Published online DD MM YYYY

Nucleic Acids Research, YYYY, Vol. xx, No. xx ***1–16***
doi:10.1093/nar/gkn000

Best practices for perturbation MPRA–a computational evaluation framework of sequence design strategies

Jiayi Liu [1,2,3], Tal Ashuach [4], Fumitaka Inoue [5], Nadav Ahituv [6,7], Nir Yosef [8,9,10], and Anat Kreimer [2,3], *

[1]Graduate Programs in Molecular Biosciences, Rutgers, The State University of New Jersey, 604 Allison Rd, Piscataway, NJ, 08854, USA [2]Department of Biochemistry and Molecular Biology, Rutgers, The State University of New Jersey, 604 Allison Road, Piscataway, NJ, 08854, USA [3]Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, 679 Hoes Lane West, Piscataway, Piscataway, NJ, 08854, USA [4]Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California, Berkeley, 387 Soda Hall, Berkeley, CA, 94720, USA [5]Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Faculty of Medicine Building B, Yoshidatachibanacho, Sakyo Ward, Kyoto, 606-8303, Japan [6]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, 513 Parnassus Ave, CA, 94143, USA [7]Institute for Human Genetics, University of California, San Francisco, 513 Parnassus Ave, CA, 94143, USA [8]Department of Systems Immunology, Weizmann Institute of Science, 234 Herzl Street, Rehovot 7610001 Israel [9]Chan-Zuckerberg Biohub, 499 Illinois St, San Francisco, CA, 94158, USA [10]Department of Systems Immunology, Ragon Institute of MGH, MIT, and Harvard Institute of Science, 400 Technology Square, Cambridge, MA, 02139, USA

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

The advent of the perturbation-based massively parallel reporter assays (MPRAs) technique has enabled delineating of the roles of non-coding regulatory elements in orchestrating gene expression. However, computational efforts remain scant to evaluate and establish guidelines for sequence design strategies for perturbation MPRAs. