Supplementary Methods

Perturbation MPRA library design

Choosing region and motif combinations

General description:
In our previous analysis [1], we identified 1,547 regulatory regions of interest and multiple occurrences of motifs within these regions. Then, our primary objective was to curate the most informative set of [region × motif] combinations, each corresponding to a specific motif instance, to fit within the confines of a single Massively Parallel Reporter Assay (MPRA) design. To address this challenge, we developed a sophisticated selection methodology that accounts for various biological factors within our system and accommodates the experimental limitations concerning the number of sequences that can be assayed.

To do this, we formalized the existing knowledge regarding motifs and genomic regions into a tripartite graph structure, where three distinct layers of nodes are employed to represent critical elements of our analysis.

The first layer of nodes pertains to DNA regions, specifically the 1,547 genomic regions we previously identified [1]. These regions exhibit temporal activity, as confirmed through lentiMPRA assays conducted across seven different time points.

The second layer comprises nodes representing motifs, computed computationally via the Fimo algorithm (*P values* $< 5 \times 10^{-5}$, [2]), using two sets of transcription factor motifs [3, 4]. Edges between the first two layers establish connections between each motif and the regions in which it is found.

The third layer consists of property nodes, each characterizing a unique property of motifs or regions based on genomic assays conducted in our previous work, which included ATAC-seq, H3K27ac and H3K27me3 ChIP-seq, and RNA-seq data gathered across the same seven time points [1]. For instance, we identified distinct temporal patterns associated with each of these data modalities, and we designated each pattern as a node in this layer (e.g., a node for "regions exhibiting a transient peak in H3K27ac at 48 hours post-induction"). We established connections between a region and a property node if the corresponding pattern is observed in that region within the endogenous genome. Similarly, connections were formed between a motif node and a property node based on the presence of the designated pattern for that motif in the endogenous genome (e.g., a node for "motifs associated with a transcription factor that is expressed 24 hours post-induction").

In total, our graph comprises 1,547 region nodes, 4,393 motif nodes, and 68 property nodes, all interconnected by a complex web of 99,165 edges. Our current objective revolves around identifying the minimum number of [region × motif] combinations, each representing a distinct motif instance (or equivalently, an edge in our graph), which will ensure adequate coverage of each property within the third layer. In essence, we seek to pinpoint a minimal subset of motif-region pairs such that every "property node" in our third layer is linked to a sufficient number of motifs and regions, as elaborated further below.

The utilization of this tripartite graph structure has empowered us to rephrase our objective as a constrained optimization problem, aiming to guarantee a minimum level of connectivity for the third layer while simultaneously minimizing the selection of nodes and edges within the first

two layers. Given that this problem falls into the NP-hard category, we have adopted the common practice of formulating it as an integer linear program (ILP). Solving this ILP efficiently is possible through various heuristics with readily available solvers.

Utilizing this ILP formulation, we have successfully identified 591 regulatory regions and 255 motifs, organized into 2144 [region × motif] pairs, meeting our criteria. Below, we offer a more comprehensive description of this intricate process.

Defining the property layer:

We compiled a list of biological attributes by integrating data from published literature and our own analysis of ATAC-seq, H3K27ac, H3K27me3 ChIP-seq, and RNA-seq data, as detailed in our prior publication [1]. These biological attributes associated with transcription factors (TFs) and genomic regions encompass the following criteria:

(i) Activation of TFsregions at specific temporal points;

(ii) Significant co-occurrence of TFs/regions with temporal MPRA, H3K27ac, ATAC-seq, and RNA-seq signals, delineated by overlapping sub-clusters as we defined in previous studies[1];

iii) Association of TFs or proximal genes with known neural factors or relevant pathways. Known neural factors include: POU3F1, MYT1L, SOX2, POU3F2, LHX2, PAX6, ASCL1, SOX1, OTX2, ZNF521, NEUROG1, NEUROG2, NEUROG3, NEUROD1, NEUROD2. Pathways were extracted from KEGG72, comprising FGFMAPK signaling pathway (hsa04010), IGF-1mTOR signaling pathway (hsa04150), WntCa+PCP signaling pathway (hsa04310), and Sonic Hedgehog signaling pathway (hsa04340);

(iv) Selection of specific TFs, including those with documented roles in neural differentiation from our previous results [1] or based on their expression in the neuroectoderm of mouse embryos, or high "TF activity scores" at relevant time points in our previous research [1].

