Title

Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction

Permalink

https://escholarship.org/uc/item/2135j1pk

Journal

Cell Stem Cell, 25(5)

ISSN

1934-5909

Authors

Inoue, Fumitaka
Kreimer, Anat
Ashuach, Tal
et al.

Publication Date

2019-11-01

DOI

10.1016/j.stem.2019.09.010

Peer reviewed

Published in final edited form as:
Cell Stem Cell. 2019 November 07; 25(5): 713–727.e10. doi:10.1016/j.stem.2019.09.010.

Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction

Fumitaka Inoue[1,2,5], **Anat Kreimer**[1,2,3,5], **Tal Ashuach**[3], **Nadav Ahituv**[1,2,*], **Nir Yosef**[3,4,6,*]

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA 94158, USA

[3]Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA

[4]Chan-Zuckerberg Biohub, San Francisco, CA 94158, USA

[5]These authors contributed equally

[6]Lead Contact

SUMMARY

Epigenomic regulation and lineage-specific gene expression act in concert to drive cellular differentiation, but the temporal interplay between these processes is largely unknown. Using neural induction from human pluripotent stem cells (hPSCs) as a paradigm, we interrogated these dynamics by performing RNA sequencing (RNA-seq), chromatin immunoprecipitation sequencing (ChIP-seq), and assay for transposase accessible chromatin using sequencing (ATAC-seq) at seven time points during early neural differentiation. We found that changes in DNA accessibility precede H3K27ac, which is followed by gene expression changes. Using massively parallel reporter assays (MPRAs) to test the activity of 2,464 candidate regulatory sequences at all seven time points, we show that many of these sequences have temporal activity patterns that correlate with their respective cell-endogenous gene expression and chromatin changes. A prioritization method incorporating all genomic and MPRA data further identified key transcription factors involved in driving neural fate. These results provide a comprehensive resource of genes and regulatory elements that orchestrate neural induction and illuminate temporal frameworks during differentiation.

[*]Correspondence: nadav.ahituv@ucsf.edu (N.A.), niryosef@berkeley.edu (N.Y.).

AUTHOR CONTRIBUTIONS

F.I., A.K., N.A., and N.Y. conceived and designed the study. F.I. performed the RNA-seq, ChIP-seq, ATAC-seq, lentiMPRA, luciferase assays, immunocytochemistry, CRISPRa/i, and overexpression experiments. A.K. performed the computational analyses. T.A. contributed to the analysis of the MPRA data. F.I., A.K., N.A., and N.Y. interpreted the data. N.A. and N.Y. provided resources, and F.I., A.K., N.A., and N.Y. wrote the manuscript.

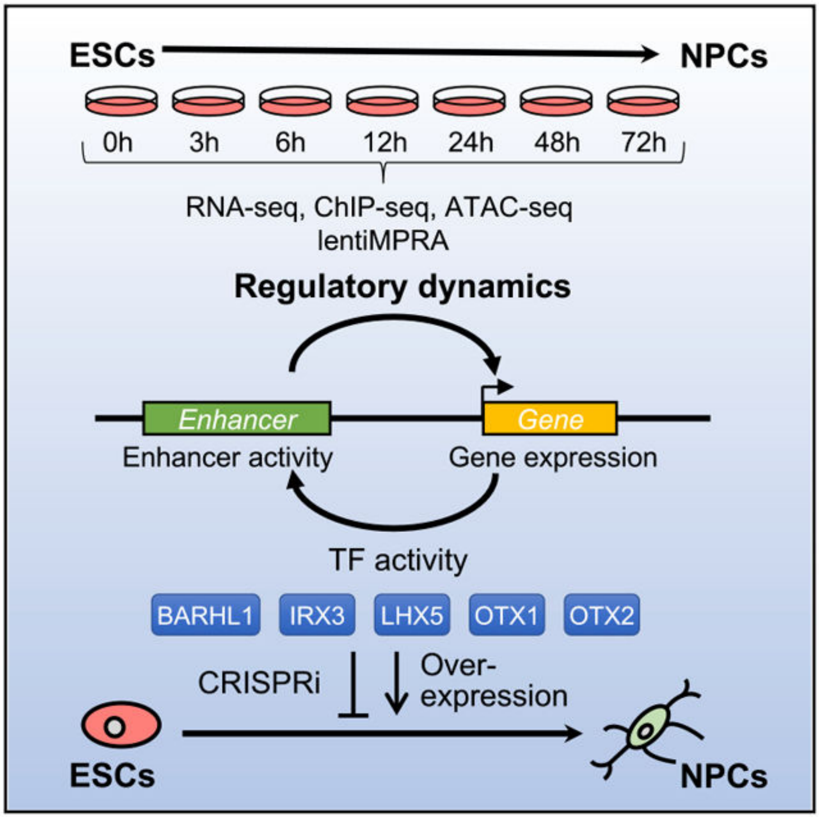
DECLARATION OF INTERESTS

N.A. is an equity holder and is part of the scientific advisory board for Encoded Therapeutics, a gene regulation therapeutics company and co-founder of Enhancer Therapeutics Inc.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.Org/10.1016/j.stem.2019.09.010.

Graphical Abstract



In Brief

To reveal regulatory dynamics during neural induction, we performed RNA-seq, ChIP-seq, ATAC-seq, and lentiMPRA at seven time points during early neural differentiation. We incorporated all information and identified TFs that play important roles in this process. We demonstrated overexpression or CRISPRi of five TFs affected ESC-NPC differentiation.

INTRODUCTION

Global changes in gene expression are an essential part of cellular differentiation (Yosef and Regev, 2016). To date, many genome-scale maps of epigenetic properties in progenitor and differentiated cells have been used in comparative studies, demonstrating the importance of modifications of the epigenome to the pertaining changes in gene expression and shedding light on the mechanisms involved in this process (Andersson et al., 2014; Arner et al., 2015; Bernstein et al.,2012). For instance, in human embryonic stem cells (hESCs), the regulatory regions marked by histone modifications and binding of key regulators associated with gene expression were globally reorganized in accordance with multilineage differentiation (Dixon et al., 2015; Gifford et al., 2013; Tsankov et al., 2015; Xie et al., 2013). However, the majority of these studies provide descriptive genome-wide maps without large-scale functional analyses of candidate sequences.

Furthermore, although a few studies used functional validation following large-scale genomic studies (Kheradpour et al., 2013; Kwasnieski et al., 2014; Ulirsch et al., 2016; Wang et al., 2018), these studies did not focus on differentiation processes.

The differentiation of hESCs into neural cells provides an exceptional model to study this. During early neural induction, the cells exhibit marked changes in gene expression as pluripotency-associated genes are rapidly downregulated and neural-associated genes are induced. These changes are then maintained for a duration of several weeks until the establishment of a neural progenitor cells (NPCs) population (Ziller et al., 2015). Several large-scale mapping efforts have characterized in a genome-wide manner the transcriptional and epigenetic landscape of hESC-derived NPCs or neural tissues and have annotated numerous genes and potential regulatory elements that could be important in neural differentiation (Andersson et al., 2014; Bernstein et al., 2012; Dixon et al., 2015; Fort et al., 2014; Gifford et al., 2013; Tsankov et al., 2015; Xie et al., 2013). However, although these studies have identified putative regulatory elements, they have not comprehensively analyzed them for their function. Furthermore, none of these genomic studies focused on the early stages of neural differentiation when neural induction takes place. Thus, the intrinsic mechanism that governs neural induction remains largely unknown.

The differentiation of hESCs to neuronal cells also provides an important model system for studying the etiology of neurodevelopmental diseases. Mutations in genes and regulatory elements involved in neural induction and development have been associated with numerous human diseases. For example, dysfunction of cortical GABA neurons in schizophrenia begins during prenatal development (Volk and Lewis, 2013). Similarly, autism spectrum disorders (ASDs) are associated with *de novo* mutations in developmental genes (Samocha et al., 2014) and alterations in canonical Wnt signaling in developing embryos (Kalkman, 2012). In addition, the majority of disease-risk loci discovered through genome-wide association studies (GWASs) in general and specifically for neuropsychiatric and neurodevelopmental disorders reside in noncoding regions (Hindorff et al., 2009; Maurano et al., 2012; Sanders et al., 2017; https://paperpile.com/c/Q8FO7P/iuTR+IK6Y+IWCK), suggesting an important role for enhancers in disease susceptibility.

Here, we set out to generate a genomic map of the transcriptional (RNA sequencing [RNA-seq]) and epigenetic landscape (H3K27ac/me3 chromatin immunoprecipitation [ChIP]-seq and ATAC-seq) of neural induction and then coupled these observations with comprehensive functional assays (massively parallel reporter assays [MPRAs]). We integrated all of the resulting data modalities (genomics maps and MPRAs) to computationally infer the activity of transcription factors (TFs) over time and characterize candidate TFs that could be important drivers of neural induction. Our work provides a comprehensive resource of genes and regulatory elements and a blueprint for the interplay between them during neural differentiation.

RESULTS

The Neural-Induction-Associated Transcriptome

We performed deep RNA-seq (average of 200 million reads per replicate) on undifferentiated H1-ESCs (0 h) and six different time points of early neural differentiation (3, 6, 12, 24, 48, and 72 h) following dual-Smad inhibition (Chambers et al., 2009). Principal-component analysis (PCA) of the RNA-seq data showed consistency between the three replicates and a clear separation between the earlier and later time points (Figure S1A). As expected, we observed neural marker genes, such as *SOX1*, to be upregulated after 12 h (Figure 1A), with limited expression changes in mesendoderm (*EOMES*), mesoderm (*T* and *TBX6*), endoderm (*SOX17* and *GATA4*), and neural crest markers (*FOXD3* and *SNAI1/2*). Pluripotent markers (*NANOG* and *POU5F1*) and direct targets of transforming growth factor β (TGF-β) and bone morphogenetic protein (BMP) signaling (*SMAD7*, *ID1*, and *LEFTY2*) were downregulated, and immediate early genes (*ATF3*, *FOS*, *FOSB*, and *EGR1/2/3*) were transiently upregulated at 3 h, corresponding to the cell's stress response against differentiation stimuli. For a more general analysis, we used a conservative approach to identify genes whose expression differed significantly over time, using a consensus over two methods—ImpulseDE (Sander et al., 2017) and DESeq2 (Love et al., 2014). Altogether, we detected 2,172 genes as differentially expressed over time (henceforth referred to as temporal genes), with 85% of them being induced at some point in time (Figure 1B; the remaining genes show monotonic decrease of expression). Gene set enrichment analysis (Subramanian et al., 2005) of the resulting clusters of temporal profiles found that genes that are more strongly expressed at the early time points (0–12 h; false discovery rate [FDR] < 0.05; hypergeometric test) are enriched for regulation of multicellular organismal development, indicating an association with pluripotency. Conversely, genes induced at later time points (>24 h; FDR < 0.05) are enriched for neurogenesis processes, consistent with the progression of the cells toward a neural lineage fate (Table S1). Combined, our transcriptomic analyses validated the ability of the dual-Smad inhibition protocol to obtain the expected neural trajectory and provide a catalog of genes involved in neural induction.

The Neural-Induction-Associated Regulome

To identify candidate enhancers involved in the differentiation process that could be driving neural induction, we performed ATAC-seq as well as ChIP-seq for the active histone mark H3K27ac and the silencing mark H3K27me3 at all seven time points. We then identified regions that are enriched (i.e., peak regions) in each of these assays (Feng et al., 2011; FDR < 0.05) by analyzing each time point separately and then taking the merged set of peaks over all time points. To establish peak calling quality, we compared our 0-h time point H3K27ac and ATAC-seq peaks to H1-ESC H3K27ac peaks and DNase I hypersensitive sites (DHSs) from ENCODE (Bernstein et al., 2012) and observed a substantial overlap of 80% and 90%, respectively. Overall, we identified 40,486 ATAC-seq peaks, 40,170 H3K27ac peaks, and 4,446 H3K27me3 peaks that are present in at least one time point. To exclude potentially inactive regions from further analysis, we removed H3K27ac peaks that overlap an H3K27me3 peak at all the time points in which that peak was detected. This resulted in a filtered set of 40,042 H3K27ac peaks, indicating that the two chromatin marks have little overlap in our data. Conversely, we observed a substantial overlap between the H3K27ac peaks and the ATAC-seq peaks, with an overall 60% (23,294) of the H3K27ac peaks overlapping an accessible region at the same point in time. These results correspond to a similar overlap of 61% between H3K27ac peaks and DHSs in H1-ESC from ENCODE (Bernstein et al., 2012). Using a strict procedure, similar to the gene expression analysis (Love et al., 2014; Sander et al., 2017), we found 2,435 ATAC-seq and 2,024 H3K27ac peaks that were differentially enriched between time points, henceforth referred to as temporal H3K27ac or ATAC-seq peaks (Table S1). Similar analysis of H3K27me3 peaks showed weaker temporal signal (STAR Methods) with a smaller number of 248 temporal peaks (Figure S1B), possibly due to histone methylation being less dynamic than acetylation (Donnard et al., 2018; Garber et al., 2012; Luizon et al., 2016; Smith et al., 2014).

We next set out to study the association between the temporal changes observed at the epigenome level and those observed at the gene expression level. We clustered the two sets of temporal regions (in terms of accessibility and H3K27ac) into several prototypical patterns (Figures 1C and 1D), as we have done for the temporal genes (Figure 1B). Functional enrichment analysis (using GREAT; McLean et al., 2010; with FDR < 0.05) on the accessibility and H3K27ac clusters was overall consistent with the results observed with the gene expression clusters, with an enrichment for pluripotent factors and nervous system development processes in early- and late-response regions, respectively (Table S1). Interestingly, the observed temporal changes of accessibility, histone acetylation, and proximal gene expression were highly correlated to each other (Figures 1E and S1C; exact overlaps are displayed in Table S1). Furthermore, for a significant fraction of the genes induced at the late stages (RNA-seq clusters 4–6), chromatin accessibility was found to be acquired first (ATAC-seq clusters 4 and 5) or simultaneously (ATAC-seq cluster 6) with H3K27ac modification followed by an increase in mRNA expression of the nearest gene (Figure 1E), and at early stages (RNA-seq clusters 1–3), this was a less obvious trend. For example, the DNA accessibility cluster 4 that peaks at 24 h showed the strongest overlap with H3K27ac clusters 5 and 6 that peak at 48–72 h, and this cluster significantly overlaps (in terms of genes; $p < 0.0014$; hypergeometric test) with gene expression cluster 6, which peaks at 72 h (Figure 1E). Specifically, examination of potential enhancers within these clusters that are located near neural marker genes, *MAP2* (Herzog and Weber, 1978) and *ROR2* (Endo et al., 2012), found them to be enriched for ATAC-seq signal at 12–24 h, H3K27ac signal at 48–72 h, and their expression to peak at 72 h (Figure S2A and S2B). Combined, these results suggest that regions that are associated with changes to chromatin structure during neural induction are statistically related to changes in gene expression.

Neurological-Disorder-Associated Variants Are Enriched in Temporal H3K27ac Peaks

As genes and regulatory elements involved in neural development may be associated with neurological disorders, we tested whether our neural induction regulome overlaps with disease-associated variants. We first tested whether the temporal loci (in terms of accessibility or H3K27ac) are enriched for GWAS variants associated with neurological disorders. To this end, we used the complete set of peaks (temporal and non-temporal) as background and added variants associated with height as negative controls. We observed a significant enrichment for H3K27ac (but not accessibility) temporal peaks with neurological disorders (Table S1; STAR Methods; $p < 0.05$; Fisher's exact test), but not with height variants. Specifically, we observe significant enrichment when examining variants associated with a combined set of neuropsychological disorders (schizophrenia, attention-deficit hyperactivity disorder [ADHD], ASD, bipolar disorder, and major depressive disorder) as well as enrichment when examining for individual disorders (i.e., bipolar and psychosis disorders). As the smaller size of ATAC-seq peaks might account for the lack of enrichment in ATAC-seq temporal peaks, we expanded the ATAC-seq peaks to the average size of H3K27ac peaks but observed similar results.

