Massively parallel reporter perturbation assay uncovers temporal regulatory architecture during neural differentiation

Anat Kreimer,[1,2,3,4,#,*], Tal Ashuach [3,#], Fumitaka Inoue [1,2,5,#], Alex Khodaverdian [3], Nir Yosef [3,6,7,*], Nadav Ahituv,[1,2,*]

[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, CA 94720, USA

[4] Department of Biochemistry and Molecular Biology, Center for Advanced Biotechnology and Medicine, Rutgers–Robert Wood Johnson Medical School, Piscataway, NJ 08854, USA

[5] Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, 606-8501, Japan

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

[7] Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA

[#] These authors contributed equally

*Correspondence: kreimer@cabm.rutgers.edu (A.K.), niryosef@berkeley.edu (N.Y.), nadav.ahituv@ucsf.edu (N.A.)

Abstract

Gene regulatory elements play a key role in orchestrating gene expression during cellular differentiation, but what determines their function over time remains largely unknown. Here, we performed perturbation-based massively parallel reporter assays at seven early time points of neural differentiation to systematically characterize how regulatory elements and motifs within them guide cellular differentiation. By perturbing over 2,000 putative DNA binding motifs in active regulatory regions, we delineated four categories of functional elements, and observed that activity direction is mostly determined by the sequence itself, while the magnitude of effect depends on the cellular environment. We also find that fine-tuning transcription rates is often achieved by a combined activity of adjacent activating and repressing elements. Our work provides a blueprint for the sequence components needed to induce different transcriptional patterns in general and specifically during neural differentiation.

Introduction

Enhancers are DNA sequences containing clustered recognition sites (i.e. motifs) for transcription factors (TFs) that play a pivotal role in transcriptional regulation of gene expression during numerous biological processes, including cellular differentiation [1]. This is evident by the abundance of disease associated variants discovered through genome-wide association studies (GWAS) and expression quantitative trait loci (eQTLs) residing in noncoding regions [2]. Despite their importance, our understanding of the regulatory grammar of enhancers, namely the manner by which their DNA sequences pertain to their function remains largely unknown, thus limiting our ability to infer how changes in these sequences affect their functionality and lead to higher- level consequences.

Various biochemical assays (e.g. ChIP-seq, DNase-seq, ATAC-seq) have enabled genome-wide identification and characterization of candidate regulatory sequences such as enhancers, across different cell types [3], providing descriptive maps of the human genome. Complementary studies use genome modification approaches, such as CRISPR-Cas9, to functionally characterize enhancer elements by targeting their locations in the genome [4]. Such assays capture both direct and indirect causal relationships between the tested regulatory elements and cellular phenotype (e.g. gene expression) and in many cases target regions that are bound by specific transcription factors of interest [5]. Massively parallel reporter assays (MPRAs) provide an alternative approach that enables simultaneously testing the regulatory activity of thousands of regulatory sequences and their variants. In MPRA, a copy of any sequence of interest is synthesized in front of a transcribed barcode. There are many variants to this technology [6], including one that utilizes lentivirus to integrate into the genome (hereafter we refer to this as lentiMPRA; [7]) used for these assays. The ratio between the abundance of transcribed barcode (read with RNA-seq) and the number of coding sequences (evaluated with DNA- sequencing) provides a quantitative readout for the regulatory activity of the assayed sequence [6,8–12].

Approaches to understanding the roles of TFs in determining the activity of a given enhancer and the interplay between TFs in an enhancer [13] are generally limited by the number of causal relationships they can study directly (e.g., via gene knockdown), primarily due to cost and availability of efficient perturbing agents. Therefore large scale studies often use correlational inference, e.g. associating TF binding with changes in gene expression based on motif- gene association [14]. These, however, are confounded by a slew of observations whereby only a small fraction of potential TF-binding sites (TFBSs) are actually occupied in any given cell type, and these sites vary substantially across cell types and conditions [15–18]. Another caveat of perturbing endogenous factors that affect gene expression (e.g., TFs, enhancer regions) is the abundance of indirect effects, which are difficult to discern from the direct ones. These two issues are mitigated by MPRA, as it provides a cost-effective approach to investigate thousands of candidate enhancer sequences along with variants of these sequences in which certain DNA binding motifs are perturbed. The concern for indirect effects is mitigated to some extent as well due to the synthetic nature of the assay (i.e., the transcribed barcode is non- functional).

Several studies have utilized MPRAs to characterize how TF binding may affect regulatory activity. For example, placing TFBSs at different numbers, order, spacing and orientation on

'neutral' background sequences allowed the dissection of regulatory grammar in a human hepatocellular carcinoma cell line [19]. One common finding is that the number of TFBSs (i.e. homotypic clusters of TFBSs [20]) largely determines expression and this relationship follows a non-linear increase with an eventual plateauing of expression [19,21–24]. Grossman et al. [25] used both synthetic and endogenous sequences to specifically test the effect of PPARγ binding motifs and show that distinct sets of features govern PPARγ binding vs. enhancer activity. Specifically, they found that PPARγ binding is largely governed by the affinity of the specific motif binding site while the enhancer activity of PPARγ binding sites depends on varying contributions from dozens of TFs in the immediate vicinity, including interactions between combinations of these TFs. Kheradpour et al. [9] examined five predicted TF activators and two predicted repressors and measured effects of their motif disruption in regulatory elements using MPRA. Their findings indicate that disrupting predicted activator motifs abolishes enhancer function, while changes in repressors maintain enhancer activity and that evolutionary conservation, nucleosome exclusion, binding of other factors, and motif affinity are predictive of enhancer activity. However, these studies examined a small number of TF motifs and assessed their functional effects in a limited number of conditions or cell types. Altogether, they focused on a specific time point and not a temporal course or developmental process.

The differentiation of stem cells into the neural lineage provides an exemplary model for studying how gradual and non-reversible changes to the cell's phenotype may be transcriptionally regulated. During this process, stem cells rapidly differentiate both on a molecular and physiological level to generate neurons. We previously characterized the temporal dynamics of gene expression (RNA-seq) and gene regulation (ATAC-seq, H3K27ac and H3K27me3 ChIP-seq and lentiMPRA) at seven time points (0-72 hours) during the early parts of this process [26]. Using lentiMPRA, we identified numerous endogenous sequences that had temporal enhancer activity (i.e. the expression of their target barcode was well over the background levels and significantly changed over time). This activity tended to correlate with cell-endogenous changes to the expression of their target gene and to the structure of their surrounding chromatin. In addition, the genomic positions of the validated temporal sequences significantly overlapped with loci that have been associated with neurodevelopmental disorders, in particular autism spectrum disorder (ASD). Combining all our genomic data, we developed a prioritization method to select TFs that are putatively involved in driving a neural fate, and validated the role of several candidates with direct genetic perturbations. This study, however, is still limited to validations of a handful of TFs and lacks in understanding of the way by which these TFs may drive changes in transcription over time.

To more comprehensively identify DNA binding motifs that may affect transcription and characterize the timing in which they carry out their effect, we utilized a 'perturbation MPRA' approach. Based on our previous data, we compiled a list of 591 regulatory sequences whose activity differed over time (considering different temporal patterns) as well as a selected set of 255 motifs within those regions. We then prioritized for testing 2,144 instances of the selected motifs in the selected regions. We used leniMPRA to perturb, via three different approaches, the selected instances, and evaluated their effect over at the same seven time points (0-72 hours) during the neural differentiation process. Using this approach, we found that 27% (598) of the perturbations had a significant effect on the transcription of the target gene. We divided these

motif instances into several subtypes based on the direction (suppressing or inducing transcription) and magnitude (fold change, compared to the unperturbed and negative control sequences) of their effect. We observed that the magnitude of the effects often varied over time (indicating that it depends on the cellular environment), while the direction of the effect is independent of time and is broadly determined by the DNA sequence (i.e the combination of perturbed motif and the surrounding region). Furthermore, we observed cases of activating and repressing motif instances that are harbored within the same regulatory region, suggesting that in those cases fine-tuning of transcription levels may be achieved by a combination of opposing effects. Finally, by perturbing pairs of motifs in a select set of sequences, we found evidence for different patterns of cooperation between motifs, and that both fundamental models, namely the 'enhanceosome' model of an all-or-nothing machinery, and the 'billboard' model of independent contribution [15,27] are supported by our data. Overall, our findings suggest that the regulatory grammar of enhancers that changed in gene expression in our system is an amalgam of a wide variety of different mechanisms. It also helps establish perturbation MPRA as a powerful approach for high-throughput investigation of such mechanisms in different cellular contexts.

Results

Selection of regions and motifs for perturbation MPRA

To characterize the effect of DNA binding motifs on gene expression over time, we first set out to choose a set of regulatory regions that showed temporal activity during early neural differentiation using lentiMPRA data from our previous study (0, 3, 6, 12, 24, 48, and 72 hours post induction; [26]). Our initial candidate set consisted of 1,547 171bp sequences that were identified as temporally active (i.e. the expression of their target barcode varied significantly, both over time and in comparison to a control sequence; **Methods;** [28]). We then used FIMO [29] to computationally identify occurences of known DNA binding motifs in each sequence (using motifs identified by Kheradpour and Kellis [30] and Weirauch et al. [31]).

Following these analyses, we chose specific regions and motifs for perturbation lentiMPRA. As we are limited by the number of sequences that can be included in a single lentiMPRA library due to low integration rate in ESCs, we developed an optimization framework to select the combination of regions and motifs that maximizes the representation of relevant genomic properties (**Fig. 1a-b; Methods**). To this end, we wanted to include regions and motifs that are associated with different temporal patterns of chromatin and gene expression signals, derived from our previous analysis of H3K27ac ChIP-seq, ATAC-seq and RNA-seq data in the same time points [26]. We made sure to include a sufficient number of regions in which H3K27ac is induced early in the differentiation, as well as regions that gain this mark later on, and closer to the neural progenitor (NP) phase. Similarly, we selected a minimal number of motifs whose corresponding TFs are induced early in the differentiation process, as well as motifs associated with late- induced TFs. We also chose to provide explicit preference for a curated list of regions and TFs that have been previously associated with neural induction pathways. Finally, we required that every selected motif will be perturbed in at least 20 regions (thus allowing us to observe the motif in multiple contexts), and every selected region will have at least two different perturbations (for two different motifs). With these considerations taken together, the respective experimental design problem can be represented as an optimization problem: selecting the minimal number of motif instances [(region x motif) pairs] while satisfying all of our design constraints above. In the methods section, we describe how we represent this as a connectivity problem in graphs and how we derive a solution for it using Integer Linear Programming. Applying this scheme to our data resulted in a selection of 2,144 motif instances over 591 regions and 255 motifs (**Fig. 1a-c**).

We considered the 2,144 motif instances both in their wild type (***WT***) and in a perturbed form (***PERT***) where the sequence of the motif instance is modified in order to estimate its effect. For 100 of our genomic regions, chosen by the motifs they harbor and their importance for neural differentiation ([26]; see **Methods**), we also perturbed pairs of motifs (including two appearances of the same motif in the sequence or two different motifs), to analyze cooperative effects (**Fig. 1c; Methods**). We perturbed each of the selected motifs using three different designs that rely on (**Fig. 1d**), two approaches: In designs 1 and 2 we identified two fixed "non-motif" sequences (i.e sequences with minimal number of predicted motif hits - details in **Methods**) and replaced the motif with the prefix of these sequences, adjusting to the motif length. In the third design we randomly shuffled the nucleotides of the motif (**Methods**). We also included two sets of negative

controls: 1) scrambled sequences (*SCRAM*) – where we shuffle all nucleotides of each of the 591 WT sequences; 2) random sequence alterations (*RAND*) – where we randomly shuffled a small region (length of the median motif size) at a random location in each region. In total, 10,041 sequences were included in our lentiMPRA library (**Fig. 1c-d**).

lentiMPRA perturbation

The designed sequences were synthesized and cloned upstream of a minimal promoter (mP) into the lentiMPRA vector (**Fig. 1d; Methods**). During the cloning process, 15-bp random barcodes were placed in the 5'UTR of the EGFP reporter gene [32]. The association between the cloned sequences and barcodes was determined via DNA-seq (**Methods**). Lentivirus was generated and human embryonic stem cells (hESCs) were infected with the library (**Fig. 1e**). Following three days, to allow for viral integration and degradation of unintegrated virus, the hESC were differentiated to a neural lineage using the dual-Smad inhibition protocol [33]. Integrated DNA barcodes and transcribed RNA barcodes were quantified by DNA-seq and RNA-seq, respectively, at seven time points of neural differentiation (0, 3, 6, 12, 24, 48, and 72 hours) (**Fig. 1f**). The library infections were carried out using three biological replicates (two replicates were infected with the same lentivirus batch, while the other replicate was infected with another lentivirus batch).

Using a computational pipeline developed in our group, MPRAflow [34], we took a stringent approach to associate barcodes with the cloned sequences. For each barcode, we required at least 80% of the reads associated with the barcode to map it to a single sequence, and a minimum of 3 reads supporting that assignment, resulting in over 1.4 million confidently assigned barcodes, and averaging 139 barcodes per sequence (**Methods**). We then analyzed the barcodes from the lentiMPRA infected cells and matched them with the confidently assigned barcodes of the library. Across biological replicates, we were able to confidently assign an average of 61.6% of the barcodes (**Methods** and **Supplementary Fig. 1**). Considering only confidently assigned barcodes that have a representation both in RNA and DNA from infected cells, we observed an average of 134.4 barcodes per sequence in each replicate (**Supplementary Fig. 2**), corresponding to 9,948 out of the 10,041 designed sequences (2,082, 2,086, and 2,114 sequences for perturbation methods 1-3 respectively (**Supplementary Table 1**)). We then used MPRAnalyze [28] to aggregate the barcodes and quantify the transcription rate induced by each tested sequence (dubbed 'alpha'). We observed reproducible results between replicates (average Pearson correlation 0.98) in every time-point (**Supplementary Fig. 3**), and results were highly concordant with our previously characterized lentiMPRA in the same system [26] (mean Pearson correlation 0.79, **Supplementary Fig. 4**). Comparing the four categories of sequences that we tested (**Supplementary Fig. 5**), we observe as expected, that overall, the scrambled negative controls (SCRAM) had the lowest transcriptional activity, while the unperturbed sequences (WT) had the highest. We also observed that sequences with a perturbed binding site (PERT) had a generally lower level of activity than sequences with a perturbation of random sites (RAND), confirming that perturbing known motifs has an effect larger than expected by chance. We next quantified the magnitude of deviation between PERT and WT transcription rates (Log(WT/PERT)) and compared the results between all three perturbation methods. Overall, we observed correlated results between the three methods, both in terms of the estimated transcription rate of

the perturbed sequences (average Pearson correlation 0.81) and the differential activity between the perturbed sequences and their corresponding WT sequence (Log(FC), average Pearson correlation 0.71) (**Supplementary Fig. 6**).
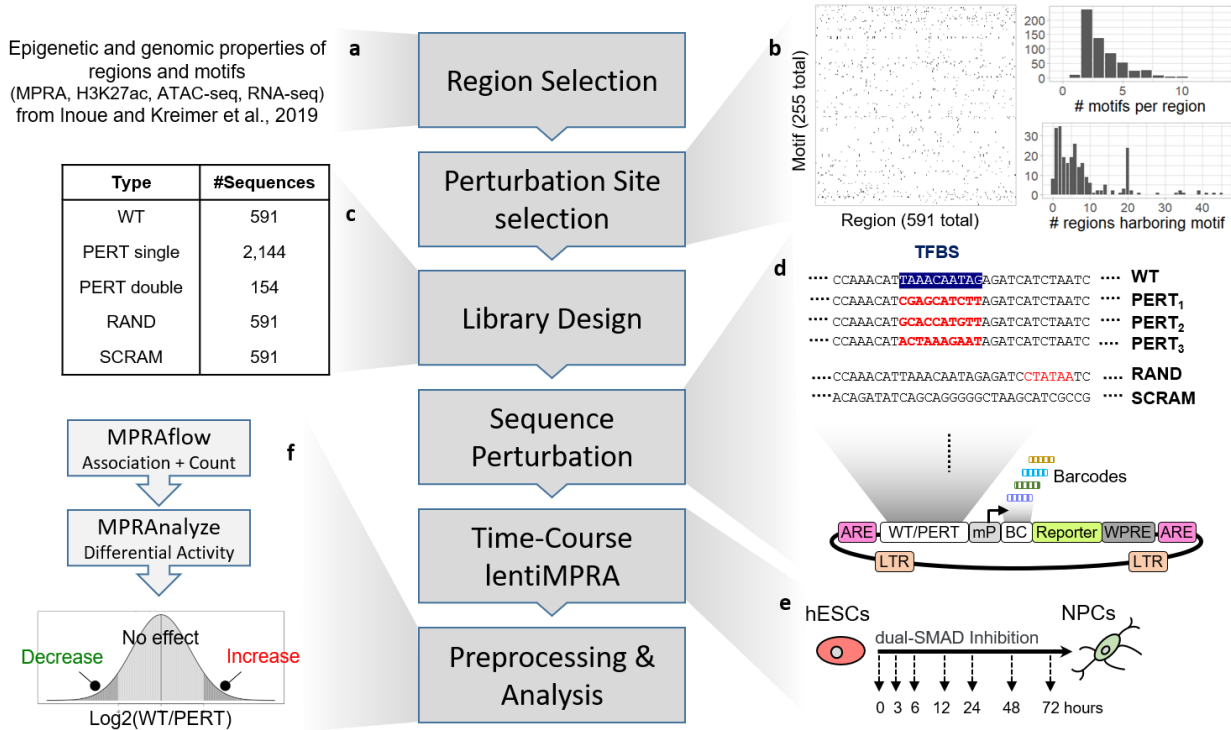

Fig. 1. Experimental design. **a**, Computational framework to select regions and perturbation sites. **b**, hitmap of motif instances in the assayed regions (left); distribution of the number of motifs perturbed in each region (top right); distribution of the number of regions harboring each motif (bottom right). **c**, Library design; selected regions were included in their wild type (WT) form, selected motifs were perturbed (by altering the sequence in the predicted motif site) using three perturbation methods individually (PERT single) as well as in combination with other perturbations in selected cases (PERT double). Random sites were perturbed (RAND) and the entire WT sequence was scrambled as negative controls (SCRAM) for each WT sequence. **d**, The designed sequences were synthesized and cloned into the lentiMPRA vector and associated with 15-bp barcodes. Abbreviations: ARE, antirepressor element. BC, barcode. Reporter, EGFP, enhanced green fluorescent protein. LTR, long terminal repeat. mP, minimal promoter. WPRE, Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element. **e**, lentiMPRA libraries were infected into hESCs and following three days, we induced neural differentiation via dual-SMAD inhibition and obtained DNA and RNA at seven time points (0, 3, 6, 12, 24, 48, and 72 hours). **f**, Association between barcodes and designed sequences, and the number of barcodes observed in DNA and RNA sequencing was determined using MPRAflow [34]. Differential analysis between WT and PERT activity to determine motif regulatory effect over time was assessed using MPRAnalyze [28].

We next set out to identify which of the DNA binding motifs we assayed is a functional site, i.e a site that causes a significant change in regulatory activity when perturbed. To this end, we initially focused on sequences with a single perturbed site (rather than deletion of two sites) and used

MPRAnalyze [28] to apply a set of four filters (illustrated in **Fig. 2a; Supplementary Table 1a**), requiring that each tested sequence passes all four filters: 1) the PERT sequence activity significantly deviates from that of the WT sequence in at least one time point (likelihood ratio test (LRT); FDR<0.05; **Methods**); 2) the time course of PERT activity significantly deviates from that of the WT sequence (LRT; FDR<0.05; **Methods**); 3) either the PERT or WT sequences are significantly more active than the SCRAM negative controls in all the time points (MAD-based z-test; FDR<0.05; **Methods**); 4) either the PERT or the WT sequence temporal activity significantly deviate from the temporal activity observed among the SCRAM negative control sequences (LRT; FDR<0.05; **Methods**). We applied these filters to each perturbation method separately, which resulted in 747, 775, and 749 sequences in perturbation methods 1, 2 and 3 respectively (**Fig. 2b; Supplementary Table 1a**). We observed an overall similar level of concordance between the different methods, with an average overlap of ~70% between the three methods (**Fig. 2b**; **Supplementary Fig. 6** and **Supplementary Table 1**).

In the subsequent analysis, we took a conservative approach to aggregate the evidence from the three ways of perturbing motif instances. We focused on instances that had strong evidence from both approaches for perturbing a motif (i.e., random shuffle or replacement by a fixed "non-motif" sequence). To this end, we consider only instances that passed all four filters above in perturbation method 3 and in either perturbation methods 1 or 2. We also require that the direction of effect (increasing or decreasing expression) to be consistent between the different methods. This resulted in 598 motif instances that had a significant and consistent effect. We refer to this set as *functional regulatory sites* (**FRS**s).

We examined the FRSs by conducting our analysis in three different axes: (i) the FRS level, i.e. perturbation of a specific motif in a specific region; (ii) the motif level, across different regions the motif appears in; (iii) the region level, taking into account the various functional sites that appear in it. For each axis we also examined how the perturbation effect may change over the different time points. While our analysis is based on the consensus set of 598 motif instances, we repeated it based on sites found by each of our three perturbation methods individually, where we observe largely consistent results (**Fig. 2b; Supplementary Tables 1-2**;**4-7; Methods**).

Delineating major categories of functional regulatory sites

We first analyzed the general effect of our perturbations in all 598 FRSs. Comparing the MPRA signal of WT to PERT sequences in each time point, we generally observed a reduction in activity, indicating that perturbing the predicted motif disrupts the function of an activating TF . For a smaller portion of the sequences we observed the opposite effect, i.e. increased activity, indicating that these sequences harbor binding sites with a repressive function (**Supplementary Fig. 7**). We thus divided the perturbation effects into two main categories (**Fig. 2c**): 1) *activators*, identified by perturbations resulting in reduced transcription (WT > PERT); 2) *repressors*, identified by perturbations resulting in increased transcription (PERT > WT). Notably, unlike transcriptional repressors that inhibit the transcription of their target gene, none of the repressor FRSs we identified lowered the barcode expression to a level below that of the negative control sequences (SCRAM). These repressor sites therefore attenuate the activating function of the enhancer region in which they reside, but not lower than baseline.

Out of the 598 FRSs, we observed 526 (87.9%) that had activating effects in at least one time point (and non-significant effects in the rest of the time points), and 70 (11.7%) that had repressing effects in at least one time point (and non-significant effects in the rest of the time points) (**Fig. 2d, Supplementary Tables 3-4**), with only two FRSs alternating between activating and repressing effects at different time points (DMRTA2 motif DMRTA2_M0629_1.02 and Interferon Regulatory Factor 4 motif IRF4_M5573_1.02; **Supplementary Tables 2-3**). This time invariance suggests that the direction of the effect (activating or repressing) of an FRS primarily depends on DNA sequence, and less so on the protein milieu or on other epigenetic properties that change during differentiation. Of note, as lentivirus randomly integrates into the genome, our results consider a cumulative signal from different integration locations in many cells, which essentially controls for the effects of local chromatin properties that may be present around the FRS.