The direct edges from motifs and regions to these attributes represent the fulfillment of the aforementioned biological criteria.

The optimization program:

1. Minimize:

$$\left(\sum_{r\in R}\theta_r\right)+3\left(\sum_{(t,r)\in E;t\in T;r\in R}e_{t,r}\right)$$

Subject to:

2. $$\sum_{(t,p')\in E;t\in T}\theta_t\geq 12 \qquad \forall p'\in P$$

3. $$\sum_{(r,p')\in E;r\in R}\theta_r\geq \min\{17,\deg_R(p')\} \qquad \forall p'\in P$$

4. $$\sum_{(t,r')\in E;t\in T}\theta_t\geq \theta_{r'}\min\{3,\deg_T(r')\} \qquad \forall r'\in R$$

5. $$\sum_{(t',r)\in E;r\in R}\theta_r\geq \theta_{t'}\min\{20,\deg_R(t')\} \qquad \forall t'\in T$$

6. $e_{t,r} \geq \theta_t + \theta_r - 1 \qquad \forall (t,r) \in E; t \in T; r \in R$

7. $\sum_{t \in T_i} \theta_t \leq 2 \qquad \forall T_i$

8. $\sum_{t \in T_i} \theta_t \geq 1 \qquad \forall T_i \in \textit{HandPicked}$

9. $\sum_{r \in R} \theta_r \geq 0.4 \cdot |R|$

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

11. $\sum_{t \in T_B} \theta_t \geq 1.5 \cdot \sum_{t \in T_S} \theta_t$

12. $\theta_t, \theta_r, e_{t,r} \in \{0, 1\}$

The decision variables represent the following: θ_t is a binary variable that indicates whether we chose the motif t; θ_r is a binary variable that represents whether the region r was selected. $e_{t,r}$ is a binary variable that denotes whether a [region × motif] pair (r and t) has been selected.

Parameters include:

P represents the properties.

R represents 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 is defined 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 five regions.

Constrains:

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 16 of the total edges used are low-confidence edges; (10) At least 60% of the motifs chosen are the motifs connected with many regions (T_B), such that 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 the number of unperturbed (WT) regions plus the number of perturbations (i.e., [region × motif] combinations). We multiply by 3 since we have three perturbation methods (i.e., we need three MPRA sequences for every pair).

Different categories of sequences designed on the array:

Overall, the solver selected 591 regions, encompassing 255 unique motifs corresponding to 166 distinct transcription factors (TFs). We employed the combinations of regions and motifs chosen by the solver to represent the following sequence categories on the array:

1. **hit1**: One motif is perturbed in the sequence. For [region × motif] combinations where the motif is detected once in the sequence (N = 1620).

2. **hit2**: Two motifs of the same kind are perturbed in the sequence. For [region × motif] combinations where the motif is detected twice in the sequence if the +/- strand carries exactly the same motif, we replace the motif only once 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 three or more occurrences of the same motif are observed, we discard those [region × motif] combinations (N=52).
   In addition to the combinations selected by the solver, we considered the 591 wild-type (WT) regions and added more combinations (not chosen by the solver) containing motifs of the following 11 TFs. These TFs were selected (LHX5, MEIS2, PAX6, FOXB1, SOX1, IRX3, OTX2, ZIC2, SP8, POU3F1, HOMEZ) based on their high "TF activity scores" at relevant time points in our data [1] and their mRNA expression in the neuroectoderm of the mouse embryo.

3. One motif is perturbed in the sequence. For [region × motif] combinations where the motif is detected once in the sequence (**Overexpressed_hit1**, N=221, and **Overexpressed_permutation**, N=58).

4. Two variations of the same kind are perturbed in the sequence. For [region × motif] combinations where the motif is detected twice in the sequence if the +/- strand carries exactly the same motif, we replace the motif only once in the + strand (**Overexpressed_hit2**, N=3). Otherwise (**Overexpressed_hit2diff**, N=1), we perturb each motif separately and then both of them, starting with the + strand.

