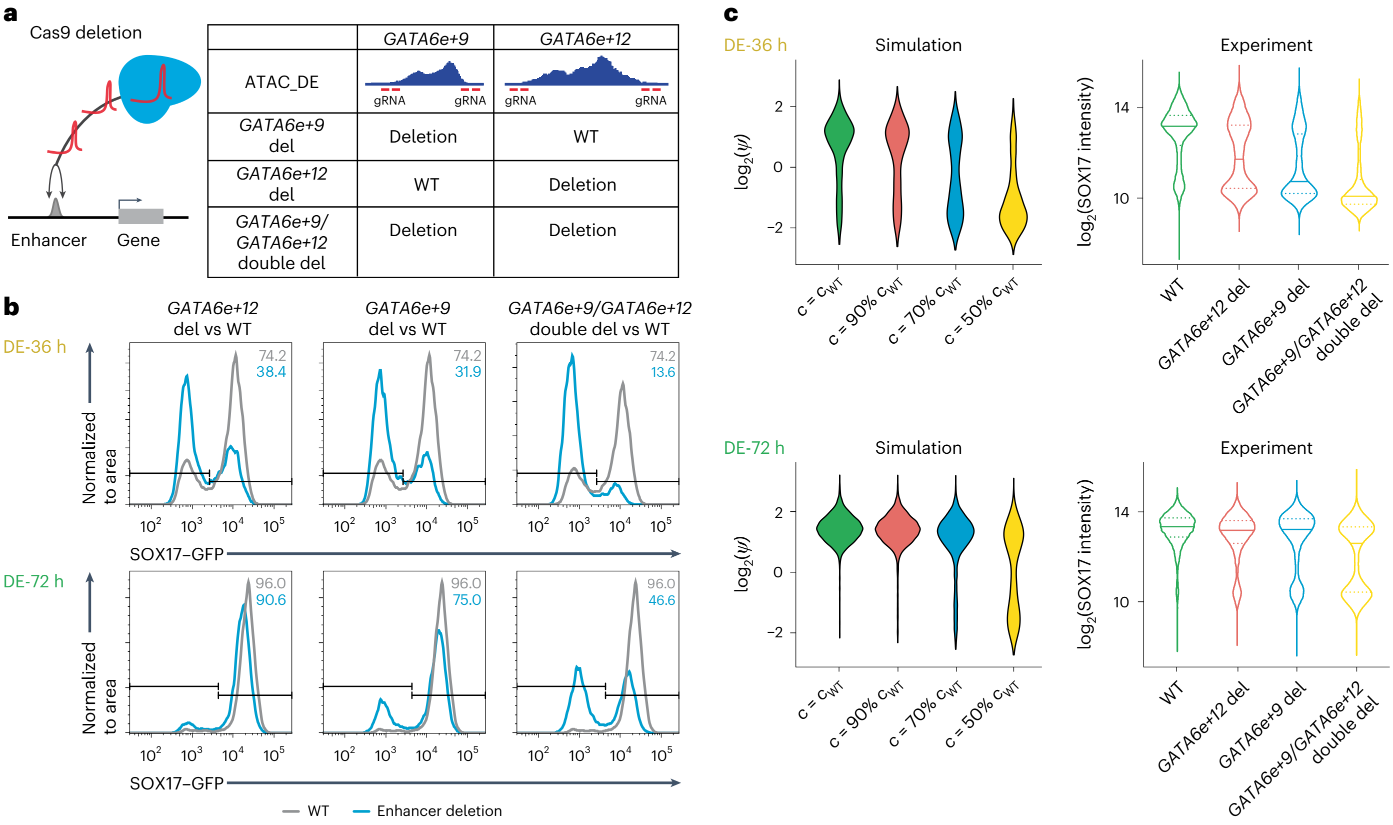


Fig. 6 | Validation of identified core enhancers using CRISPR–Cas9-mediated deletion. a, Schematics of *GATA6e+9* del, *GATA6e+12* del and *GATA6e+9*/*GATA6e+12* double del hESC lines generation using two pairs of gRNAs with Cas9. **b**, Flow plots showing *GATA6* core enhancer del reduced hESC-DE transition efficiency at DE-36 h and DE-72 h. **c**, The comparison between the simulation and experiment results of the impact of different levels of enhancer perturbation on cell state transition. Solid lines indicate median. Dashed lines indicate quartiles.

and ABC models, we integrated all three datasets (current study and refs. 9,40). We scaled $\log_2$(FC) to 'effect size' (Methods) and show the effect size versus distance from TSS for all enhancer–promoter pairs tested (Fig. 7f). For an equal number of positive predictions by CIA and ABC in each dataset (24 + 36 + 69, fixed recall), many strong effect enhancers are predicted by both models (yellow). However, the predictions made by CIA alone (red) have larger effects than predictions made by ABC alone (blue). Adding TF binding slightly improves the CIA model (Fig. 7d). In the absence of TF binding information, gkm-SVM scores can reproduce this small improvement (Extended Data Fig. 8d, Extended Data Fig. 9c and Extended Data Fig. 10c). In summary, using three large functional enhancer screening datasets, we show that enhancer impact is more reliably predicted by enhancer location within a CTCF loop than by direct enhancer–promoter Hi-C contact frequency, and we constructed a simple predictive model of enhancer impact by combining CTCF-loop-constrained interaction and enhancer activity.

Discussion

We performed a CRISPRi screen for enhancers controlling the stimulated hESC-DE transition and discovered many enhancers flanking each core DE TF gene. ChIP–seq results show that each of these enhancers is bound by multiple core TFs, supporting that cooperative autoregulation through multiple TFs at enhancers is a prevalent feature of GRNs controlling cell state. Interestingly, enhancer perturbation delays the transition to DE, and the effect of enhancer perturbation is almost negligible after the transition to DE has been completed. A similar delayed phenotype in response to enhancer perturbation has been observed at the *Hox* cluster in flies[41] and mice[42,43]. Simulations of the hESC-DE transition using our dynamic GRN model agree with the observation of a delayed transcriptional response to enhancer perturbation and show that nonlinear saturation of enhancer activity post-transition is responsible for this robustness once the GRN has fully activated the core TFs. In our simulations, the hysteresis of the GRN occurs via autoregulatory enhancer activity coupled with the translation of the TFs, which is at a longer time scale than simulations of nonlinearities that may also be present at the time scale of transcriptional activation[44]. Together, these observations provide a plausible explanation for low validation rates for intergenic genome-wide association study (GWAS) variants when tested individually in the steady state. These enhancers may contribute to cell state transitions and cell abundance while not strongly affecting transcript levels when tested in the established state. This suggests a direct mechanism by which enhancers may contribute to human disease even in the absence of strong effects in post-transition cells, and recent reports have emphasized that cell type abundance quantitative trait loci (QTLs) can have a profound impact on human phenotypes[45–47]. Our results predict that screens designed to target enhancers during a cell state transition will have a greater sensitivity. In addition, one may explore ways to weaken the GRN (for example, through manipulating a core TF or enhancer) to increase its vulnerability to further enhancer perturbation, as suggested by the *GATA6* enhancer double deletion studies. Beyond developmental cell fate changes, we envision that stimulating physiological or pathological cell state transitions can accelerate the discovery of disease-relevant enhancers or variants.

The sensitivity of our screen also allowed us to find that most functional enhancers fell within CTCF loops, leading us to propose an interaction model whereby CTCF loops constrain enhancer interactions and activity (CIA model). Our large number of validated enhancers

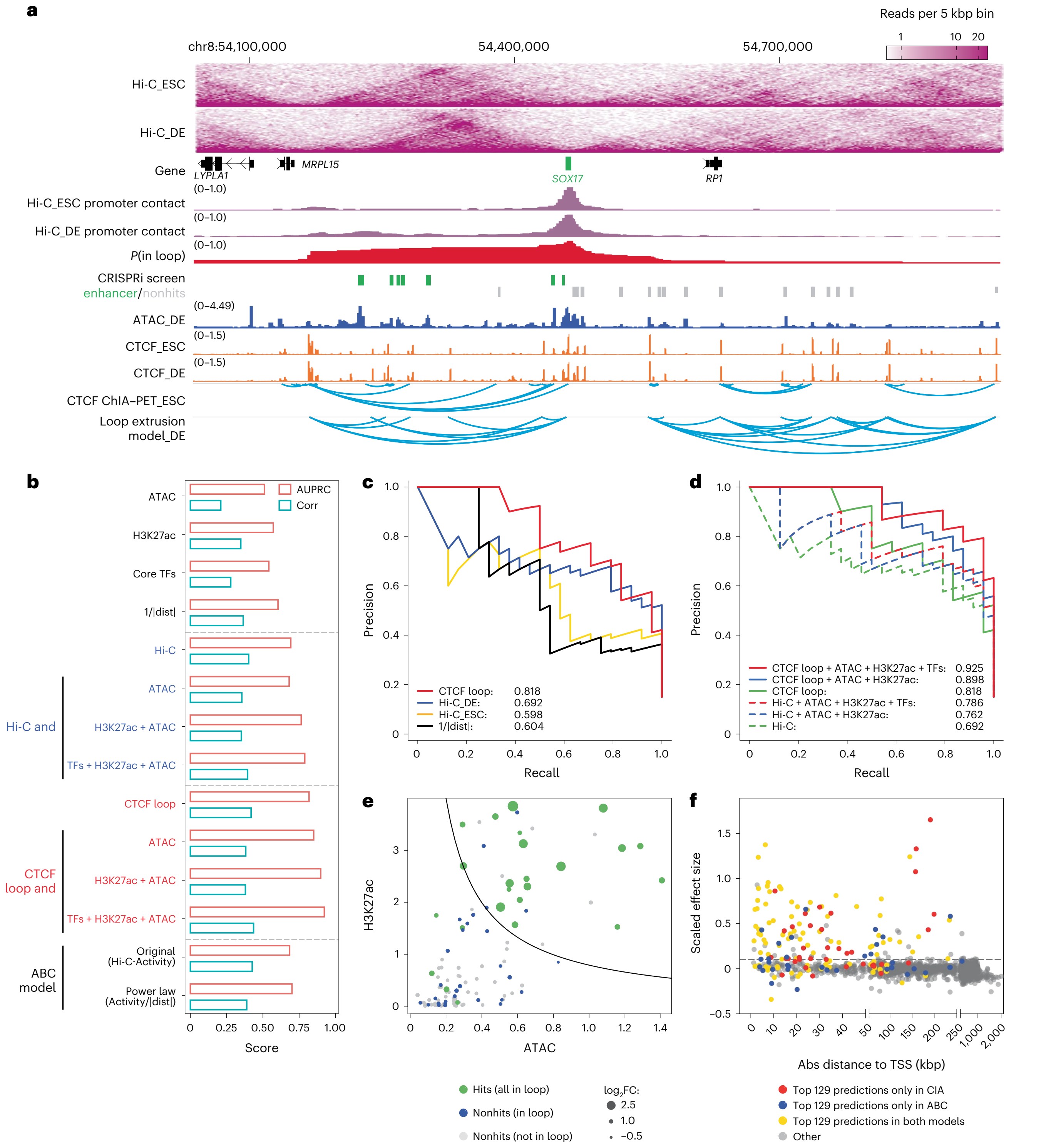


Fig. 7 | The CIA model provides improved enhancer prediction. a, Hi-C-based and CTCF-loop-based chromatin conformation analysis at the *SOX17* locus. **b**, Bar plots comparing the AUPRC and Spearman correlation scores between single chromatin feature-based, Hi-C-based or CTCF-loop-based enhancer prediction model with the ABC model. **c**, A precision–recall plot comparing the performance for prediction of enhancer hits from the screen using *P*(in loop, red), DE Hi-C (blue), ESC Hi-C (yellow) and enhancer–promoter distance (black). **d**, A precision–recall plot comparing the performance for prediction of enhancer hits from the screen using CTCF loop (solid line) and Hi-C (dash line) with or without additional chromatin feature combination. **e**, A scatter plot showing the combinatory criteria of *P*(in loop), H3K27ac and ATAC can clearly separate the hits (green) and nonhits (blue and gray). *P*(in loop) > 0.5 is used to justify the enhancers, and targeting promoters are in the same CTCF loop (green and blue). The solid line represents the threshold value (criteria) of Activity = 1. The size of each dot represents the $\log_2$(FC) of each enhancer from the screen. **f**, A scatter plot showing the comparison between the CTCF loop-CIA and activity-by-contact (ABC) model with all three datasets (ESC-DE, K562 Reilly[9] and K562 Nasser[40]). The numbers of selected top predictions are based on the hits identified from each dataset, including top 24 predictions from ESC-DE, top 36 predictions from K562 Reilly and top 69 predictions from K562 Nasser. Yellow dots represent top predictions in both models. Red dots represent top predictions only in the CIA model. Blue dots represent top predictions only in the ABC model. Gray dots represent regions that do not belong to top predictions. The effect sizes from each dataset are scaled to reach the same threshold (-0.1 dashed line).

allowed comparisons between alternative hypotheses and showed that the CTCF constraint is substantially more predictive than Hi-C-based measurements of contact frequency between the enhancer and target promoter. Many distal enhancer hits within CTCF loops have a large transcriptional impact but relatively low enhancer–promoter contact frequency as measured by Hi-C. Adjusting for the distance-dependency in Hi-C data through simple power law distance corrections did not substantially improve AUPRC (Methods), but it may be possible to improve the predictions by developing newer methods based on Hi-C data or combining Hi-C with CTCF and other datasets. Mechanistically, while the prevailing model supports that transcriptional activity depends on enhancer–promoter contact, our observation of multiple enhancers implies a highly nonlinear relationship between contact frequency and transcription. This notion is supported by recent studies suggesting that contact probability has a nonlinear impact on transcription[44,48]. Through simulations, multiple plausible mechanisms have been shown to generate this nonlinear relationship, including accumulation of promoter-bound factors or post-transcriptional modifications at the promoter[44]. Instead of relying on direct enhancer–promoter contact frequency measurements from Hi-C data, our CIA model uses CTCF loop information to constrain enhancer prediction. This simple CIA model should be quite useful for prioritizing and understanding SNPs implicated in expression QTLs or GWAS-associated loci.

Our CIA model is a quantitative improvement over the 'insulated neighborhood' hypothesis[49] and is consistent with many studies showing that enhancer activity is largely restricted within CTCF loops and TADs[48,50,51] and that TAD disruption can lead to oncogenic misexpression and developmental diseases[52,53]. The CTCF loops are visually consistent with Hi-C TADs. However, TAD calling can be unstable and sensitive to parameters and methods. In comparison, *P*(in loop) calculates the probability that a genomic region is enclosed within a CTCF loop, thus providing a more reliable measurement. However, there are also seemingly conflicting findings that auxin degradation of CTCF can only have a modest effect on transcription on a short time scale[54]. We speculate that during cell state transitions, CTCF-mediated enhancer–promoter proximity is necessary for establishing de novo enhancer–protein–promoter complexes, while in post-transition steady states, this proximity may be maintained for short times without the CTCF loop. Our dynamic GRN model shows a similar result that once the active transcriptional state is established, the transcriptional response to a perturbation of enhancer activity occurs at a much longer time scale (Fig. 2e). The hysteresis in our model thus provides a plausible explanation for the long-standing paradox that CTCF is required to restrict enhancer activity, yet removing CTCF does not immediately affect transcription.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01450-7.