To gain a better understanding of perturbation effects, we further divided our sites into four subcategories (**Fig. 2c-d**, **Methods**): 1) *Essential*: activating sites that when perturbed, reduced the expression level to that of the control (SCRAM) sequence ; 2) *Contributing*: activating sites that upon perturbation reduce the expression but not to baseline levels; 3) *Inhibiting*: sites that when perturbed lead to increased activity suggesting that they encompass repressive sites that fine-tune transcription levels; 4) *Silencing*: repressive sites that block a sequence from regulating transcription, i.e. WT levels are similar to control (SCRAM) and when perturbed make the sequence active. (**Fig. 2c**; **Methods**).
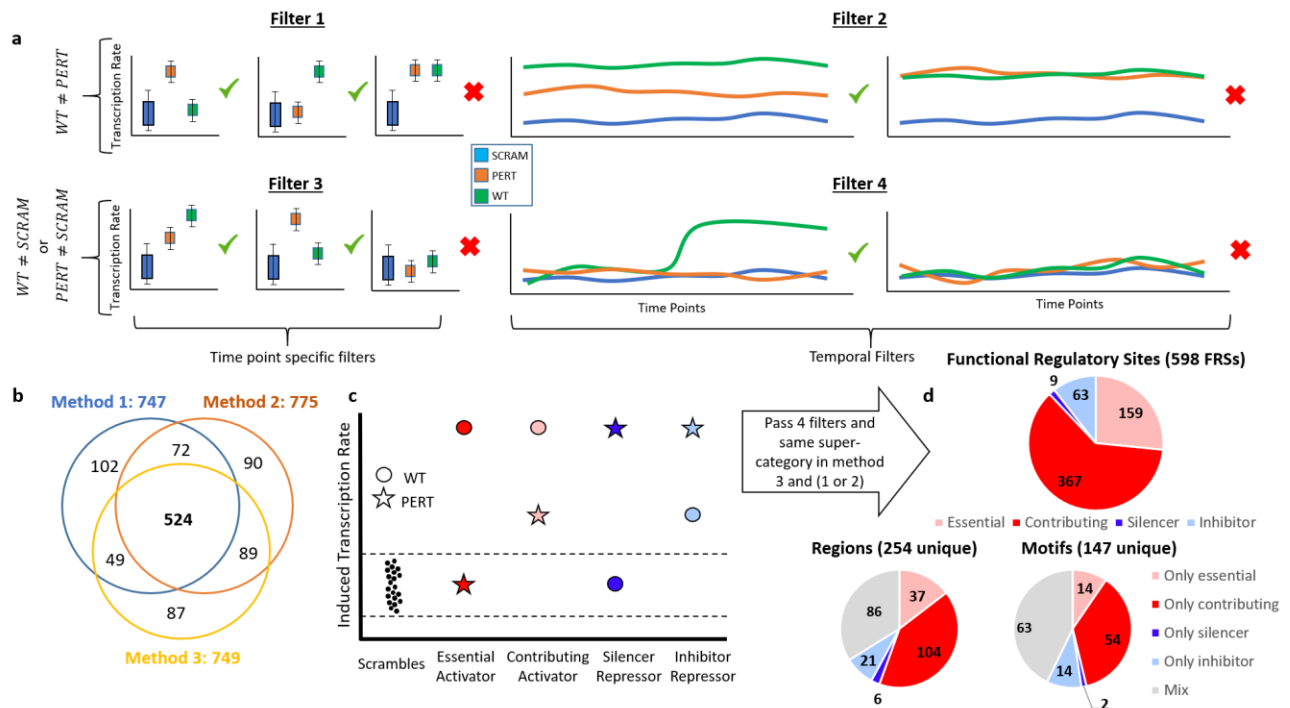

Fig. 2. Preprocessing, consistency, and categorization of FRSs. **a**, Four filters applied to perturbed sequences to remove inactive and non-functional sites, both at each time-point and across time-points (Methods). **b**, number of sequences that passed all three filters for each perturbation method. **c**, Definition of main and sub-categories of motif binding effects based on their effect on transcription. **d**, Distribution of

categories across FRSs that pass the 4 filters and are under the same main-category (activating or repressive) in perturbation method 3 and at least one of the perturbation methods 1 or 2. The distribution is shown across FRSs (top) and across the unique motifs and regions composing the FRSs in this study (bottom).

Considering this refined division, we found that 159 and 367 out of the 526 activating FRSs, correspond to categories *essential* and *contributing* respectively (**Fig. 2d, Supplementary Table 2**). Out of 72 repressor sequences we find 12 *silencers* and 60 *inhibitors* (**Fig. 2d; Supplementary Table 1**). These results represent the distribution of FRSs categories in our dataset.

We next examined how the strength of the mutation effects caused by perturbing activator sites (WT - PERT) depends on the strength of the expression generated from their respective unperturbed sequence (WT). We found that these effects scale linearly with the WT activity levels (*WT - PERT ~ a + b* WT)*. While this is trivial for essential FRSs, we found that this linear relationship still holds among contributing activators as well (median R-squared 0.95, methods, **Supplementary Fig. 8a-c**). This suggests that the fold change (WT/PERT) saturates for sufficiently-high levels of unperturbed (WT) expression, and reaches a constant level (**Methods**; **Supplementary Fig. 8d-e)**. Notably, we observed that different FRSs within a given region often have different constants, and the same motif has different constants when harbored in different regions, suggesting that this constant is not context- or factor-specific, but rather a combination of both.

Characterization of activating and repressive motif effects

Overall, our 598 FRSs include 147 unique motifs. Out of these 147, we observed 68 motifs that are strictly activators, 16 motifs that are strictly repressors and 63 motifs that show either activating or repressing effects in different genomic contexts (**Fig. 2d; Supplementary Tables 2-3**). We set out to examine the aforementioned subcategories of specific motifs. Within the activating FRSs, we observed that motifs associated with the SRY-Box Transcription Factor SOX1 are enriched in the set of *essential* FRSs (i.e over-representation that is unlikely to occur by chance; hypergeometric test, FDR<0.05). Both SOX1 and its homolog SOX2 are thought to function as pioneer factors that enable subsequent binding by other TFs [35]. This is in line with our observation that the enhancer activity is completely disrupted when these motifs are perturbed. Among the motifs that were enriched in the second category of having a contributing binding effect, we observed ZIC factors, which play important roles in neuroectoderm cell development [36].

Among the transcription factors whose motifs are associated primarily with a silencing effect is the Neuronal Differentiation factor NEUROD2. For example, perturbing a NEUROD2 binding site in a late-response regulatory element (chr15:75409661-75409832 (hg19); Fig 4a) increases the transcription induced by that sequence at the later time points (48-72hr) (**Fig. 3a**). While NEUROD2 is thought to be a transcriptional activator, our results accord with its previously reported role as a repressor of REELIN gene expression in primary cortical neurons by interacting

with CTCF that is known to function as transcriptional repressor in a context dependent manner [37].

Considering the set of *Inhibitor* motifs, which could fine-tune regulatory activity by partially reducing it, we saw enrichment for the P53-Like Transcription Factor TP73. For example, perturbing a TP73 binding site in region chr6:167854597-167854768 (hg19) substantially increases the activity of that enhancer across all time points. Notably, this region also contains two functional binding sites that activate transcription, and harbor NANOG (NANOG_disc2) and SOX1 (SOX1_M3910_1.02) binding motifs (**Fig. 3b**). Interestingly, we also found six instances where TP73 binding motifs function as activators (**Fig. 3c, Supplementary Table 1**). TP73 has been shown to regulate NPC proliferation in the developing and adult mouse central nervous system [38,39] and is known to interact via its subdomains with many different partner proteins, including POU [40] which has corresponding motifs in this region and YAP1, which is known to function as both an activator or repressor [41] in a context dependent manner [42]. These instances demonstrate that FRSs can achieve their desired transcriptional rate by combining both activating and repressive motif sites, and that using our perturbation MPRA approach allowed us to distinguish the functionality in each specific context.

When examining the distribution of sub-categories effects across motifs, we observed 84 (57%) motifs that appear in only one subcategory and 63 (43%) motifs with mixed effects (**Fig. 2d; Supplementary Table 1**). These results indicate that enhancer activity is influenced both by the motif sequence and the surrounding sequence of the region harboring the motif.
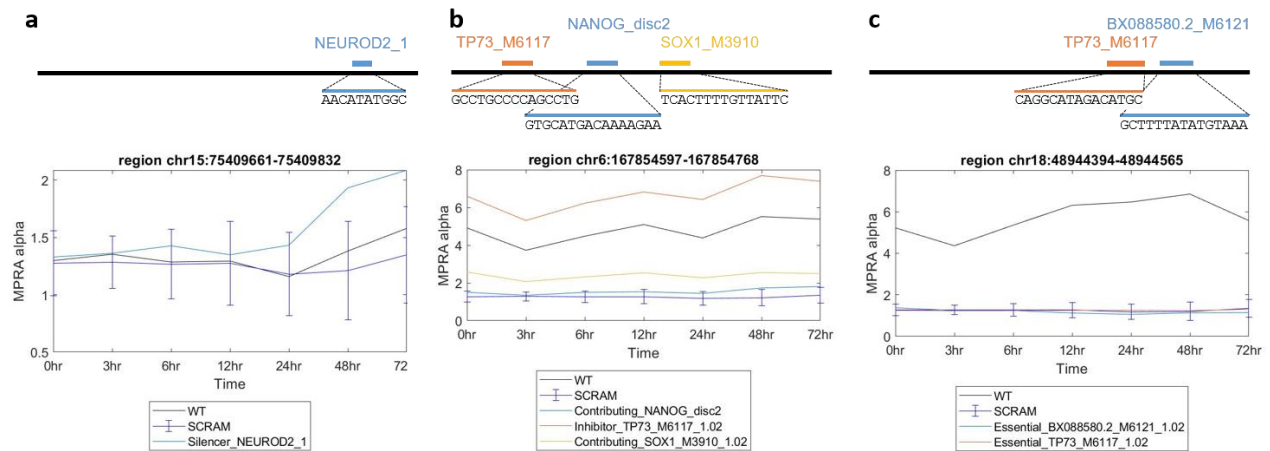

Fig. 3 Examples of the 4 subcategories of FRSs. All genomic coordinates are hg19.