5. Combinations of two or more variations of one motif are perturbed in the sequence. For [region × motif] combinations, we observe two or more different motifs in the sequence (**Overexpressed_permutation**, N=125). We examined combinations of motif hits from these 11 TFs in our regions.

In summary, most of the data involve a single motif perturbation per region (N=2144), with a smaller portion having two or more motif perturbations per region (N=216, comprising N=154 with two motifs and N=62 with more than two motifs), totaling 2360 designed region and motif sequences.

In this study, we used the single motif perturbation sequences (N = 2144) for our analyses. We also assayed WT and control sequences:

1. We assayed 591 **WT** sequences. WT sequences are the endogenous 171-bp 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 selected a location in the region and **perturbed** the median motif size (12 bp), starting at that location, creating an additional set of negative controls.

We perturbed one single motif in each of the 2,144 perturbation sequences. We use three respective perturbation approaches for each motif perturbation: the first two methods replace the predicted binding site with a "non-motif" sequence; the third method randomly shuffles the nucleotides of the predicted binding site described in the next section. For the **RAND** sequence category, we used the same three perturbation approaches.

Different Motif Scrambling (Perturbation) Approaches

Approaches 1 & 2: Replacing the motifs with "non-motif" sequences

For these two approaches, we first generated a set of "non-motif" sequences and filtered them using two strategies:

Step 1: Creating "scrambled motifs":

To generate scrambled motifs, we followed these workflows:

1. We first utilized all the 2,464 MPRA sequences from our previous work [1] based on their potential for neural differentiation activity.

2. We determined the frequency and percentage of di-nucleotides within these sequences.

3. We generated di-nucleotide scrambled sequences of maximal motif length, referred to as "scrambled motifs."

4. We then created 1,000 such "scrambled motifs."

5. Using Fimo[2] to analyze these 1,000 "scrambled motifs" against two sets of TF motifs [5, 4], we selected those with the fewest motif hits $P < 10^{-4}$. In this study, we ended up with 13 "scrambled motifs" yielding zero hits.

Step 2: Generating "non-motif" sequences using "scrambled motifs":

Within each chosen [region × motif] combination (as described in the previous section), we replaced the original motif occurrences with one of the 13 "scrambled motif" prefix sequences, adapting it to the motif's length. This is done using these two strategies to prevent motif creation at the sequence edges:

- **Approach 1**: incorporating a 3-bp downstream and upstream sequence of the original motif, which is in the format of "3bp_scrambled motif prefix_3bp." Using the example in Figure 1, the sequences generated using this approach are in the format of:

  1 TGTXXXXXXX XXXACA,

  where the red-colored sequence is one of the 13 scrambled motifs (of which the length has been adjusted) created by the previous step.

- **Approach 2**: using the entire original sequence, which follows the format "original_sequence_start_scrambled motif prefix_original_sequence_end." Using the example in Figure 1, the sequences generated using this approach are in the format of:

```
  1 CCCTCCCTGC CAGGTCGGCC CAGGCAGCCT GAGGTCAGGA GGGATTTGTX XXXXXXXXXA
 61 CAGGCCCCTC GTTGCCCTGG CAACAGGCCC CGCCCCTCCA GCCTGGCCCG GGAAGGGGGG
121 ACAGCCCCTC CTCCGCCCTG CTGCCCTCCA CACCCCGCCC CTCCTTGCCT C,
```

where the red-colored sequence is one of the 13 scrambled motifs (of which the length has been adjusted) created by the previous step.

These two strategies are repeated 13 times respectively using each of the 13 0-hit "scrambled motifs."

Step 3: Selecting the "non-motif" sequences generated by each approach:

Using each approach described above, we have generated a total of 27,872 "non-motif" sequences, resulting from 13 of 0-hit "scrambled motifs" sequences each with 2,144 [region × motif] combinations. Subsequently, we applied the Fimo algorithm [2] to evaluate these sequences. We selected the sequence with the fewest motif hits, as indicated by the median rank for each approach.