References

1. Luo, Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
2. Korkmaz, G. et al. Functional genetic screens for enhancer elements in the human genome using CRISPR–Cas9. *Nat. Biotechnol.* **34**, 192–198 (2016).
3. Fulco, C. P. et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
4. Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
5. Sanjana, N. E. et al. High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545–1549 (2016).
6. Klann, T. S. et al. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).
7. Diao, Y. et al. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.* **26**, 397–405 (2016).
8. Diao, Y. et al. A tiling-deletion-based genetic screen for *cis*-regulatory element identification in mammalian cells. *Nat. Methods* **14**, 629–635 (2017).
9. Reilly, S. K. et al. Direct characterization of *cis*-regulatory elements and functional dissection of complex genetic associations using HCR-FlowFISH. *Nat. Genet.* **53**, 1166–1176 (2021).
10. Gasperini, M. et al. CRISPR/Cas9-mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. *Am. J. Hum. Genet.* **101**, 192–205 (2017).
11. Huang, J. et al. Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat. Commun.* **9**, 943 (2018).
12. Thakore, P. I. et al. Highly specific epigenome editing by CRISPR–Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).
13. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
14. Wakabayashi, A. et al. Insight into GATA1 transcriptional activity through interrogation of *cis* elements disrupted in human erythroid disorders. *Proc. Natl Acad. Sci. USA* **113**, 4434–4439 (2016).
15. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol. Cell* **66**, 285–299 (2017).
16. Hong, J. W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314 (2008).
17. Barolo, S. Shadow enhancers: frequently asked questions about distributed *cis*-regulatory information and enhancer redundancy. *Bioessays* **34**, 135–141 (2012).
18. Cannavò, E. et al. Shadow enhancers are pervasive features of developmental regulatory networks. *Curr. Biol.* **26**, 38–51 (2016).
19. Perry, M. W., Boettiger, A. N., Bothma, J. P. & Levine, M. Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol.* **20**, 1562–1567 (2010).
20. Frankel, N. et al. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490–493 (2010).
21. Osterwalder, M. et al. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
22. Dailey, L. High throughput technologies for the functional discovery of mammalian enhancers: new approaches for understanding transcriptional regulatory network dynamics. *Genomics* **106**, 151–158 (2015).
23. White, M. A. Understanding how *cis*-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. *Genomics* **106**, 165–170 (2015).
24. Beer, M. A., Shigaki, D. & Huangfu, D. Enhancer predictions and genome-wide regulatory circuits. *Annu. Rev. Genomics Hum. Genet.* **21**, 37–54 (2020).
25. Li, Q. V. et al. Genome-scale screens identify JNK-JUN signaling as a barrier for pluripotency exit and endoderm differentiation. *Nat. Genet.* **51**, 999–1010 (2019).
26. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).

27. Cao, Q. et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* **49**, 1428–1436 (2017).
28. Xi, W. & Beer, M. A. Local epigenomic state cannot discriminate interacting and non-interacting enhancer-promoter pairs with high accuracy. *PLoS Comput. Biol.* **14**, e1006625 (2018).
29. Maston, G. A., Landt, S. G., Snyder, M. & Green, M. R. Characterization of enhancer function from genome-wide analyses. *Annu. Rev. Genomics Hum. Genet.* **13**, 29–57 (2012).
30. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Academic Press, 2006).
31. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).
32. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
33. Cai, X. Exact stochastic simulation of coupled chemical reactions with delays. *J. Chem. Phys.* **126**, 124108 (2007).
34. Walczak, A. M., Mugler, A. & Wiggins, C. H. Analytic methods for modeling stochastic regulatory networks. *Methods Mol. Biol.* **880**, 273–322 (2012).
35. Loh, K. M. et al. Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. *Cell Stem Cell* **14**, 237–252 (2014).
36. Boyer, L. A. et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
37. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
38. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
39. Xi, W. & Beer, M. A. Loop competition and extrusion model predicts CTCF interaction specificity. *Nat. Commun.* **12**, 1046 (2021).
40. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
41. Batut, P. J. et al. Genome organization controls transcriptional dynamics during development. *Science* **375**, 566–570 (2022).
42. Juan, A. H. & Ruddle, F. H. Enhancer timing of Hox gene expression: deletion of the endogenous Hoxc8 early enhancer. *Development* **130**, 4823–4834 (2003).
43. Rodriguez-Carballo, E. et al. Chromatin topology and the timing of enhancer function at the HoxD locus. *Proc. Natl Acad. Sci. USA* **117**, 31231–31241 (2020).
44. Xiao, J. Y., Hafner, A. & Boettiger, A. N. How subtle changes in 3D structure can create large changes in transcription. *eLife* **10**, e64320 (2021).
45. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
46. Marderstein, A. R. et al. Demographic and genetic factors influence the abundance of infiltrating immune cells in human tissues. *Nat. Commun.* **11**, 2213 (2020).
47. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* **11**, 955 (2020).
48. Zuin, J. et al. Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, 571–577 (2022).
49. Dowen, J. M. et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).
50. Anderson, E., Devenney, P. S., Hill, R. E. & Lettice, L. A. Mapping the Shh long-range regulatory domain. *Development* **141**, 3934–3943 (2014).
51. Symmons, O. et al. The Shh topological domain facilitates the action of remote enhancers by reducing the effects of genomic distances. *Dev. Cell* **39**, 529–543 (2016).
52. Wang, X. et al. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods* **18**, 661–668 (2021).
53. Yan, J. & Huangfu, D. Epigenome rewiring in human pluripotent stem cells. *Trends Cell Biol.* **32**, 259–271 (2022).
54. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **162**, 687–688 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Ethics statement

Experiments with hESCs were conducted per National Institute of Health (NIH) guidelines and approved by the Tri-SCI Embryonic Stem Cell Research Oversight Committee.

Cell lines and culture conditions

iCas9 *SOX17*$^{eGFP/+}$ HUES8 and idCas9–KRAB *SOX17*$^{eGFP/+}$ HUES8 hESCs were cultured on vitronectin-coated (Gibco, A14700) plates and maintained in E8 medium (Gibco, A1517001). Cells were dissociated by 3–5 min treatment with 0.5 mM EDTA in 1× DPBS without calcium and magnesium at room temperature. A total of 10 μM ROCK inhibitor Y-27632 (Selleck Chemicals, S1049) was added into the E8 medium for the first 24 h after passage. The medium was changed every day. hESCs were passaged every 3–4 d depending on the cell growth speed and confluency. All cell lines were routinely tested by Memorial Sloan Kettering Cancer Center (MSKCC) Antibody and Bioresource Core Facility to confirm there was no mycoplasma contamination and by MSKCC Molecular Cytogenetics Core to confirm there were no karyotyping abnormalities.

hESC-DE differentiation

We found that DE differentiation is more robust when using hESCs during the logarithmic growth phase. To optimize the differentiation, cells were maintained in the logarithmic phase before seeding for DE differentiation with a recovery passage as follows. Please note that the cell numbers described below were optimized based on the growth of the HUES8 hESCs in our laboratory, and they may need to be adjusted for other hESC lines. hESCs were dissociated with 1× TrypLE Select (Gibco, 12563029) for 3 min at room temperature. Then TrypLE was removed, and cells were washed and resuspended into E8 medium with 10 μM ROCK inhibitor Y-27632. Cells were counted by using Vi-CELL XR Cell Viability Analyzer (Beckman Coulter), and 2–3 million cells were seeded per 10-cm plate with E8 medium containing 10 μM ROCK inhibitor Y-27632. The E8 medium was refreshed daily, and the cell number should reach 10–12 million per 10-cm plate 2 d after seeding. hESCs after the recovery passage were collected and counted again, as described above, and 1 million cells per well were seeded in a six-well plate in E8 medium with 10 μM ROCK inhibitor Y-27632; and 18 h later, hESCs were changed into fresh E8 medium. At 24 h after seeding, hESCs should reach 2–2.4 million per well and be ready for differentiation. For DE differentiation, hESCs were first washed with 1× DPBS once and then cultured in 2.5 ml S1/2 differentiation medium daily, as previously described[25,55]—cells were treated with 50 ng ml^{-1} Activin A (Bon-Opus Biosciences, C687-1MG) for 3 d and 5 μM CHIR99021 (Tocris Bioscience, 4423) for the first day.

scRNA-seq

scRNA-seq was performed as previously described[56]. Cells were collected every 12 h during DE differentiation for scRNA-seq experiments with a targeted collection ranging from 3,000 to 6,000 cells. Single-cell 3′ RNA-seq libraries were generated with 10x Genomics Chromium Single Cell 3′ Reagent Kit v.3 following the manufacturer's guidelines using 10x chromium controller firmware v5.0. The libraries were sequenced on NovaSeq 6000 platform following the manufacturer's guidelines.

scRNA-seq PCA analysis

Chromium v3 analysis software Cell Ranger (version 3.1.0) was run using 'cellranger count –expect-cells 1000.' Genes with fewer than ten reads summed over all cells were removed, yielding 21,099 transcripts detected across 35,988 cells, approximately 5,000 cells per time point. The cells ranked in the bottom 10% of the total transcript number per cell were removed from further analysis. For PCA, we further restricted the analysis to transcripts whose s.d. across the seven-time points was greater than 0.2.

scRNA-seq gene expression correlation analysis

Per best practices of the Seurat package (version 4.1.1) (ref. 57), the quality control of cells included the number of features found in a cell and the percentage of tags mapped to the mitochondrial chromosome. For ESC, DE-12 h, DE-24 h, DE-36 h, DE-48 h, DE-60 h and DE-72 h cells with the following number of features were kept, respectively, 1 K–4 K, 1.5 K–5 K, 1.8 K–5 K, 2 K–5 K, 1.5 K–4 K, 1.8 K–5 K and 2 K–6 K. Additional mitochondrial gene percentage filters were applied as follows: 5–20%, 5–25%, 5–20%, 5–20%, 5–1%, 5–20% and 0–20%, respectively.

The data slot in the assay of the Seurat object, equivalent to the scaled expression of genes in cells, was used for gene-wise visualization by UMAP. For selecting the relevant expressed genes, each gene was applied a filter where the percentage of nonzero expression in cells was at least 55% at any time point. We focused on relevant expressed genes by removing mitochondrial, ribosomal, miRNAs, lincRNAs, antisense transcripts and genes that were not yet named (for example, those containing orf and starting with AC or AL). The matrix from the 2,606 relevant expressed genes, after transposition for gene-centric embedding, was embedded in UMAP using the R package umap, using the Pearson correlation as the distance metric of umap.config. Each point was labeled with the color scale defined from the expression FC from ES to DE-72 h.

GRN model

The network state in our GRN model is described by a vector of core TF concentrations $\psi = (\psi_1, \psi_2, \ldots, \psi_n) = (\psi_{i=1\cdot\cdot n})$. For example, a network with three genes and $\psi = (A, B, C)$ is shown in Fig. 1a. Each component TF gene *i* is described by an equation similar to Eq (1):

$$\frac{d\psi_i}{dt} = -r_i\psi_i + \frac{e_{i1}\left(c_i f_i\left(\psi_{i=1.n}\right) + \delta_i\left(t\right)\right)}{b_i + c_i f_i\left(\psi_{i=1.n}\right) + \delta_i\left(t\right)} + \frac{e_{i0} b_i}{b_i + c_i f_i\left(\psi_{i=1.n}\right) + \delta_i\left(t\right)} + \xi\left(t\right)\psi_i.$$

The activation probabilities for gene *i* are $p_{on} \sim c_i f\left(\psi_{i=1.n}\right) + \delta_i\left(t\right)$ and $p_{off} \sim b_i$, so together the probability that gene *i* is activated and transcribed at rate e_{i1} is $\frac{\left(c_i f_i\left(\psi_{i=1.n}\right)+\delta_i(t)\right)}{b_i + c_i f\left(\psi_{i=1.n}\right)+\delta_i(t)}$, and the probability that gene *i* is transcribed at basal rate e_{i0} is $\frac{b_i}{b_i + c_i f\left(\psi_{i=1.n}\right)+\delta_i(t)}$, with parameters b_i, c_i and $\delta_i(t)$ specific to each core TF in the network. Each term $f_i\left(\psi_{i=1.n}\right)$ will be a function of the activity of the core TFs binding gene *i*'s enhancers. As a consequence of strong nonlinear co-operativity, we now make the simplifying assumption that the core TFs turn on synchronously $(\alpha_1\psi_1 \approx \alpha_2\psi_2 \approx \ldots \approx \alpha_n\psi_n)$ where $\alpha_{i=1\cdot\cdot n}$ are constants. This allows us to approximately represent the entire network activity with the scalar $\psi\left(t\right)$ and replace the gene-specific enhancer activities with an average enhancer activity for each core TF gene, $f_i\left(\psi_{i=1.n}\right) \approx c\psi^n$, where *n* now represents an average degree of co-operativity and could arise from either individual TF co-operativity at an enhancer or from multiple enhancers interacting nonadditively. Because the probability that the gene is activated quickly saturates at large ψ, the precise form of this nonlinearity does not strongly affect the results, as long as $n \geq 3$. We similarly replace gene-specific rates b_i, c_i and $\delta_i(t)$ with weighted averages, leading to the model equation in Fig. 1b. This equation was solved by the Euler–Maruyama method using parameters $(b, c, e_0, e_1, r, n) = (.5, 1, .1, 3, 1, 3)$ unless stated otherwise, and with normally distributed noise $\xi(t)$ with amplitude $\xi_0 = 0.4$. For $e_1 \gg e_0$, the OFF fixed point is at $\psi \approx e_0/r$ and the ON fixed point is at $\psi \approx e_1/r$ and both are stable for $c \gtrsim br/e_1$ and $n \geq 3$. The stimulus $\delta(t)$ was modeled as $\delta\left(t\right) = \delta_0\left(1 + \mathrm{erf}\left(\left(t - t_0\right)/2\sqrt{2}\right)\right)/2$, with $\delta_0 = 0.035$, which turns on smoothly at $t = t_0$, to model the sustained effect of Activin A in the experiments, which is known to act through WNT/nodal signaling. The stimulus produces a weak increase in enhancer activity (p_{on}), which is independent of ψ, and is therefore independent of core TF network activity. Although the bifurcation and stability results are valid for any $e_1 \gg e_0$ and $c \gtrsim br/e_1$ and $n \geq 3$, the parameters δ_0, ξ_0, e_1, and e_0 were chosen to approximately match the distributions of *SOX17* expression levels and population variation in the FACS data. A model with a similar mathematical form was previously derived to model cellular differentiation induced by MAPK signaling and was also shown to admit bistable solutions[58,59].