Expression quantitative trait loci (eQTLs) mark variants that can be associated with modulating the regulation of nearby genes. We tested for overlap between eQTLs found in various tissues (GTEx Consortium, 2015) and our temporal ATAC-seq or H3K27ac peaks We found the temporal H3K27ac peaks to be significantly enriched for eQTL variants (Leslie et al., 2014) in general and specifically for those from brain tissues (GTEx Consortium, 2015; Table S1; STAR Methods; $p < 0.05$, Fisher's exact test). Similarly to GWAS variants, we did not observe an enrichment of eQTLs in temporal ATAC-seq peaks, even upon their expansion. When restricting H3K27ac peaks to not overlap with H3K27ac ChIP-seq peaks obtained from different cell types (GM12878, K562, and HepG2) via the ENCODE project (Bernstein et al., 2012), we observe similar results, indicating that our signal is not biased by constitutive peaks. When restricting variant enrichment analysis to H3K27ac temporal peaks and assessing the enrichment in each temporal cluster (Figure 1C), we observe that the late response cluster 6 is significantly enriched in nervous system disease and specifically with ASD-associated variants (Fisher's exact test; $p < 0.005$). Combined, these results suggest that our temporal H3K27ac regions could be functional enhancers that harbor neurological disease risk variants. They also suggest that temporal changes to the chromatin early in the differentiation process can facilitate the identification of potentially functional regions more so than data from a single time point.

lentiMPRA Identifies Regulatory Regions that Are Active during Neural Induction

In order to test whether our candidate regulatory sequences can in fact induce temporal transcriptional response, we carried out lentiMPRA at all seven time points. Overall, we investigated 2,464 candidate sequences, covering both promoters (n = 386; 15.7%) and putative enhancers (n = 2,078; 84.3%), termed henceforth as candidate regulatory sequences (CRSs). As the number of potential CRSs is large, we developed a prioritization scheme to select the set of assayed regions (Figure 2A; Table S2; STAR Methods) using the following criteria: (1) manually curated list of enhancers that are next to genes involved in neural differentiation (n = 102; Table S2); (2) sequences that overlap a temporal H3K27ac ChIP-seq peak that also overlap an ATAC-seq peak (not necessarily temporal) and that their closest gene shows increased expression due to neural induction (n = 1,596); (3) sequences that overlap non-temporal H3K27ac peaks and temporal ATAC-seq peaks and their closest gene shows increased expression due to neural induction (n = 441); (4) among the regions not included in the first three groups, we select sequences that showed the strongest difference in signal of either H3K27ac ChIP-seq, ATAC-seq, or mRNA of the closest genes (n = 132; comparing either 0 versus 3 h or 0 versus 72 h); and (5) positive control sequences (n = 193) that included previously reported sequences that were validated forebrain enhancers in the VISTA Enhancer Browser (Visel et al., 2007; n = 105), sequences near pluripotent factors (n = 42), and commonly used positive controls from the ENCODE project (n = 46; Table S2). For negative controls, we randomly selected 200 of our candidate sequences and shuffled their nucleotides obtaining scrambled sequences. Overall, we chose 2,664 sequences using this process. As our assayed sequences were 171 bp long, due to oligonucleotide synthesis limitations, we chose the 171-bp window within a peak of interest by maximizing the number of motifs in it (Grant et al., 2011; Kheradpour and Kellis, 2014).

The selected oligonucleotides were generated and cloned upstream of a minimal promoter (mP) and EGFP reporter gene into a lentivirus-based enhancer assay vector (Figure 2B) as previously described (Inoue et al., 2017). Although the sequences are assayed in the context of enhancer activity in this assay, previous work has shown that it also provides a good indication for promoter activity (Kreimer et al., 2017, 2019; Melnikov et al.,2014). Each individual CRS was designed to be associated with 90 different 15-bp barcodes, thus allowing robust evaluation of the pertaining expression output and to correct for site of integration biases (Ashuach et al., 2019; Inoue et al., 2017). In total, 239,760 sequences (2,664 CRSs and negative controls × 90 barcodes) were included in the library (Figure 2B). The cloned library was sequenced in order to evaluate the quality of the designed oligonucleotides and the representation of individual barcodes (STAR Methods; Figure S3A–S3D).

hESCs were infected with the library with an average of 5–8 integrations per cell (Figure S3E), cultured for 3 days to clean out for unintegrated lentivirus, and then subsequently induced into a neural lineage via dual-Smad inhibition. lentiMPRA was performed at all seven time points of neural differentiation with three replicates (two biological replicates, one of which was split into two technical replicates; Table S3). Due to the short time spans between some conditions, we collected nuclear RNA in all time points to detect their immediate expression. We observed an average of 70 barcodes out of 90 per CRS in each replicate (Table S3). By aggregating these barcodes (STAR Methods), we were able to get highly reproducible results across replicates (Figure S3F) with similar magnitude to a previously characterized lentiMPRA in another cell type (Inoue et al., 2017). We then combined replicates to produce a normalized RNA/DNA ratio for each CRS (henceforth referred to as MPRA signal). Examination of the signal observed for regions nominated by the different experimental design criteria found that temporal H3K27ac signal (criterion 2) provides a highly effective predictor of functional enhancer activity, and as expected, the negative controls showed the lowest activity (Figures 3A and S3G).

lentiMPRA Identifies Temporal CRS

We next set out to examine whether the enhancer activity observed in our assay changes over time and then characterize these changes with respect to the cell-endogenous temporal processes depicted in Figure 1. As a starting point, we considered each time point separately and applied MPRAnalyze (Ashuach et al., 2019), a method and Bioconductor package for statistical analysis of MPRA data developed in our group, to identify active enhancers, namely enhancers whose activity significantly deviates from that of the negative controls (median-based Z score; FDR < 0.05; STAR Methods). Of note, the dynamic ranges we observed were comparable to a previous library generated in a similar manner (Inoue et al., 2017). From the 2,464 CRSs that we tested via lentiMPRA, 1,681 (68%) were called significant in at least one of the time points, and on average, 1,141 (46.3%) sequences were active per each individual time point.

Although we saw similar levels of activity at each time point, the respective sets of active CRSs may differ greatly over time. Reassuringly, we observed substantial overlaps between the sets of active CRSs at nearby time points (Figure 3B), along with a marked decrease in overlap as the distance between the respective time points increases (Figure 3C). This indicated that regulatory programs carried out by enhancers are far from fixed but instead change over the course of neural induction. As an example, we observed that a known enhancer of *NANOG* as well as its promoter (Rodda et al., 2005; Wu et al., 2006) both have activity only at the early time points (Figure 3D), as expected. We also found novel enhancers near *SOX1* that showed increased temporal activity at 24–48 h while being less active at 72 h and further away (140 kb) an enhancer that has strong enhancer activity at 72 h, suggesting a complex temporal regulation pattern for this gene (Figure 3E). To validate our temporal observations with lentiMPRA, we individually tested five ESC enhancers, four immediate response enhancers (12–24 h), and four NPC enhancers (48–72 h) using luciferase assays. We observed the expected temporal activities for these sequences, which were consistent with our MPRA results (Figures 4A and 4B). As an additional validation, we used CRISPR activation (CRIPSRa) (Gilbert et al., 2013) to target three CRSs detected in our study at the *SOX1*, *IRX3*, and *OTX2* loci in hESCs. We found that all three CRSs upregulated the expression of their predicted target gene (*SOX1*, *IRX3*, and *OTX2*) following CRISPRa (Figures 4C-4F), further suggesting that the enhancers we identified are functional and can affect gene expression.

We next carried out a more global analysis that aims to identify enhancers whose MPRA signal significantly changed over time (Ashuach et al., 2019). This alternative approach pools together information from all time points, rather than considering each time point individually, and therefore has the potential to identify effects that may otherwise be missed. In this analysis, the temporal activity of each CRS was compared with a null temporal behavior displayed by the set of negative controls. Regions with significantly different temporal activity were called temporally active using a likelihood ratio test (FDR < 0.05; Ashuach et al., 2019). We found that 1,547 sequences out of the 2,464 we tested (63%) showed temporal regulatory activity (henceforth referred to as temporal CRS). Out of these temporal CRSs, 1,261 (82%) were also detected by the per-time-point analysis. In the following analyses, we focused on the complete set of temporal CRSs. Importantly, we observed consistent results when limiting our analyses to the smaller and more stringent set of 1,261 regions.

Comparative analysis of temporal versus non-temporal CRSs for differences in TF binding motifs (Kheradpour and Kellis, 2014) found an enrichment for pluripotency-related regulators, such as POU5F1, SOX2, SALL4, NANOG, and SMAD1 (largely targeting CRSs with a marked decrease in activity over time), as well as the NPC-associated TF, SOX1 (largely targeting CRSs with a marked increase in activity over time; Figure 3F). Of note, previous reports have shown that SOX2-POU5F1 and SOX2-POU3F2 regulate ESC and NPC genes, respectively (Lodato et al., 2013), suggesting that SOX and POU motifs not only function in a pluripotent state but also in a neural state. We also observed an enrichment for immediate early response factors (e.g., AP-1, ATF3, and EGR3), corresponding to the cell's response to differentiation stimuli. Finally, we found that regulators of chromatin conformation, including regulators of histone acetylation (EP300 and HDAC) and chromatin boundary and looping (CTCF), were also enriched in temporal CRSs, indicating that changes in activity over time may also be mediated by a more direct regulation of the epigenome and not only by state-specific TFs.

Enhancer Activity Is Consistent with the Endogenous Temporal Profiles

To further evaluate the significance of the temporal CRSs, we turned to estimate the extent to which changes to their MPRA signal correlates with the respective changes to gene expression and the endogenous chromatin. To this end, we clustered the temporal CRSs into four patterns of activity: (1) early (mainly active at 0–6 h); (2) mid-early (primarily active at 12–24 h); (3) mid-late (mainly active at 24–48 h); and (4) late response (primarily active at 48–72 h; Figure 5A). To facilitate direct comparison to temporal profiles in the endogenous genome, we quantified for each temporal CRS the expression of its closest gene and the epigenetic signal (accessibility, H3K27ac) in the respective endogenous position. We then stratified the resulting profiles into clusters, in a similar way to that of the MPRA, and tested the overlap between the resulting endogenous clusters (Figures 5B–5D; Table S4) and the MPRA-based clusters (Figure 5A). Starting with the expression of the closest gene, we find significant levels of overlap between the respective clusters (lentiMPRA and RNA-seq; Bonferroni-corrected hypergeometric $p < 0.05$; Figure 5E). The significant overlap is observed primarily in time-matched clusters, indicating that an overall trend in the data is that the temporal CRSs are capable of inducing reporter gene expression that is similar to the (postulated) endogenous target gene. Indeed, in an alternative analysis, we defined the maximal segment of each enhancer as the two adjacent time points in which it reaches its maximal expression. Comparing the MPRA and the endogenous mRNA, we found that, in 48% (752/1,547) of the temporal CRSs, the respective maximal segments overlap. Gene ontology enrichment analyses for the genes associated by proximity with the regions in the different clusters also found gene categories fitting with the temporal expression (Figure S4; Table S4). For instance, the early cluster is enriched for recruitment of histone acetyltransferases (HATs) and expression of pluripotent genes (e.g., *KLF4*), indicative of stem cell differentiation processes that take place during these early time points (Tsankov et al., 2015). The mid-early and mid-late clusters are enriched with early chromatin response and genes that are involved in developmental processes. The late cluster is enriched for open chromatin, HAT recruitment, and expression of neural genes (e.g., *OTX1*). We observe similar results for the more restricted set of regions that are temporal and active in at least one time point (Figure S5).

We next set out to compare the temporal patterns observed with MPRA to those observed at the chromatin level. As expected, we find that the temporal CRSs rarely overlap with H3K27me3 peaks (Figure S1D). Conversely, we find significant overlaps between the MPRA clusters and their time-matched H3K27ac clusters (Figure 5E). Using the concept of maximal segments, we find that 50% of the temporal CRSs reach their maximum level around the same time as the respective H3K27ac peak (i.e., 604 out of 1,208 temporal CRSs that intersect with an H3K27ac peak; Figure S4). With ATAC-seq, we observe a less coordinated pattern of overlap between clusters yet a similar level of overlap in maximal segments, which allows for some discrepancy in peak times (48% of CRSs). This is possibly due to chromatin accessibility preceding enhancer activation and open chromatin not necessarily being synonymous with active regions (for example, *SOX1* downstream enhancers were found to be accessible at an undifferentiated state but only active at a later stage).

Overall, we observe a substantial level of agreement between MPRA and the endogenous transcriptional changes, with 67% of the temporal CRSs (1,038/1,547) consistent with the temporal patterns of at least one of the endogenous signals (H3K27ac, accessibility, or mRNA expression). These results suggest that the signal captured by lentiMPRA could be relevant for neural induction and that the activity of the endogenous counterparts of the temporal CRSs may be functional during this process. For example, a MPRA temporal enhancer on chr2:58023768-58023938 (hg19) was part of the late MPRA cluster and associated with the late H3K27ac cluster and mid-late ATAC-seq cluster. The H3K27ac peak overlapping this region harbors ASD-associated SNPs that are in linkage disequilibrium ($r2 \geq 0.8$) with the lead SNP rs2176546 (rs6545663, rs6545664, and rs6545665; Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, 2017), and its two closest genes, *VRK2* and *FANCL*, are both associated with ASD. *VRK2* and *FANCL* belong to the mid-early and late gene expression clusters, respectively (Figure S2C). Although we observed an overall strong correlation between temporal CRSs and gene expression, it is important to note that this overlap was not obtained for all sequences (see STAR Methods for discussion of these cases).

TF Binding Site Analyses Identify Important Neural Induction Genes

As the RNA product of MPRA is non-endogenous, it provides an effective way for directly estimating the effects of TFs on transcription. We utilized this property to pinpoint which TFs could be driving neural induction at the different time points. To this end, we used experimental data from the public domain along with DNA binding motifs to determine the potential binding landscape of a large cohort of TFs across our tested regions. We recorded, for each temporal CRS, (1) its predicted binding sites using FIMO (Grant et al., 2011) with two sets of TF motifs (Kheradpour and Kellis, 2014; Weirauch et al., 2014) and (2) its overlap with TF ChIP-seq peaks in hESCs (Gifford et al., 2013) or in hESC-derived neuroectoderm (Tsankov et al., 2015). The result of this analysis is a binary binding matrix of TFs by CRSs with entries indicating either potential binding using $FDR < 10^{-4}$ for TF motifs or overlap with TF ChIP-seq peaks.

We next employed a strict enrichment analysis based on comparing the number of putative binding sites in regions within each temporal MPRA cluster versus the set of all regions in our MPRA design ($FDR < 0.05$; hypergeometric test; Table S5). This analysis was designed to nominate candidate TFs whose activity is specific to certain phases of the differentiation process. Accordingly, we found that motifs of pluripotent factors (e.g., NANOG, POU5F1, and SOX2; Boyer et al., 2005) were enriched in the early cluster. Furthermore, immediate early response factors (ATF, JUN, and FOS) were enriched in mid-early enhancers (Table

S5). These observations suggest that early- and mid-early clusters may respond to TFs that function in pluripotency maintenance and the cell's acute response, such as apoptosis (Herschman, 1991), respectively. We also found that both mid-late and late clusters were enriched for cell fate commitment and specification factor binding. Specifically, SOX, OTX, and class III POU factor motifs were enriched in both mid-late and late enhancers, suggesting that enhancers in these groups were the direct targets of these key neural factors (Table S5).