Focusing on motifs that are consistently associated primarily with one direction of effect (activating or repressing), we next set out to analyze the effects of motifs on transcription during our time course, by aggregating the results from all their respective instances. We summarized the signals of motifs that show activating or repressing cumulative effect (**Fig. 4**). Among the TFs associated with activator motifs, we observe the neural markers SOX, LHX, ZIC, and FOX families (**Fig. 4a**) [35][364326][26,44–50], as well as motifs associated with factors known to be involved in neural induction,

such as OTX2 [51,52] and PAX6 [46,53]. Consistent with our recent characterization of neural induction associated TFs [26], we also identified Iroquois Homeobox Protein 3 (IRX3) to be one of the strongest activating motifs. Among the TFs associated with repressive activity (**Fig. 4b**), we observed factors from the HOXD gene family, which are thought to function as repressors when bound in monomeric form [54]. We also found an enrichment for a SIN3A motif, which is generally known to interact with histone deacetylase (HDAC) and function as a transcriptional co-repressor [55]. It was also reported that the SIN3A/HDAC co-repressor complex was involved in the maintenance of ESC pluripotency [55,56].
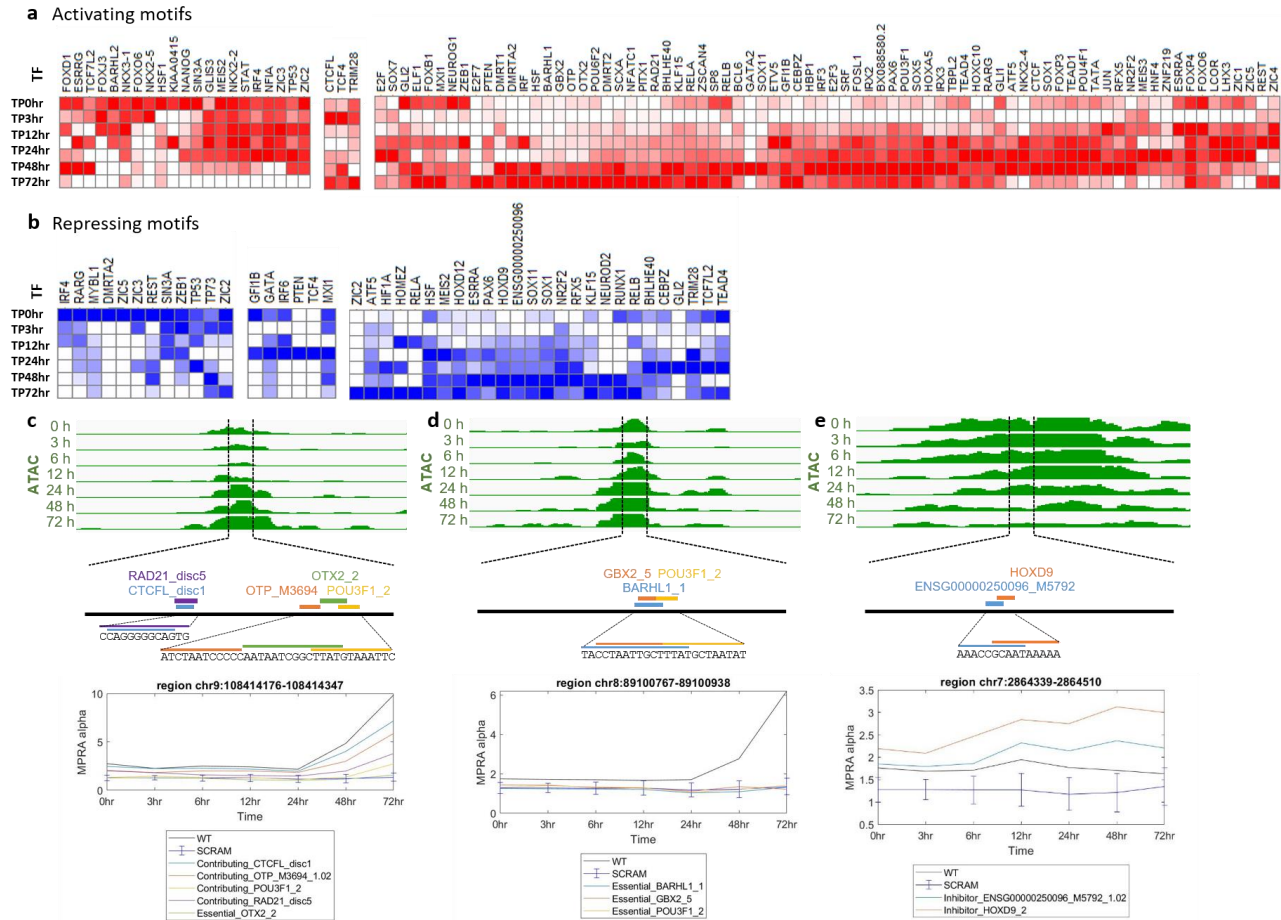

Fig. 4. Temporal motif effects. The **a**. activating or **b**. repressing motifs in at least one time point. red - activator motifs, blue - repressor motifs. Color scale indicates the average of the perturbation signal (LogFC(WT/PERT)) across all significant instances of motifs for a specific TF (row normalized). Data is organized using hierarchical clustering. Genome browser snapshots of assaysed sequences near predicted sequence motifs that are associated with **c.** *OTX2 and* **d.** *BARHL1* TFs, showing the motifs that were perturbed and their effect on activity across time points. **e.** HOXD9 repressor motif example. All coordinates are hg19.

To examine how the effects of motifs change over time, we clustered the signal of all activating and repressing motifs. We observed that the magnitude of effects often changes over time in a

manner proportional to the unperturbed expression level (**Supplementary Fig. 8)**. These effects range from perturbations that are effective only at the ESC stage to those that influence late induced regions (**Fig. 4**). For example, enhancers that have OTX2 binding sites reach their peak activity during the neural progenitor cell (NPC) stage. When the OTX2 sites are perturbed, the activity at later time points (48-72hr) was decreased (**Figs. 4a,c**). Similarly, NPC enhancers harboring IRX2/3 (**Fig. 4a**) or BARHL1 (**Fig. 4d**) motifs decreased in activity when the binding sequences were mutated. Correspondingly, we observe that OTX2, IRX2/3 and BARHL1 mRNA levels peak at later time points (48-72hr) (based on data published in [26]). When HOXD sites (HOXD12_M5560_1.02, HOXD9_2) were mutated, the activity at later time points (48-72hr) (**Fig. 4b,e; Supplementary Tables 2-3**) was increased. These findings indicate that these binding sites have different levels of induced activity at distinct time points of neural differentiation. This suggests that the abundance of the binding TF (i.e. the TF's mRNA levels) at a given time point and the abundance of additional cell-state specific factors (e.g. expression of other TFs) play a significant role in proper enhancer activity.

Interestingly, we also observed TFs whose corresponding motifs show both activating and repressing effects in different regions (**Fig. 4a,b**; **Supplementary Table 3**). For example, different motifs for the Zinc finger protein (ZIC) family have repressing and activating effects across different regions (ZIC2 and ZIC3). Members of the ZIC family are involved in neurogenesis and are known to function as both transcriptional activators and repressors in a context-dependent manner during embryogenesis [57]. Additionally, we observed both effects for the ZEB1 motif in different regions (**Fig. 4a,b**; **Supplementary Table 3**) in concordance with the role of ZEB1, acting as both a transcriptional activator and repressor during neurogenesis [58,59]. We saw similar effects for the RARG motif (**Fig. 4a,b**; **Supplementary Table 3**). RARG is a retinoic acid receptor (RAR), a family of factors that plays a role in developmental processes and acts as a ligand-dependent transcriptional regulator. When bound to ligands, RARs activate transcription, whereas in their unbound form they repress transcription of their target genes [60].

Characterization of genomic regions harboring sequence motifs

We next examined the activity of the assayed regions as composite functional units consisting of multiple FRSs. Our 598 FRSs include 254 unique genomic regions. We observe complexity in these regions in terms of having sites with different direction of effect and different sub-categorization. Specifically, when examining the set of significant perturbation effects in those regions, we observed 141 cases with only activating effects (**Fig. 2d**; **Supplementary Table 2**), which is consistent with our analysis being focused on regions that were previously identified as enhancers during neural induction [26]. We found 86 regions (>30%) that harbor both activating and repressing motif instances. This suggests that regulatory activity within these enhancers can be achieved by fine tuning of binding effects, including both activating and repressing motifs to achieve the desired regulatory function. This phenomena of context-dependent repression by transcriptional activators is consistent with what was previously reported in yeast [21], drosophila [61] and mammalian cells [62].

Specifically, when examining the distribution of sub-categories of effects across regions, out of these 254 regions, we observed 168 (66%) regions with only *essential*, *contributing*, *silencing* or *inhibiting* effects respectively and 86 (34%) regions with a mix of effects (**Fig. 2d; Supplementary Table 2**). Regions with multiple *essential* FRSs, all required for regulatory activity, offer support to the 'enhanceosome model' of a specific combination of factors being required in an all-or-nothing machinery [15]. In contrast, regions with multiple *contributing* FRSs are evidence of the 'billboard model', of a flexible modular machinery that fine tunes the induced transcription levels by having independently contributing factors [19]. These results demonstrate that different regulatory sequences may be governed by either the enhanceosome or the billboard model, and some appear to be governed by a combination of both.

Perturbation of motif pairs identifies different modes of motif interaction

We wanted to further examine how pairs of motifs interact in regulatory sequences. To that end, we examined the results of perturbing pairs of motifs, both individually and in combination, to determine how different binding sites interact in a single FRS (**Fig. 5a**). We conservatively considered only pairs of sites for which both single-site perturbations and the joint perturbation belong to the same main category and the double perturbation is functional (i.e passed all four filters as previously described) in both perturbation approaches (perturbation method 3 and either of methods 1 or 2; **Methods**). This approach resulted in 22 pairs of motifs, which we refer to as *interaction* FRSs (IFRSs) (**Fig. 5b**). Amongst these IFRSs, 21 led to a reduction in activity, indicating an overall activating effect, and one IFRS had a repressing effect, leading to increased activity when both motifs were perturbed (**Supplementary Table 8**).
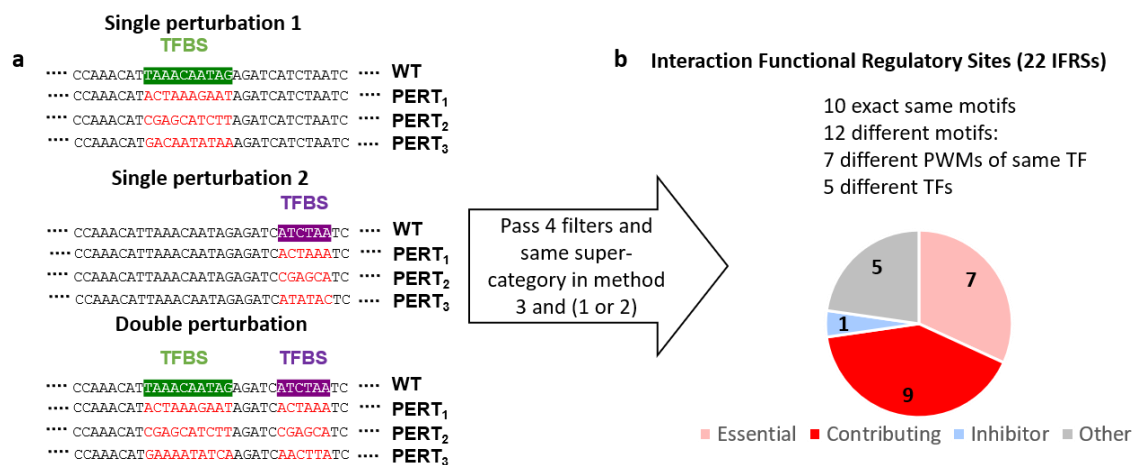

Fig. 5 Double perturbation scheme. **a.** Experimental design for perturbing two single motifs separately and then a double perturbation of both simultaneously. **b.** Category distribution of IFRSs: Essential (pink) - if one of the single perturbations is essential, then the double perturbation is also essential; Contributing (red) - If the two single perturbations are contributing then the outcome is contributing; Inhibitor (light blue) If the two single perturbations are inhibiting, then the outcome is inhibiting; Other: the 5 remaining IFRSs are categorized as follows: 4 [contributing, essential, contributing], 1 [contributing, inhibiting, contributing].