Consequently, we have identified a single "non-motif" sequence for each approach. These selected sequences have been employed to replace the target motif sequences within the 2,144 [region × motif] perturbations associated with each approach. As a result, we now possess 2,144 sequences representing each of the 2,144 distinct [region × motif] perturbations for both approaches.

Approach 3: Shuffling the motif sequences

In each chosen combination of region and motif, we scrambled the target motif by randomly shuffling its nucleotides.

Supplementary Notes

The DNA sequences in illustrative examples

• Figure 1:

– In the illustration part of Figure 1 (top left), we used the motif "GATA_known9" as an example (the target motif is colored green in the WT sequence, and perturbed ones are colored red):

*>**WT**_region211_chr9:126122582-126122753*:

```
   1 CCCTCCCTGC CAGGTCGGCC CAGGCAGCCT GAGGTCAGGA GGGATTTGTA AAGATAAGCA
  61 CAGGCCCCTC GTTGCCCTGG CAACAGGCCC CGCCCCTCCA GCCTGGCCCG GGAAGGGGGG
 121 ACAGCCCCTC CTCCGCCCTG CTGCCCTCCA CACCCCGCCC CTCCTTGCCT C
```

*>hit1_**PERT1**_chr9:126122582-126122753_GATA_known9*:

```
   1 CCCTCCCTGC CAGGTCGGCC CAGGCAGCCT GAGGTCAGGA GGGATTTGTA CTAAAGAATA
  61 CAGGCCCCTC GTTGCCCTGG CAACAGGCCC CGCCCCTCCA GCCTGGCCCG GGAAGGGGGG
 121 ACAGCCCCTC CTCCGCCCTG CTGCCCTCCA CACCCCGCCC CTCCTTGCCT C
```

*>hit1_**PERT2**_chr9:126122582-126122753_GATA_known9*:

```
   1 CCCTCCCTGC CAGGTCGGCC CAGGCAGCCT GAGGTCAGGA GGGATTTGTC GAGCATCTTA
  61 CAGGCCCCTC GTTGCCCTGG CAACAGGCCC CGCCCCTCCA GCCTGGCCCG GGAAGGGGGG
 121 ACAGCCCCTC CTCCGCCCTG CTGCCCTCCA CACCCCGCCC CTCCTTGCCT C
```

*>hit1_**PERT3**_chr9:126122582-126122753_GATA_known9*:

```
   1 CCCTCCCTGC CAGGTCGGCC CAGGCAGCCT GAGGTCAGGA GGGATTTGTC GTGAAAAAAA
  61 CAGGCCCCTC GTTGCCCTGG CAACAGGCCC CGCCCCTCCA GCCTGGCCCG GGAAGGGGGG
 121 ACAGCCCCTC CTCCGCCCTG CTGCCCTCCA CACCCCGCCC CTCCTTGCCT C
```

• Figure 2:

– In Figure 2A, we use the motif BCL6_M6136_1.02 on chromosome 1 as an example to illustrate "Hit" and "Fail" sequences:

*>**WT**_region327_chr1:38736205-38736376*:

```
   1 CCCTGAGAAA CAACTCCCAG CGTAGACAAT GGGCAAAACA AAGGGCAGGA AGGAAAGGAA
  61 GTGGTCTGTC AGCACCTGGG GCTGCTGTGG AGCGCAGACC CACCTGAGCT CTGCAGGTAG
 121 GCAATCCTGG CTGGGATTCT GCAGACAGGC CTGGAGCAGA GCCCCCCCCA A
```

*>hit1_**PERT1**_chr1:38736205-38736376_BCL6_M6136_1.02*:

```
   1 CCCTGAGAAA CAACTCCCAG CGTAGACAAT GACTAAAGAA TCTACAAATG ATCTCTCAAA
  61 TTGGTCTGTC AGCACCTGGG GCTGCTGTGG AGCGCAGACC CACCTGAGCT CTGCAGGTAG
 121 GCAATCCTGG CTGGGATTCT GCAGACAGGC CTGGAGCAGA GCCCCCCCCA A
```