Stochastic Gillespie simulations

The stochastic Gillespie algorithm was used to simulate a strongly nonlinear cooperative autoregulatory gene circuit model[33,34], analogous to the rate equation Fig. 1b but also valid for arbitrarily small numbers of molecules. In this model, up to three molecules of the TF *A* sequentially bind to an enhancer (*E*), and the differentially bound TF-enhancer complexes are denoted as *EA, EAA and EAAA*. The eight possible reactions, their probabilities, a_i, and their impact on molecular species numbers are given as follows, where X_M indicates the number of molecules of species *M*:

Production of A :	X_A+1	$a_1=c_{bt}+c_{et_1}\cdot X_{EA}$ $+c_{et_2}\cdot X_{EAA}+c_{et_3}\cdot X_{EAAA}$
Degradation of A :	X_A-1	$a_2=c_d\cdot X_A$
Binding of A to E :	$X_{EA}+1, X_E-1, X_A-1$	$a_3=c_{f_1}\cdot X_A\cdot X_E$
Unbinding of A from EA :	$X_{EA}-1, X_E+1, X_A+1$	$a_4=c_{r_1}\cdot X_{EA}$
Binding of A to EA :	$X_{EAA}+1, X_{EA}-1, X_A-1$	$a_5=c_{f_2}\cdot X_A\cdot X_{EA}$
Unbinding of A from EAA :	$X_{EAA}-1, X_{EA}+1, X_A+1$	$a_6=c_{r_2}\cdot X_{EAA}$
Binding of A to EAA :	$X_{EAAA}+1, X_{EAA}-1, X_A-1$	$a_7=c_{f_3}\cdot X_A\cdot X_{EAA}$
Unbinding of A from $EAAA$:	$X_{EAAA}-1, X_{EAA}+1, X_A+1$	$a_8=c_{r_3}\cdot X_{EAAA}$

Unless noted otherwise, the rates used are $(c_{bt}, c_{et1}, c_{et2}, c_{et3}, c_{f1}, c_{r1}, c_{f2}, c_{r2}, c_{f3}, c_{r3}, c_d)$ =(0.04, 0.04, 0.04,2.4, 0.1, 15, 0.1, 15, 0.1, 15, 0.003). The strong co-operativity of enhancer activity is reflected in the fact that we chose $c_{et3} \gg c_{bt}, c_{et1}, c_{et2}$. Fifty independent runs, each representing a single cell, were performed for each datapoint shown. The initial enhancer state at $t = 0$ is given by $(X_E, X_{EA}, X_{EAA}, X_{EAAA}) = (1, 0, 0,0)$, and the initial number of TFs, *XA*, was sampled from a uniform distribution. The stimulus is modeled by a transitory increase in the initial rate of *A* binding to *E*, c_{f1}.

Generation of the idCas9–KRAB *SOX17*$^{eGFP/+}$ HUES8 hESC line

idCas9–KRAB *SOX17*$^{eGFP/+}$ HUES8 hESC line was generated using a cassette switch strategy based on previously established iCas9 *SOX17*$^{eGFP/+}$ HUES8 hESC line[25,60]. iCas9 *SOX17*$^{eGFP/+}$ HUES8 hESCs were treated with 10 µM ROCK inhibitor Y-27632 and 2 µg ml^{-1} doxycycline 1 d before transfection. crRNAs that specifically targeted on *AAVS1–iCas9* allele and tracrRNA were ordered from IDT and cotransfected with AAVS1–idCas9–KRAB donor vector (Addgene, 199621) into the cells by using Lipofectamine Stem Transfection Reagent (Thermo Fisher Scientific, STEM00001) following manufacturer's guidelines. Transfected cells were treated by Hygromycin selection for 7 d, and single-cell colonies were picked for genotyping. Inducible dCas9–KRAB were validated by RT–qPCR and flow cytometry analysis. gRNAs and genotyping primer sequences are listed in Supplementary Tables 7 and 8.

ChIP–MS and analysis

ChIP–MS and analysis were performed, as previously described[61], with minor modifications. Antibodies used for immunoprecipitation are listed in Supplementary Table 9. Briefly, ChIP–MS was performed using the same ChIP protocol as in ChIP–seq. In total, 30–40 million cells per sample were crosslinked with 1% formaldehyde and then lysed and sonicated. Clear supernatant was collected for chromatin immunoprecipitation. Total protein was eluted after immunoprecipitation by incubation with 5% SDS and 5 mM DTT at 98 °C for 5 min. Then protein samples were alkylated, trypsinized and desalted for LC–MS/MS acquisition. For LC–MS/MS acquisition, samples were resuspended in 10 µl of 0.1% TFA and loaded onto a Dionex RSLC Ultimate 300 (Thermo Scientific), coupled with Orbitrap Fusion Lumos (Thermo Fisher Scientific). Chromatographic separation was performed with a two-column system, consisting of a C18 trap cartridge (300 µm ID, 5 mm length) and a picofrit analytical column (75 µm ID, 25 cm length) packed in-house with reversed-phase Repro-Sil Pur C18-AQ 3 µm resin. Peptides were separated using a 60 min gradient from 4% to 30% buffer B (buffer A: 0.1% formic acid, buffer B: 80% acetonitrile + 0.1% formic acid) at a flow rate of 300 nl min^{-1}. The mass spectrometer was set to acquire spectra in data-dependent acquisition mode. Briefly, the full MS scan was set to 300–1200 m/z in the orbitrap with a resolution of 120,000 (at 200 m/z) and an AGC target of 5×10^5. MS/MS was performed in the ion trap using the top speed mode (2 s), an AGC target of 10×10^4 and an HCD collision energy of 35.

Each experimental condition was performed with two biological replicates. Protein levels were $\log_2$ transformed, normalized by the average value of each sample and missing values were imputed using a normal distribution 2 s.d. lower than the mean before statistical analysis. Statistical significance was calculated using a *t*-test. Specifically, we used the *F* test to assess if the replicates for each protein are homoscedastic or heteroscedastic (equal or unequal variance). If the *F* test resulted significantly, that is, $P < 0.05$, we applied the heteroscedastic *t*-test; if not, the homoscedastic. Protein levels from EOMES, GATA6 and SOX17 ChIP–MS were further compared to IgG controls, and the common interacting TFs with $\log_2(\mathrm{FC}) > 2$ and $-\log_{10}(P) > 2$ were used to plot Fig. 3d.

RNA isolation, reverse transcription and RT–qPCR

Total RNA was extracted using Quick-RNA MiniPrep kits (ZYMO Research, R1055) following the manufacturer's guidelines. cDNA was produced by using a High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, 4368817) with 2 µg of total RNA per reaction measured by Nanodrop 2000 (Thermo Fisher Scientific). RT–qPCR reaction was performed with SYBR Green Master mix (Applied Biosystems, A25742) in the 7500 or QuantStudio 6 Flex Real-Time PCR system (Applied Biosystems). GAPDH was used as an internal control. Primers used in RT–qPCR are listed in Supplementary Table 8.

RNA-seq

After RiboGreen quantification and quality control by Agilent BioAnalyzer, 500 ng of total RNA with RIN values of 6.5–10 underwent polyA selection and TruSeq library preparation according to instructions provided by Illumina (TruSeq Stranded mRNA LT Kit, RS-122-2102), with eight cycles of PCR. Samples were barcoded and run on a HiSeq 4000 or NovaSeq 6000 platform in a PE50 run, using the HiSeq 3000/4000 SBS Kit or NovaSeq 6000 SP or S2 Reagent Kit (100 cycles; Illumina).

RNA-seq analysis

We followed the ENCODE RNA-seq processing pipeline, aligning reads to hg38 with STAR_2.5.1b and parameters '--outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000.' Transcripts were quantified with RSEM v1.2.23.

ATAC-seq

Profiling of chromatin was performed by ATAC-seq as previously described[62]. In total, 50 K cryopreserved cells were washed in cold PBS and lysed. The transposition reaction containing TDE1 Tagment DNA Enzyme (Illumina, 20034198) was incubated at 37 °C for 30 min. The DNA was purified with the MinElute PCR Purification Kit (Qiagen, 28004) and amplified for five cycles using NEBNext High-Fidelity 2× PCR Master Mix (New England Biolabs, M0541L). After evaluation by real-time PCR, 3–14 additional PCR cycles were done. The final product was cleaned by AMPure XP beads (Beckman Coulter, A63882) at a 1× ratio, and size selection was performed at a 0.5× ratio.

Libraries were sequenced on a HiSeq 4000 or NovaSeq 6000 platform in a PE50 run, using the HiSeq 3000/4000 SBS Kit or NovaSeq 6000 S1 Reagent Kit (100 cycles; Illumina).

ATAC-seq analysis

Paired-end reads were mapped to hg38 with bowtie2 version 2.2.5 using default parameters, duplicate reads were removed with Picard version 2.23.3 and peaks were called using macs2 version 2.2.7.1 and parameters 'macs2 callpeak --nomodel -g hs.' gkm-SVM (R-package version 0.82.0 with $L = 11$, $k = 7$, $d = 3$ and truncated filter) was run on the top 10,000 300 bp distal peaks (>2 kbp from TSS) and five independent GC and repeat matched negative sequence sets following[24,31,32]. The five replicate gkm-weight vectors were averaged and motifs were extracted using gkm-PWM.

gRNA design of core enhancer perturbation screen

In total, 2 Mbp upstream/downstream of the ten core TFs were selected for further chromatin accessibility filtering. Only the regions showing accessibility in either the ESC stage or the DE-48 h stage were kept. Regions that overlapped with promoters or exons were removed from the list, resulting in 394 putative enhancers being selected in total. gRNA design was performed by using CHOPCHOP[55] to achieve full-tiled coverage of the selected regions. We further added three gRNAs targeting *SOX17* promoter and 1,100 safe-targeting gRNAs[63] as positive and negative controls, respectively. gRNA sequences of the library are listed in Supplementary Table 2.

Oligo synthesis and library cloning

gRNA oligos were synthesized on-Chip (Agilent). Synthesized oligos were amplified and restriction cloned into lentiGuide-puro (Addgene, 52963) by the MSKCC Gene Editing and Screening Core Facility. Cloned plasmid library was PCR amplified to incorporate adapters for NGS. Samples were purified and sequenced using Illumina HiSeq 2500 platform. FASTQ files were clipped by position and reads were mapped back to the reference library file to show relative abundance of reads per gRNA. Reads within each sample were normalized to total number of mapped reads and library size. The overall representation of the library was charted over a one-log FC to evaluate if any gRNA was over- or under-represented in the final library. Primers used for PCR are listed in Supplementary Table 8.

Lentiviral library generation

The core enhancer perturbation lentiviral library generation was performed, as previously described[25], with minor modifications. Briefly, a total of 13.6 µg core enhancer perturbation library plasmids with 5.44 µg lentiviral packaging vector psPAX2 and 1.36 µg vesicular stomatitis virus G envelope expressing plasmid pMD2.G (Addgene plasmids, 12260 and 12259) were transfected with the JetPRIME (VMR, 89137972) reagent into 293T cells to produce the lentivirus. Fresh medium was changed 24 h after transfection, and viral supernatant was collected, filtered and stored at −80 °C for 72 h after transfection.

Core enhancer perturbation screen

We aimed for a ~1,000-fold coverage per gRNA to maximize sensitivity. A total of 35 million idCas9–KRAB *SOX17*$^{eGFP/+}$ HUES8 hESCs were collected and infected with the lentiviral library at a low MOI of ~0.3 on day 0 in 15-cm plates. A total of 6 µg ml^{-1} protamine sulfate per plate was added during the first 24 h of infection to improve the infection efficiency. One day after infection (day 1), cells were treated with 2 µg ml^{-1} doxycycline to induce dCas9–KRAB expression, which continues till the end of the screen at DE-36 h. Infected cells were selected with 1 µg ml^{-1} puromycin from day 2 to day 4 and collected on day 5 for recovery passage. Two days after recovery passage, 60 million cells were collected and seeded into 15-cm plates for DE differentiation, as described above. Thirty-six hours after differentiation, cells were dissociated using 1X TrypLE Select and sorted using FACS Aria according to GFP expression. Cells whose GFP expression levels were in the top or bottom 20% were pelleted individually, with each pellet containing 15 million cells.

gRNA enrichment sequencing and data analysis

The gRNA enrichment sequencing was manipulated by MSKCC Gene Editing and Screening Core Facility. Genomic DNA from sorted cell pellets was extracted using the QIAGEN Blood and Cell Culture DNA Maxi Kit (Qiagen, 13362) and quantified by Qubit (Thermo Fisher Scientific, Q32850) following the manufacturer's guidelines. A quantity of gDNA covering 1000x representation of gRNAs was PCR amplified to add Illumina adapters and multiplexing barcodes. Primer sequences to amplify lentiGuide-puro are shown in Supplementary Table 8. Amplicons were quantified by Qubit and Bioanalyzer (Agilent) and sequenced on the Illumina HiSeq 2500 platform. Sequencing reads were aligned to the gRNA library sequences, and counts were obtained for each gRNA. The read counts were normalized to total reads of each sample to offset differences in read depth. To calculate the *z* score of each gRNA, we subtracted the mean $\log_2$(FC) of all negative control safe-targeting gRNAs from the $\log_2$(FC) of each gRNA and then divided the result by the s.d. of $\log_2$(FC) from the negative control gRNAs (Supplementary Table 2). Off-targets of each gRNA were further assessed by CRISPOR[64]. gRNAs with 0 mismatch (MM) = 1, 1 MM < 10, 2 MM < 30, 3 MM < 100 and total raw reads > 00 were kept for calculating average *z* score of each putative enhancer region. Putative enhancer regions with less than three qualified gRNAs were filtered out (Supplementary Table 3).

Hit validation

Selected gRNAs for each enhancer hit were cloned into lentiGuide-puro (for single perturbations) and lentiGuide-blast (when a second perturbation was used in combination; Addgene, 199622). In total, 1.36 µg lentiGuide-puro (or lentiGuide-blast), 0.1 µg pMD2.G and 0.4 µg psPAX2 plasmids were transfected with the JetPRIME (VMR, 89137972) reagent into 293T cells to pack lentiviruses. Viral supernatant was made and collected, as described above. idCas9–KRAB SOX17$^{eGFP/+}$ HUES8 hESCs were then infected with viruses containing different gRNAs individually following the same process as described above for the screen. One day after infection (day 1), cells were treated with 2 µg ml^{-1} doxycycline to induce dCas9–KRAB expression, which continues till the end of the experiment (DE-36 h or DE-72 h). Infected cells were selected with 1 µg ml^{-1} puromycin from day 2 to day 4 and collected on day 5 for recovery passage and followed by DE differentiation described above. For dual selection, cells were selected with 1 µg ml^{-1} puromycin and 10 µg ml^{-1} blasticidin together for 5 d. Cells were collected at both 36 h and 72 h for flow cytometry analysis. gRNA sequences selected from the core enhancer validation are listed in Supplementary Table 10.

Flow cytometry

Antibodies used for flow cytometry are listed in Supplementary Table 9. For live GFP and surface marker data collection, cells were dissociated and stained with DAPI and corresponding antibodies at room temperature for 15 min. For TF data collection, cells were first stained with LIVE-DEAD Fixable Violet Dead Cell Stain (Invitrogen, L34955) at room temperature for 15 min and then fixed with Fixation/Permeabilization reagent (Invitrogen, 00-5223-56/00-5123-43) and stained with corresponding antibodies using Permeabilization buffer (Invitrogen, 00-8333-56). Flow cytometry data were collected using BD LSRFortessa or BD LSRII with BD FACSDIVA. Flow cytometry analysis and figures were generated in FlowJo v10.

Generation of clonal enhancer KO hESC lines

Enhancer KO hESC lines were generated by using two paired crRNAs surrounding targeted enhancers to increase knockout efficiency. crRNAs and tracrRNA were ordered from IDT. iCas9 *SOX17*$^{eGFP/+}$ HUES8 hESCs were treated with 2 µg ml^{-1} doxycycline and 10 µM ROCK inhibitor

Y-27632 for 24 h, dissociated with 1X TrypLE Select and transfected with 0.15 μM of each crRNA and 0.6 μM of tracrRNA by using Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher Scientific, 13778100) following the manufacturer's guidelines. Transfected cells were further cultured in E8 with 10 μM ROCK inhibitor Y-27632 for 48 h and ~2,000 cells were seeded into 100-mm plate to raise colonies. Then, the genomic DNA of the individual colony was extracted by using DNeasy Blood and Tissue DNA Kit (Qiagen, 69506) for genotyping. crRNA and genotyping primer sequences are listed in Supplementary Tables 7 and 8.

ChIP–seq

ChIP–seq was performed, as previously described[25], with minor modifications. Antibodies used for immunoprecipitation are listed in Supplementary Table 9. For each sample, around 30 million cells were crosslinked with 1% formaldehyde and quenched with 0.125 M glycine. Fixed cells were then lysed in 700 μl SDS buffer (1% SDS, 10 mM EDTA, 50 mM Tris–HCl, pH 8), and incubated for 10 min on ice. Sonication was performed on a Branson Sonifier 150 set at 30% amplitude for 5 min and 30 s total on (10-s on/10-s off pulsing). Clear supernatant was collected for antibody binding overnight, followed by Dynabeads (Thermo Fisher Scientific, 10004D) incubation for 6 h at 4 °C. Then the beads were pelleted and washed twice with low salt (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl, pH 8, 150 mM NaCl), high salt (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl, pH 8, 500 mM NaCl) and TE buffer (10 mM Tris–HCl, pH 8, 1 mM EDTA), respectively. The DNA was eluted from the beads by incubating in elution buffer (1% SDS, 0.1 M $NaHCO_3$) at 65 °C for 15 min and decrosslinked with 5 M NaCl at 65 °C overnight. A total of 10 μl (0.5 M) EDTA, 20 μl (1 M) Tris–HCl (pH 6.5) and 1 μl Proteinase K (20 mg ml^{-1}) were added to decrosslinked product and incubated for 1 h at 45 °C. DNA was isolated by using QIAquick PCR purification kit (Qiagen, 28104). Then the sequencing library was generated by using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, E7103S) and NEBNext Multiplex Oligos for Illumina (New England Biolabs Index Primers Set 1; NEB, E7335S). Samples were pooled and submitted to MSKCC Integrated Genomics Operation core for quality control and sequencing on Illumina HiSeq 4000 platform.

ChIP–seq/ChIA–PET analysis

Paired-end reads were mapped to hg38 with bowtie2 version 2.2.5 using default parameters, and peaks were called using macs2 version 2.2.7.1 using parameters 'macs2 callpeak --nomodel -g hs.' Processed ChIA–PET bedpe files for K562, GM12878 (ref. 65), H1 and HUVEC[38] were downloaded from encodeproject.org.

Hi-C

Two million cells were collected and fixed with 1% formaldehyde. The subsequent steps of Hi-C were then performed using the Arima-Hi-C kit (Arima, A510008), while libraries for sequencing were prepared with the KAPA Hyper Prep Kit (KAPA, KK8502) following the manufacturers' guidelines. Samples were pooled and submitted to MSKCC Integrated Genomics Operation core for quality control and sequencing on Illumina HiSeq 4000 platform.

Hi-C analysis

Before alignment, Hi-C Pro 2.11.4 was used to fragment, with GATC and GANTC as the restriction sites, with all alternative haplotypes removed. Reads were aligned with default settings for Bowtie 2.4.1 in Hi-C Pro, for both the global and local alignment steps. After both alignment steps, reads with a MAPQ of at least 30 were retained for further analysis and duplicates were removed. Sample level Hi-C maps were converted to .hic file format with JuicerTools 1.22.01. Condition-specific Hi-C maps were generated by combining all sample level.allValidPairs files and then converting them to the .hic file format with JuicerTools. Hi-C datafile ENCFF080DPJ.hic (ref. 54) for K562 was downloaded from encodeproject.org. Hi-C.hic datafiles for HUES64 and differentiated to EC, MS and EN were downloaded from NCBI GEO accession GSE130085 (ref. 66). Contact frequency .bedpe files were generated from Juicer Tools version 1.22.01 or 2.13.05 (K562).

Predictive modeling of screen hits

Scores for ATAC, DHS, ChIP–seq and H3K27ac signal features were mapped onto uniform 1,000 bins centered on each target enhancer. For discriminative AUPRC analysis, positive hits were defined as $\log_2$(FC) > 0.15 for hESC-DE (24 positive hits of 160 DE enhancers tested; Supplementary Table 4). For K562 Reilly, we downloaded all 'Flow-Fish CRISPR Screen' 'tsv' or 'tsv guide quantification' files from https://www.encodeproject.org, which yielded experiments at 20 loci[9]. For analysis of the Reilly data, we mapped gRNAs to nonpromoter (>2 kbp from TSS) K562 DHS peaks and normalized high and low expression gRNA counts. Twelve genes had an enhancer hit with $\log_2$(FC) > 0.8 (36 positive hits of 450 K562 regions; Supplementary Table 5), and we included these loci in model evaluation (Extended Data Fig. 9a). For the Nasser data, we downloaded Supplementary Table 5 from ref. 40 and mapped all tested regions to nonpromoter K562 DHS peaks within 1 Mbp of a tested gene. Hits were defined by Regulated=TRUE in their Supplementary Table 5 (69 positive hits of 1,931 K562 regions; Supplementary Table 6), flanking 65 genes. For modeling, Hi-C contact frequency was calculated in 5 kbp bins from the .hic files normalized to one for the most promoter proximal bin using our data in ESC and DE, and in ENCFF080DPJ in K562. We also tried as a feature a distance-corrected promoter Hi-C contact frequency, Hi-C × $|\text{dist}|^{k}$, but found no improvement in AUPRC over $k = 0$ for k between 0 and 2. *P*(in loop) for each enhancer–promoter pair was calculated from ChIA–PET reads for all loops in a 2 Mbp window spanning the promoter. Total *P*(in loop) for each distal enhancer–promoter pair is given by the ratio of total ChIA–PET read counts of all loops spanning both the enhancer and the promoter divided by the total counts of all loops containing the promoter (but not necessarily also containing the enhancer). The minimum threshold for loop calls in the ChIA–PET data is either 3 or 4 reads, and we also used this threshold for loop counts. To reduce variability in the ChIA–PET data, we averaged counts for multiple datasets. For hESC-DE and Reilly, H1, HUVEC and GM12878 ChIA–PET were used, while for K562 Nasser, K562 ChIA–PET was used, but different ChIA–PET datasets yield very similar *P*(in loop) and predictive performance. *P*(in loop) was normalized to one for the most promoter proximal bin. Logistic regression was used to combine features and predict performance. For these very low dimensional logistic regression models, test set performance is reduced by <2% compared to using the full dataset, which we used to reduce statistical variation when comparing all models. Spearman correlation was calculated between the probability of being in positive class and enhancer effect ($\log_2$(FC)). To compare all models in Fig. 7f, $\log_2$(FC) in ESC-DE and K562 Reilly were scaled to 'effect size' by dividing $\log_2$(FC) by 1.5 for ESC-DE and 8.0 for K562 Reilly so all datasets had a similar effect size threshold of 0.1 for hits. We took the top predictions from each model to compare performance at constant recall (24 in ESC-DE, 36 in K562 Reilly and 69 in K562 Nasser, 129 total).

Statistics and reproducibility

All datapoints refer to biological replicates. No statistical method was used to predetermine sample sizes. The investigators were not blinded to allocation during experiments and outcome assessment. Stochastic simulations used randomized noise and averaged over 4,000 cells for Fig. 1b and 50 independent runs for Gillespie simulations. No data were excluded from the analyses unless the differentiation experiment itself failed. The number of biological replicates is reported in the legend of each figure. Flow cytometry analysis and RT–qPCR experiments were derived from at least three independent biological experiments. For bulk ATAC-seq, ChIP–seq, ChIP–MS, Hi-C and bulk RNA-seq, quantification and statistics were derived from at least two independent biological experiments except for one biological replicate of DED1–GATA6 ChIP–seq and DED2–GATA6 ChIP–seq. Screen and scRNA-seq experiments were performed once. All the statistical analysis methods are indicated in the figure legends of Figs. 3–5 and 7 and Methods. Quantification of flow cytometry and RT–qPCR data are shown as the mean ± s.d. Student's *t*-test was used for comparison between two groups. Analysis of variance was used for multiple comparisons. Statistical significance (exact *P* value) is indicated in each figure.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The parental HUES8 hESC line was obtained from Harvard University under a material transfer agreement. Sequencing data are available at GEO under accession GSE213394 (new data from this study) and GSE114102 (published DE-72 h H3K4me1 ChIP–seq data), GSE63525 (published K562 Hi-C data) and GSE72816, GSE177081, GSE177471 (published ChIA–PET data). The Hi-C data are available in the 4D Nucleome Data Portal (https://data.4dnucleome.org/) under accession numbers 4DNESDO2ZYBM, 4DNESQMUTYXH, 4DNESFL8KDMT, 4DNESW8SIXN7, 4DNESW9GVC97, and 4DNESI1DNSGF. Mass spectrometry data are available in the PRIDE database under ProteomeXchange accession PXD043070. Source data are provided with this paper.

Code availability

Publicly available software and packages were used throughout this study according to each developer's instructions. The MATLAB codes are provided in the Supplementary Code.

References

55. Rezania, A. et al. Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. *Nat. Biotechnol.* **32**, 1121–1133 (2014).
56. Yang, D. et al. CRISPR screening uncovers a central requirement for HHEX in pancreatic lineage commitment and plasticity restriction. *Nat. Cell Biol.* **24**, 1064–1076 (2022).
57. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
58. Xiong, W. & Ferrell, J. E. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature* **426**, 460–465 (2003).
59. Wang, L. et al. Bistable switches control memory and plasticity in cellular differentiation. *Proc. Natl Acad. Sci. USA* **106**, 6638–6643 (2009).
60. Zhu, Z. et al. Genome editing of lineage determinants in human pluripotent stem cells reveals mechanisms of pancreatic development and diabetes. *Cell Stem Cell* **18**, 755–768 (2016).
61. Dixon, G. et al. QSER1 protects DNA methylation valleys from de novo methylation. *Science* **372**, eabd0875 (2021).
62. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
63. Morgens, D. W. et al. Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat. Commun.* **8**, 15178 (2017).
64. Concordet, J. P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018).
65. Tang, Z. et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
66. Wu, H. J. et al. Topological isolation of developmental regulators in mammalian genomes. *Nat. Commun.* **12**, 4897 (2021).
67. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).

Acknowledgements

We acknowledge the assistance from the following MSKCC Cores: Antibody & Bioresource, Flow Cytometry, Gene Editing & Screening, Integrated Genomics Operation and Molecular Cytogenetics. We thank R. Garippa, H. Liu and S. Mehta for assistance with CRISPR library generation and HiSeq for quantifying gRNA abundance after CRISPRi screen; Y. Furuta, J. Liu, Y. Lan and C. Schwarz for assisting with additional experiments not included in the manuscript and C.S. Leslie, L. Studer, A. Ventura, A.-K. Hadjantonakis, T. Evans and members of D.H.'s laboratory for insightful advice. We acknowledge the funding resources from the following: National Institutes of Health grant U01HG012051 (to D.H., M.B. and T.V.), National Institutes of Health grant U01DK128852 (to D.H. and E.A.), National Institutes of Health grant R01HG012367 (to M.B.), National Institutes of Health grant U01HG009380 (to M.B.), National Institutes of Health grant R56HG012110 (to M.B.), National Institutes of Health grant R01DK096239 (to D.H.), National Institutes of Health grant 1S10OD030286-01 (to S.S.), National Institutes of Health grant 1S10OD030286-01 (to S.S.), Starr Tri-I Stem Cell Initiative #2019-001 (to D.H. and E.A.), American Federation for Aging Research Sagol Network GerOmics award (to S.S.), Deerfield Xseed award (to S.S.), National Institutes of Health MSKCC Cancer Center Support Grant P30CA008748, Einstein Cancer Center grant P30CA013330 (to S.S.), NYSTEM postdoctoral fellowship (to D.Y.) from the MSKCC Center for Stem Cell Biology DOH01-TRAIN3-2015-2016-00006 and National Institutes of Health T32 training grant T32GM008539 (to B.P.R.).

Author contributions

R.L., D.H. and M.A.B. devised experiments and interpreted results. R.L. performed most experiments and analyzed the results. M.A.B. developed the mathematical models and performed computational data analysis, with contributions from J.W.O., W.X. and D.S. J.Y., B.P.R., D.Y. and Q.V.L. assisted with ChIP–seq. S.S. supervised and J.Y. and R.C. assisted with ChIP–MS. T.V. and D.H. supervised and R.A.G. and T.C. assisted with validation. E.A. and D.H. supervised and J.P., D.M. and W.W. assisted with Hi-C and subsequent data analysis. H.S.C. performed gene expression correlation analysis. R.L., D.H. and M.A.B. wrote the manuscript; all other authors provided editorial advice.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01450-7.

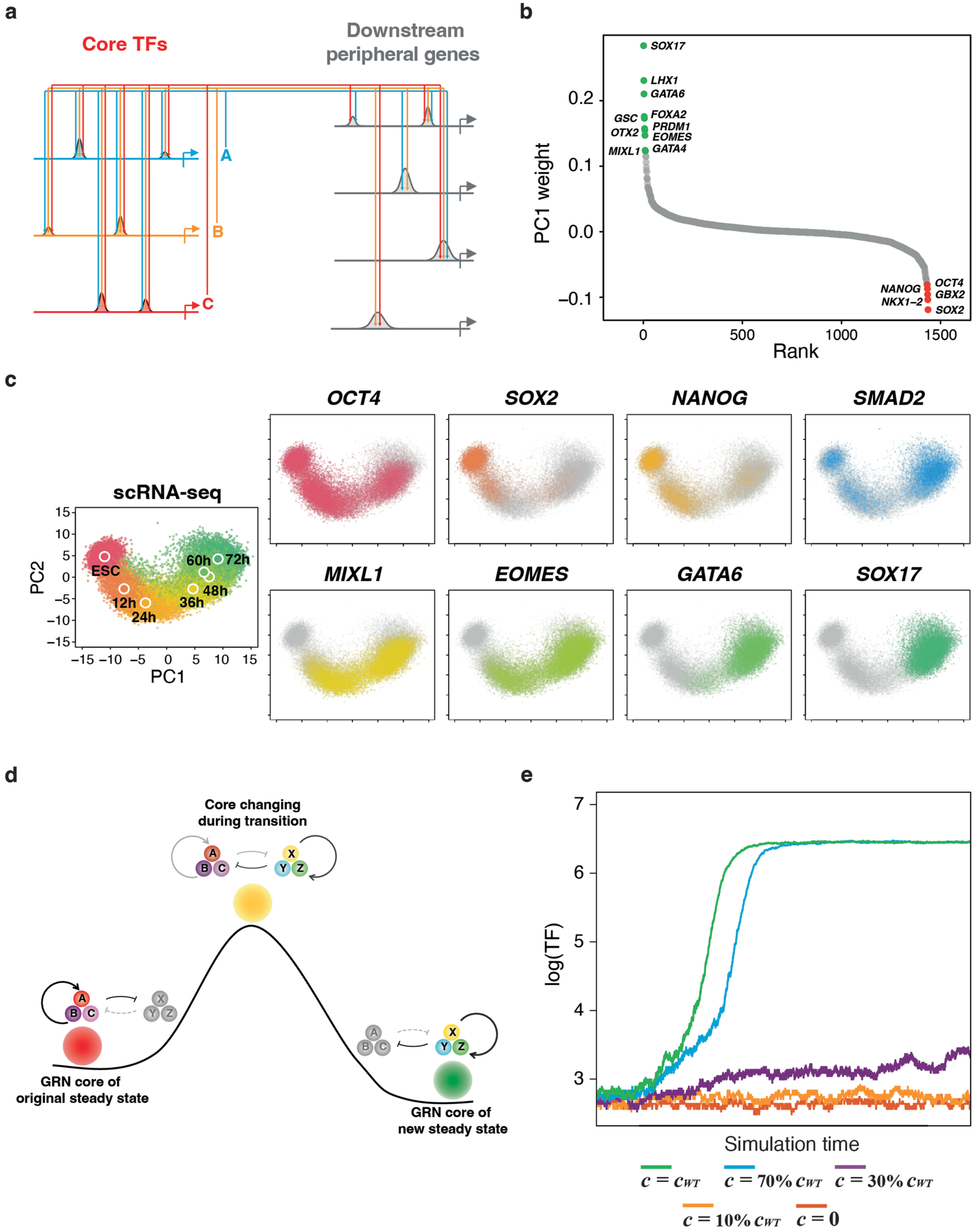
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01450-7.

Correspondence and requests for materials should be addressed to Danwei Huangfu or Michael A. Beer.

Peer review information *Nature Genetics* thanks Kyle Loh and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

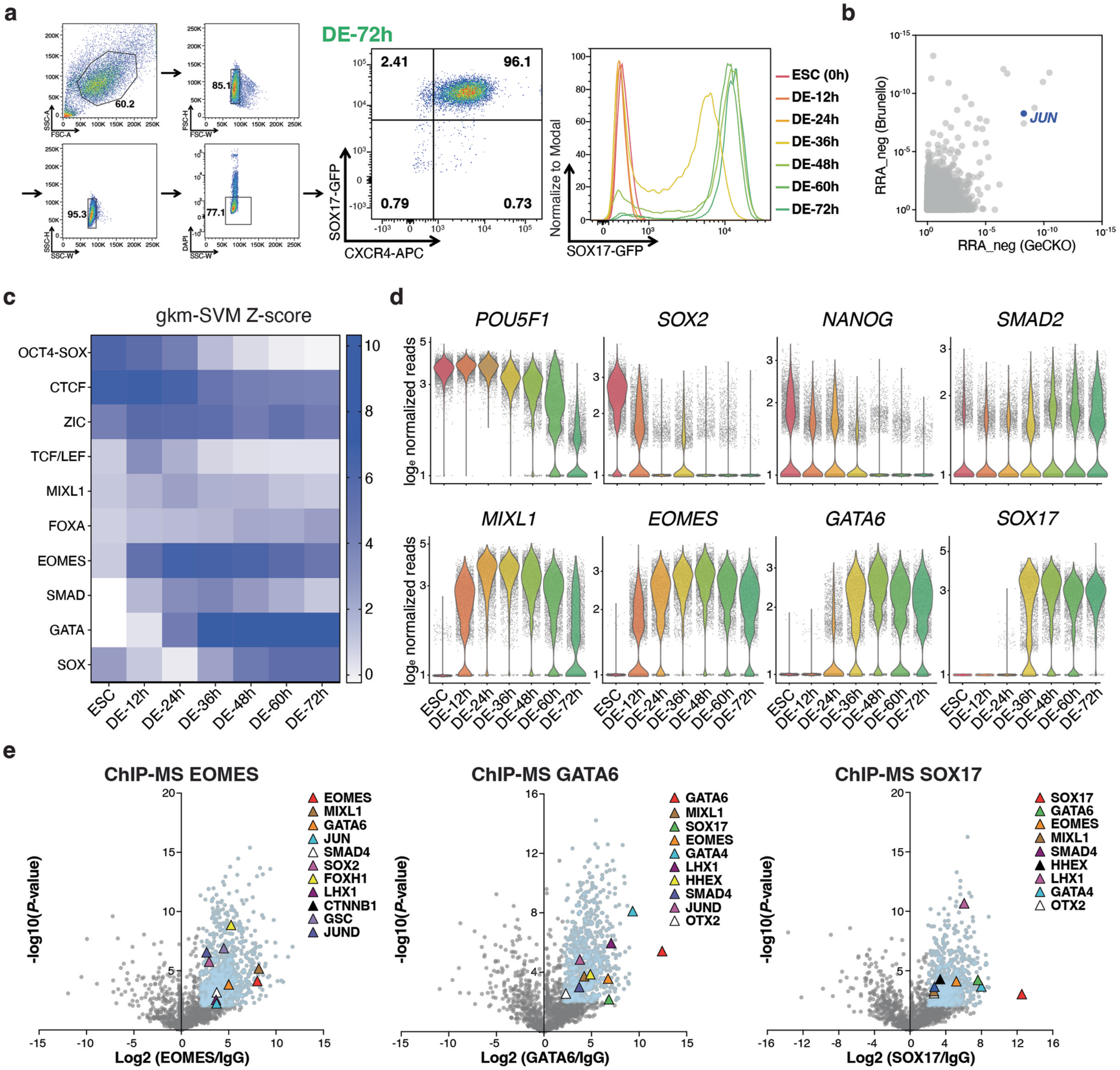
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41588-023-01450-7

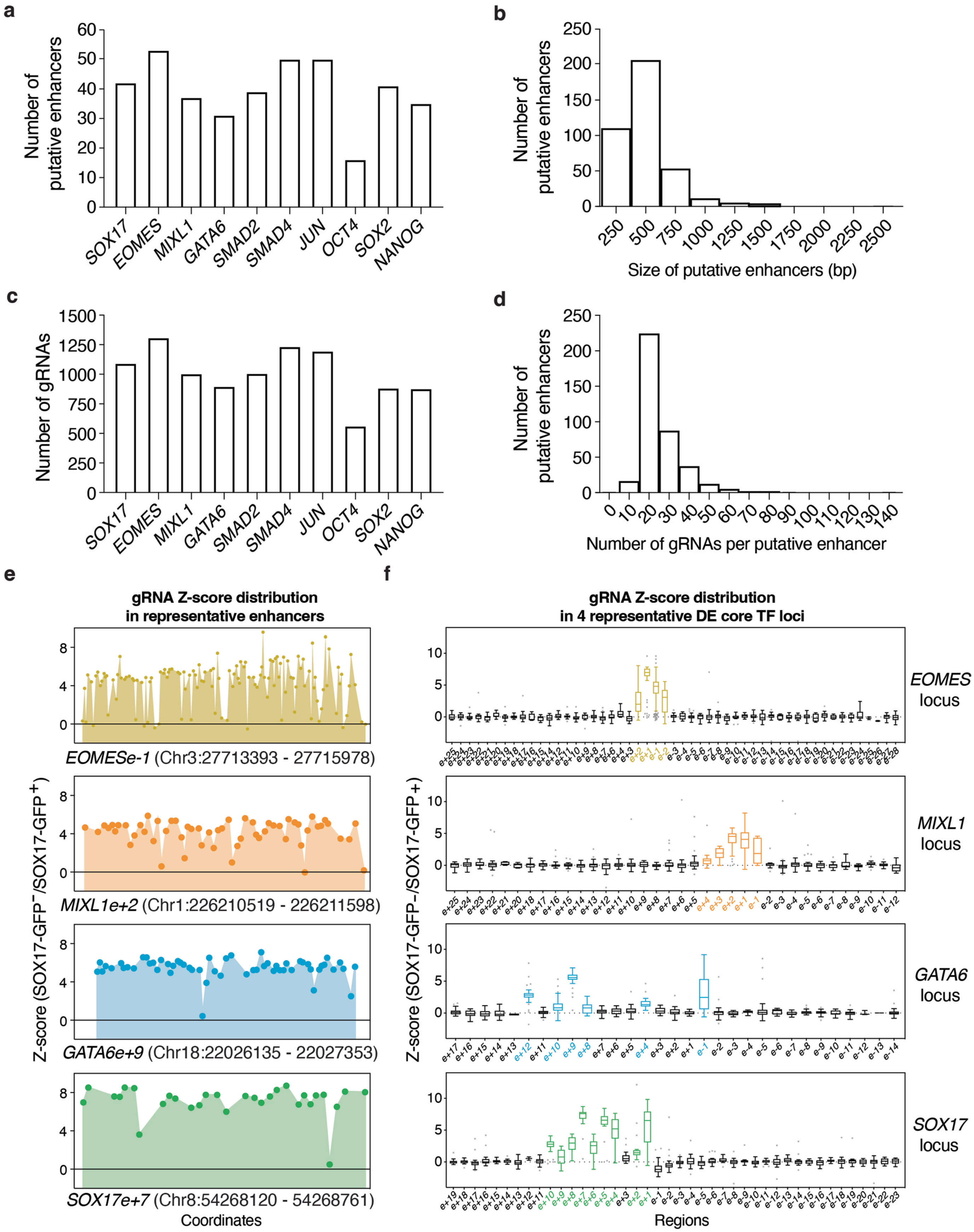


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Supporting data for the dynamic GRN model. a, The detailed schematic of core circuit in the gene regulatory network (GRN). The core transcription factors (TFs) cooperatively auto-regulate each other by binding to core enhancers and co-regulate downstream peripheral genes by binding to peripheral enhancers. **b**, The ranking plot of principle component 1 (PC1) weight of all TFs in PCA analysis from scRNA-seq data during human embryonic stem cell to definitive endoderm (hESC-DE) transition. **c**, The principle component analysis (PCA) plots showing selective TFs from the PCA component 1 (Extended Data Fig. 1b) of single-cell RNA-seq (scRNA-seq) sampled every 12 hours during hESC-DE transition. **d**, The schematic of core circuit establishment during cell state transition, similar to *Moris* et al.[67]. The transition of a cell from one steady state to another is accompanied by the deconstruction of the original core circuit (A, B, C) and the establishment of core circuit of the new state (X, Y, Z). **e**, Stochastic Gillespie simulations of the dynamic GRN network model. The green, cyan, purple, yellow and orange lines represent 100%, 70%, 30%, 10% and 0% of original total enhancer strength respectively.