We found that our double perturbations provide support for both regulatory models: billboard and enhanceosome. For instance, in chr7:76302788-76302959 (hg19) we observed a SP8 (SP8_1) and a POU3F1 (POU3F1_2) motif instance which both reduce enhancer activity upon perturbation. Perturbing both motifs reduces the activity further than either single perturbation, consistent with the flexible billboard model of independent contribution by different factors (**Fig. 6a**). Conversely, we observed cases where the different elements act in a fully-dependent manner, consistent with an all-or-nothing enhacesome model. For example, chr9:108414176-108414347 (hg19) contains two essential functional sites: an OTX2 (OTX2_2) and a POU3F1 (POU3F1_2) motif, both necessary for activity. Perturbing either one, and concordingly both, reduces induced transcription to SCRAM levels (**Fig. 6b**). Both factors are known to have a key role in determining neural fate [63].

Interestingly, our results also exemplify a hybrid model [27], where in a single regulatory region and at the same time, some factors are strictly required (as in an enhanceosome) while others are independently contributing (as in a billboard). For instance, an FRS residing in chr14:65801655-65801826 (hg19) is active throughout differentiation and its activity increases at later time points. This region contains two functional binding sites: an essential SOX1 motif (SOX1_M6134_1.02) that reduces transcription down to SCRAM levels when perturbed in all time points, regardless of the other functional site; and a contributing FOXB1 motif (FOXB1_4) that has no effect when perturbed in the early time points but prevents the increased activity in later time points, consistent with FOXB1's temporal expression at 24-72h [26] and its known temporal role in NPCs [43]. Perturbing both sites results in an essential effect, reducing induced transcription to SCRAM levels (**Fig. 6c**). In another case, the regulatory region, chr10:100206539-100206710 (hg19), residing in an intron of the *HPS1* gene, contains two FRSs each containing a motif instance of ELF1 (ELF1_known3), a transcription factor known for its binding near prefrontal cortex splicing QTL SNPs [64] and for its role in brain development [65]. Both FRSs are activators, but do not have an identical effect: with one driving down transcription to SCRAM levels when perturbed (*essential*), and the other having a milder effect (*contributing*). Perturbing both FRSs in this region further reduces the expression to levels significantly below the SCRAM baseline (**Fig. 6d**).

Our results also include activator and repressor FRSs interacting to achieve a specific transcription level. For example, chr4:152405951-152406122 (hg19), an intronic region in the FAM106A1 gene body, contains an FRS with a SOX1 motif that has a contributing effect and an FRS with a ZIC2 motif that has an inhibiting effect. Perturbing both sites results in a cumulative effect: transcription levels that are lower than WT, but higher than those obtained when perturbing the SOX1 motif alone (**Fig. 6e**). Finally, our results include anecdotal evidence of more complex modes of dependency between FRSs. For example, the regulatory region chr1:27933863-27934034 (hg19), downstream of the promoter of the *AHDC1* gene, which has been implicated in the neurological disorder Xia-Gibbs syndrome (OMIM 615829). This region contains two sites of the NR2F2 binding motif (NR2F2_M5676_1.02), a factor that is known to have a crucial role in the activation of neural genes during early differentiation in humans[66]. The perturbation of one NR2F2 motif does not have a significant effect on the transcription rate, individually perturbing the second motif has a modest effect, but perturbation of both sites causes a substantial reduction in activity (**Fig. 6f**), suggesting the FRSs are partially redundant, and their cumulative effect is not strictly additive. Overall our results demonstrate the power and potential of perturbation MPRA in

uncovering a variety of different patterns of interaction, and elucidating the complex regulatory grammar governing these behaviors.
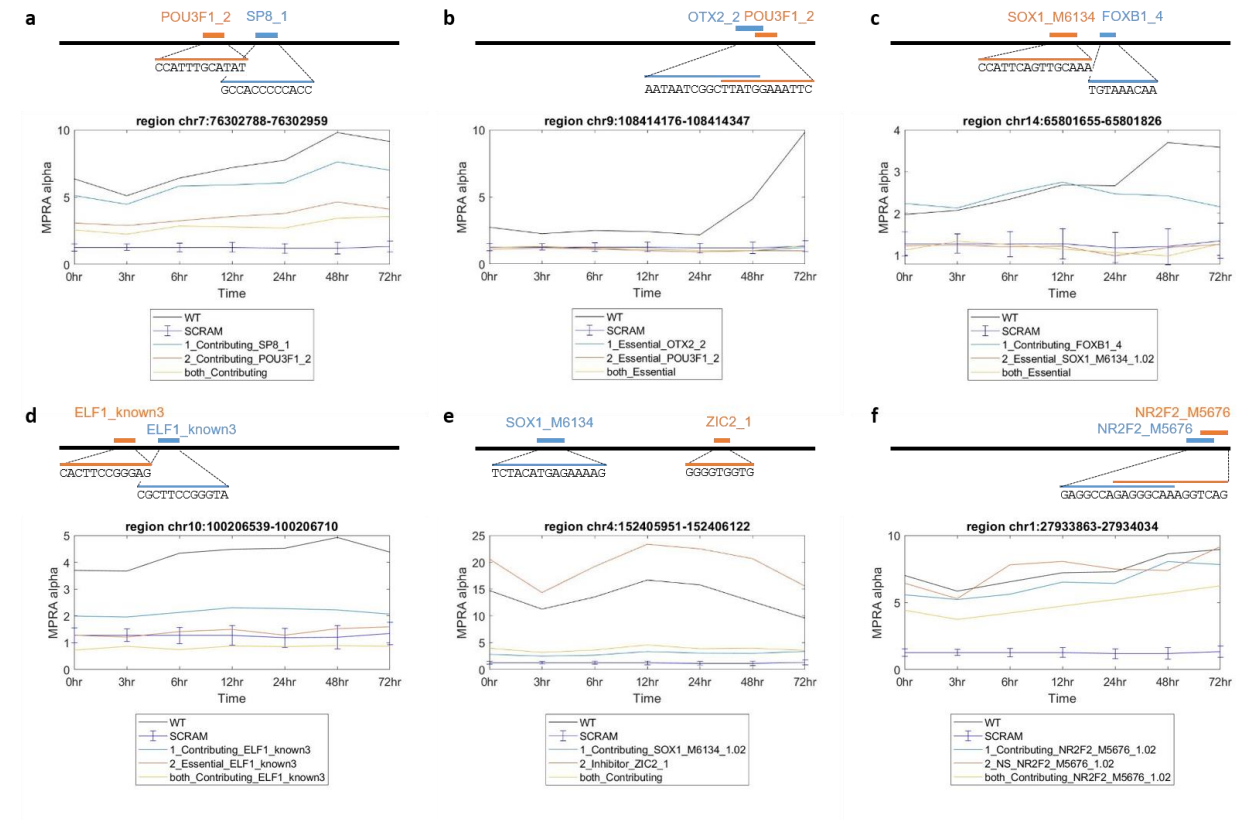

Figure 6. Cooperation models and interaction examples (different and same motifs). **a.** Billboard model. **b.** Enhanceosome model. **c.** Hybrid model. **d.** Same motifs example ELF1. **d.** NR2F2 motifs example. **e.** Different motifs example - active and repressive motifs in the same enhancer. **f.** Same motifs example NR2F2. All genomic coordinates are hg19.

Discussion

Regulatory elements play a major role in cell-type specific response to environmental conditions and perturbations. Teasing out the regulatory rules and sequences responsible for these responses could lead to a better understanding of how variations in these sequences alter their activity, and allow the accurate design or targeting of specific sequences for therapeutic purposes. Here, we used perturbation MPRA across seven time points of neural differentiation to characterize the regulatory grammar during early stages of neural induction. Our work allowed us to evaluate the effect of intact motif instances over time and annotate these instances into four major categories (*essential*, *contributing*, *inhibiting* or *silencing*). We observe that generally a FRS either has an activating or repressive effect across all time points, suggesting that the binding

motif and surrounding region largely determine the direction of effect, and that the magnitude of this effect changes over time, in a manner proportional to the activity of the WT sequence, in different cellular environments, indicating earlier and later functional motifs in this process. Finally, by carrying out two motif perturbations in a single sequence, we observed different modes of interaction between pairs of motifs.

Previous approaches to perturbation MPRA for sequence motifs were limited to several factors and a specific cell-type or condition. For example, a previous study [9,25] explored the activity of five activator motifs and two repressor motifs in K562 and HepG2 cells by introducing different variations to the motif sequence and another study [25] disrupted a single motif (PPARγ) in mouse adipocytes. Here, we analyzed the effect of over 250 motifs with three different perturbations using two approaches. In the first approach, we replaced the motif with two different 'non-motif' sequences and in the second approach, we scrambled the motif's nucleotides. All these perturbations showed high reproducibility between replicates (r>0.95). Analyzing and comparing the three perturbation methods, we observed a similar level of overlap between the different methods, but we do not observe more consistency between perturbation methods 1 and 2 than either one is with perturbation method 3 (**Fig. 2b,2d; Supplementary Fig. 6; Supplementary Tables 1;4-7**). This may indicate that at least one of the fixed-sequence perturbation methods potentially introduces bias that separates it from the other, e.g by forming de novo binding sites with endogenous sequences adjacent to the perturbed sites. Since methods 1 and 2 insert a fixed sequence, this introduced bias could be systemic across the assayed regions and skew downstream results. For future experimental designs, we suggest using a more robust perturbation approach which randomly shuffles the nucleotides of the perturbed site and is less likely to introduce systemic biases.

We cataloged the function of 598 FRSs representing 254 unique endogenous regions and 147 unique motifs. Approximately 90% of FRSs act as activators with ~30% of them as essential and the rest as contributors. This finding is also in line with a saturation based MPRA that analyzed ten disease-associated promoters and enhancers, finding that the majority of mutations lead to a reduction in activity (i.e. act as activators that when mutated reduce activity) [67]. Additionally, while our data does not contain FRSs that repress transcription below the baseline rate, we found many instances of binding sites that have a repressive effect on the function of the enhancer itself: reducing the level of induced transcription, or even completely blocking the enhancer's activity. These instances suggest that enhancers can be kept in a pseudo-poised state: residing in open chromatin but being blocked from activity by TF binding, and that repressive factors are often bound to functional enhancers as a mechanism for fine-tuning transcription levels.

Finally, a smaller subset of sequences were perturbed in two locations, where we perturbed two single motifs separately and jointly to assess their interaction, as a proof of concept. We found that interactions involving at least one essential motif mostly result in a double perturbation that is *essential*, and interactions involving contributing motifs mostly result in a double perturbation that is *contributing* (**Fig. 5b**). For some of the cases we observe additive effects as previously described in the literature [25]. By perturbing two motifs in over 100 sequences, we were able to characterize cooperativity between two factors (**Fig. 6**). We observed two major models of TF cooperation to be supported by our data: 1) we identified FRSs that contain multiple essential

binding sites, consistent with the all-of-nothing 'enhanceosome model' in which a strict composition of TFs are required for an enhancer's function; 2) we also found FRSs that contain several seemingly independent contributing binding sites, consistent with the flexible 'billboard model'. Notably, for FRSs containing two instances of the same motif, the single perturbations did not have identical effects, consistent with the growing body of work showing that the function of an enhancer depends on the specific locations and distances between binding sites, and not only of their presence [19,21–24].

Examining whether we can gain a better understanding on the determinants of time-point specific regulatory activity using this model system, revealed complex results, suggesting that motif sequence alone is less likely to determine temporality without the context of the surrounding region and other bound factors (**Supplementary Note 1, Supplementary Fig. 9, Supplementary Table 10**). Therefore, future challenges following our work will include developing strategies to further understand regulatory logic and its determinants across different conditions. For example, using endogenous manipulations via CRISPR to examine the function of specific motifs and their combinations across different cellular conditions.