*>hit1_**PERT2**_chr1:38736205-38736376_BCL6_M6136_1.02*:

```
   1 CCCTGAGAAA CAACTCCCAG CGTAGACAAT GCGAGCATCT TTAAGAGTTA GAGTAGGCAA
  61 ATGGTCTGTC AGCACCTGGG GCTGCTGTGG AGCGCAGACC CACCTGAGCT CTGCAGGTAG
 121 GCAATCCTGG CTGGGATTCT GCAGACAGGC CTGGAGCAGA GCCCCCCCCA A
```

*>hit1_**PERT3**_chr1:38736205-38736376_BCL6_M6136_1.02*:

```
   1 CCCTGAGAAA CAACTCCCAG CGTAGACAAT GAAAAAGGCG AGAAAGAAGG CGAAAGGGAG
  61 CTGGTCTGTC AGCACCTGGG GCTGCTGTGG AGCGCAGACC CACCTGAGCT CTGCAGGTAG
 121 GCAATCCTGG CTGGGATTCT GCAGACAGGC CTGGAGCAGA GCCCCCCCCA A
```

Notes:
○ In the WT sequence, the target motif is BCL6_M6136_1.02:
GGCAAAACAA AGGGCAGGAA GGAAAGGAAG
○ In PERT1 and PERT2 sequences: BCL6_M6136_1.02 are removed respectively → both sequences are annotated as "Hit"
○ PERT3: A variation of BCL6_M6136_1.02 is found at the same genomic position 1
AAAAAGGCGA GAAAGAAGGC GAAAGGGAGC → this sequence is annotated as "Fail"

– In Figure 2B, we use the motif BHLHE40_disc2 on chromosome chr14 as an example to illustrate "Perturbed" and "Not perturbed" sequences:
>***WT***_*region462_chr14:77428196-77428367*:
```
   1 GGGCCGGAAG CCcCGCTGCG GCCCCAGCCC GCGGTTAGCC CCTGTTTGTC ATTTTGCAAA
  61 TCTGGTCAAC CCACCTCCGG TGAAAACTCC CAACCTCACC CCAGGGGGCA ATGACTAATT
 121 ACAAAACACA TTTTCTCTCG TTTTCGTCAA GCTCGCTGTC CCGCCCACAC A
```
>*hit2diff_2_**PERT1**_chr14:77428196-77428367_BHLHE40_disc2*:
```
   1 GGGCCGGAAG CCCACTAAAG AATCTACAAA GCGGTTAGCC CCTGTTTGTC ATTTTGCAAA
  61 TCTGGTCAAC CCACCTCCGG TGAAAACTCC CAACCTCACC CCAGGGGGCA ATGACTAATT
 121 ACAAAACACA TTTTCTCTCG TTTTCGTCAA GCTCGCTGTC CCGCCCACAC A
```
>*hit2diff_2_**PERT2**_chr14:77428196-77428367_BHLHE40_disc2*:
```
   1 GGGCCGGAAG CCCCGAGCAT CTTTAAGAGT GCGGTTAGCC CCTGTTTGTC ATTTTGCAAA
  61 TCTGGTCAAC CCACCTCCGG TGAAAACTCC CAACCTCACC CCAGGGGGCA ATGACTAATT
 121 ACAAAACACA TTTTCTCTCG TTTTCGTCAA GCTCGCTGTC CCGCCCACAC A
```
>*hit2diff_2_**PERT3**_chr14:77428196-77428367_BHLHE40_disc2*:
```
   1 GGGCCGGAAG CCC***GGCTCGA CGCCCGCCC***C GCGGTTAGCC CCTGTTTGTC ATTTTGCAAA
  61 TCTGGTCAAC CCACCTCCGG TGAAAACTCC CAACCTCACC CCAGGGGGCA ATGACTAATT
 121 ACAAAACACA TTTTCTCTCG TTTTCGTCAA GCTCGCTGTC CCGCCCACAC A
```
Notes:
○ The WT sequence contains two variations of the target motif BHLHE40_disc2:
   1 CGCTGCGGCC CCAGCCC (bolded in the sequence),
   1 cCGCTGCGGC CCCAGCC (small-sized in the sequence)
○ In PERT1 and PERT2 sequences: BHLHE40_disc2 are removed respectively → both sequences are annotated as "Perturbed"
○ PERT3: Two variations of BHLHE40_disc2 are found:
   1 *C**GGCTCGACG CCCGCCC*** (italic in the sequence)
   1 GGCTCGACGC CCGCCC**C** (bolded in the sequence)
→ this sequence is annotated as "Not perturbed"