Activity Score Identifies Novel TFs that Are Important for Neural Induction

To narrow down the list of candidate TFs for a follow-up investigation of their effect on neural induction, we defined a TF activity score, which represents the potential to affect transcription at each time point (STAR Methods). We considered two factors that influence this score at each time point: (1) the extent of deviation from the null-expected amount of active enhancers at that time point that are potentially bound by the TF (Grossman et al., 2017), suggesting that these TFs may provide a parsimonious explanation for the MPRA signal (Kashtan and Alon, 2005), and (2) an added requirement that the mRNA that codes for the TF is induced compared to previous time points, which may also suggest functional importance (Rosenfeld et al., 2005; Setty et al., 2003; Yosef et al., 2013). For the former, we focused our attention to enhancers in which the temporal MPRA pattern significantly overlaps with the endogenous patterns, namely those CRSs pertaining to significant entries in Figure 5E. Each of these "consistent" entries represents a certain mode of temporal relationship between MPRA and the endogenous genome—e.g., early induction with matched timing of mRNA expression or late induction that appears after the establishment of chromatin accessibility. Although other CRSs in our data can be of additional interest, we postulate that focusing on temporal regions that conform with the major patterns of overlap with the endogenous processes is desirable when integrating additional genomic readouts (TF binding potential in this case) and may also increase the odds that the respective endogenous region is indeed functional.

The resulting activity matrix (Figure 6A; Table S5) provided a catalog of 107 TFs that could potentially function as regulators of neural induction. Repeating this analysis with the stricter set of temporal regions that were also detected by the per-time-point analysis yielded largely similar results (94 out of 107 cataloged TFs were detected). Similar to previous analyses, we clustered the TF activity score to four representative patterns of activity, early, mid-early, mid-late, and late response, and ranked it by the strength of induction of the respective TF's mRNA expression and the extent of overlap between TF's targets and the significant sub-clusters of MPRA activity. Overall, we observed an agreement between known hESCs and neural-induction-associated TFs and their temporal time points. For example, in the early cluster, the pluripotent marker NANOG showed high TF activity score at 0 h, and immediate-early gene products, ATF3, MYC, and EGR1, showed high score at 3 h, as expected (Herschman, 1991). TFs that had a high score at later time points (24–72 h) included several neural TFs, such as SOX1, OTX2, and PAX6.

Overexpression and CRISPRi Identify Novel Neural-Induction-Associated TFs

To test whether our identified TFs are indeed involved in neural induction, we selected for follow-up overexpression studies 26 highly ranked TFs that were predicted as active during different time points of the induction process: top six in mid-early (FOXL2, BACH2, NR3C1, SMAD1, ELF3, and HOMEZ; primarily active at 12–24 h) and top ten in mid-late (SOX1, NFE2, OTX2, SP5, MAF, ID4, TCF7L2, IRX3, SMAD4, and SOX2; 24–48 h) and late response (DMBX1, OTX1, BARHL1, POU3F2, FOXB1, NR2F2, SOX11, LHX5, SOX5, and PAX6; 48–72 h) clusters. In this follow-up analysis, we used PAX6 as a positive control, because overexpression of PAX6a (short isoform of PAX6) is known to function as a neuroectoderm fate determinant and was previously shown to induce hESCs into a neural lineage (Zhang et al., 2010). In addition, we used EGFP as a negative control.

The chosen TFs were individually overexpressed in hESCs via lentivirus. 4 days post-infection, cells were harvested and examined for various lineage marker genes by qRT-PCR (Figure 6B). We found that overexpression of *BARHL1*, *IRX3*, *LHX5*, *OTX1*, and *OTX2* were sufficient to induce *PAX6* expression, suggesting that these TFs may play a role in neural fate specification. Overexpression of these TFs also induced other neural marker genes directly or indirectly via *PAX6* (Figure 6B). Previous studies have shown that *OTX2* overexpression promotes *PAX6* expression in hESCs upon treatment with the TGF-β inhibitor SB431542 and FGF2 (Greber et al., 2011), and its paralogous gene *OTX1* is known to function similarly (Acampora et al., 2003). It was also reported that *LHX2*, a paralog of *LHX5*, promotes *PAX6* expression and neural differentiation in hESCs (Hou et al., 2013). However, despite *LHX5* being expressed in NPCs, its overexpression has yet to be associated with neural induction. The same holds true for *IRX3* and *BARHL1*, which are known to be expressed in the neuroectoderm and CNS, respectively, in the mouse embryo (Bosse et al., 1997) but whose function in neural induction has not been evaluated. Consistent with these findings, analysis of the *PAX6* promoter region identified binding sites of OTX, IRX3, SOX, and POU that are evolutionarily conserved (Figure 6D). These results are in line with the observations that the respective region is active around 12–24 h, when these TFs are significantly expressed (Figure 2A), and starts to gain a high TF activity score (Figure 6A) at those time points. We found several additional examples of functional neural enhancers that contain conserved OTX, SOX, IRX, and/or homeo-domain (recognized by both LHX and BARHL) binding motifs upstream of the *LHX5*, *POU3F2*, and *OTX2* genes (Figure S6).

We next tested whether these five TFs could lead to a more established neural lineage by analyzing the expression of late neural marker genes (e.g., *FABP7* and *CDH2*) 9 and 14 days after infection. We observed continuous upregulation of the late neural markers at day 9 and 14 (Figure 6C), consistent with the observation that these five factors activated the neural lineage determinant *PAX6* at an early stage. Further examination of these cells at day 14 via bright-field microscopy and immunocyto-chemistry for the neuronal marker MAP2 indicated that the TF-activated cells have acquired neuronal hallmarks (Figure 6E).

To gain a broader understanding of the changes to the transcriptional landscape following overexpression of the five TFs, we carried out RNA-seq analysis at day 14 post-infection, using three replicates per condition. As before, we used *PAX6* as a positive control as well as EGFP as a negative control. As reference, we also sequenced NPCs that were induced by dual-Smad inhibition for 72 h followed by 72-h culture in N2B27 medium supplemented with fibroblast growth factor (FGF) and epidermal growth factor (EGF), termed here as "dSi." PCA of the resulting data validated the reproducibility among three replicates (Figure 7A). Interestingly, the first principal component captured the dichotomy between the two reference states (ESC and NPC, represented by EGFP and dSi, respectively), as can be observed by marker genes and by a more systematic analysis of gene set enrichment (Table S6). The assayed TFs spanned a spectrum between the reference states, where *PAX6* overexpression has the most similar effect to that of dSi as expected and *LHX5* overexpression has the least amount of similarity.

To assess cell lineage, we examined overlaps of differential expression (DE) genes (Love et al., 2014) between each of TF-overexpressed cells and previously published hESC-derived mesendoderm (ME), trophoblast-like cells (TBL), mesenchymal stem cells (MSCs), and NPCs (Xie et al., 2013). This analysis confirmed that the overlaps are most significant for NPCs (Figure 7B) than the other non-neural cells, supporting the role of all the six TFs in neural lineage specification. Gene set enrichment analysis of Gene Ontology (GO) annotations (Subramanian et al., 2005) confirmed, for all overexpressed TFs, significant enrichment in CNS development and neurogenesis processes (Table S6). This analysis also validates that all overexpressed TFs lead to transcriptional changes that significantly overlap with those induced in dSi (Figure 7C; $p < 1e{-}20$; hypergeometric test). Indeed, specific neural marker genes (e.g., *CDH2* and *FABP7*) in TF-overexpressed cells were upregulated at a similar level to PAX6-overexpressed cells (Figure 7D), recapitulating the qPCR results (Figure 6C), and mesoderm and endoderm markers were expressed in a more limited manner.

To explore potential regional characteristic of the TF-overexpressed cells, we focused on anterior-posterior brain marker genes and found that *BARHL1* induced more posterior markers (hindbrain marker *GBX2* and hindbrain-spinal cord markers *HOXB2* and *HOXD4*), although, as expected, *OTX1* and *OTX2* induced anterior identity (fore-midbrain markers *FOXG1*, *SIX3*, *OTX1*, *EMX2*, *PAX2*, *EN1*, and *EN2*; Figure 7E). However, we should note that these TF-overexpressed cells are likely to comprise a heterogeneous regional identity.

To further validate the role of these five genes in neural induction, we set out to test whether knocking them down via CRISPR interference (CRISPRi) (Gilbert et al., 2013) will affect neural differentiation. We introduced dCas9-KRAB and single guide RNAs (sgRNAs) that target the promoters of the five genes into hESCs, followed by neural differentiation and qRT-PCR analyses for various markers at 72 h post-neural induction. We found that CRISPRi of each of the five TFs decreased the expression of early neural genes, such as *PAX6* and *POUF3F1*, and increased the expression of the pluripotent marker *NANOG* (Figure 7F). At 6 days post-induction, later neural markers (i.e., *CDH2* and *FABP7*) were also decreased, although other neural markers, such as *MEIS2* and *DLK1*, showed normal expression and NANOG was downregulated. These results suggest that knockdown of these genes leads to impairment or possibly a delay in neural differentiation and therefore associate these genes as potential players in the regulation of neural differentiation.

DISCUSSION

Genomic analyses of multiple time points during early neural induction provided several findings. We confirmed that neural induction first involves the silencing of pluripotent markers and upregulation of immediate early genes, corresponding to the cell's stress response against differentiation stimuli. This is then followed by the upregulation of genes involved in neural lineage fate specification. We also observed that this process is first controlled by chromatin accessibility or simultaneously with H3K27ac modification followed by an increase in mRNA expression. These results support previous reports about the importance of H3K27ac as an active promoter and enhancer mark that correlates with (and possibly affects) temporal changes in transcription levels, which are not captured by accessibility alone (Heintzman et al., 2009). Finally, our work provides a comprehensive catalog of dynamically changing genes and regulatory elements during neural induction.

Analysis of temporal genes and DNA regions is important not only to understand the regulatory network underlying neural induction but also to dissect neurological disease. Indeed, a large body of evidence suggests that the temporal alteration of genes and regulatory elements involved in neural development can affect neurological phenotypes (Grove et al., 2019), such as cognition and brain size (de la Torre-Ubieta et al., 2018). Fitting with these studies, we observed significant overlap between regions with induced H3K27ac histone modification and neurological disorder GWAS variants. We also observed a significant overlap between the set of loci that had temporal H3K27ac signal in our data and the set of loci found to have an eQTL in the brain and in other tissues. Our null model for computing this statistic was the observed overlap between the set of all H3K27ac peaks in our data (regardless of how they change over time) and the eQTL hits. Finding significant enrichments beyond this baseline suggests that the temporal aspect adds important information, pointing at phenotypically important regions.

The use of lentiMPRA allowed us to functionally test thousands of CRSs and identify 63% that have temporal activity. Although we observed an overall strong correlation between the temporal patterns of these regions and their respective gene expression and chromatin features, it is important to note that this overlap was not obtained for all sequences. Although this may result from inaccuracies in the various assays, it may also point to biologically driven causes (described in detail in STAR Methods).

By integrating information from across all of our large-scale assays, we proposed a scheme to identify and rank TFs based on their predicted activity during the course of development. After an initial screen, we identified BARHL1, IRX3, LHX5, OTX1, and OTX2 as important regulators of neural induction, as both overexpression and knockdown of these factors up- and downregulated *PAX6* and other neural markers, respectively. Although *PAX6* expression in hESCs was shown to be upregulated via *OTX2* (Greber et al., 2011), this finding was novel for the other TFs. Although *LHX5* is a commonly used neural marker, its ability to induce neural induction was not tested. *IRX3* and *BARHL1* were of particular interest. In our study, we found their expression to increase at 24–72 h, and it obtained a high activity score at these time points, suggesting an important role in neural induction. In our overexpression experiments, we observed that they could by themselves control several neural markers in the hESC culture condition, including *PAX6*. To our knowledge, this is the first report demonstrating a potential role for either *IRX3* or *BARHL1* in neural induction. Although we identified important regulators of neural induction, we also observed that both over-expression and knockdown of different TFs perturbed different sets of neural markers or even non-neural markers. This observation suggests that the orchestration of multiple TFs is necessary to fine-tune neural differentiation. Assays that target the molecular function or regulatory grammar of these regulators will be necessary in order to further understand this regulatory network.

STAR★METHODS

LEAD CONTACT AND MATERIALS AVAILABILITY

Plasmids generated in this study have been deposited to Addgene. Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Nadav Ahituv (nadav.ahituv@ucsf.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

hESC culture and neural differentiation—H1 hESCs (WiCell WA-01, RRID:CVCL_9771) were cultured on Matrigel (Corning) in mTeSR1 media (STEMCELL Technologies). Medium was changed daily. For passaging, cells were dissociated using StemPro Accutase (Fisher Scientific), washed and replated on Matrigel-coated dish at a dilution of 1:5 to 1:10 in mTeSR1 media supplemented with 10 μM Y-27632 (Selleck Chemicals). For genomic assays, hESCs were allowed to expand until they were nearly confluent and harvested to obtain undifferentiated hESCs (0 hour). Neural differentiation was performed using a dual-Smad inhibition protocol (Chambers et al., 2009). Briefly, the mTeSR1 media were replaced by neural differentiation media (Knockout DMEM; Life technologies) supplemented with knockout serum replacement (Life technologies), 2 mM L-glutamine, 1× MEM-NEAA (Life technologies), 1x beta-mercaptoethanol (Life technologies), 200 ng/mL Recombinant mouse Noggin (R&D systems), and 10 μM SB431542 (EMD Millipore), and harvested at 3, 6, 12, and 24 hours. At these time points, the cells were 50%–90% confluent in 6-well plates (for RNA-seq and ATAC-seq) or 10 cm dishes (for ChIP-seq and lentiMPRA). The cells were further cultured by refreshing the neural differentiation media daily and harvested at 48 and 72 hours, when the cells were 100% confluent.

METHOD DETAILS

RNA-seq—hESCs were plated in 6-well plates and induced to neural differentiation as described above. Cells from all time points were lysed in RLT buffer (QIAGEN) supplemented with beta-mercaptoethanol and stored in −20°C. Total RNA were extracted using the RNeasy mini kit (QIAGEN) following the manufacturer’s protocol. RNA was quantified with Qubit RNA HS assay kit (Thermo Fisher Scientific). Sequencing library preparation was carried out using Illumina TruSeq Stranded Total RNA Kit. Massively parallel sequencing was performed on an Illumina NextSeq500 with 75 bp paired-end reads. RNA-seq was done with three biological replicates for each of the seven time points and sequenced deeply with an average of 200M reads per replicate.

ChIP-seq—ChIP-seq was performed using LowCell# ChIP kit (Diagenode) according to manufacturer's instruction with modifications. Briefly, cells cultured in 10 cm dishes were crosslinked in 1% formaldehyde (Thermo Fisher Scientific) for 5 minutes. Crosslinking was quenched with 125 mM Glycine. The cells were washed with PBS and precipitated with centrifugation at 6000 rpm for 5 minutes. The cell pellet was stored in −80°C for each time point, so that all the samples were processed together. The pellet was lysed in 250 μL of Buffer B (LowCell# ChIP kit) supplemented with complete protease inhibitor (Roche) and 20 mM Na-butyrate (Sigma). 130 μL of lysed chromatin was sheared using a Covaris S2 sonicator to obtain on average 250 bp size fragments. 870 μL of Buffer A (LowCell# ChIP kit) supplemented with complete protease inhibitor (Roche) and 20mM Na-butyrate (Sigma) was added to the shared chromatin. 20 μL of the chromatin solution was saved as an input control. To obtain magnetic bead-antibody complexes, a mixture of 40 μL of Dynabeads protein A and 40 μL of Dynabeads protein G was washed twice with Buffer A (LowCell# ChIP kit) and resuspended in 800 μL of Buffer A. 10 μg of H3K27ac (Abcam Cat# ab4729, RRID:AB_2118291) or H3K27me3 antibodies (Millipore Cat# 07-449, RRID:AB_310624) were added to the washed beads, and gently agitated at 4°C for 2 hours. The beads-antibody complex was precipitated with a magnet and the supernatant was removed. 800 μL of shared chromatin was added to the beads-antibody complex and rotated at 4°C overnight. The immobilized chromatin was then washed with Buffer A three times and Buffer C once, and eluted in 100 μL of IPure elution buffer (IPure kit; Diagenode). In addition, 80 μL of IPure elution buffer was added to the 20 μL input that were saved before immunoprecipitation, and purified using the IPure kit. Purified DNA was sheared using a Covaris S2 sonicator once again to obtain on average 250 bp fragments. Sequencing libraries were generated using ThruPLEX DNA-seq kit (Rubicon Genomics) according to manufacturer's protocol. The DNA was size-selected using SPRIselect (Beckman Coulter). 0.7× and 0.9× volume of SPRIselect was used for right side and left side selection, respectively. DNA was quantified with Qubit DNA HS assay kit and Bioanalyzer using the DNA High Sensitivity kit (Agilent). Massively parallel sequencing was performed on an Illumina HiSeq4000 with 50 bp single-end read. ChIP-seq was done with two biological replicates for each time point.