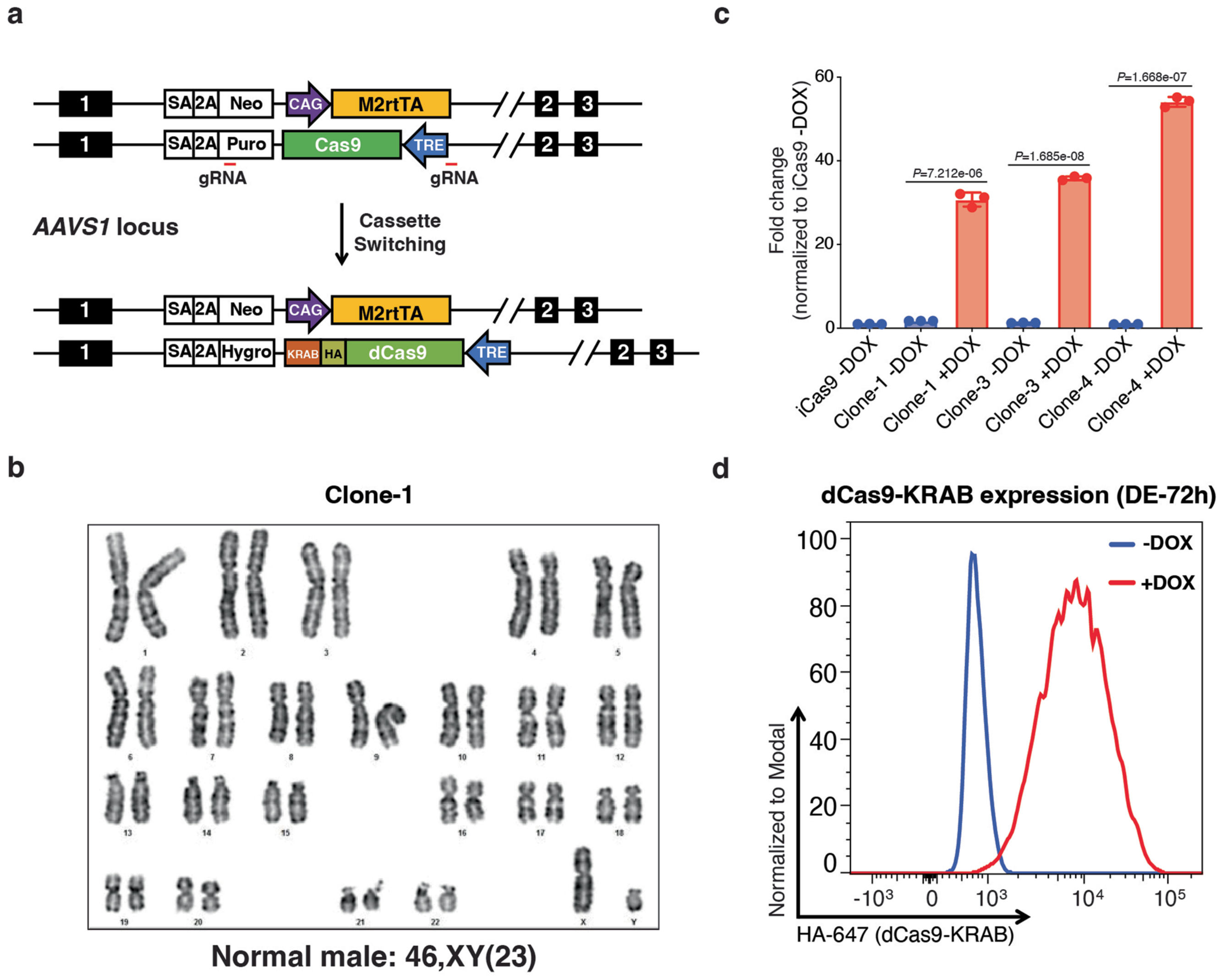


Extended Data Fig. 2 | Core TFs identification and characterization during hESC-DE transition. a, Flow cytometry analysis showing the gating strategy (left), differentiation efficiency at DE-72h measured by DE markers SOX17 and CXCR4 (middle) or transition efficiency every 12 h measured by SOX17 (right). **b**, MAGeCK robust ranking aggregation (RRA) scores for negative hits in two genome-scale DE screens from Li et al.[25]. JUN is the only identified TF among the negative hits. **c**, Motif z score of ATAC-seq by gkm-SVM at each time point during hESC-DE transition. **d**, Feature violin plots from scRNA-seq data showing core TFs expression changing during hESC-DE transition at single cell resolution. **e**, Volcano plots showing protein-protein interactions identified by ChIP-MS using EOMES as the bait at DE-24h, GATA6 and SOX17 as baits at DE-48h. Blue dots represent the significantly enriched proteins with log2FC > 2 and -log10 (*P*-value) > 2. Selective TFs enriched in ESC and endoderm GO terms are labeled by triangles.

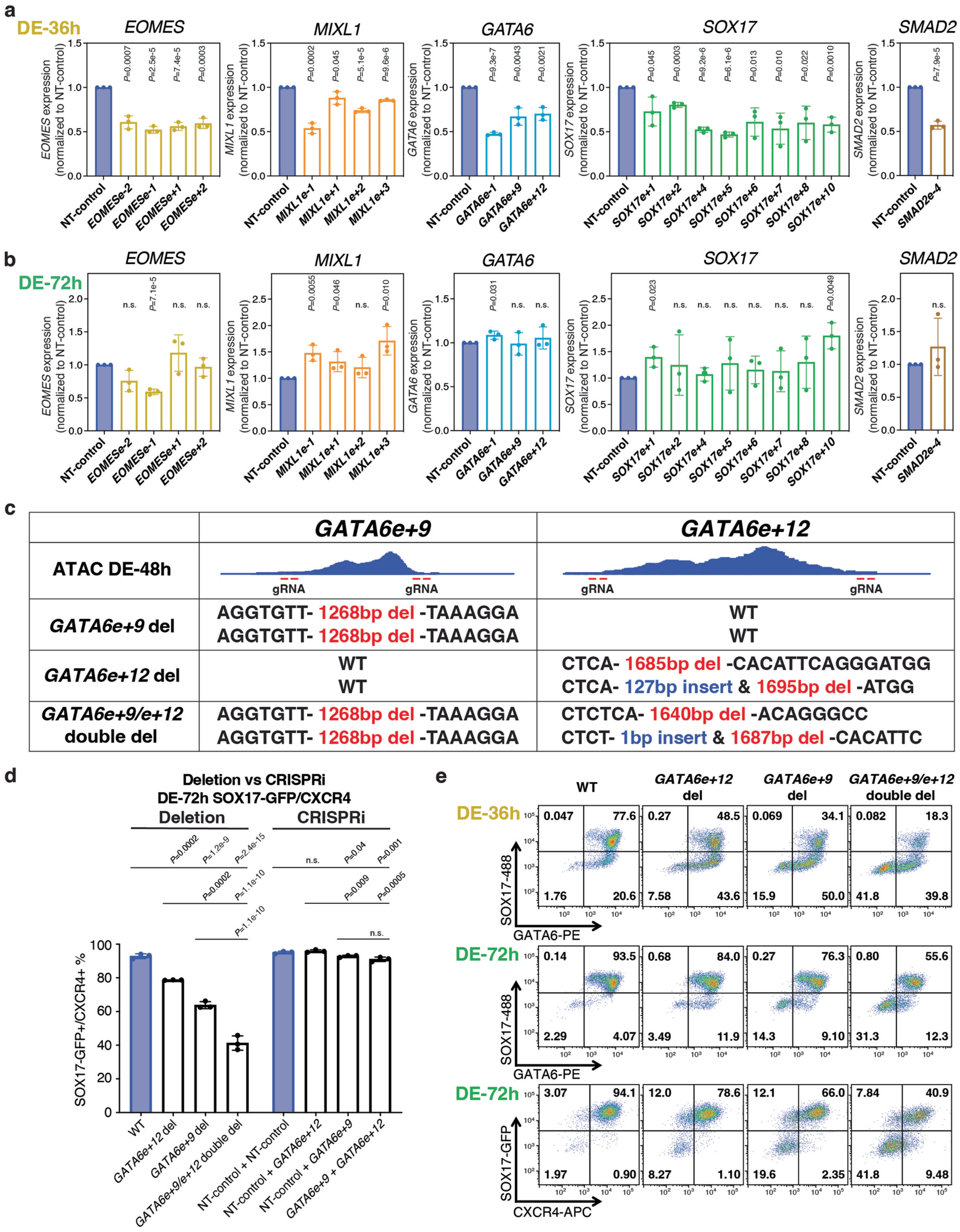


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Supporting data for the screen design and gRNA enrichment analysis. a to **d**, Statistics of putative enhancers selection and gRNAs design. The number of putative enhancers selected for each core TFs (**a**), the size of putative enhancers (**b**), the total number of gRNAs targeted on putative enhancers of each core TF (**c**), the number of gRNAs targeted on each putative enhancer (**d**). **e**, The gRNA z score distribution at representative enhancers showing gRNAs targeting the same enhancer have similar perturbation effect. **f**, Box plots showing the gRNA z score distribution in all putative enhancers of *EOMES*, *MIXL1, GATA6* and *SOX17* loci. All box plots follow the following format: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers.



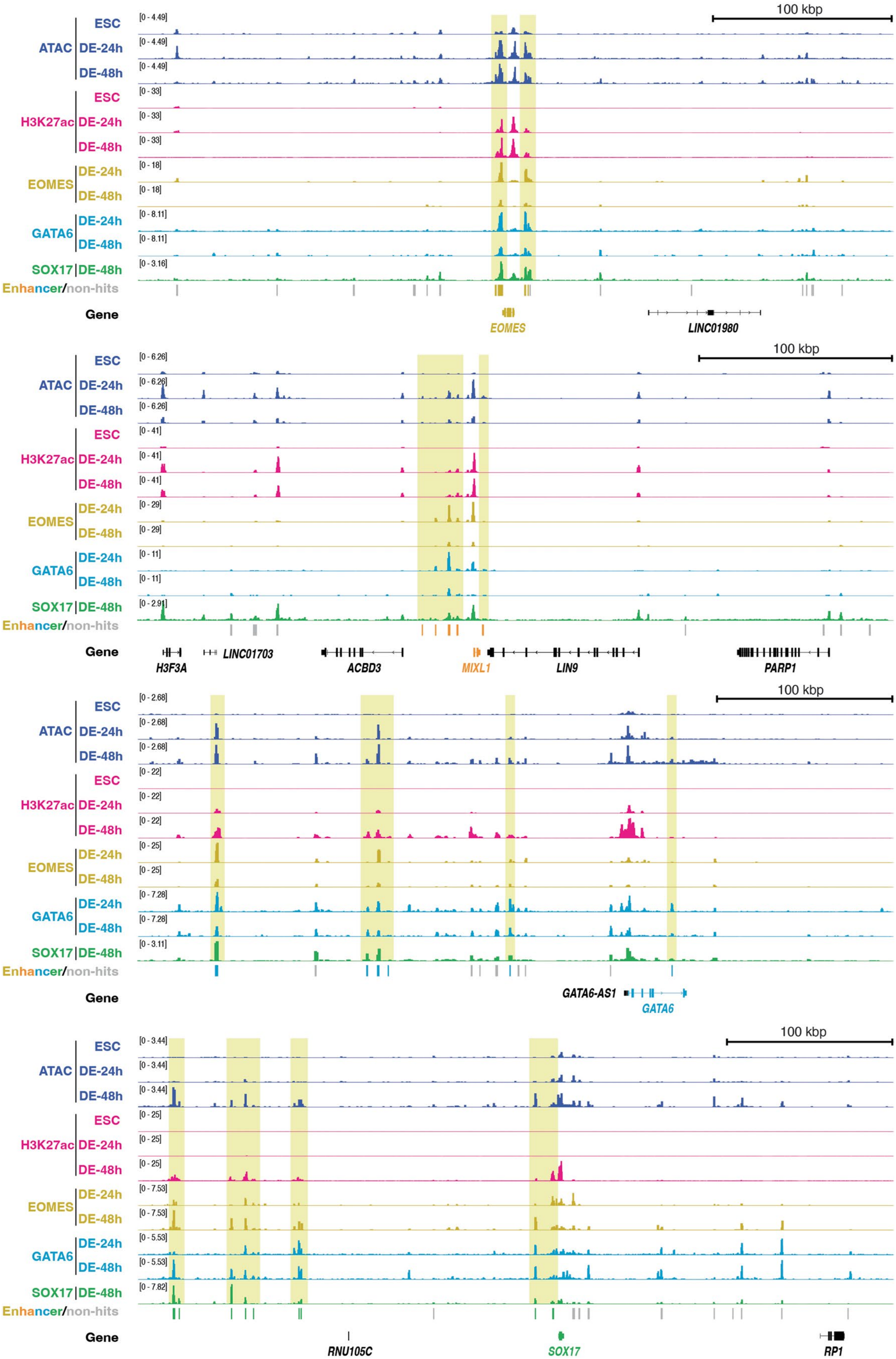
Extended Data Fig. 4 | idCas9-KRAB *SOX17*$^{GFP/+}$ hESC line generation using cassette switching. **a**, The schematics of idCas9-KRAB *SOX17*$^{GFP/+}$ hESC line generation using cassette switching. gRNAs targeting the puromycin selection cassette and the 5′ sequence outside TRE are designed for inducing double-strand break for homology repair. **b**, Karyotyping results of the idCas9-KRAB *SOX17*$^{GFP/+}$ hESC line. **c**, RT-qPCR results showing the inducible expression of dCas9-KRAB with doxycycline treatment. n = 3 biologically independent experiments. Error bars indicate mean ± SD. Statistical analysis was performed by two-tailed unpaired student *t*-test. **d**, Flow cytometry results showing the inducible expression of dCas9-KRAB with doxycycline treatment.



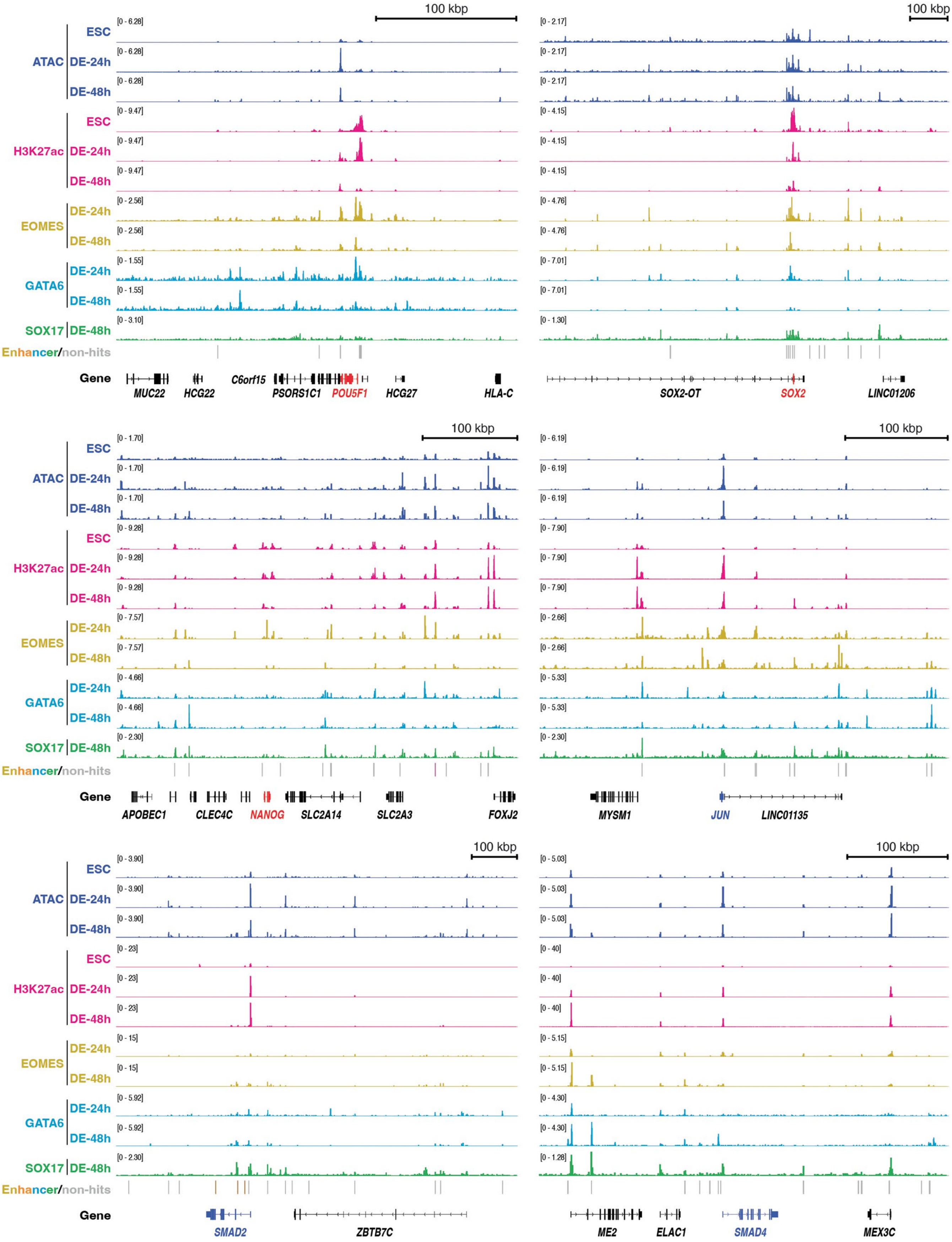
	GATA6e+9	*GATA6e+12*
ATAC DE-48h	gRNA gRNA	gRNA gRNA
GATA6e+9 del	AGGTGTT- 1268bp del -TAAAGGA AGGTGTT- 1268bp del -TAAAGGA	WT WT
GATA6e+12 del	WT WT	CTCA- 1685bp del -CACATTCAGGGATGG CTCA- 127bp insert & 1695bp del -ATGG
GATA6e+9/e+12 double del	AGGTGTT- 1268bp del -TAAAGGA AGGTGTT- 1268bp del -TAAAGGA	CTCTCA- 1640bp del -ACAGGGCC CTCT- 1bp insert & 1687bp del -CACATTC

Extended Data Fig. 5 | See next page for caption.