During early neural induction, pluripotency-associated genes are rapidly downregulated and neural associated genes are induced by a variety of factors [33,46]. As such, the rapid differentiation of hESCs into neural cells provides an exceptional model to study motif effects and how they change across developmental time points. Using this model, we previously interrogated [26] the temporal dynamics of gene expression (RNA-seq) and gene regulation (ATAC-seq, H3K27ac and H3K27me3 ChIP-seq and LentiMPRA) at seven time points during early neural differentiation. Our current work further validated the novel motifs and TFs identified in our previous report to have temporal effects across neural induction. For example, we find that FRSs harboring BARHL1 and IRX3 motifs exhibit time point specific activating effects and show changes in magnitude over time, with higher signal at the NPC state - supporting their suggested role in neural induction.

Overall, our results provide an atlas of motif function across early time points of neural differentiation by directly testing hundreds of regulatory regions for the function of the motifs they harbor. To the best of our knowledge this provides the first comprehensive perturbation MPRA study across a developmental time course, showing clear changes in regulatory activity over time. This system provides a model for how perturbation MPRA can be leveraged to identify and characterize in a high throughput manner the functional effects of regulatory sequences across different cellular conditions/perturbations.

Methods

Computational analysis

Perturbation MPRA library design

<u>Choosing region and motif combinations</u>

General description. Our previous analysis [26] points to a large number of regulatory regions of interest as well as multiple motif hits within those regions. Our goal is to select the most informative set of [region x motif] combinations (each corresponding to a motif instance) so as to fit within a single MPRA design. To address this, we developed a selection scheme to represent various biological aspects of our system and account for experimental limitations for the number of assyed sequences.

To do this, we formalize the information that we have about the motifs and regions as a tripartite graph, with one layer of nodes corresponding to DNA regions, another layer of nodes that represent motifs and a third layer of nodes, each representing a different property of motifs or regions (**Supplementary Fig. 10**). The region layer consists of the 1,547 genomic regions we identified in our previous work [26] that show temporal activity when tested using lentiMPRA in the same seven time points. The motif layer consists of motif hits found in those regions computationally (using Fimo (p-value<10^{-5}; Grant et al., 2011) with two sets of TF motifs [30,31]). Edges between the first two layers connect every motif with the regions in which it occurs. Each node in the third layer corresponds to a property of interest which characterizes a subset of the motifs and regions that are represented in the first two layers. These properties are based on genomics assays from our previous work [26] (based on ATAC-seq, H3K27ac and H3K27me3 ChIP-seq and RNA-seq data from these 7 time points). For instance, we identified several temporal patterns associated with each data modality and designate each of these patterns as a node (e.g., a node for "regions that have a transient peak in H3K27ac 48 hour post induction"). We then connect a region to a node if that respective pattern is observed in that region in the endogenous genome. Similarly we connect a motif node to a property node. For example, a node for "motifs with associated TF that is expressed 24 hour post induction"). We then connect a motif to a node if that respective pattern is observed for that motif in the endogenous genome. We describe the "property layer" and its edges with the "motif" and "region" layers in greater detail below.

Altogether our graph now has 1,547 region nodes, 4,393 motif nodes and 68 property nodes. These nodes are connected by a total of 99,165 edges. Our goal now becomes to find the minimum number of [region x motif] combinations (each representing a specific motif instance, or - equivalently- an edge in our graph) that will guarantee a sufficient coverage of each property. In other words, we want to select a minimal number of motif-region pairs such that every "property node" in our third layer is connected by an edge to a sufficient number of motifs and regions (as detailed below). Having staged our data in a tripartite graph helps us re-state our goal as a constrained optimization problem - guaranteeing minimal level of connectivity for the third layer, while minimizing the number of selected nodes and edges in the first two layers. Since this problem is NP- hard, we followed the common practice and formulated it as an integer linear program (ILP), which can be solved efficiently through a range of heuristic with available solvers. With this ILP, we were able to select 591 regulatory regions and 255 motifs that are organized into 2,144 region-motif pairs. Below, we provide a more in-depth description of this process.

Defining the property layer: We composed a list of biological properties based on published literature and on ATAC-seq, H3K27ac and H3K27me3 ChIP-seq and RNA-seq data we produced and analyzed in our previous paper [26]. The biological properties of TFs associated with motifs and regions include: (i)TF/region is induced/active at a specific time point. (ii) TF/region binds/belongs to significantly overlapping sub-clusters (as defined in [26]) of temporal MPRA and H3K27ac/ATAC-seq/RNA-seq signals (iii) The TF/the proximal gene for the region is a known neural factor or belongs to one of the pathways defined below. Known neural factors: POU3F1, MYT1L, SOX2, POU3F2, LHX2, PAX6, ASCL1, SOX1, OTX2, ZNF521, NEUROG1, NEUROG2, NEUROG3, NEUROD1, NEUROD2. Pathways taken from KEGG [68]: FGF/MAPK signaling pathway hsa04010, IGF-1/mTOR signaling pathway hsa04150, Wnt/Ca+/PCP signaling pathway hsa04310, Sonic Hedgehog signaling pathway hsa04340. (iv) Hand picked TFs (POU3F1, POU3F2, SOX2, SOX1, PAX6, OTX2, LHX2, NEUROG1, NEUROG2, NEUROD2, SP8, IRX3, SOX10, PKNOX2, HHEX, LMX1A, BARHL1, LHX5, NR2F2, DMBX1, MEIS2, OTX1, SOX21, FOXB1, SOX5, MEIS3, HOMEZ, TCF3, TCF4, ZIC1, ZIC2, ZIC3, ZIC4, ZIC5), including factors known to have a role in neural differentiation based on previous literature [26,44–50], or based on their expression in neuroectoderm in mouse embryo, or show high 'TF activity score' in the relevant time points in our data [26]. The direct edges from motifs and regions to properties, represent the biological properties a region or a motif satisfies as described above.

The optimization program:

Minimize $\left(\sum_{r \in R} \boldsymbol{\theta}_r\right) + 3 * \left(\sum_{(t,r) \in E;\, t \in T;\, r \in R} e_{t,r}\right)$.

Subject to:

1. $\sum_{(t,p\prime) \in E;\, t \in T} \boldsymbol{\theta}_t \geq 12$ $\qquad\qquad \forall p\prime \in P$

2. $\sum_{(r,p\prime) \in E;\, r \in R} \boldsymbol{\theta}_r \geq min\{17, deg_R(p\prime)\}$ $\qquad \forall p\prime \in P$

3. $\sum_{(t,r\prime) \in E;\, t \in T} \boldsymbol{\theta}_t \geq \boldsymbol{\theta}_{r\prime} min\{3, deg_T(r\prime)\}$ $\qquad \forall r\prime \in R$

4. $\sum_{(t\prime,r) \in E;\, r \in R} \boldsymbol{\theta}_r \geq \boldsymbol{\theta}_{t\prime} min\{20, deg_R(t\prime)\}$ $\qquad \forall t\prime \in T$

5. $e_{t,r} \geq \boldsymbol{\theta}_t + \boldsymbol{\theta}_r - 1$ $\qquad\qquad \forall(t,r) \in E;\, t \in T;\, r \in R$

6. $\sum_{t \in T_i} \boldsymbol{\theta}_t \leq 2$ $\qquad\qquad \forall T_i$

7. $\sum_{t \in T_i} \boldsymbol{\theta}_t \geq 1$ $\qquad\qquad \forall T_i \in Hand\ Picked$

8. $\sum_{r \in R} \boldsymbol{\theta}_r \geq 0.4 * |R|$

9. $\sum_{(t,r) \in E;\, t \in T;\, r \in R} e_{t,r} \geq 5 * \sum_{(t,r) \in E_p;\, t \in T;\, r \in R} e_{t,r}$

10. $\sum_{t \in T_B} \boldsymbol{\theta}_t \geq 1.5 * \sum_{t \in T_S} \boldsymbol{\theta}_t$

11. $\boldsymbol{\theta}_t, \boldsymbol{\theta}_r, e_{t,r} \in \{0,1\}$

The decision variables represent the following: $\boldsymbol{\theta}_t$ is a binary variable that indicates whether we chose the motif t; $\boldsymbol{\theta}_r$ is a binary variable that represents whether the region r was selected. $e_{t,r}$ is a binary variable that denotes whether a motif x region pair (t and r) has been selected.

Parameters include:

P - represent the properties.

R - represent the regions.

T - represents the motifs.

deg_R(p) - represents the number of edges connecting property p to regions.

deg_R(t) - represents the number of edges connecting motif t to regions.

deg_T(r) - represents the number of edges connecting region r to motifs.

T_i is a subset of *T* that contains all the motifs corresponding to TF i.

E_p as a subset of the edges with lower confidence (i.e. edges that connect to properties representing non significantly overlapping sub-clusters of temporal MPRA and H3K27ac/ATAC-seq/RNA-seq signals),

We define T_B as the subset of motifs connected to at least 5 regions, and T_S as the subset of motifs connected to fewer than 5 regions.

Constraints: The constraints described in the equations above ensure that: (1) Each property is connected to at least 12 motifs. (2) Each property is connected to at least 17 regions (or all regions if it's below 17). (3) Each region is connected to at least 3 motifs. (4) Each motif is connected to at least 20 regions. (5) An edge is active if both nodes of the edge are active. (6) For each TF, no more than two motifs are chosen. (7) All hand picked TFs are used at least once. (8) At least 40% of all regions are used. (9) At most 1/6 of the total edges used are low confidence edges. (10) At least 60% of motifs chosen are motifs connected with many regions (T_B), s.t. the solver does not bias towards lowly connected motifs. (11) All variables are binary.

For each $T_i \in$ *Hand picked* - one representative motif must be in the solution.

Our objective is to **minimize** the overall number of MPRA sequences to design. It is a sum that accounts for corresponding to the number of unperturbed (WT) regions plus the number of perturbations (i,e, regions and motif combinations). We multiply by 3 since we have 3 perturbation methods (i.e., we need 3 MPRA sequences for every pair).

<u>Different categories of sequences designed on the array</u>

Overall, the solver picked 591 regions, 255 unique motifs which correspond to 166 unique TFs. We used the combinations of region and motifs chosen by the solver to represent the following sequence categories on the array (**Supplementary file 1**):

1) One motif is perturbed in the sequence. For combinations of regions and motifs where the motif is detected once in the sequence (**hit1** N=1620).
2) Two motifs of the same motif are perturbed in the sequence. For combinations of regions and motif s where the motif is detected twice in the sequence: if the +/- strand carry exactly the same motif we only replace the motif one time in the + strand (**hit2** N=62), otherwise (**hit2diff** N=90) we perturbed each motif separately and then both of them - starting with the + strand. If 3 or more hits of the same motif are observed – we discard those region-motif combinations (N=52).

Additionally to the combinations picked by the solver, we considered the 591 WT regions and added more combinations (not chosen by the solver) that contain motifs of the following 11 TFs. These TFs were chosen (LHX5, MEIS2, PAX6, FOXB1, SOX1, IRX3, OTX2, ZIC2, SP8, POU3F1, HOMEZ) based on their high 'TF activity score' in the relevant time points in our data [26] and their mRNA expression in neuroectoderm in mouse embryo.

3) One motif is perturbed in the sequence. For combinations of region and motif where the motif is detected once in the sequence (**Overexpressed_hit1** N=221 and **Overexpressed_permutation** N=58).
4) Two motifs of the same motif are perturbed in the sequence. For combinations of regions and motifs where the motif is detected twice in the sequence: if the +/- strand carry exactly the same motif we only replace the motif one time in the + strand (**Overexpressed_hit2** N=3), otherwise (**Overexpressed_hit2diff** N=1) we perturbed each motif separately and then both of them - starting with the + strand.
5) Combinations of two or more motifs are perturbed in the sequence. For combinations of regions and motifs where we observe two or more different motifs in the sequence (**Overexpressed_permutation** N=125). We examined combinations of motif hits of these 11 TFs in our regions.