ATAC-seq—ATAC-seq was performed according to previously described protocol (Buenrostro et al., 2013) with modifications. Briefly, 50,000 cells were dissociated using Accutase and precipitated with centrifugation at 500 g for 5 minutes. The cell pellet was washed with PBS, resuspended in 50 μL lysis buffer (10 mM Tris·Cl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% Igepal CA-630), and precipitated with centrifugation at 500 g for 10 minutes. The nuclei pellet was resuspended in 50 μL transposition reaction mixture which includes 25 μL Tagment DNA buffer (Nextera DNA sample preparation kit; Illumina), 2.5 μL Tagment DNA enzyme (Nextera DNA sample preparation kit; Illumina), and 22.5 μL nuclease-free water, and incubated at 37°C for 30 minutes. Tagmented DNA was purified with MinElute reaction cleanup kit (QIAGEN). The DNA was size-selected using SPRIselect (Beckman Coulter) according to the manufacturer's protocol. 0.6× and 1.5× volume of SPRIselect was used for right and left side selection, respectively. Library amplification was performed as previously described (Buenrostro et al., 2013). Amplified library was further purified with SPRIselect as described above. DNA was quantified on a Bioanalyzer using the DNA High Sensitivity kit (Agilent). Massively parallel sequencing was performed on an Illumina HiSeq2500 or HiSeq4000 with PE150. ATAC-seq was done in 2 biological replicates for each time point.

lentiMPRA library generation—The lentiMPRA plasmid library was constructed as previously described (Inoue et al., 2017) with minor modifications. Briefly, array-synthesized oligos were amplified with two sets of adaptor primers mentioned previously (pLSmP-AG-f01/r02 and pLSmP-AG-f03/r04, Table S7, sheet 1). The amplified fragments were cloned into pLS-mP vector (Addgene_81225, RRID:Addgene_81225) following its digestion with *SbfI* and EcoRI using In-Fusion HD cloning kit (Takara). The reaction products were transformed into electrocompetent cells (NEB C3020). The pre-library was purified using Plasmid plus midi kit (QIAGEN) and tested for its quality via sequencing on a MiSeq (see below section). The minimal promoter and EGFP fragment (mP-EGFP) was inserted into the *SbfI* and EcoRI restriction sites contained between the enhancer and barcode sequence in the pre-library using T4 DNA Ligase (NEB M0202). The ligation products were then transformed and midi-prepped as mentioned above to obtain the final lentiMPRA library.

Before inserting the mP-EGFP, the plasmid pre-library was examined for the quality of the designed oligos and the representation of individual barcodes via sequencing as previously described (Inoue et al., 2017). CRS-barcode fragments were amplified using pLSmP-ass-F and pLSmP-ass-R-i# primers (Table S7, sheet 1), and purified using MinElute PCR cleanup kit (QIAGEN). The DNA was sequenced with MiSeq (PE150). Two sets of sequencing primers (pLSmP-AG-seqR1 and pLSmP-AG-seqR1_2 for read 1, pLSmP-AG-seqR2 and pLSmP-AG-seqR2_2 for read 2, and pLSmP-AG-seqIndx and pLSmP-AG-seqIndx_2 for index read) were mixed at 1:1 ratio and used for the sequencing. We sequenced the CRS, spacer, and barcode sequences from both read ends and called a consensus sequence from the two reads using PEAR (Quinlan and Hall, 2010; Zhang et al., 2014). We obtained 16.4 million paired-end consensus sequences from this sequencing experiment, 43% of which had the expected length, 30% of sequences were 1 bp short, and 13% were 2 bp short (summing up to 86%), similar to previously reported results (Inoue et al., 2017). Only 0.9% of sequences showed an insertion of 1 bp (Figure S3A). These results are in line with expected dominance of small deletion errors in oligo synthesis. We aligned all consensus sequences back to all designed sequences using BWA MEM (Li and Durbin, 2009) with parameters penalizing soft-clipping of alignment ends (−L 80). We consensus called reads aligning with the same outer alignment coordinates and SAM-format CIGAR string to reduce the effects of sequencing errors. We analyzed all those consensus sequences based on at least three sequences with a mapping quality above 0. Figure S3B shows the distribution of alignment differences (as a proxy for synthesis errors) along the designed oligo sequences. Errors are distributed evenly along the designed sequence, with deletions dominating the observed differences, similar to previous libraries generated in a similar manner (Inoue et al., 2017). We characterized the abundance of oligos further by focusing only on the barcode sequences. Barcode sequences were identified from the respective alignment positions of the alignments created above. To match the RNA/DNA count data analysis (see below), we only considered barcodes of 15-bp length. The number of barcodes per CRS are shown in Figure S3C. The distribution of the abundance of barcodes is available in Figure S3D.

The lentiMPRA library was packaged into lentivirus using Lenti-Pac HIV expression packaging kit (GeneCopoeia) and the lentivirus was concentrated using Lenti-Pac lentivirus concentration solution (GeneCopoeia) according to manufacturer's protocol. The lentivirus was titrated as described previously (Inoue et al., 2017). In brief, H1-hESCs were plated at 1-2 × 10^5 cells/well in 24-well plates and incubated for 24 hours. Serial volume (0, 2, 4, 8, 16, 32 μl) of the lentivirus was added with 8 μg/mL polybrene. The infected cells were cultured for 3 days and washed with PBS three times. Genomic DNA was extracted using the Wizard SV genomic DNA purification kit (Promega). Copy number of viral particle per cell was measured by qPCR as previously described (Inoue et al., 2017).

Lentiviral infection and DNA and RNA extraction—H1 hESCs cultured in a 10 cm dish at 80%-90% confluency were split at 1:4 ratio and re-plated on Matrigel-coated 10 cm dishes in mTeSR1 media supplemented with 10 uM Y-27632. After 24 hours, the cells were infected with the lentivirus library with a multiplicity of infection (MOI) of 5-8 along with 8 μg/mL polybrene (Sigma) and incubated for 3 days with a daily change of the media. Three independent replicate cultures were infected. The infected cells were harvested right before differentiation (0 hours), or differentiated into neural lineage as described previously until appropriate time points (3, 6, 12, 24, 48, and 72 hours). In order to distinguish the barcode expression level between short time gaps (i.e., 0 versus 3 hours, and 3 versus 6 hours), we collected nuclear RNA from all time points and analyzed the nascent state of barcode RNA expression as below. The cells were washed with PBS three times and dissociated with Accutase. The cells were then precipitated with centrifugation at 500 g for 3 minutes and washed with PBS. To isolate cell nuclei, the pellet was lysed in 500 μl lysis buffer [(10 mM Tris-HCL, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$,0.1% (v/v) Igepal CA-630,1U/μl RiboLock RNase inhibitor Thermo Fisher Scientific)]. The cell nuclei were precipitated with centrifugation at 500 g for 10 minutes, and lysed in RLT plus lysis buffer (QIAGEN) supplemented with 2-mercaptoethanol. Genomic DNA and nuclear RNA was purified using an AllPrep DNA/RNA mini kit (QIAGEN). Copy number of viral particle per cell was confirmed by qPCR and shown in Figure S3E. RNA was treated with Turbo DNase (Thermo Fisher Scientific) to remove contaminating DNA.

RT-PCR, amplification, and sequencing of RNA/DNA—Sequencing libraries were prepared as previously described (Inoue et al., 2017). Briefly, 25 μg nuclear RNA was used for reverse transcription with SuperScript II (Invitrogen) using a primer downstream of the barcode (pLSmP-ass-R-UMI-i#, Table S7, sheet 1), which contained a sample index, unique molecular identifier (UMI), and a P7 Illumina adaptor sequence. Barcode sequence was amplified with NEBNext high-fidelity 2X PCR master mix (New England Biolabs) for three cycles using this same reverse primer paired with a forward primer complementary to the 3′ end of EGFP with a P5 adaptor sequence (BARCODE_ lentiF_v4, Table S7, sheet 1). PCR products were purified with 1.8× volume of SPRIselect and underwent a second round of amplification for 22 cycles with the same forward primer and a P7 primer. The PCR products were gel-extracted and purified with MinElute reaction cleanup kit (QIAGEN). To amplify barcode sequence integrated into the genome, 4 μg of genomic DNA was used for PCR amplification as the RNA. The amplified DNA was quantified on a Bioanalyzer (Agilent) using the DNA High Sensitivity kit, and sequenced with an Illumina HiSeq4000 with 100 bp paired-end read. The BARCODE-SEQ-R1-v4 primer was used for read 1 (Table S7, sheet 1). The same primers as pre-library sequencing with MiSeq were used for read 2 and index reads.

The forward and reverse reads on this run each sequenced the designed 15-bp barcodes as well as an adjacent sequence to correctly trim and consensus call barcodes. We obtained a total of 398.9, 406.5 and 415.1 million reads for replicate 1, 2 and 3 respectively; full statistics of read counts is presented in Table S3, sheet1. Across replicates and time points, 95% of barcodes were of the correct length of 15 bp when matched against the designed barcode; full counts and barcode statistics is presented in Table S3, sheet 1-3.

Luciferase assays—To analyze enhancer temporal changes via luciferase assays, we engineered a lentiviral reporter vector pLS-mP-Luc (Addgene, #106253, RRID:Addgene_106253) to have a destabilized form of the luciferase gene by placing a degeneration sequence, hPEST, downstream of the luciferase gene. The hPEST sequence (including 3′ partial sequence of the luciferase gene) was amplified from pGL4.11 (Promega) using the following primers (forward, TTCGAGGCTAAGGTGGTGGA; reverse TACGAAGTTATTAGGTCCCTC GACGAATTCTTAGACGTTGATCCTGGCGC), and inserted into *AgeI* and EcoRI sites of the pLS-mP-Luc. OTX2 ESC (chr14:57385639-57386304; hg19), NANOG ESC (chr12:7940151-7940848; hg19), ENSA ESC (chr1:150613721-150614409; hg19), EDNRB ESC (chr13:78427691-78428404; hg19), TMEM132D ESC (chr12:130255696-130256309; hg19), PAX6 IR (chr11:31832507-31832981; hg19), PARD3 IR (chr10:34716118-34717054; hg19), CRIM1 IR (chr2:36688268-36688907; hg19), PCLO IR (chr7:82434703-82435321; hg19), OTX2 NPC (chr14:57313599-57314370; hg19), CTNNA2 NPC (chr2:80235184-80235754; hg19), DLK1 NPC (chr14:101306429-101307069; hg19), MYB NPC (chr6:135571820-135572468; hg19) enhancers were amplified from the human genome and inserted into XbaI site of the vector. Primers used for cloning are shown in Table S7, sheet 2. The empty pLS-mP-Luc vector was used as a negative control. The plasmids were individually packaged into lentivirus together with Renilla luciferase vector, pLS-SV40-mP-Rluc (Addgene, #106292, RRID:Addgene_106292) at 1:1 molar ratio using Lenti-Pac HIV expression packaging kit (GeneCopoeia) and the lentivirus was concentrated using Lenti-X concentrator (Takara) according to the manufacturer’s protocol. H1 hESCs seeded 24 hours before were infected with the lentivirus along with 8 μg/mL polybrene (Sigma). Three independent replicate cultures were infected. After 48 hours, the cells were induced into a neural lineage using the dual-Smad inhibition method described above. The cells were lysed in buffer PLB (Promega) at 0 (before neural induction), 12 and 72 hours after neural induction. Firefly and renilla luciferase activities were measured on a Glomax microplate reader (Promega) using the Dual-Luciferase Reporter Assay System (Promega). Enhancer activity was calculated as the fold change of each plasmid’s firefly luciferase activity normalized to Renilla luciferase activity.

CRISPRa—To generate a H1 hESC line that stably expresses dCas9-VP64, the lenti dCAS-VP64_Blast vector (Addgene, #61425, RRID:Addgene_61425) was transduced into H1 hESCs via lentivirus at a MOI of 0.2 along with 8 μg/mL polybrene (Sigma) and incubated for 2 days to allow genomic integration. The cells were further cultured for 5 days in a media supplemented with 2 μg/mL blasticidin for selection. Individual colonies were isolated to obtain clonal cell populations and expanded for 2 weeks in blasticidin media. sgRNA sequences for *SOX1*, *IRX3*, *OTX2* enhancers, *PAX6* promoter and non-targeting negative control sequence were amplified as a part of PCR primers (Table S7, sheet 3) using the pLG1 plasmid (gift from Prof. Jonathan Weissman) as a template and cloned into *XhoI* and *BstXI* site of the pLG1. The sgRNA plasmids were transduced into dCas9-VP64 ESCs via lentivirus at a MOI of 5 along with 8 μg/mL polybrene (Sigma) and incubated for 2 days to allow genomic integration. The cells were further cultured for 2 days in a media supplemented with 2 μg/mL puromycin for selection. Total RNA was collected using RNeasy mini kit (QIAGEN). Reverse-transcription was carried out using SuperScript III first-strand synthesis system (Invitrogen). qPCR was performed using SsoFast EvaGreen supermix (Bio Rad) according to the manufacturer's protocol. Primer sequences used for qPCR are shown in Table S7, sheet 5.

Immunocytochemistry—TF-overexpressed cells were fixed using 4% paraformaldehyde (Thermo Fisher Scientific) for 10 minutes and washed three times with PBS. Blocking was performed using blocking/staining solution (0.05% sodium azide, 0.1% NP40, 0.4% BSA, 4% normal goat serum in PBS) for 1 hour. Mouse anti-MAP2 antibody (Thermo Fisher Scientific, catalog# 13-1500, RRID: AB2533001) and Donkey anti-Mouse IgG conjugated with Alexa Fluor 488 (Thermo Fisher Scientific, catalog# R37114, RRID:AB2556542) were used for immunostaining.

Overexpression and RT-qPCR—Total RNA was collected from neural progenitor cells differentiated from hESCs by dual-Smad inhibition as described above. The total RNA was reverse-transcribed using SuperScript III first-strand synthesis system (Invitrogen) according to manufacturer's protocol. cDNA of *ELF3*, *FOXB1*, *HOMEZ*, *ID4*, *IRX3*, *LHX5*, *OTX2*, *PAX6a*, *SMAD1*, *SMAD4* and *SOX1* were amplified. *BARHL1* (catalog#, MHS6278-213245170), *MAF* (catalog#, MHS6278-202806268), *NR2F2* (catalog#, MHS6278-202800802), *NR3C1* (catalog#, MHS6278-202832263), *POU3F2* (catalog#, OHS6271-213587035) and *SOX2* (catalog#, MHS6278-202826163) cDNA clones were obtained from Dharmacon. *SOX11* (clone ID, OHu15579D) and *SP5* (clone ID, OHu03497D) cDNA clones were obtained from Genscript. *BACH2*, *DMBX1*, *FOSL2*, *NFE2*, *OTX1*, *SOX5* and *TCF7L2* cDNA sequences were synthesized by Twist Bioscience. The EGFP gene (negative control) and T2A fragment were also amplified using the pJA291 vector (Addgene #74487, RRID:Addgene_74487) as a template. Sequences synthesized by Twist Bioscience and primers used for the cloning are shown in Table S7, sheet 4. The gene's cDNA fragment and T2A fragment were assembled into pJA291 vector that had been digested with EcoRI and XcmI to generate overexpression vectors that expresses PuroR-mCherry-T2A-cDNA under the control of EF1-alpha promoter. For *BACH2*, *SOX5* and *TCF7L2*, as their cDNA are quite long (i.e., 3 kb), 5′ and 3′ parts of the sequences that overlap each other were separately synthesized by Twist Biosciences and assembled when cloned into the vector. The sequences of cloned cDNA were confirmed by Sanger sequencing. The overexpression vectors were individually packaged into lentivirus using Lenti-Pac HIV expression packaging kit (GeneCopoeia) and the lentivirus was concentrated using Lenti-X concentrator (Takara) according to the manufacturer's protocol. The lentivirus were titrated with H1-ESCs by qPCR, as described above. H1-ESCs cultured in a 24-well plate for 24 hours were infected with the lentivirus with a MOI of 5 along with 8 μg/mL polybrene (Sigma) and incubated for 2 days to allow genomic integration. The cells were further cultured for 2-7 days in a media supplemented with 2 μg/mL puromycin for selection. Three independent replicate cultures were infected. Total RNA was collected using RNeasy mini kit (QIAGEN). Reverse-transcribed using Superscript III first-strand synthesis system (Invitrogen). qPCR was performed using SsoFast EvaGreen supermix (Bio Rad) according to manufacturer's instruction. Primer sequences used for qPCR are shown in Table S7, sheet 5.

CRISPRi—sgRNA sequences for *BARHL1*, *IRX3*, *LHX5*, *OTX1*, *OTX2*, and *PAX6* promoters were cloned into pLG1 as described above. sgRNA plasmids and pHR-SFFV-KRAB-dCas9-P2A-mCherry (Addgene, #60954, RRID:Addgene_60954) were co-packaged and transduced into H1 hESCs via lentivirus at a MOI of 5 along with 8 μg/mL polybrene (Sigma). Cells were incubated for 2 days to allow genomic integration and further cultured for 2 days in mTeSR media supplemented with 2 μg/mL puromycin for selection. At day 4 after infection, the cells were replated and cultured in mTeSR supplemented with puromycin. At day six, cells were induced into a neural lineage by dual-Smad inhibition. At day 9 and 12 (72 hours and 6 days post neural induction), total RNA was collected using RNeasy mini kit (QIAGEN) and reverse-transcribed using SuperScript III first-strand synthesis system (Invitrogen). qPCR was performed using SsoFast EvaGreen supermix (Bio Rad) according to the manufacturer's protocol. sgRNA sequences and primers used for the plasmid construction and RT-qPCR are shown in Table S7, sheet 5.

QUANTIFICATION AND STATISTICAL ANALYSIS

Computational pipeline for RNA-seq, ChIP-seq and ATAC-seq—For RNA-seq, reads were aligned to the hg19 human genome assembly with Tophat2 (Version 2.1.1) (Kim et al., 2013), and low quality reads were trimmed or removed with Trimmomatic (Version 0.3.2) (Bolger et al., 2014). Reads that aligned to more than one gene as well as chimeric fragments were excluded. We also removed genes that failed to be quantified in at least one sample by Cufflinks (Trapnell et al., 2010). We implemented a quality control (QC) pipeline that computes an extensive set of quality metrics, relying in part on FASTQC (Version 0.3.2; Babraham Bioinformatics) and the PICARD suite of alignment metrics (Version 2.5.0 with samtools 1.3.1). Transcript levels were determined using RefSeq transcript annotations, and counting the number of reads aligning to every gene (defined as the union of all splice forms) with featureCounts (Version 1.5.0-p3) (Liao et al., 2014).

For both ChIP-seq and ATAC-seq, we used the FASTQC pipeline (Version 0.3.2; Babraham Bioinformatics) on our reads, and aligned them to the reference genome (hg19) with bowtie version 1.1.1 (Langmead et al., 2009) (for ChIP-seq) and bowtie2 version 2.2.9 (Langmead and Salzberg, 2012) for ATAC-seq, retaining only reads that mapped to a unique position in the genome [“−m 1”]. We marked duplicate reads in the bam files using PICARD and checked for contamination of primer sequences using Trimmomatic (Version 0.3.2) (Bolger et al., 2014).

For each of our H3K27ac and H3K27me3 ChIP-seq replicate pairs per time points peaks were called using MACS2 version 2.1.0 (Zhang et al., 2008), with the relevant control input file with default parameters (setting the FDR to 0.05 and default hg19 human genome size). For each mark in each time point, we intersected the peaks from the two replicates. We then took the union of all of these peaks from all time points per mark while merging regions with maximum distance of 1,000 bp using bedtools (Quinlan and Hall, 2010). This resulted in 40,170 H3K27ac peaks and 4,446 H3K27me3 peaks (Table S1, sheet 1).

For ATAC-seq, after alignment of the reads to the reference genome, reads aligned to the positive strand were moved +4 bp, and reads aligning the negative strand were moved −5bp. For each of our two replicate pairs per time point, we called peaks using MACS2 version 2.1.0 (Zhang et al., 2008) with default parameters (setting the FDR to 0.05 and default hg19 human genome size). For each time point, we intersected the narrow peaks from the two replicates. To generate our final universe of peaks for the LentiMPRA experimental design, we took the union of all peaks from all time points while merging regions with maximum distance of 100 bp using bedtools (Quinlan and Hall, 2010), resulting in 40,486 peaks (Table S1, sheet 1).

Differential activity analyses—We used the 40,042 H3K27ac peaks that did not intersect with H3K27me3 peaks (excluding intersecting peaks per time point) to test for differential activity. We counted the number of H3K27ac reads that fell inside each peak region for each time point, for each of the two replicates. We extracted the shifted ATAC-seq cut sites from our data and counted the number of cut sites that fell inside each of the 40,486 ATAC-seq peak regions for each time point, for each of the two replicates. We used the read count for each transcript across time points and replicates for the RNA-seq. We then used DESeq2 (Love et al., 2014) for all three assays to identify the differential abundance of reads and provide normalized reads (by a scaling factor) matrices. We performed all pairwise comparisons of the seven time points and recorded the FDR for each such comparison for every region/gene.

As an additional analysis of DE/activity over time, we used ImpulseDE (Sander et al., 2017), a package that fits impulse like functions to temporal data and reports differential signals across a time course by assigning an FDR value to each region/gene. To call differential H3K27ac, ATAC-seq or RNA-seq signal we used a cutoff of FDR <0.01 from ImpulseDE and FDR <0.05 for DESeq2, while taking a maximum of 500 top regions/genes per every two time point comparison. This resulted in 2,435 H3K27ac regions, 2,024 ATAC-seq regions (for the 7 time points experiment) and 2,172 genes that showed differential and temporal activity (Table S1, sheet 2). To call differential H3K27me3 signal, we used a more relaxed cutoff of FDR < 0.1 from ImpulseDE and FDR < 0.05 for DESeq2. This resulted in 248 H3K27me3 regions that showed differential and temporal activity.

ChIP-seq, ATAC-seq and RNA-seq clustering—Considering those regions/genes defined as differential in the previous section, we created for each assay, a matrix with the number of reads in each peak region or gene scaled according to the DESeq2 scaling factor and averaged between the two replicates for each time point. We clustered these matrices using the K-means clustering algorithm with 6 clusters (Figures 1B–1D and S1B; Table S1, sheet 3). We also computed for each cluster the significance of its intersection with a cluster from a different assay (Figure 1E) using the hypergeometric test with Bonferroni correction for the p value. For the intersection between H3K27ac and ATAC-seq peaks, we used bedtools (Quinlan and Hall, 2010) to determine if two peaks share at least 1 bp. For the intersection between H3K27ac/ATAC-seq peak region and a gene, we assigned the closest gene to a region (up to 1MB) using GREAT (McLean et al., 2010).

Enrichment of genomic variants—We compared enrichment of variant groups in H3K27ac and ATAC-seq temporal regions to their respective full set of peaks using Fisher's exact test. Variant groups included: the full GWAS catalog as downloaded in February 2018 (MacArthur et al., 2017), relevant disorder subsets of the catalog: alcohol, alzheimer, anxiety, autism, bipolar, borderline, brain volume, cognitive, depression, epilepsy, major depression, OCD, psychosis, schizophrenia, a combined list of disorders (schizophrenia, attention deficit disorder (ADHD), autism, bipolar and major depressive disorder), nervous system disorders obtained using the Experimental Factor Ontology (EFO) (Malone et al., 2010) and negative control height variants. For nervous system disorders and height variants we extracted all variants in linkage disequilibrium ($r^2 > 0.8$) using the SNP Annotation and Proxy Search (SNAP) tool Version 2.2 (Johnson et al., 2008). We also examined eQTLs from different studies (Leslie et al., 2014) and eQTLs from brain tissues (GTEx Consortium, 2015). All variants were converted to hg19 genomic location using the LiftOver tool available on the human genome browser (Kent et al., 2002).

lentiMPRA library design—We devised five criteria to nominate a set of CRS to be tested for their function during neural induction. For criterion 1 we selected sequences that are next to genes involved in neural differentiation or known enhancers that were validated (Table S2, sources of manually curated enhancers were shown in the column "References"). Criteria 2 and 3 require the closest gene to be induced upon neural induction; to satisfy this, we require the gene to be included in one of clusters 2 to 6 in Figure 1B. For criterion 4, we selected the most significant 20 (sorted by FDR) of each of the following tests: 1) RNA-seq differential expression over time (using ImpulseDE); differential signal in one of the following: 2) ATAC-seq 3hr versus 0hr (using Deseq2); 3) ATAC-seq 72hr versus 0hr (Deseq2); 4) H3K27ac 3hr versus 0hr (Deseq2); 5) H3K27ac 72hr versus 0hr (Deseq2); 6) RNA-seq of nearest gene 3hr versus 0hr (Deseq2); 7) RNA-seq of nearest gene 72hr versus 0hr (Deseq2). The criteria were applied sequentially (in the order in which they were described), and the respective sets of candidate enhancers are mutually exclusive. To focus on neural induction, we excluded regions that are adjacent to the pluripotent factors SOX2, KLF4, MYC, NANOG, and POU5F1.

Notably, all selection criteria use a subset of our ATAC-seq data (0, 3, and 72 hours), which was available during the design of the library. Furthermore, we excluded from the design sequences that overlap with regions in the hg19/ ENCODE blacklist (https://sites.google.com/site/anshulkundaje/projects/blacklists) (Table S2).

Due to limitations of the procedure of oligonucleotide synthesis, the assayed sequence are required to be 171 bp long. If a selected candidate region is shorter than 171 bp, we extended it equally from each side. If it is longer than 171 bp - we record all 171 bp options with a sliding window of 1 bp. For each such 171 bp sequence candidate we recorded motif hits using Fimo (Grant et al., 2011) with FDR < 10^{-4} cutoff using the motif list from ENCODE (Kheradpour and Kellis, 2014) and chose the candidate sequence that has the maximal number of hits and satisfies the following criteria: 1) The sequence should not contain EcoRI (GAATTC) and *SbfI* (CCTGCAGG), because these sites were later used for inserting the minimal promoter (mP) and EGFP gene between the candidate regulatory sequence and barcode; 2) We discarded sequences with homopolymers longer than 8bp, since homopolymers can affect oligo synthesis; 3) There should be no more than 25% overlap (of the 171bp) with simpleRepeats regions from ENCODE (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/simpleRepeat.txt.gz).

We added to these 171 bp sequences a 5′ primer sequence (AGGACCGGATCAACT), along with a 14 bp spacer sequence (CCTGCAGGGAATTC) that contains two restriction enzyme sites (*SbfI* and EcoRI), to allow for the subsequent insertion of the minimal promoter and EGFP gene followed by a 15 bp designed barcode sequences and a 3′ primer sequence (CATTGCGTGAACCGA) (Figure 2B) (Inoue et al., 2017). In our final array design, we included 2,664 different target sequences (2,271 – sequences after filtering, 193 controls, 200 scrambled sequences), each with 90 different barcodes to provide a robust readout (Figure 2B). Barcode sequences of 15 bp length were designed to have at least two substitutions and one 1 bp insertion distance to each other. Homopolymers of length 3 bp and longer were excluded in the design of these sequences, and so were ACA/CAC and GTG/TGT trinucleotides (bases excited with the same laser during Illumina sequencing). More than 556,000 such barcodes were designed using a greedy approach. The barcodes were then checked for the creation of *SbfI* and EcoRI restriction sites when adding the spacer and 3′ flanking sequences. From the remaining sequences, a random subset of 239,760 barcodes was chosen for the design. The final designed oligo sequences are available in Data S1.

Replicates, normalization and RNA/DNA ratios—We used both the forward and reverse reads to sequence the 15 bp reporter barcodes and obtain consensus sequences. We matched the observed barcodes against the designed barcodes, and noticed that across replicates and sample types, ~95% of barcodes had the correct 15 bp length. Only correct size barcodes that are observed at least once in both RNA and DNA of the same sample were subsequently used for analysis (assuming basal levels of transcription through the minimal promoter).

To estimate the RNA to DNA ratio per barcode in each replicate, we first scaled the RNA and DNA read counts using the number of reads as scaling factor.

$$\text{RNA / DNA ratio per barcode:}\frac{RNA\ reads}{(sum\ RNA\ reads)}\Big/\frac{DNA\ reads}{(sum\ DNA\ reads)}$$

Although the DNA and RNA counts of individual barcodes are highly correlated between experiments (Table S3, sheet 4), the noise of each measure results in a poor correlation of RNA/DNA ratios (Table S3, sheet 4). However, there are on average 68-72 barcodes per CRS in each replicate (out of 90 barcodes programmed on the array; Table S3, sheet 5). To reduce noise, we aggregated the RNA or DNA counts across all associated barcodes for each CRS.

To estimate the abundance of DNA or RNA per CRS and for each replicate (in order to compare replicates and time point, we use a simple averaging scheme:

D(R)NA per CRS = $(10^6 * \sum_{i=1}^{\#BC} D(R)NA_i / \#BC * (sum\ D(R)NA\ reads))$ where D(R)NA$_i$ denotes the reads of a specific barcode I among the #BC barcodes that belong to the respective CRS.

To determine the RNA/DNA ratios per CRS and for each replicate we used two strategies:

$$\text{Ratio of sums:}\frac{\sum_{i=1}^{\#BC}\frac{RNA_i}{(sum\ RNA\ reads)}}{\sum_{i=1}^{\#BC}\frac{DNA_i}{(sum\ DNA\ reads)}} \quad (1)$$

$$\text{Sum of ratio:}\frac{\sum_{i=1}^{\#BC}\left(\frac{RNA_i}{(sum\ RNA\ reads)}\Big/\frac{DNA_i}{(sum\ DNA\ reads)}\right)}{\#BC} \quad (2)$$

We added a pseudo count of 1 to the numerator and denominator to stabilize signal from CRS with low numbers of reads. Table S3, sheet 6 shows DNA or RNA abundance and RNA/DNA ratios per CRS between every two replicates per time point. Table S3, sheet 7 shows DNA or RNA abundance and RNA/DNA ratio (using the two schemes) per CRS for each replicate, comparing every two time points. Notably, the two schemes are largely consistent. In the remainder of this study, we used the second scheme (sum of ratios). We also compared between DNA, RNA and ratio per time point for each replicate. We observed low correlation between RNA/DNA ratios and DNA counts, indicating that enhancer activity was not influenced by the number of DNA integrations (Table S3, sheet 8).

Although normalized individually, the three replicates do not seem to be on the exact same scale (Table S3). To combine replicates, we therefore first divided the RNA/DNA ratios observed in each sample (time point/ replicate) by the median ratio and then obtained the final RNA/DNA ratio by averaging the normalized values across replicates.