• **Figure 3:**

– In Figure 3A (Perturbation specificity), we used a sequence from PERT2 to illustrate the perturbation specificity:
>***WT***_*region255_chr14:63163083-63163254*:
```
   1 CACCGTACTC ACAACTACAG TGGCCAATGA AACGCACTTG GAAGTCACTT GGGTTTCTGG
```

61 CAGGACTCAC TGAAAGCTGA TCTGATCCTT TTGTCT**TC**TG CCCTGCCCCT TCTTCCTTTC
121 TGAAAGGCAA GTGTGGCCCT GCAGGTGGCC CATACTTCTT GCAACTACAG G

>*hit1_**PERT2**_chr14:25412372-25412543_ATF5_M2977_1.02* :

1 CACCGTACTC ACAACTACAG TGGCCAATGA AACGCACTTG GAAGTCACTT GGGTTTCTGG
61 CAGGACTCAC TGAAAGCTGA TCTGATCCTT TTGTCT***TCTG CCCTGCCCCC GAGCATCTTC***
121 TGAAAGGCAA GTGTGGCCCT GCAGGTGGCC CATACTTCTT GCAACTACAG G

Notes:

○ In the WT sequence, the target motif is ATF5_M2977_1.02:

1 TTCTTCCTTT

○ Four motifs in the WT sequence are overlapped with the target motif:

△ NF4_known12 (small-sized in the sequence):

1 TGCCCTGCCC CTTCTTCCTT TCTG

△ VDR_3 (italic in the sequence):

1 *CCTGCCCCTT CTTCCTTTCT*

△ IRF_disc4 (bolded in the sequence):

1 TGCCCTGCCC CTT

○ In the perturbation sequence, two motifs are removed (NF4_known12 and VDR_3). However, a variation of motif IRF_disc4 is found:

1 ***TCTGCCCTGC CCCCG***

– In Figure 3B (Newly introduced target motifs per sequence), we used a sequence from PERT3 to illustrate the newly introduced target motif:

>***WT**_region350_chr10:100206539-100206710*:

1 AGGACCGGAT CAACTCTCCG GGACCCCTAG GGACCCTACC TCACTTCCGG GAGGGTTGAA
61 GGGGGGCTCC GGAGGGAGGA TCGCCGCCCC CAGAGGGGAC AGCCCGGAGG CCCACGTACC
121 GGATCGCGGC GCGCACAGCG CCCCGCCTGC AGGAGCCCGG GCGCGCTTCC GGGTAGGACC
181 GCGTGACATT GCGTGAACCG A

>*hit2diff_1_**PERT3**_chr10:100206539-100206710_BHLHE40_disc2*:

1 AGGACCGGAT CAACTCTCCG GGACCCCTAG GGACCCTACC TCACTTCCGG GAGGGTTGAA
61 GGGGGGCTCC GGAGGGAGGA TCGCCGCCCC CAGAGGGGAC AGCCCGGAGG CCCACGTACC
121 GGATCGCGGC GCGCACAGCA TGAGCGGCCC CCGCGCCCGG ***GCGCGC**TTCC GGGTAGGACC
181 GCGTGACATT GCGTGAACCG A

Notes:

○ In the WT sequence, the target motif is BHLHE40_disc2:

CGCCCCGCCT GCAGGAG

○ In the perturbation sequence, two variations of motif BHLHE40_disc2 is introduced:

1 GCGGCCCCCG CGCCCGG (small-sized in the sequence)
1 *CCCGCGCCCG GGCGCGC* (bold italic in the sequence)

In this study, the sequences of non-motifs are:

○ non-motif_PERT1:

1 ACTAAAGAAT CTACAAATGA TCTCTCAAAT

◦ non-motif_PERT2:

```
1 CGAGCATCTT TAAGAGTTAG AGTAGGCAAA
```

◦ non-motif-RAND_PERT2:

```
1 ACTAAAGAAT CT
```

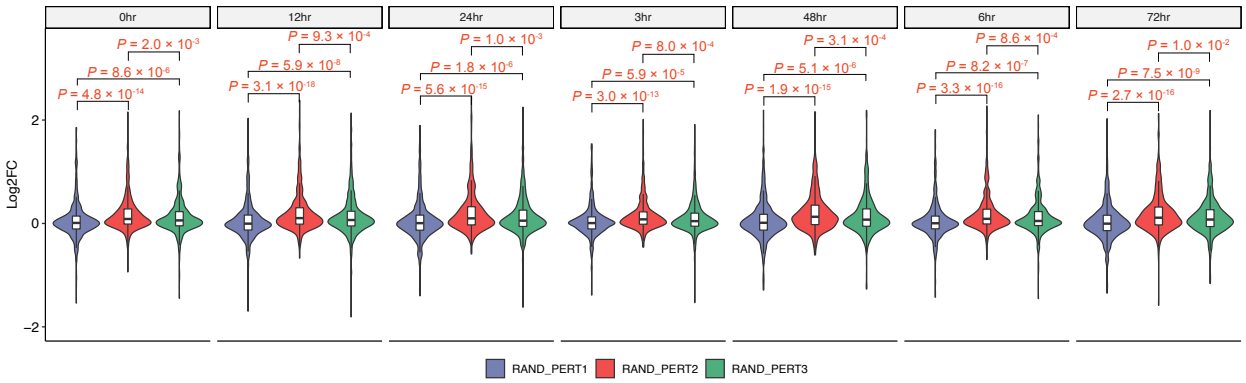
◦ non-motif-RAND_PERT2:

```
1 CGAGCATCTT TA
```

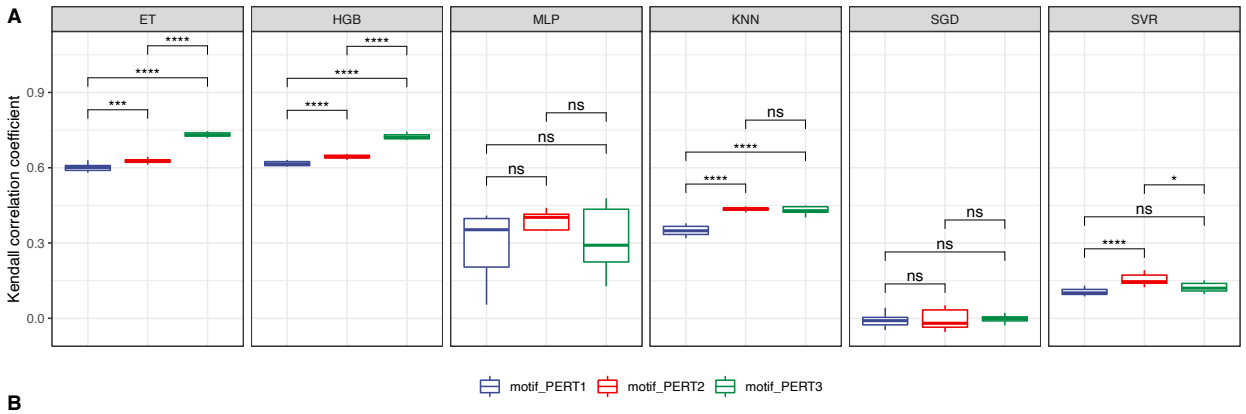
The exclusion criteria of low-quality sequences

In the barcode association step of this study, 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. The number of discarded sequences of each perturbation approach: PERT1 = 35, PERT2 = 32, PERT3 = 2. These sequences are marked as N/A in Figure 2A.

Supplementary Figures and Tables



Supplementary Figure 1: Performance of regression models. Comparison of the Log2FC of RAND sequences among three perturbation approaches. The Log2FC values are separated by time point before being compared among three perturbation approaches.