Overall, most of the data includes a single motif perturbation per region (N=2,144) and a smaller part with two or more motif perturbations per region (N=216 out of those: N=154 two motifs; N=62 more than two motifs) comprising a total of 2,360 designed region and motif sequences.

We also assayed WT and control sequences:

1) We assayed 591 **WT** sequences. WT sequences are the endogenous 171bp sequences.
2) We assayed 591 scrambled sequences (**SCRAM**). Scrambled sequences are based on WT sequences with shuffled nucleotides, creating a set of negative controls.

3) We assayed 591 sequences with random alterations (**RAND**) – where we randomly chose a location in the region and perturbed the median motif size (12bp) starting in that location, creating an additional set of negative controls.

We perturbed predicted motifs within each genomic region (2,360 combinations) using three perturbation approaches: the first two replace the predicted binding site with a "non-motif" sequence whereas the third one shuffles the nucleotides of the predicted binding site described in the next section. For the **RAND** sequence category, we used the same 3 perturbation approaches.

Different motif scrambling (perturbation) approaches

Approach 1 - create "non-motif" sequences following these steps:

1) Use all the 2,464 MPRA sequences we designed in our previous work [26] based on their potential to be active during neural differentiation.
2) Count #di-nucleotides and calculate their percentage of appearance in those sequences.
3) Create a di-nucleotide scrambled sequence in the length of the maximal motif, i.e. "scrambled motif".
4) Create 1,000 maximal length "scrambled motifs".
5) Run these 1,000 "scrambled motifs" through Fimo [29] with the two sets of TF motifs [30,31] and choose the ones with the lowest number of motif hits (p-val<10^{-4}) - 13 "scarmbled motifs" had 0 hits.
6) In each chosen combination of region and motif (described in the previous section) - replace the motif appearance with the prefix of the "scrambled motif" (adjusting to each motif length) using these two strategies that avoid motifs creation in the edges of the sequences: (1) use 3bp downstream and upstream of the motif in the original sequence (2) use the original sequence. Repeat this 13 times using each one of the "scrambled motifs".
7) Run the sequences created using the two strategies: (1) 3bp"scrambled motif prefix"3bp (2) original_sequence_start"scrambled motif prefix"original_sequence_end, through Fimo [29] with the two sets of TF motifs [30,31] (p-val<10^{-4}).
8) Choose the 2 "scrambled motifs" that result in the lowest number of motif hits indicated by the median rank across the two strategies, i.e. "non-motif sequences" that would be used on the array.

Approach 2 - shuffle the motif:

In each chosen combination of region and motif - scramble the motif by shuffling its nucleotides.

Library processing: replicates, association, barcode count, ratio

Association

Reads from the association library were aligned to the reference set of sequences using bowtie2 [69] with the *--very-sensitive* preset parameters for maximal accuracy. A barcode was confidently

assigned to a sequence if at least 3 unique UMIs supported that assignment and at least 80% of the UMIs associated with that barcode were aligned to the sequence. Barcodes that were not confidently assigned were considered ambiguous and discarded from downstream analyses. Overall, 7,004,354 barcodes were observed, of which 1,447,874 (20%) were confidently assigned, averaging 139.2 barcodes per sequence **Supplementary Figs. 1-2**)

MPRA Barcode Counting

Reads from the MPRA libraries were processed against the set of confidently assigned barcodes, requiring a perfect match. Of the barcodes observed in the MPRA libraries, an average of 61.6% were confidently assigned, 37.4% were ambiguous (observed in the association library but were not confidently assigned), and 0.9% were unobserved in the association library (**Supplementary Fig. 2**). Only barcodes that appeared in at least two corresponding libraries (DNA and RNA libraries from the same time point and replicate) were included in downstream analyses, resulting in an average of 134.4 barcodes per sequence.

Quantification of Induced Transcription Rate with MPRAnalyze

Quantification of induced transcriptional rates (`alpha` values) was performed using MPRAnalyze [28]. Briefly, MPRAnalyze fits two nested generalized linear models (GLMs): the first estimates the latent construct counts from the observed DNA counts, and the second estimates the latent rate of transcription from the latent construct estimates and observed RNA counts. The models are optimized using likelihood maximization, with a gamma likelihood for the DNA counts and a negative binomial likelihood for the RNA counts. MPRAnalyze includes library-size normalization factors, which were computed once using the entire dataset and then used across all analyses, including per-time-point analyses, to maintain consistency. For quantification of alpha values, the full experimental design was included in the design matrix for the DNA model (~ timepoint + replicate + barcode), and an alpha value was extracted for each time point and replicate (RNA model: ~ timepoint + replicate).

Classification of active sequences with MPRAnalyze

Classification of active sequences was performed using the standard MPRAnalyze classification analysis, in which alpha values are mad-normalized (a median-based variant of z-normalization) and tested each value against the null distribution, estimated from the alpha values from the negative control scrambled sequences.

Comparative Analyses with MPRAnalyze

The GLM structure of MPRAnalyze allows for a flexible framework to perform comparative analyses by using various design matrices for the different analyses (detailed below). Since the models are optimized using likelihood maximization, a likelihood ratio testing can be used for statistical significance, and was used throughout all analyses in the manuscript. P-values were computed for each comparison and corrected within each analysis using Benjamini-Hochberg FDR correction [70].

For the per time point comparative analyses, each PERT and RAND sequence was compared with the corresponding WT sequence within each time point (DNA design: ~ replicate + barcode + sequence; Full RNA design: ~ sequence; reduced RNA design: ~ 1). The resulting p-values were corrected jointly across all timepoints.

For temporal analyses, aimed at determining which sequences had temporal activity, we set the null behavior to be the temporal behavior exhibited by the scrambled sequences, by fitting a joint model to all SCRAM sequences and using the model coefficients as normalization factors for the comparative models (DNA design: ~ timepoint + replicate + barcode; Full RNA design: ~ timepoint; reduced RNA design: ~ 1).

For the comparative temporal analyses, we compared the temporal activity of each PERT or RNAD sequence with the corresponding WT sequence, using an interaction term in the design (DNA design: ~ timepoint + replicate + barcode; full RNA design: ~ time * sequence; reduced RNA design: ~ time). Note that the barcode covariate in the allele-comparative analyses (per-time point comparative analysis and temporal comparative analysis) is sequence-specific, so the barcode factor is confounded by the sequence variable.

Calculating RNA/DNA ratios

The calculation of RNA to DNA ratio is explained in detail in our previous work [19,26]. Briefly, to estimate the abundance of DNA or RNA per sequence and for each replicate (in order to compare replicates and time point), we use a simple averaging scheme:

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

To determine the RNA/DNA ratios per sequence and for each replicate we the sum of ratios:

$$(\sum_{i=1}^{\#BC} \frac{RNA_i}{sum\ (RNA\ reads)} / \frac{DNA_i}{sum\ (DNA\ reads)}) / \#BC$$

We added a pseudo count of 1 to the numerator and denominator to stabilize the signal from sequences with low numbers of reads. To combine replicates, we 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. We use the ratio calculation to compare the MPRA signal in this work to our previous work [26] (**Supplementary Fig 4.**).

Filtering sequences

Filtering sequences per time point

We use MPRAnalyze to determine differential activity (explained in the previous section), for each perturbation method and each time point, comparing the following:
(PERT,WT), (RAND,WT), (PERT,RAND), (WT,SCRAM) and (PERT,SCRAM).
We use the following filters:

Filtering sequences per time point

1. We consider only sequences where WT (at each of the 7 time points) **or** PERT have significantly different (MAD-score) regulatory activity than the null (SCRAM): length(FDR(WT, SCRAM)<0.05)==nof_TPs || FDR(PERT, SCRAM)<0.05
2. We only consider sequences where PERT has significantly different regulatory activity than its matching WT: FDR(PERT, WT)<0.05.

1008, 1042, 998 out of (2082, 2086, 2114) sequences for perturbation methods 1, 2 and 3 respectively, pass these filters in at least one time point.

Filtering sequences across time

3. We consider only sequences where WT **or** PERT have significantly temporally different regulatory activity than the null (SCRAM).
FDR(temporal(PERT,SCRAM))<0.05 || FDR(temporal(WT,SCRAM))<0.05)
4. **and** PERT has significantly temporally different regulatory activity than its matching WT
FDR(temporal(PERT,WT))<0.05

1189, 1224, 1354 out of (2082, 2086, 2114) sequences pass the temporal filtering for perturbation methods 1, 2 and 3 respectively.

We consider only sequences that are significant (pass all filtering steps per time point) in at least one time point and follow the temporal constraints, after filtering for duplicates, resulting in overall 747, 775, 749 sequences for perturbation methods 1, 2 and 3 respectively. Duplicates, i.e. sequences with motifs perturbed in the exact same locations (corresponding to different PWMs), were filtered, by picking the sequence with the lowest temporal FDR. FRSs are defined as sequences that pass all 4 filters and belong to the same main category (as described in the next section) in perturbation method 3 and either perturbation methods 1 or 2.

Filtering for pairs of motifs (double perturbation)

IFRSs are defined as sequences with pairs of sites for which both single-site perturbations and the joint perturbation belong to the same main category and the double perturbation is functional (i.e passed all four filters as described above) in both perturbation approaches (perturbation method 3 and either of methods 1 or 2).

Motif effect - main and sub categories

Activators - when this motif is perturbed in a region, the regulatory activity of PERT compared to WT is significantly reduced in at least one time point.

(i) Essential - this motif is essential for the regulatory activity of the region - i.e. scrambling this motif reduces the regulatory activity to null (SCRAM) or for all time points - the regulatory activity of PERT is similar to SCRAM.
FDR(temporal(PERT,SCRAM))>0.05 || length(FDR_MAD(PERT,SCRAM)>0.05))==nof_TP

(ii) Contributing - this motif is contributing to the regulatory activity of the region - i.e. if we scramble this motif, the region is still regulatory active and its activity is different from null (SCRAM). If a motif is not essential, it is deemed contributing.

Repressors - when this motif is perturbed in a region, the regulatory activity of PERT compared to WT is significantly increased in at least one time point.

(iii) Silencing- this motif has a silencing effect on the regulatory activity of the region - i.e. the regulatory activity of the WT region is not temporarily different from SCRAM <u>or</u> for all time points - the regulatory activity of WT is similar to SCRAM.
FDR(temporal(WT,SCRAM))>0.05 || length(FDR_MAD(WT,SCRAM)>FDR_thresh))==nof_TP) scrambling this motif increases the regulatory activity in at least one time point.

(iv) Inhibiting- this motif is reducing the regulatory activity of the region. If a motif is not silencing, it is deemed inhibiting.

Activation dynamics analysis

To examine the activation dynamics of activating FRSs, we looks at activators that are active in all 7 time points and fit a linear regression line to each FRS, modeling the absolute effect (WT - PERT) as a function of the WT activity level (delta ~ wt), using the *lm* function in R. the model parameters were then extracted and used for the extrapolation in **Supplementary Fig. 8e**).