Determining differential and temporal CRS activity using MPRAnalyze— MPRAnalyze is a statistical framework for analyzing MPRA data (Ashuach et al., 2019), using a parametric graphical model to infer the enhancer induced transcription rate. The model assumes a linear relationship between the latent plasmid (DNA) and transcript (RNA) counts, relating them through scaling by the transcription rate α. The plasmid counts are assumed to follow a log-Normal distribution, and the RNA counts are assumed to follow a Negative Binomial distribution. This model incorporates external covariates, such as batch effect, barcode-specific effect and conditions of interest, by fitting two nested generalized linear models, one fitting the latent plasmid counts from the DNA counts, and the other fitting the transcription rate from the latent plasmid counts and the observed transcript counts. This model was designed to leverage the statistical power of multiple barcodes. More details are provided in Ashuach et al. (2019).

<u>Quantification and classification of active enhancers:</u> to classify active CRS, estimates of α were extracted for each time point from the model described above. The α values corresponding to control enhancers are used as the baseline, and a modified z-score is computed for each CRS. The scores are computed as the distance from the median of the control α values, normalized by the median absolute deviation (MAD): $score_i = \alpha_i - median(\overrightarrow{\alpha_C}) / K \cdot MAD(\overrightarrow{\alpha_C})$, where the constant K is set to ensure that the scores behave asymptotically normal, and a_C is the vector of values corresponding to control enhancers. P values are produced based on these scores compared with the standard normal distribution.

Identifying temporal CRS— To test for temporal activity, we incorporate control enhancers to define the null temporal behavior and use a likelihood ratio testing to detect significant temporal behavior. For a given CRS, the null assumption is that it behaves according to the null temporal behavior. We evaluate this assumption by fitting a joint model for the time course data of this enhancers, together with the set of negative controls. In the alternative model, the CRS has a temporal profile that is different form the null. To evaluate it, we fit a separate model for the controls and the CRS. In this scheme, a CRS with temporal behavior that significantly deviates from the null will have a clear benefit to the likelihood under the alternative model. The score is therefore computed by a likelihood ratio test between the two models.

Clustering MPRA data and association with other genomic assays— We clustered temporal MPRA regions into four rough patterns of expression, namely early, mid-early, mid-late and late response (Table S4, sheet 1). Using the genomic location of each region, we retrieved the normalized number of reads using DESeq2 (Love et al., 2014) from an overlapping H3K27ac and ATAC-seq peaks (if any), as well as the expression of the nearest gene. We clustered each genomic assay separately to four clusters (similar to MPRA signal clustering). We then compared MPRA temporal profile to that of each genomic assay by measuring the overlap between the resulting clusters using a hypergeometric test and Bonferroni corrected FDR < 0.05 (Table S4, sheet 2).

TF activity score computation—To compute the activity score of each TF (represented by a motif or a ChIP-seq experiment) at each time point, we look for consistent sub-clusters that peak during that time point (in terms of MPRA signal) and that significantly overlap with the putative target regions of the TF (p value < 0.005, Hypergeometric test). We then count the number of putative target regions that appear in at least one significantly overlapping sub-cluster. The final score is defined by the number of regions found at each time point divided by the total number of regions found across all time points. As an additional constraint, we only consider time points in which the mRNA that encodes for the TF is highly expressed (6th or higher quantile of expressed genes) and significantly induced compared to the preceding time point [p value < 10^{-5}; for the first time point (0 hour), we compare to the subsequent time point (3 hours) (Love et al., 2014)].

TF activity score ranking—The TF activity score was ranked per each one of the 4 clusters with an unbiased approach that is based only on the data produced for this paper. It uses two components: (i) the p value of the overlap between the TF's targets and significant sub clusters of MPRA activity (Figure 5E; Table S5) – taking the minimum p value. (ii) log fold of the TF's mRNA induction according to cluster: 0-12hr for cluster 2, 0-48hr for cluster 3, 0-72hr for cluster 4. We rank (i) and (ii) per cluster and use their average as the final ranking score.

RNA-seq following TF overexpression—RNA-seq was performed by Novogene for the eight cell populations across three replicates, including: overexpression of *BARHL1*, *IRX3*, *LHX5*, *OTX1*, *OTX2* and *PAX6*, negative control EGFP (corresponding to hESC state) and dSi, which were induced from hESCs via dual-Smad inhibition for 72 hours followed by further 72h-culture in N2B27 medium supplemented with 20 ng/mL bFGF (R&D systems) and 20 ng/mL EGF (MilliporeSigma). The RNA-sequencing data was processed similarly to the procedure described above. PCA analysis of RNA-sequencing these eight cell populations across three replicates, is based on the 1000 most variable genes (Figure 7A). For the overlap between DE genes (EGFP, dSi) and (EGFP, factor) or (EGFP, factor_i) and (EGFP, factor_j) we used a jaccrad score of: (|∩ upregulated genes| + |∩ downregulated genes|)/(|∪ upregulated enes| + | ∪ downregulated genes|).

The hypergeometric test of the overlap used a background of all genes with TMP > 1, resulting in p value = 0 for all the tests (Figure 7C). We used DESeq2 (Love et al., 2014) for differential expression (DE) analysis for comparing each of the six overexpressed TFs to controls (EGFP and dSi) and comparing EGFP to dSi. Upregulated and downregulated genes were defined based on the cutoff of FDR < 0.05; |logFC| > 1 (Figures 7D and 7E). For the cell lineage analysis, we used data on lineage-restricted genes of four hESC-derived cell types (i.e., trophoblast-like cells (TBL), mesendoderm (ME), mesenchymal stem cells (MSCs) and neural progenitor cells (NPCs)), and restricted the lineage-restricted genes to have FPKM > 1 only in that lineage based on Table S1 in Xie et al. (2013). We examined their overlaps with our DE genes (EGFP, factor_i) in a similar way to the jaccard score described above (Figure 7B).

Characterizing MPRA and chromatin/mRNA inconsistencies—To investigate the inconsistency phenomenon at the chromatin level, we turned to the cluster- level analysis

(Figures 5E and S6). This analysis was designed to identify cases where regions that exhibit a certain temporal pattern with MPRA are likely to exhibit another pattern in their accessibility or H3K27 acetylation (adjusted p value < 0.05). As a general trend, the results indicate that 'inconsistent' temporal regions tend to become induced (when assayed by MPRA) after the occurrence of chromatin changes in their respective endogenous loci. For instance, we observe a significant overlap between the set of regions that become induced after 24 hours when examined by MPRA (MPRA cluster 3; Figure 5A), and the set of regions that become (or remain) accessible during the preceding time points (ATAC-seq cluster 1; Figure 5D). Furthermore, this pattern of delay is observed more often with chromatin accessibility, compared with H3K27ac (Figure S4). These results could potentially be explained by our previous observations that DNA accessibility precedes H3K27ac during neural induction, which is followed by gene expression changes (Figure 1E), and that the temporal H3K27ac signal is a stronger indicator for MPRA enhancer activity (Figure 3A). In addition to inconsistency with the chromatin readouts, we also observe temporal CRS that show inconsistency with their postulated target genes.

Specifically, we observe temporal CRS that were active several time points before their postulated target genes (Figure S4B) and the opposite, where genes were active before the CRS (Figure S4C). The pattern of MPRA induction before the endogenous mRNA can be rationalized by additional constraints that may exist in the endogenous regions, but not necessarily in the (random) integration sites such as dependence on a wider chromatin context, which may be required to enable transcription. Additional technical factors of the assay, including the length of the assayed sequence (171 bp), may also underlie these discrepancies. It is also worth noting that our assays only find potential enhancers but not their target gene/s. Conversely, the latter pattern (mRNA before MPRA) is harder to rationalize and is more likely a result of the assay's inaccuracy.

We investigated the cases in which the MPRA data of temporal CRS and the mRNA data of their respective genes did not match. To account for cases of CRS- gene miss-assignment, we removed from this analysis cases where there was another nearby gene (looking at the closest four) that was more correlated with the MPRA but showed inconsistency with the closest gene mRNA signal. To this end we first separately clustered each of the two sets (closest genes and the most correlated neighboring genes) to four temporal clusters. We declared two clusters c1 (from the set of closest genes) and c2 (from the set of most correlated genes) as sufficiently matching if the median of the Pearson correlation coefficient across all pairwise comparisons of the respective genes was larger than 0.5. We considered a region for further analysis if the clusters that contain its closest gene and its most correlated neighboring gene are sufficiently matching. Counting the number of occurrences of each of the two patterns, we find that the second one (mRNA before MPRA) is of a substantially lower abundance (137 versus 358 enhancers), and that its size is in fact at the level of overlap between random sets (adjusted Hypergeometric p value > 0.05) (Figure 5E). The resulting regions are depicted in Figure S4.

DATA AND CODE AVAILABILITY

The datasets generated during this study are available at the NCBI Gene Expression Omnibus (GEO) as accession number GEO: GSE115046. The published article includes all code generated or analyzed during this study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We want to thank Encoded Therapeutics for assistance on the RNA-seq studies and David Fischer, Martin Kircher, Beth Martin, and Jay Shendure (UW) for assistance with the MPRA studies and analyses. This work was supported in part by the National Institute of Mental Health 1K99MH117393-01 (A.K.), 1R01MH109907, and U01MH116438 (N.A.) and the Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation (N.A.). N.A. is also supported in part by the National Human Genome Research Institute (NHGRI)1UM1HG009408 and R01HG010333, the National Institute of Child Health and Human Development 1P01HD084387, the National Heart Lung and Blood Institute R35HL145235 and R01HL117004, and the National Institute of Diabetes and Digestive Kidney Diseases R01DK116738. N.Y. and A.K. were supported by NHGRI U01HG007910.

REFERENCES

Acampora D, Annino A, Puelles E, Alfano I, Tuorto F, and Simeone A (2003). OTX1 compensatesfor OTX2 requirement in regionalisation of anterior neuroectoderm. Gene Expr. Patterns 3, 497–501. [PubMed: 12915318]

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461. [PubMed: 24670763]

Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, DrablØs F, Lennartsson A, Rönnerblad M, Hrydziuszko O, Vitezic M, et al. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science 347, 1010–1014. [PubMed: 25678556]

Ashuach T, Fischer DS, Kreimer A, Ahituv N, Theis FJ, and Yosef N (2019). MPRAnalyze: statistical framework for massively parallel reporter assays. Genome Biol. 20, 183. [PubMed: 31477158]

Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Mol. Autism 8, 21. [PubMed: 28540026]

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, and Snyder M; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. [PubMed: 22955616]

Bolger AM, Lohse M, and Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. [PubMed: 24695404]

Bosse A, Zülch A, Becker MB, Torres M, Gómez-Skarmeta JL, Modolell J, and Gruss P (1997). Identification of the vertebrate Iroquois homeobox gene family with overlapping expression during early development of the nervous system. Mech. Dev 69, 169–181. [PubMed: 9486539]

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122, 947–956. [PubMed: 16153702]

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, and Greenleaf WJ (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218. [PubMed: 24097267]

Chambers SM, Fasano CA, Papapetrou EP, Tomishima M, Sadelain M, and Studer L (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat. Biotechnol 27, 275–280. [PubMed: 19252484]

de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, and Geschwind DH (2018). The dynamic landscape of open chromatin during human cortical neurogenesis. Cell 172, 289–304.e18. [PubMed: 29307494]

Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature 518, 331–336. [PubMed: 25693564]

Donnard E, Vangala P, Afik S, McCauley S, Nowosielska A, Kucukural A, Tabak B, Zhu X, Diehl W, McDonel P, et al. (2018). Comparative analysis of immune cells reveals a conserved regulatory lexicon. Cell Syst. 6, 381–394.e7. [PubMed: 29454939]

Endo M, Doi R, Nishita M, and Minami Y (2012). Ror family receptor tyrosine kinases regulate the maintenance of neural progenitor cells in the developing neocortex. J. Cell Sci 125, 2017–2029. [PubMed: 22328498]

Feng J, Liu T, and Zhang Y (2011). Using MACS to identify peaks from ChIP-seq data. Curr. Protoc. Bioinformatics *Chapter 2*, Unit 2.14.

Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, et al.; FANTOM Consortium (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat. Genet 46, 558–566. [PubMed: 24777452]

Garber M, Yosef N, Goren A, Raychowdhury R, Thielke A, Guttman M, Robinson J, Minie B, Chevrier N, Itzhaki Z, et al. (2012).A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell 47, 810–822. [PubMed: 22940246]

Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, Tsankov A, Shalek AK, Kelley DR, Shishkin AA, Issner R, et al. (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell 153, 1149–1163. [PubMed: 23664763]

Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154, 442–451. [PubMed: 23849981]

Grant CE, Bailey TL, and Noble WS (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018. [PubMed: 21330290]

Greber B, Coulon P, Zhang M, Moritz S, Frank S, Müller-Molina AJ, Araúzo-Bravo MJ, Han DW, Pape HC, and Schöler HR (2011). FGF signalling inhibits neural induction in human embryonic stem cells. EMBO J. 30, 4874–4884. [PubMed: 22085933]

Grossman SR, Zhang X, Wang L, Engreitz J, Melnikov A, Rogov P, Tewhey R, Isakova A, Deplancke B, Bernstein BE, et al. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proc. Natl. Acad. Sci. USA 114, E1291–E1300. [PubMed: 28137873]

Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, Pallesen J, Agerbo E, Andreassen OA, Anney R, et al.; Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium; BUPGEN; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium; 23andMe Research Team (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet 51, 431–444. [PubMed: 30804558]

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648–660. [PubMed: 25954001]

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459, 108–112. [PubMed: 19295514]

Herschman HR (1991). Primary response genes induced by growth factors and tumor promoters. Annu. Rev. Biochem 60, 281–319. [PubMed: 1883198]

Herzog W, and Weber K (1978). Fractionation of brain microtubule-associated proteins. Isolation of two different proteins which stimulate tubulin polymerization in vitro. Eur. J. Biochem 92, 1–8. [PubMed: 729584]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 706, 9362–9367.

Hou PS, Chuang CY, Kao CF, Chou SJ, Stone L, Ho HN, Chien CL, and Kuo HC (2013). LHX2 regulates the neural differentiation of human embryonic stem cells via transcriptional modulation of PAX6 and CER1. Nucleic Acids Res. 47, 7753–7770.

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, and Shendure J (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. 27, 38–52. [PubMed: 27831498]

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, and de Bakker PIW (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24, 2938–2939. [PubMed: 18974171]

Kalkman HO (2012). A review of the evidence for the canonical Wnt pathway in autism spectrum disorders. Mol. Autism 3, 10. [PubMed: 23083465]

Kashtan N, and Alon U (2005). Spontaneous evolution of modularity and network motifs. Proc. Natl. Acad. Sci. USA 702, 13773–13778.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D (2002). The human genome browser at UCSC. Genome Res. 72, 996–1006.

Kheradpour P, and Kellis M (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 42, 2976–2987. [PubMed: 24335146]

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, and Kellis M (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 23, 800–811. [PubMed: 23512712]

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 74, R36.

Kreimer A, Zeng H, Edwards MD, Guo Y, Tian K, Shin S, Welch R, Wainberg M, Mohan R, Sinnott-Armstrong NA, et al. (2017). Predicting gene expression in massively parallel reporter assays: A comparative study. Hum. Mutat. 38, 1240–1250. [PubMed: 28220625]

Kreimer A, Yan Z, Ahituv N, and Yosef N (2019). Meta-analysis of massively parallel reporter assays enables prediction of regulatory function across cell types. Hum. Mutat. 40, 1299–1313. [PubMed: 31131957]

Kwasnieski JC, Fiore C, Chaudhari HG, and Cohen BA (2014). High-throughput functional testing of ENCODE segmentation predictions. Genome Res. 24, 1595–1602. [PubMed: 25035418]

Langmead B, and Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. [PubMed: 22388286]

Langmead B, Trapnell C, Pop M, and Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 70, R25.