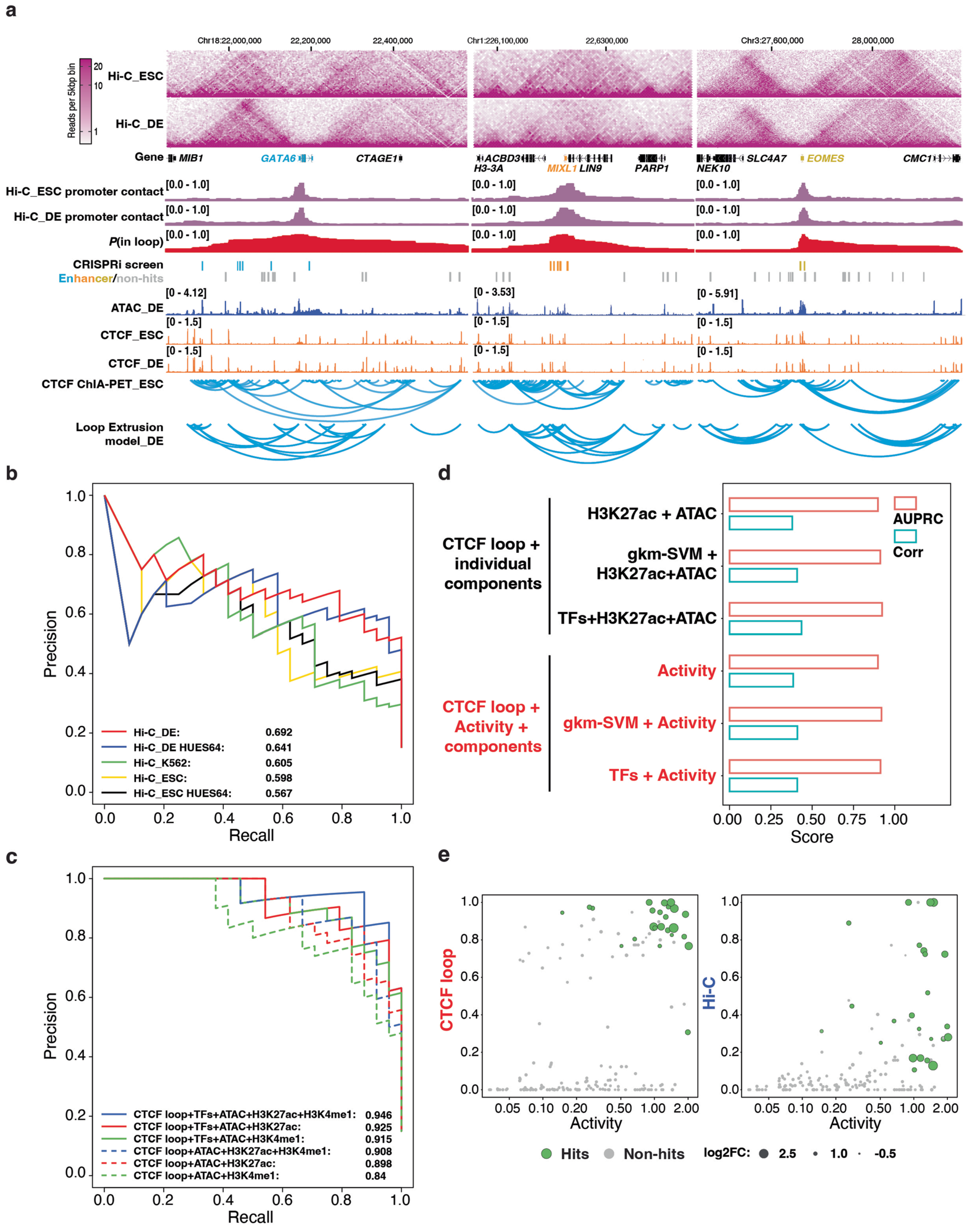
Extended Data Fig. 5 | Supporting data for validation of core enhancers. **a**, **b**, RT-qPCR showing the expression of the cognate genes decreases by enhancer perturbations at DE-36h (**a**) but mostly restored at DE-72h (**b**). n = 3 biologically independent experiments. Error bars indicate mean ± SD. Statistical analysis was performed by two-tailed unpaired student *t*-test. n.s.: not significant. **c**, Illustration of the enhancer deletion experiments that resulted in the *GATA6e + 9* deletion (del), *GATA6e + 12* del and *GATA6e + 9/e + 12* double del hESC lines. **d**, Statistics of SOX17-GFP/CXCR4 double positive cells at DE-72h in WT, *GATA6e + 9* del, *GATA6e + 12* del, *GATA6e + 9/e + 12* double del cells, as well as cells with non-targeting control, *GATA6e + 9* perturbation, *GATA6e + 12* perturbation and *GATA6e + 9/GATA6e + 12* dual-perturbation. n = 3 biologically independent replicates. Error bars indicate mean ± SD. Statistical analysis was performed by two-tailed unpaired multiple comparison test with Dunnett correction. n.s.: not significant. **e**, Flow plots showing SOX17/GATA6 expression at DE-36h, DE-72h and SOX17-GFP/CXCR4 expression at DE-72h of WT, *GATA6e + 9* del, *GATA6e + 12* del, *GATA6e + 9/e + 12* double del.



Extended Data Fig. 6 | Epigenetic features of DE core enhancers. Relevant ATAC-seq and ChIP-seq tracks of 4 DE core TF loci. Yellow boxes highlight the DE core TFs (EOMES, GATA6 and SOX17) bind to DE core enhancers. Genomic coordinates from GRCh38 (human hg38) for each gene are labeled. kbp, kilobase pair.

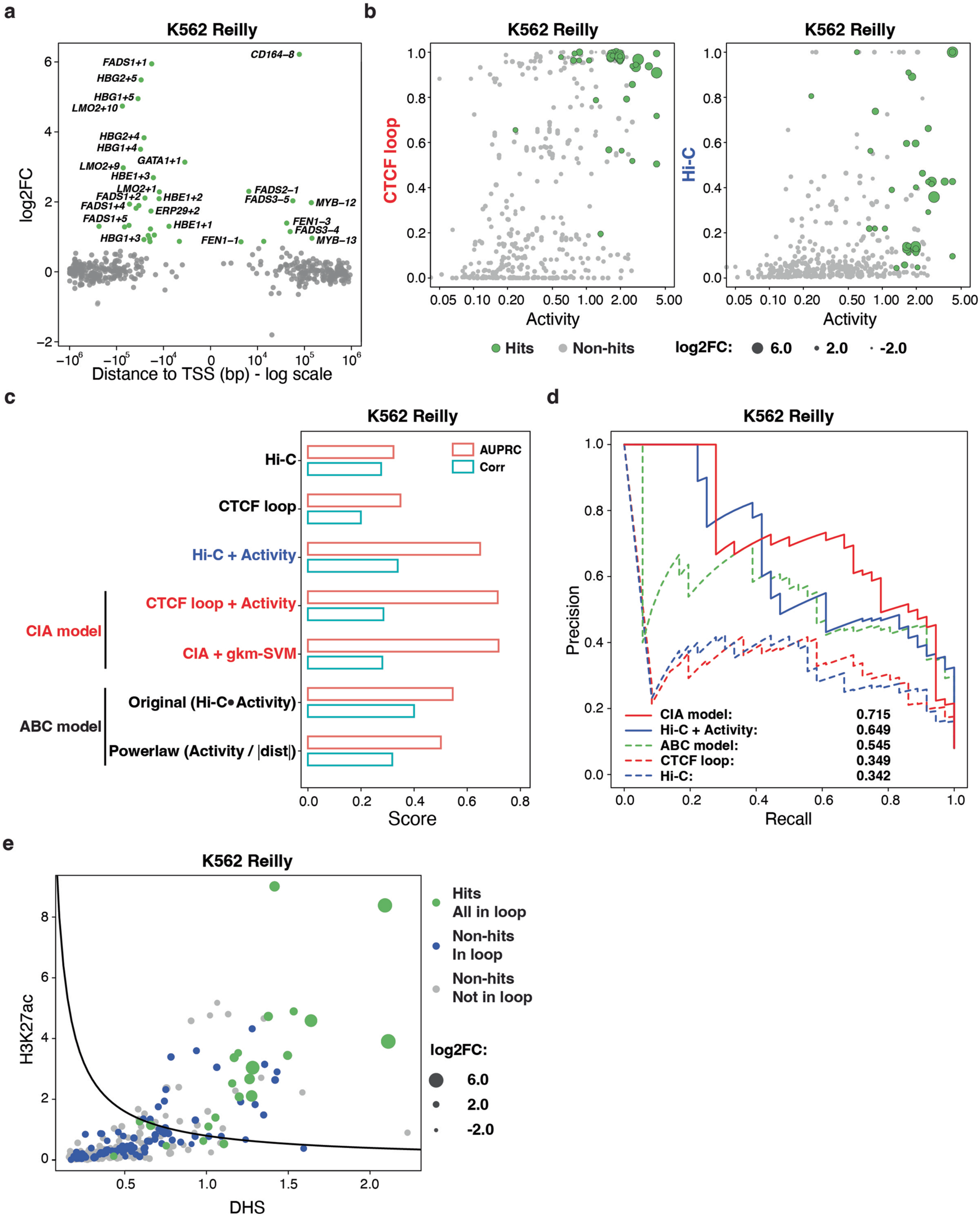


Extended Data Fig. 7 | Epigenetic features of ESC and signaling core TF loci. Relevant ATAC-seq and ChIP-seq tracks of ESC and signaling core TF loci. Genomic coordinates from GRCh38 (human hg38) for each gene are labeled. kbp, kilobase pair.



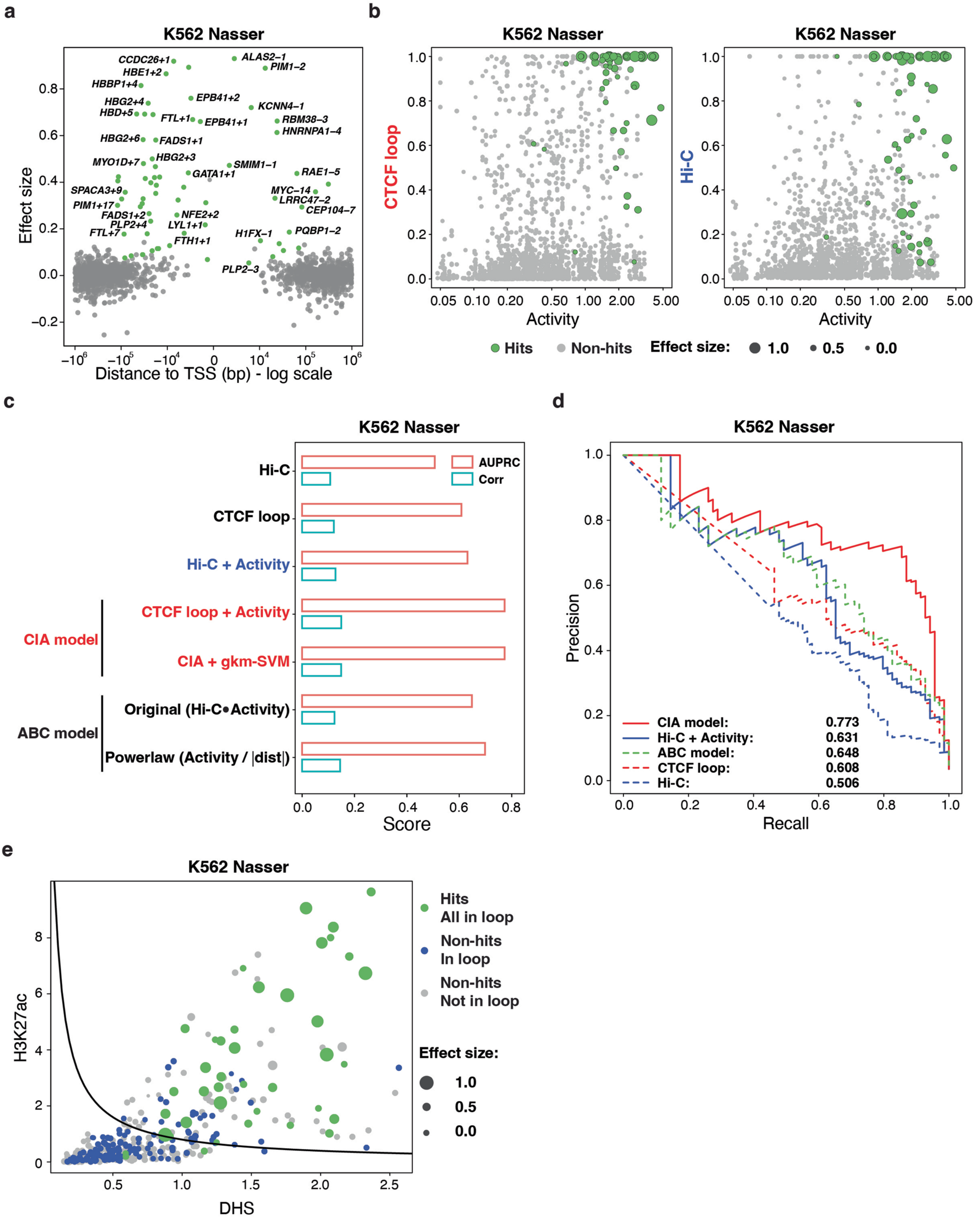
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Supporting data for enhancer prediction using CIA model. a, Hi-C-based and CTCF loop-based chromatin conformation analysis at the *GATA6*, *MIXL1* and *EOMES* loci. **b**, Precision-recall plot comparing the performance for prediction of enhancer hits from the screen using different Hi-C datasets. **c**, Precision-recall plot comparing the performance for prediction of enhancer hits from the screen using CIA model with additional H3K4me1 chromatin feature. **d**, Bar plot comparing the area under precision recall curve (AUPRC) and correlation scores between logistic regression of chromatin feature combination and Activity $= \sqrt{\text{ATAC} * \text{H3K27ac}}$ in CTCF loop-constrained Interaction Activity (CIA) model. **e**, A scatter plot showing the *P*(in loop) can classify hits (green) and non-hits (gray) more clearly than Hi-C-based enhancer-promoter contact frequency. The size of each dot represents the log2FC of each enhancer from the screen.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | The CIA model predicts active enhancers in different scenarios (K562 Reilly). **a**, gRNA enrichment analysis identified 36 hits from the HCR-FF (Hybridization Chain Reaction Fluorescent In-Situ Hybridization coupled with Flow Cytometry) screen in K562 cells from *Reilly* et al.[9] **b**, The scatter plot showing the P(in loop) can classify hits (green) and non-hits (gray) in K562 HCR-FF screen more clearly than Hi-C-based enhancer-promoter contact frequency. The size of each dot represents the log2FC of each enhancer from the screen. **c**, Bar plot comparing the AUPRC and correlation scores between Hi-C-based enhancer prediction with CIA model and ABC model using K562 HCR-FF screen results. **d**, Precision-recall plot comparing the performance for prediction of enhancer hits from the K562 HCR-FF screen using CTCF loop-based model and Hi-C-based model. **e**, The scatter plot showing the combinatory criteria of P(in loop), H3K27ac and ATAC can clearly separate the hits (green) and non-hits (blue and gray) from the K562 HCR-FF screen. P(in loop) > 0.5 is used to highlight enhancers and targeting promoters in the same CTCF loop (green and blue). The solid line represents the same threshold criterion of $\text{Activity} = \sqrt{\text{ATAC} * \text{H3K27ac}}$ in Fig. 7e. The size of each dot represents the log2FC of each enhancer from the K562 HCR-FF screen.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | The CIA model predicts active enhancers in different scenarios (K562 Nasser). a, 69 identified hits in the K562 cells from *Nasser et al* are plotted[40]. **b**, The scatter plot showing the *P*(in loop) can classify hits (green) and non-hits (gray) in K562 Nasser more clearly than Hi-C-based enhancer-promoter contact frequency. The size of each dot represents the effect size of each enhancer from *Nasser* et al. **c**, Bar plot comparing the AUPRC and correlation scores between Hi-C-based enhancer prediction with CIA model and ABC model using K562 Nasser results. **d**, Precision-recall plot comparing the performance for prediction of enhancer hits from the K562 Nasser using CTCF loop-based model and Hi-C-based model. **e**, The scatter plot showing the combinatory criteria of *P*(in loop), H3K27ac and ATAC can clearly separate the hits (green) and non-hits (blue and gray) from the K562 Nasser. *P*(in loop) > 0.5 is used to highlight enhancers and targeting promoters in the same CTCF loop (green and blue). The solid line represents the same threshold criterion of $\mathrm{Activity} = \sqrt{\mathrm{ATAC} * \mathrm{H3K27ac}}$ in Fig. 7e. The size of each dot represents the effect size of each enhancer from the K562 Nasser.