Experimental procedures

LentiMPRA library cloning and sequence-barcode association

The lentiMPRA library construction was performed as previously described (Gordon et al., 2020). In brief, array-synthesized oligo pool was amplified by 5-cycle PCR using forward primer (5BC-AG-f01, Table S-primer) and reverse primer (5BC-AG-r01, Table S-primer) that adds mP and spacer sequences downstream of the sequence. The amplified fragments were purified with 1.8x AMPure XP (Beckman coulter), and proceeded to second round 11-cycle PCR using the same forward primer (5BC-AG-f01) and reverse primer (5BC-AG-r02, Table S-primer) to add 15-nt random sequence that serves as a barcode. The amplified fragments were then inserted into *Sbf*I/*Age*I site of the pLS-SceI vector (Addgene, 137725) using NEBuilder HiFi DNA Assembly mix (NEB), followed by transformation into 10beta competent cells (NEB, C3020) using the Gemini X2 machine (BTX). Colonies were allowed to grow up overnight on Carbenicillin plates and midiprepped (Qiagen, 12945). We collected approximately 1 million colonies, so that on average 100 barcodes were associated with each sequence. To determine the sequences of the random barcodes and their association to each sequence, the sequence-mP-barcodes fragment was amplified from the plasmid library using primers that contain flowcell adapters (P7-pLSmP-ass-gfp and P5-pLSmP-ass-i#, Primer Table). The fragment was then sequenced with a NextSeq 150PE kit using custom primers (R1, pLSmP-ass-seq-R1; R2 (index read), pLSmP-ass-seq-ind1; R3, pLSmP-ass-seq-R2, Primer Table) to obtain approximately 50M total reads.

Lentiviral infection and barcode sequencing

Lentivirus was produced in twelve 15cm dishes of 293T cells using Lenti-Pac HIV expression packaging kit following the manufacture's protocol (GeneCopoeia, LT002). Lentivirus was filtered through a 0.45um PES filter system (Thermo Scientific, 165-0045) and concentrated by Lenti-X concentrator (Takara Bio, 631232). Titration of the lentiMPRA library was conducted on hESCs as described previously (Gordon et al., 2020). Lentiviral infection, DNA/RNA extraction, and barcodes sequencing were all performed as previously described (Inoue et al., 2019). Briefly, approximately 8 million cells (three 10 cm dishes) per time point were infected with the lentivirus library with a multiplicity of infection (MOI) of 5-8 along with 8 µg/mL polybrene (Sigma). Three independent replicate cultures were infected. To normalize technical bias of lentivirus preps, two of these replicates were infected with the same lentivirus batch, while the other replicate was infected with another lentivirus batch. The cells were incubated for 3 days with a daily change of the media. The infected cells were induced into neural lineage using dual-Smad inhibition and harvested at 0 (right before differentiation), 3, 6, 12, 24, 48, and 72 hours. DNA and RNA were purified using an AllPrep DNA/RNA mini kit (Qiagen). RNA was treated with Turbo DNase (Thermo Fisher Scientific) to remove contaminating DNA, and reverse-transcribed with SuperScript II (Invitrogen, 18064022) using barcodes specific primer (P7-pLSmp-assUMI-gfp, Primer Table), which has a unique molecular identifier (UMI). Barcode DNA/cDNA from each replicate of each time point were amplified with 3-cycle PCR using specific primers (P7-pLSmp-assUMI-gfp and P5-pLSmP-5bc-i#, Primer Table) to add sample index and UMI. A second round of PCR was performed for 19 cycles using P5 and P7 primers (P5, P7, Primer Table). The fragments were purified and further sequenced with NextSeq 15PE with 10-cycle dual index reads, using custom primers (R1, pLSmP-ass-seq-ind1; R2 (index read1 for UMI), pLSmP-UMI-seq; R3, pLSmP-bc-seq; R4 (index read2 for sample index), pLSmP-5bc-seqR2, Primer Table).

References

1. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics* vol. 20 437–455 (2019).

2. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* vol. 337 1190–1195 (2012).

3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

4. Kearns, N. A. *et al.* Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat. Methods* **12**, 401–403 (2015).

5. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34**, 192–198 (2016).

6. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

7. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).

8. White, M. A., Myers, C. A., Corbo, J. C. & Cohen, B. A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11952–11957 (2013).

9. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).

10. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).

11. Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530–1545 (2016).

12. Wang, X. *et al.* High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* **9**, 5380 (2018).

13. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994–1004 (2012).

14. Yosef, N. *et al.* Dynamic regulatory network controlling TH17 cell differentiation. *Nature* **496**, 461–468 (2013).

15. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

16. Whitfield, T. W. *et al.* Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* **13**, R50 (2012).

17. Biggin, M. D. Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* **21**, 611–626 (2011).

18. Landolin, J. M. *et al.* Sequence features that drive human promoter function and tissue specificity. *Genome Res.* **20**, 890–898 (2010).

19. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45**, 1021–1028 (2013).

20. Gotea, V. *et al.* Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* **20**, 565–577 (2010).

21. van Dijk, D. *et al.* Large-scale mapping of gene regulatory logic reveals context-dependent repression by transcriptional activators. *Genome Res.* **27**, 87–94 (2017).

22. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).

23. Weingarten-Gabbay, S. *et al.* Systematic interrogation of human promoters. *Genome Res.* **29**, 171–183 (2019).

24. Davis, J. E. *et al.* Dissection of c-AMP Response Element Architecture by Using Genomic and Episomal Massively Parallel Reporter Assays. *Cell Syst* **11**, 75–85.e7 (2020).

25. Grossman, S. R. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1291–E1300 (2017).

26. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell* **25**, 713–727.e10 (2019).

27. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).

28. Ashuach, T. *et al.* MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 1–17 (2019).

29. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

30. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research* vol. 42 2976–2987 (2014).

31. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

32. Klein, D., Bugl, B., Günzburg, W. H. & Salmons, B. Accurate estimation of transduction efficiency necessitates a multiplex real-time PCR. *Gene Ther.* **7**, 458–463 (2000).

33. Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature Biotechnology* vol. 27 275–280 (2009).

34. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* **15**, 2387–2412 (2020).

35. Iwafuchi-Doi, M. & Zaret, K. S. Cell fate control by pioneer transcription factors. *Development* **143**, 1833–1837 (2016).

36. Aruga, J. The role of Zic genes in neural development. *Mol. Cell. Neurosci.* **26**, 205–221 (2004).

37. Guzelsoy, G. *et al.* Terminal neuron localization to the upper cortical plate is controlled by the transcription factor NEUROD2. *Sci. Rep.* **9**, 19697 (2019).

38. Agostini, M. *et al.* p73 regulates maintenance of neural stem cell. *Biochem. Biophys. Res. Commun.* **403**, 13–17 (2010).

39. Talos, F. *et al.* p73 is an essential regulator of neural stem cell maintenance in embryonal and adult CNS neurogenesis. *Cell Death Differ.* **17**, 1816–1829 (2010).

40. Hudson, C. D. *et al.* Brn-3a/POU4F1 interacts with and differentially affects p73-mediated transcription. *Cell Death Differ.* **15**, 1266–1278 (2008).

41. Kim, M., Kim, T., Johnson, R. L. & Lim, D.-S. Transcriptional co-repressor function of the hippo pathway transducers YAP and TAZ. *Cell Rep.* **11**, 270–282 (2015).

42. Ozaki, T., Kubo, N. & Nakagawara, A. p73-Binding Partners and Their Functional Significance. *Int. J. Proteomics* **2010**, 283863 (2010).

43. Leung, A. W. *et al.* Pre-Border Gene Foxb1 Regulates the Differentiation Timing and Autonomic Neuronal Potential of Human Neural Crest Cells. *Cold Spring Harbor Laboratory* 646026 (2019) doi:10.1101/646026.

44. Hou, P.-S. *et al.* LHX2 regulates the neural differentiation of human embryonic stem cells via transcriptional modulation of PAX6 and CER1. *Nucleic Acids Res.* **41**, 7753–7770 (2013).

45. Lodato, M. A. *et al.* SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state. *PLoS Genet.* **9**, e1003288 (2013).

46. Zhang, X. *et al.* Pax6 is a human neuroectoderm cell fate determinant. *Cell Stem Cell* **7**, 90–100 (2010).

47. Kamiya, D. *et al.* Intrinsic transition of embryonic stem-cell differentiation into neural progenitors. *Nature* **470**, 503–509 (2011).

48. Matsushita, M. *et al.* Neural differentiation of human embryonic stem cells induced by the

transgene-mediated overexpression of single transcription factors. *Biochem. Biophys. Res. Commun.* **490**, 296–301 (2017).

49. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035–1041 (2010).

50. Zhang, Y. *et al.* Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* **78**, 785–798 (2013).

51. Frantz, G. D., Weimann, J. M., Levin, M. E. & McConnell, S. K. Otx1 and Otx2 define layers and regions in developing cerebral cortex and cerebellum. *J. Neurosci.* **14**, 5725–5740 (1994).

52. Su, Z. *et al.* Antagonism between the transcription factors NANOG and OTX2 specifies rostral or caudal cell fate during neural patterning transition. *J. Biol. Chem.* **293**, 4445–4455 (2018).

53. Sansom, S. N. *et al.* The level of the transcription factor Pax6 is essential for controlling the balance between neural stem cell self-renewal and neurogenesis. *PLoS Genet.* **5**, e1000511 (2009).

54. Pinsonneault, J., Florence, B., Vaessin, H. & McGinnis, W. A model for extradenticle function as a switch that changes HOX proteins from repressors to activators. *EMBO J.* **16**, 2032–2042 (1997).

55. Kadamb, R., Mittal, S., Bansal, N., Batra, H. & Saluja, D. Sin3: insight into its transcription regulatory functions. *Eur. J. Cell Biol.* **92**, 237–246 (2013).

56. Saunders, A. *et al.* The SIN3A/HDAC Corepressor Complex Functionally Cooperates with NANOG to Promote Pluripotency. *Cell Rep.* **18**, 1713–1726 (2017).

57. Ahmed, J. N., Diamand, K. E. M., Bellchambers, H. M. & Arkell, R. M. Systematized reporter assays reveal ZIC protein regulatory abilities are Subclass-specific and dependent upon transcription factor binding site context. *Sci. Rep.* **10**, 13130 (2020).

58. Singh, S. *et al.* Zeb1 controls neuron differentiation and germinal zone exit by a mesenchymal-epithelial-like transition. *Elife* **5**, (2016).

59. Lehmann, W. *et al.* ZEB1 turns into a transcriptional activator by interacting with YAP1 in aggressive cancer types. *Nat. Commun.* **7**, 10498 (2016).

60. di Masi, A. *et al.* Retinoic acid receptors: from molecular mechanisms to cancer therapy. *Mol. Aspects Med.* **41**, 1–115 (2015).

61. Stampfel, G. *et al.* Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151 (2015).

62. White, M. A. *et al.* A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. *Cell Rep.* **17**, 1247–1254 (2016).

63. Iwafuchi-Doi, M. *et al.* Transcriptional regulatory networks in epiblast cells and during anterior neural plate development as modeled in epiblast stem cells. *Development* **139**, 3926–3937 (2012).

64. Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat. Commun.* **8**, 14519 (2017).

65. Gao, P. P. *et al.* Regulation of topographic projection in the brain: Elf-1 in the hippocamposeptal system. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 11161–11166 (1996).

66. Rosa, A. & Brivanlou, A. H. A regulatory circuitry comprised of miR-302 and the transcription factors OCT4 and NR2F2 regulates human embryonic stem cell differentiation. *EMBO J.* **30**, 237–248 (2011).

67. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583. doi: 10.1038/s41467–019–11526–w. (2019).

68. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

69. Langmead, B. & Salzberg, S. L. Langmead. 2013. Bowtie2. *Nat. Methods* **9**, 357–359 (2013).

70. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* (1995).