Leslie R, O'Donnell CJ, and Johnson AD (2014). GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics 30, i185–i194. [PubMed: 24931982]

Li H, and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. [PubMed: 19451168]

Liao Y, Smyth GK, and Shi W (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930. [PubMed: 24227677]

Lodato MA, Ng CW, Wamstad JA, Cheng AW, Thai KK, Fraenkel E, Jaenisch R, and Boyer LA (2013). SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state. PLoS Genet. 9, e1003288. [PubMed: 23437007]

Love MI, Huber W, and Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 75, 550.

Luizon MR, Eckalbar WL, Wang Y, Jones SL, Smith RP, Laurance M, Lin L, Gallins PJ, Etheridge AS, Wright F, et al. (2016). Genomic characterization of metformin hepatic response. PLoS Genet. 12, e1006449. [PubMed: 27902686]

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45 (D1), D896–D901. [PubMed: 27899670]

Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, and Parkinson H (2010). Modeling sample variables with an experimental factor ontology. Bioinformatics 26, 1112–1118. [PubMed: 20200009]

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195. [PubMed: 22955828]

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, and Bejerano G (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol 28, 495–501. [PubMed: 20436461]

Melnikov A, Zhang X, Rogov P, Wang L, and Mikkelsen TS (2014). Massively parallel reporter assays in cultured mammalian cells. J. Vis. Exp 51719.

Quinlan AR, and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. [PubMed: 20110278]

Rodda DJ, Chew JL, Lim LH, Loh YH, Wang B, Ng HH, and Robson P (2005). Transcriptional regulation of nanog by OCT4 and SOX2. J. Biol. Chem 280, 24731–24737. [PubMed: 15860457]

Rosenfeld N, Young JW, Alon U, Swain PS, and Elowitz MB (2005). Gene regulation at the single-cell level. Science 307, 1962–1965. [PubMed: 15790856]

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A, et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet 46, 944–950. [PubMed: 25086666]

Sander J, Schultze JL, and Yosef N (2017). ImpulseDE: detection of differentially expressed genes in time series data using impulse models. Bioinformatics 33, 757–759. [PubMed: 27797772]

Sanders SJ, Neale BM, Huang H, Werling DM, An JY, Dong S, Abecasis G, Arguello PA, Blangero J, Boehnke M, et al.; Whole Genome Sequencing for Psychiatric Disorders (WGSPD) (2017). Whole genome sequencing in psychiatric disorders: the WGSPD consortium. Nat. Neurosci. 20, 1661–1668. [PubMed: 29184211]

Setty Y, Mayo AE, Surette MG, and Alon U (2003). Detailed map of a cisregulatory input function. Proc. Natl. Acad. Sci. USA 700, 7702–7707.

Smith RP, Eckalbar WL, Morrissey KM, Luizon MR, Hoffmann TJ, Sun X, Jones SL, Force Aldred S, Ramamoorthy A, Desta Z, et al. (2014). Genome-wide discovery of drug-dependent human liver regulatory elements. PLoS Genet. 10, e1004648. [PubMed: 25275310]

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, and Mesirov JP (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 702, 15545–15550.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, and Pachter L (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol 28, 511–515. [PubMed: 20436464]

Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, Amit I, Gnirke A, and Meissner A (2015). Transcription factor binding dynamics during human ES cell differentiation. Nature 518, 344–349. [PubMed: 25693565]

Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS, and Sankaran VG (2016). Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell 165, 1530–1545. [PubMed: 27259154]

Visel A, Minovitsky S, Dubchak I, and Pennacchio LA (2007). VISTA Enhancer Browser–a database of tissue-specific human enhancers. Nucleic Acids Res. 35, D88–D92. [PubMed: 17130149]

Volk DW, and Lewis DA (2013). Prenatal ontogeny as a susceptibility period for cortical GABA neuron disturbances in schizophrenia. Neuroscience 248, 154–164. [PubMed: 23769891]

Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, Claussnitzer M, and Kellis M (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat. Commun 9, 5380. [PubMed: 30568279]

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158, 1431–1443. [PubMed: 25215497]

Wu Q, Chen X, Zhang J, Loh YH, Low TY, Zhang W, Zhang W, Sze SK, Lim B, and Ng HH (2006). Sall4 interacts with Nanog and co-occupies Nanog genomic sites in embryonic stem cells. J. Biol. Chem 281, 24090–24094. [PubMed: 16840789]

Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, et al. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell 153, 1134–1148. [PubMed: 23664764]

Yosef N, and Regev A (2016). Writ large: genomic dissection of the effect of cellular environment on immune response. Science 354, 64–68. [PubMed: 27846493]

Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, Awasthi A, Wu C, Karwacz K, Xiao S, Jorgolli M, et al. (2013). Dynamic regulatory network controlling TH17 cell differentiation. Nature 496, 461–468. [PubMed: 23467089]

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, and Liu XS (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. 9, R137. [PubMed: 18798982]

Zhang X, Huang CT, Chen J, Pankratz MT, Xi J, Li J, Yang Y, Lavaute TM, Li XJ, Ayala M, et al. (2010). Pax6 is a human neuroectoderm cell fate determinant. Cell Stem Cell 7, 90–100. [PubMed: 20621053]

Zhang J, Kobert K, Flouri T, and Stamatakis A (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30, 614–620. [PubMed: 24142950]

Ziller MJ, Edri R, Yaffe Y, Donaghey J, Pop R, Mallard W, Issner R, Gifford CA, Goren A, Xing J, et al. (2015). Dissecting neural differentiation regulatory networks through epigenetic footprinting. Nature 518, 355–359. [PubMed: 25533951]

Highlights

- RNA-seq, ChIP-seq, and ATAC-seq reveal regulatory dynamics during neural induction
- lentiMPRA functionally characterized >1,500 temporal enhancers
- Combined genomic analyses ranked and identified key neural induction factors
- Overexpression or CRISPRi of 5 different factors affected neural differentiation

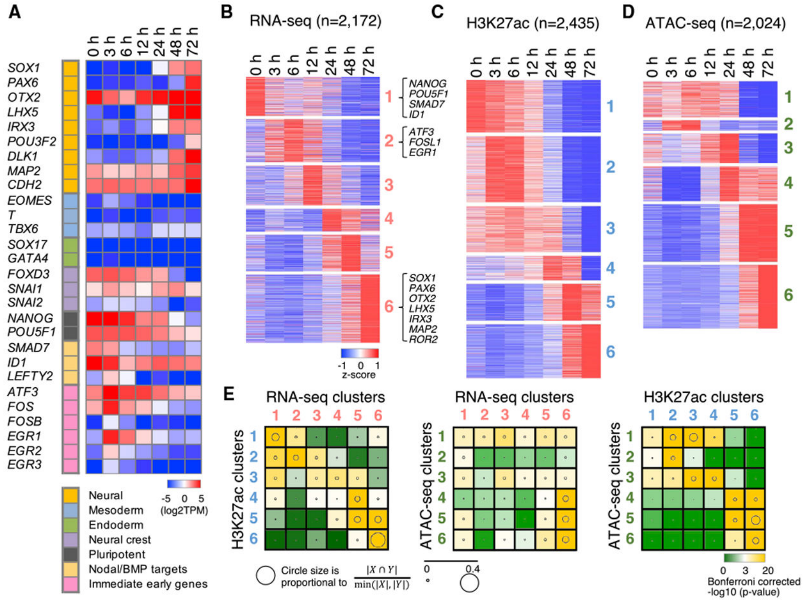


Figure 1. The Dynamic Changes of ATAC-seq, ChIP-seq, and RNA-seq Peaks Are Sequentially Correlated

(A) Transcripts per million (TPM) (log2, averaging over three biological replicates) per time point of marker genes (neural, mesoderm, endoderm, neural crest, pluripotent, nodal/BMP targets, and immediate early genes).

(B–D) Heatmap of scaled read counts (log2, averaged over three biological replicates and standardized per row) of temporal genes and genomic regions, showing data from RNA-seq (B), H3K27ac ChIP-seq (C), and ATAC-seq (D). The loci in each assay were clustered into six groups based on their temporal patterns.

(E) Overlap between the temporal clusters in the three data modalities (Bonferroni-corrected p values of a hypergeometric test). Circle sizes represent the proportion of overlap between every two clusters. The overlap is computed either at the region level (ATAC-seq versus ChIP-seq) or at the gene level (ATAC/ChIP-seq versus RNA-seq; regions in the former assays are represented by their nearest gene).

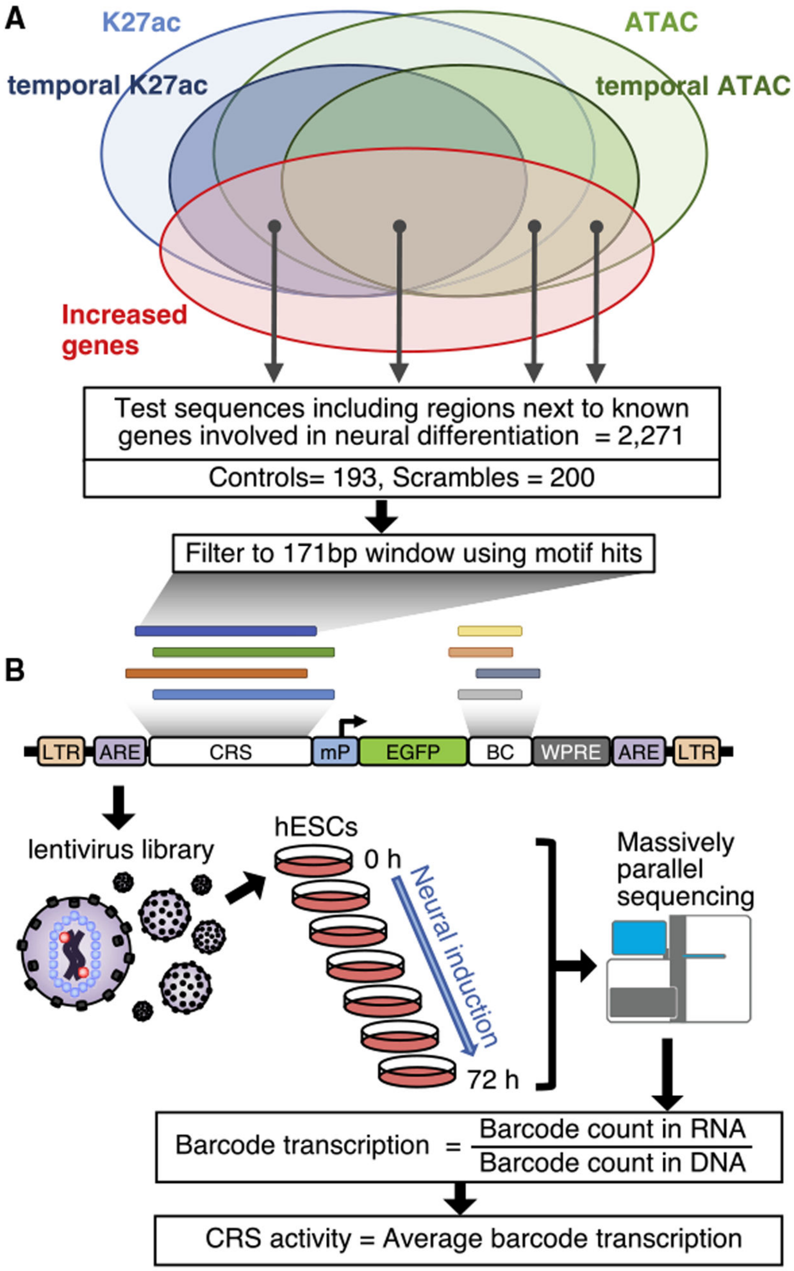

A
K27ac
ATAC
temporal K27ac
temporal ATAC
Increased genes
Test sequences including regions next to known genes involved in neural differentiation = 2,271
Controls= 193, Scrambles = 200
Filter to 171bp window using motif hits
B
LTR
ARE
CRS
mP
EGFP
BC
WPRE
ARE
LTR
lentivirus library
hESCs
0 h
Neural induction
72 h
Massively parallel sequencing
Barcode transcription = Barcode count in RNA / Barcode count in DNA
CRS activity = Average barcode transcription

Figure 2. Experimental Design of lentiMPRA

(A) Sequence selection for lentiMPRA. 2,271 candidate regulatory regions (CRSs) were selected based on RNA-seq, H3K27ac ChIP-seq, and ATAC-seq data. Curated known enhancers (Table S2), 193 positive control regions, and 200 negative controls were included as well.

(B) Schematic showing lentiMPRA design. CRSs along with 15-bp barcodes were synthesized on a custom array and cloned into a lentiMPRA vector. The library was packaged into lentivirus and infected into hESCs. The infected cells were cultured for 3 days to allow genomic integration. DNA and nuclear RNA were extracted at seven time points (0, 3, 6, 12, 24, 48, and 72 h) and subjected to sequencing followed by estimation of transcriptional activity. ARE, antirepressor element; BC, barcode; LTR, long terminal repeat; mP, minimal promoter; WPRE, woodchuck hepatitis virus posttranscriptional regulatory element.

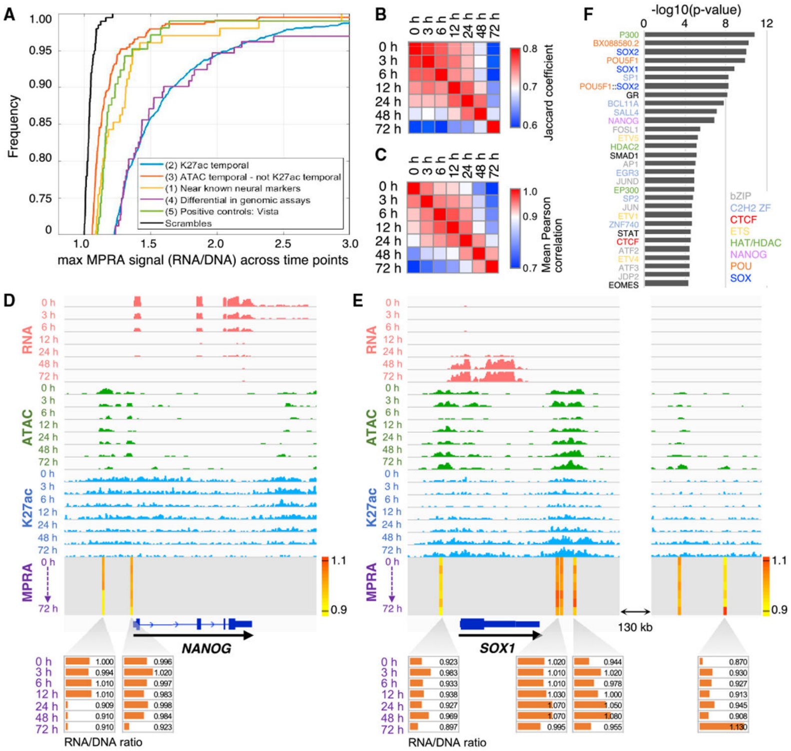


Figure 3. lentiMPRA Signal for Different Enhancer Types

(A) Cumulative distribution function indicating the frequency (y axis) of sequences with a specific MPRA signal (x axis; taking the maximum signal over time). Design criterion number (1–5) is indicated per each tested group of CRSs.

(B and C) Similarity between the MPRA signal measured at different time points, using either the intersection of the sets of significantly active regions (Jaccard coefficient; B) or the correlation of the signals (Pearson correlation; C).

(D and E) RNA-seq (red), ATAC-seq (green), H3K27ac ChIP-seq (blue), and MPRA (RNA/DNA ratio heatmap) tracks around *NANOG* (D) and *SOX1* (E). RNA/DNA ratio at each time point is shown as bar charts at the bottom.

(F) Enrichment of predicted TF binding sites in temporal CRS. Top 30 differentially enriched TF binding sites when comparing temporal and non-temporal CRSs are shown (Fisher's exact test). TF categories are indicated on the right side.