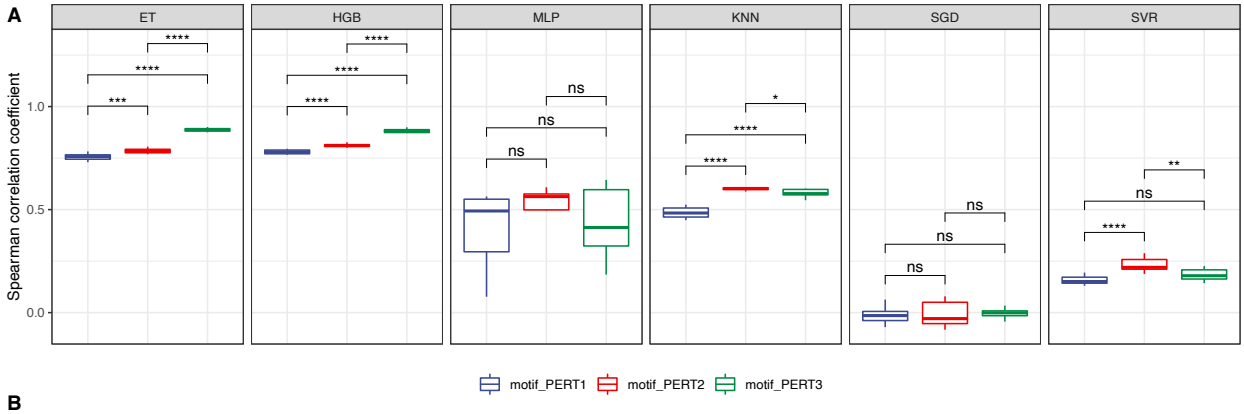


B

	ET	HGB	MLP	KNN	SGD	SVR
motif_PERT1	0.60 ± 0.02	0.62 ± 0.01	0.29 ± 0.14	0.35 ± 0.02	-0.01 ± 0.03	0.11 ± 0.01
motif_PERT2	0.63 ± 0.01	0.64 ± 0.01	0.35 ± 0.11	0.43 ± 0.01	-0.01 ± 0.04	0.15 ± 0.02
motif_PERT3	0.73 ± 0.01	0.72 ± 0.01	0.32 ± 0.12	0.43 ± 0.02	0.00 ± 0.02	0.12 ± 0.02

Supplementary Figure 2: Performance of regression models. **(A)** The Kendall correlation coefficients of different regression models. Asterisks/ns indicate levels of statistical significance, calculated by pairwise Wilcoxon rank sum tests (P-value < 0.05 *, < 0.01 **, < 0.001 ***, $<$ 0.0001 ****; ns, non-significant). **(B)** A summary of the mean $\pm$ standard deviation values for Kendall correlation coefficients of regression models.

A



B

	ET	HGB	MLP	KNN	SGD	SVR
motif_PERT1	0.76 ± 0.02	0.78 ± 0.01	0.40 ± 0.19	0.49 ± 0.03	-0.01 ± 0.04	0.16 ± 0.02
motif_PERT2	0.78 ± 0.01	0.81 ± 0.01	0.49 ± 0.15	0.60 ± 0.01	-0.01 ± 0.06	0.23 ± 0.03
motif_PERT3	0.89 ± 0.01	0.88 ± 0.01	0.44 ± 0.16	0.58 ± 0.02	0.00 ± 0.03	0.18 ± 0.03

Supplementary Figure 3: Performance of regression models. **(A)** The Spearman correlation coefficients of different regression models. Asterisks/ns indicate levels of statistical significance, calculated by pairwise Wilcoxon rank sum tests (P-value < 0.05 *, < 0.01 **, < 0.001 ***, < 0.0001 ****; ns, non-significant). **(B)** A summary of the mean $\pm$ standard deviation values for Spearman correlation coefficients of regression models.

References

[1] Fumitaka Inoue, Anat Kreimer, Tal Ashuach, Nadav Ahituv, and Nir Yosef. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell*, 25(5):713–727.e10, November 2019.

[2] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011.

[3] Pouya Kheradpour, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S. Mikkelsen, and Manolis Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 23(5):800–811, May 2013.

[4] Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J. M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431–1443, September 2014.

[5] Jason C. Klein, Vikram Agarwal, Fumitaka Inoue, Aidan Keith, Beth Martin, Martin Kircher, Nadav Ahituv, and Jay Shendure. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nature Methods*, 17(11):1083–1091, November 2020.