nature portfolio

Corresponding author(s): Beer, Huangfu

Last updated by author(s): Jun 18, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

FACS acquisition: BD LSRFortessa or BD LSRII; BD FACSDIVA.
Real-time qPCR: applied biosystem 7500 software v2.3; applied biosystem QuantStudio 6 flex software v1.3
Cell Counting: Vi-CELL CR 2.03 from Beckman Coulter
Single cell RNA-seq: 1OX chromium controller firmware v5.0
DNA/RNA concentration measurement: Thermo Nanodrop 2000 v1.6.198; Agilent BioAnalyzer
Sequencing platform: NovaSeq 6000; HiSeq 4000; HiSeq 2500
LC-MS/MS: Dionex RSLC Ultimate 300

Data analysis

Microsoft Words, Excels, Adobe illustrators were used for manuscript writing and figure productions.
Prism 9 and R package ggplot2 were used for making scatter plots, violin plots, bar graph and statistical tests.
FlowJo V10 was used for flow cytometry analysis.
Browser images were generated using the Integrated Genome Viewer (Version 2.3.92).
RNA-seq data were analyzed by STAR (Version 2.5.1b) and RSEM (Version 1.2.23)
ATAC-seq data were analyzed by bowtie2 (Version 2.2.5), picard (Version 2.23.3), macs2 (Version 2.2.7.1), and gkm-SVM (R-package version 0.82.0 with L=11, k=7, d=3 and truncated filter)
ChIP-seq data were analyzed by bowtie2 (Version 2.2.5) and macs2 (Version 2.2.7.1)
Hi-C data were analyzed by Hi-C Pro (Version 2.11.4), bowtie2 (Version 2.2.5), and JuicerTools (Version 1.22.01)
scRNA-seq data were analyzed by Cell Ranger (Version 3.1.0) and Seurat(Version 4.1.1)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The parental HUES8 hESC line was obtained from Harvard University under a material transfer agreement. Sequencing data are available at GEO, accession no. GSE213394 (new data from this study) and GSE114102 (published DE-72h H3K4me1 ChIP-seq data), GSE63525 (published K562 Hi-C data) and GSE72816, GSE177081, GSE177471 (published ChIA-PET data). The Hi-C data are also available 4D Nucleome Data Portal (https://data.4dnucleome.org/) under accession numbers 4DNESDO2ZYBM, 4DNESQMUTYXH, 4DNESFL8KDMT, 4DNESW8SIXN7, 4DNESW9GVC97, and 4DNESI1DNSGF. Mass spectrometry data are available in the PRIDE database under ProteomeXchange accession PXD043070.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No statistical method was used to predetermine the sample size. Sample size were determined based on previous studies on similar type of samples and experimental experience. At least three independent differentiation and cell culture experiments were performed for quantitative analysis (FACS and qPCR). Replication of sequencing-based methods are listed below.

Data exclusions

No data exclusions were performed unless the differentiation experiment itself failed.

Replication

All attempts of experiments at replication were successful unless the differentiation experiment itself failed.
The number of biological replicates were provided in the figure legends.
scRNA-seq: 1 biological replicate for each time point with ~5000 cells per time point.
Screen: 1 biological replicate.
Flow cytometry: 3-9 biological replicates.
RT-qPCR: 3 biological replicates.
RNA-seq: 3 biological replicates of each sample
ATAC-seq: 3 biological replicates of each samples except 2 biological replicates of ESC
Hi-C: 3 biological replicates of each sample
ChIP-MS: 2 biological replicates of each sample
ChIP-seq: 2 biological replicates of each sample except 1 biological replicate of DED1-GATA6 and DED2-GATA6.

Randomization

No randomization in molecular-biology-based experiments was involved. Majority of the molecular-biology-based results involved equipment-based quantitative measurements rather than subjective rating of data that could be affected by randomization.
Stochastic simulations used randomized noise and averaged over 4000 cells for Eq. (1) and 50 independent runs for Gillespie simulations.

Blinding

No blinding was involved. The experimenters designed, performed and analyzed all the molecular biological experiments. Majority of the molecular-biology-based results involved equipment-based quantitative measurements rather than subjective rating of data that could be affected by no blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Name Catalog # Company Dilution/Quantity Application. Also see Extended Data Table 7.
SOX17-488 562205 BD Biosciences 1:200 Flow cytometry
GATA6-PE 26452 Cell Signaling Technology 1:250 Flow cytometry
EOMES-660 50-112-8679 Fisher Scientific 1:200 Flow cytometry
CXCR4-APC FAB170A-100 R&D System 1:50 Flow cytometry
HA-Tag 3724 Cell Signaling Technology 1:500 Flow cytometry
Donkey anti-Rabbit IgG-647 A-31573 Invitrogen 1:500 Flow cytometry
EOMES 66325 Cell Signaling Technology 10ug/30-40M cells ChIP-seq
GATA6 5851 Cell Signaling Technology 10ug/30-40M cells ChIP-seq
SOX17 81778 Cell Signaling Technology 10ug/30-40M cells ChIP-seq
CTCF 07-729 EMD Millipore 3ul/30-40M cells ChIP-seq
H3K27ac 39133 active motif 5ug/30-40M cells ChIP-seq
IgG 2729 Cell Signaling Technology 10ug/30-40M cells ChIP-MS
EOMES 66325 Cell Signaling Technology 10ug/30-40M cells ChIP-MS
GATA6 5851 Cell Signaling Technology 10ug/30-40M cells ChIP-MS
SOX17 81778 Cell Signaling Technology 10ug/30-40M cells ChIP-MS

Validation

SOX17-488 562205 BD Biosciences: https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/alexa-fluor-488-mouse-anti-human-sox17.562205
GATA6-PE 26452 Cell Signaling Technology: https://www.cellsignal.com/products/antibody-conjugates/gata-6-d61e4-xp-rabbit-mab-pe-conjugate/26452
EOMES-660 50-112-8679 Fisher Scientific: https://www.fishersci.com/shop/products/eomes-monoclonal-antibody-wd1928-efluor-660-ebioscience-invitrogen/501128679
CXCR4-APC FAB170A-100 R&D System: https://www.rndsystems.com/products/human-cxcr4-apc-conjugated-antibody-12g5_fab170a
HA-Tag 3724 Cell Signaling Technology: https://www.cellsignal.com/products/primary-antibodies/ha-tag-c29f4-rabbit-mab/3724?gclid=Cj0KCQjw9deiBhC1ARIsAHLjR2ConCjYZyiEl4tS25qETDA_smSYQGNo8SPyzw1axIrQ80yfVrAMbj4aAuNeEALw_wcB&gclsrc=aw.ds
EOMES 66325 Cell Signaling Technology: https://www.cellsignal.com/products/primary-antibodies/eomes-e4z4x-rabbit-mab-chip-formulated/66325
GATA6 5851 Cell Signaling Technology: https://www.cellsignal.com/products/antibody-conjugates/gata-6-d61e4-xp-rabbit-mab-pe-conjugate/26452?N=102260
+1475042463&fromPage=plp&gclid=Cj0KCQjw9deiBhC1ARIsAHLjR2BfaSb4si8O4gXWFkJolJoOKVi6T5S8tRO5WM9u3Qoq-PuqWMyW2owaAp13EALw_wcB&gclsrc=aw.ds
SOX17 81778 Cell Signaling Technology: https://www.cellsignal.com/products/primary-antibodies/sox17-d1t8m-rabbit-mab/81778
CTCF 07-729 EMD Millipore: https://www.emdmillipore.com/US/en/product/Anti-CTCF-Antibody,MM_NF-07-729
H3K27ac 39133 active motif: https://www.activemotif.com/catalog/details/39133/histone-h3-acetyl-lys27-antibody-pab
IgG 2729 Cell Signaling Technology: https://www.cellsignal.com/products/primary-antibodies/normal-rabbit-igg/2729

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

This study used HUES8 (NIHhESC-09-0021) hESC line.
293T cell line is a gift originally from Robert Weinberg's lab at MIT. We are not able to track the commercial source of the 293T cells since they have been passed around through many labs. Given that we only used the 293T cells for virus package, we don't believe it is important to figure out the source of these cells.

Authentication

STR was performed on the parental HUES8 hESC lines.

Mycoplasma contamination

Cells were routinely confirmed to be mycoplasma-free by the MSKCC Antibody and Bioresource Core Facility.

Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/geo/query/acc.cgi?acc=GSE213394
Files in database submission	GSE213394_DED1-CTCF_hg38.bigwig GSE213394_DED1-EOMES_hg38.bigwig GSE213394_DED1-GATA6_hg38.bigwig GSE213394_DED1-H3K27ac_hg38.bigwig GSE213394_DED2-CTCF_hg38.bigwig GSE213394_DED2-EOMES_hg38.bigwig GSE213394_DED2-GATA6_hg38.bigwig GSE213394_DED2-H3K27ac_hg38.bigwig GSE213394_DED2-SOX17_hg38.bigwig GSE213394_DED3-CTCF_hg38.bigwig GSE213394_DED3-H3K27ac_hg38.bigwig GSE213394_ESC-CTCF_hg38.bigwig GSE213394_ESC-H3K27ac_hg38.bigwig Raw data are available in SRA
Genome browser session (e.g. UCSC)	IGV 2.3.92 HG38

Methodology

Replicates	2 replicates of each sample except 1 replicate of DED1-GATA6 and DED2-GATA6.
Sequencing depth	20-30M reads per sample
Antibodies	EOMES 66325 Cell Signaling Technology GATA6 5851 Cell Signaling Technology SOX17 81778 Cell Signaling Technology CTCF 07-729 EMD Millipore H3K27ac 39133 active motif
Peak calling parameters	Paired-end reads were mapped to hg38 with bowtie2 using default parameters, and peaks were called using macs2 using default parameters.
Data quality	Methods to ensure data quality are described in the Methods section (ChIP-seq/ChIA-PET analysis). Paired-end reads were mapped to hg38 with bowtie2 using default parameters, and peaks were called using macs2 using default parameters.
Software	ChIP-seq data were analyzed by bowtie2 and macs2

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells were dissociated using TrypLE Select and resuspended in FACS buffer (5% FBS, 5 mM EDTA in PBS). For surface marker staining, cells were stained with surface antibody (CXCR4-APC) with DAPI for 15 minutes at room temperature (RT). After staining, cells were washed, suspended in FACS buffer and analyzed. For intracellular marker staining, cells were stained with LIVE/DEAD violet dye (Invitrogen; L34955, 1:1,000) for 15 minutes at

	room temperature. After washing with FACS buffer, cells were fixed and permeabilized in 1X fixation/permeabilization buffer (eBioscience, 00-5523-00) for 30 minutes at RT. After fixation and permeabilization, cells were stained with intracellular antibody (SOX17-Alexa-488, GATA6-PE) in permeabilization buffer (eBioscience, 00-5523-00) for 30 minutes at RT. After staining, cells were washed, suspended in FACS buffer and analyzed.
Instrument	BD LSRFortessa or BD LSRII.
Software	FlowJO V10
Cell population abundance	We collected around 5000-10000 live cells for the final analysis.
Gating strategy	FSC/SSC gating was used to identify single cells. DAPI or live/dead staining were used for identify live cells. We used negative control cell lines (undifferentiated cells) for gating control.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.