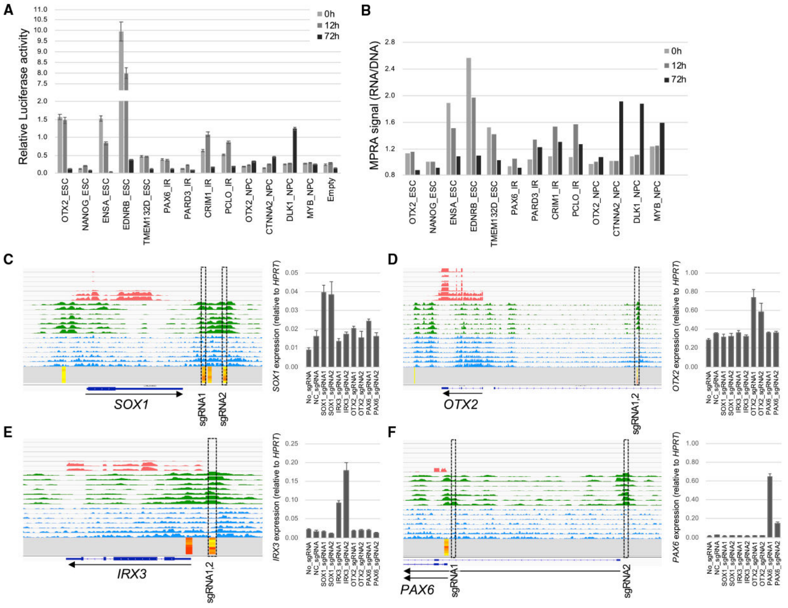


Figure 4. Validation of Enhancers by Luciferase Assays and CRISPRa

(A) Relative luciferase activity for each enhancer compared to Renilla luciferase activity at 0, 12, and 72 h post-neural induction. Five ESC enhancers, four immediate response (IR) enhancers, four NPC enhancers, and empty pLS-mP-Luc vector (negative control) were tested.

(B) MPRA signal (RNA/DNA ratio) at 0, 12, and 72 h post-neural induction.

(C–F) Functional validation of enhancers by CRISPRa. sgRNAs that target enhancers nearby *SOX1* (C), *OTX2* (D), *IRX3* (E), and *PAX6* alternative promoters (F) or negative control sgRNA (NC_sgRNA) were infected into hESCs that stably express dCas9-VP64. Upregulation of respective genes relative to *HPRT* were examined by qPCR and shown as bar charts on the right.

Data are presented as means ± SD of three independent experiments.

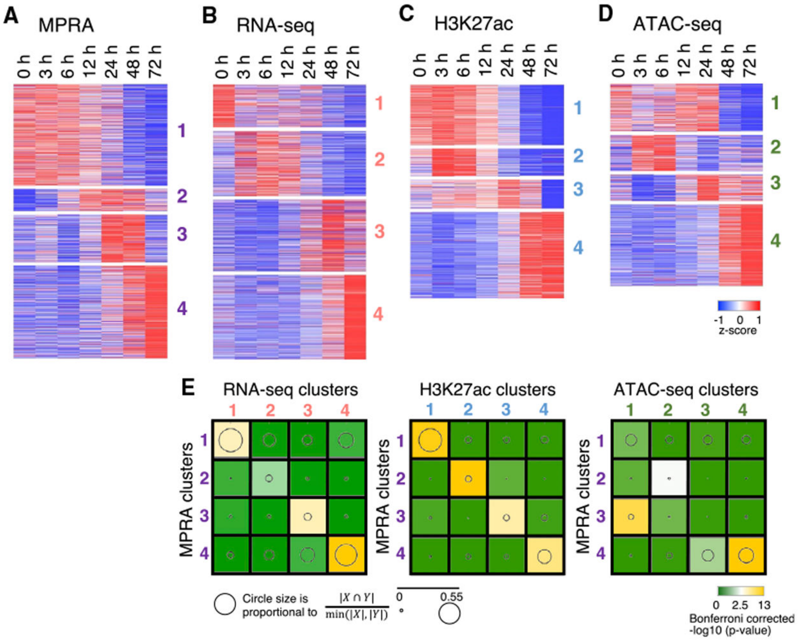


Figure 5. Activity of Temporal CRS: Comparing lentiMPRA to the Endogenous Signals

(A–D) Temporal MPRA signal (RNA/DNA ratio; A), normalized read count of the closest gene detected by RNA-seq (B), H3K27ac ChIP-seq (C), and ATAC-seq (D), clustered into four temporal groups separately. Rows are standardized.

(E) Overlap between the lentiMPRA clusters and the three genomic data modalities. Shown are Bonferroni-corrected p values of a hypergeometric test. Circle sizes represent the proportion of overlap between every two clusters. The overlap is computed either at the region level (lentiMPRA versus ATAC-seq or ChIP-seq) or at the gene level (lentiMPRA versus RNA-seq; using the nearest gene to represent each lentiMPRA region).

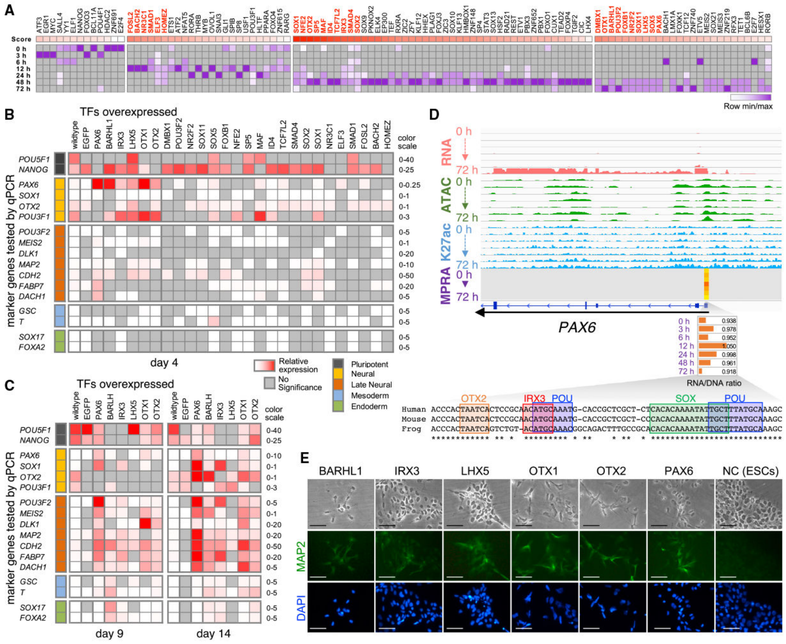


Figure 6. Activity Score Identifies Novel TFs Involved in Neural Induction

(A) Heatmap of activity scores per TF per time point. Values are normalized (minimum to maximum) per each row and sorted by considering both the induction of the TF's mRNA expression and the overlap of the TF's targets with significant sub-clusters of MPRA activity for each cluster. The 26 TFs used for overexpression are marked in red font.

(B and C) TF overexpression. Marker gene expressions (pluripotent, mesoderm, endoderm, and neural) are examined by qRT-PCR at early (B; day 4) and late (C; days 9 and 14) time points post-vector transduction. Relative expression compared to the *HPRT* gene is shown as a heatmap with the scale on the right side. Grey entries indicate no significant changes (Student's t test; $p > 0.05$).

(D) TF analyses of the *PAX6* promoter region show binding sites for OTX2, IRX3, POU, and SOX that are evolutionally conserved between human, mouse, and frog (*Xenopus tropicalis*).

(E) MAP2 immunocytochemistry. hESCs overexpressing *BARHL1*, *IRX3*, *LHX5*, *OTX1*, *OTX2*, *PAX6*, and negative control (NC) were stained with MAP2. Bright field (top), MAP2 (middle), and DAPI (bottom) are shown. Scale bars represent 200 μm.

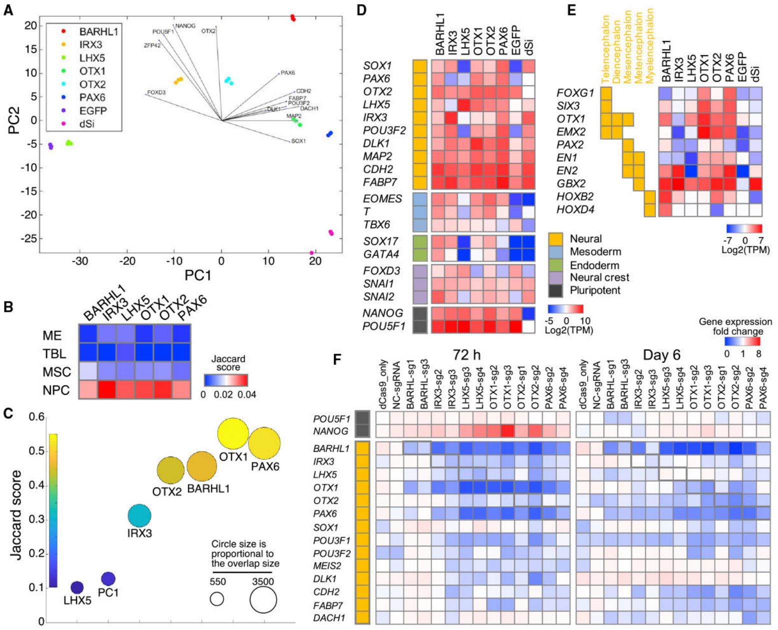


Figure 7. RNA-seq for Cell Overexpressing TFs and CRISPRi

(A) PCA analysis of RNA-seq for overexpression of *BARHL1*, *IRX3*, *LHX5*, *OTX1*, *OTX2*, *PAX6*, and EGFP (negative control) and dSi (positive control) across three replicates, based on the 1,000 most variable genes. x axis PC1; y axis PC2.

(B) Jaccard score for the overlap between lineage-restricted genes of four hESC-derived cell types (ME [mesendoderm]; MSCs [mesenchymal stem cells]; NPCs [neural progenitor cells]; TBLs, [trophoblast-like cells]) (Xie et al., 2013) and our DE genes (EGFP and factor).

(C) Overlap between genes that are differentially expressed between the reference conditions (EGFP and dSi) and genes that are differentially expressed after overexpression using a Jaccard score (intersection over union; note that only genes that had consistent direction of change [upregulated in both or down-regulated in both] were considered to be a part of the intersection set).

(D and E) TPM (log2, averaging over three biological replicates) for selected cell lineage markers (neural, mesoderm, endoderm, neural crest, and pluripotent; D) and brain regional markers (telencephalon, diencephalon, mesencephalon, metencephalon, and myelencephalon; E).

(F) TF knockdown by CRISPRi. sgRNAs that target promoters of *BARHL1*, *IRX3*, *LHX5*, *OTX1*, *OTX2*, and *PAX6* or negative control sgRNA (NC_sgRNA) were infected into hESCs along with dCas9-KRAB. Cells infected only with dCas9-KRAB (dCas9 only) were used as a negative control. Marker gene expression relative to *HPRT* was examined by qPCR at 72 h and 6 days after neural induction. Upregulation (red) or downregulation (blue) comparing to non-treated wild-type hESCs is shown as heatmap matrices.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-Histone H3 (acetyl K27) antibody	Abcam	Cat# ab4729, RRID:AB_2118291
Rabbit Anti-Histone H3, trimethyl (Lys27) Polyclonal antibody	Millipore	Cat# 07-449, RRID:AB_310624
Mouse anti-MAP2 antibody	Thermo Fisher Scientific	Cat# 13-1500, RRID:AB_2533001
Donkey anti-Mouse IgG conjugated with Alexa Fluor 488	Thermo Fisher Scientific	Cat# R37114, RRID:AB_2556542
Bacterial and Virus Strains		
Biological Samples		N/A
Chemicals, Peptides, and Recombinant Proteins		
Matrigel	Corning	Cat# 354277
Y-27632	Selleck Chemicals	Cat# S1049
knockout serum replacement	Life technologies	Cat# 10828-028
Recombinant mouse Noggin	R&D systems	Cat# 1967-NG-025/CF
SB431542	EMD Millipore	Cat# 616464-5MG
Complete protease inhibitor	Roche	Cat# 11 873 580 001
Na-butyrate	Sigma	Cat# B5887-1G
polybrene	Sigma	Cat# TR-1003-G
N-2 supplement	Life technologies	Cat# 17502048
B27 supplement w/o A	Life technologies	Cat# 12587-010
Recombinant Human FGF basic	R&D systems	Cat# 233-FB-025/CF
Epidermal Growth Factor (EGF)	MilliporeSigma	Cat# GF144
Critical Commercial Assays		
LowCell# ChIP kit	Diagenode	Cat# C01010072
IPure kit v2	Diagenode	Cat# C03010015
ThruPLEX DNA-seq kit	Rubicon Genomics	Cat# R400428
Nextera DNA sample preparation kit	Illumina	Cat# FC-121-1030
Lenti-Pac HIV expression packaging kit	GeneCopoeia	Cat# HPK-LvTR-40
Allprep DNA/RNA mini kit	QIAGEN	Cat# 80204
Dual-Luciferase Reporter Assay System	Promega	Cat# E1980
SuperScript II	Invitrogen	Cat# 18064071
SuperScript III first-strand synthesis system	Invitrogen	Cat# 18080051
Deposited Data		
Raw and analyzed data	This paper	GEO: GSE115046
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Experimental Models: Cell Lines		
H1 hESCs	WiCell	Cat# WA-01, RRID:CVCL_9771
Oligonucleotides		
BARHL1	Dharmacon	Cat# MHS6278-213245170
MAF	Dharmacon	Cat# MHS6278-202806268

REAGENT or RESOURCE	SOURCE	IDENTIFIER
NR2F2	Dharmacon	Cat# MHS6278-202800802
NR3C1	Dharmacon	Cat# MHS6278-202832263
POU3F2	Dharmacon	Cat# OHS6271-213587035
SOX2	Dharmacon	Cat# MHS6278-202826163
SOX11	Genscript	Cat# OHu15579D
SP5	Genscript	Cat# OHu03497D
Primer Sequences for lentiMPRA	See Table S7 for sequences	N/A
Primer Sequences for cDNA cloning	See Table S7 for sequences	N/A
Primer Sequences for enhancer cloning	See Table S7 for sequences	N/A
sgRNA Sequences	See Table S7 for sequences	N/A
Primer Sequences for RT-qPCR	See Table S7 for sequences	N/A
Recombinant DNA		
pLS-mP	Addgene	Cat#81225, RRID:Addgene_81225
pLS-mP-Luc	Addgene	Cat#106253, RRID:Addgene_106253
pLS-SV40-mP-Rluc	Addgene	Cat#106292, RRID:Addgene_106292
pGL4.11	Promega	Cat#E6661
lenti dCAS-VP64_Blast	Addgene	Cat#61425, RRID:Addgene_61425
pLG1	Gilbert et al., 2013	N/A
pJA291	Addgene	Cat#74487, RRID:Addgene_74487
pHR-SFFV-KRAB-dCas9-P2A-mCherry	Addgene	Cat#60954, RRID:Addgene_60954
Software and Algorithms		
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
bowtie	Langmead et al., 2009	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools	Li and Durbin, 2009	http://samtools.sourceforge.net/
Tophat2	Kim et al., 2013	https://ccb.jhu.edu/software/tophat/
Trimmomatic	Bolger et al., 2014	http://www.usadellab.org/cms/?page=trimmomatic
Cufflinks (Trapnell et al., 2010	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/
FASTQC	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/
PICARD suite		https://broadinstitute.github.io/picard/
featureCounts	Liao et al., 2014	http://subread.sourceforge.net/
MACS2	Zhang et al., 2008	https://github.com/taoliu/MACS
bedtools	Quinlan and Hall, 2010	https://bedtools.readthedocs.io/en/latest/
DESeq2	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
ImpulseDE	Sander et al., 2017	https://bioconductor.org/packages/release/bioc/html/ImpulseDE.html
GREAT	McLean et al., 2010	http://great.stanford.edu/public/html/
BWA MEM	Li and Durbin, 2009	http://bio-bwa.sourceforge.net/
MPRAnalyze	Ashuach et al., 2019	https://bioconductor.org/packages/release/bioc/html/MPRAnalyze.html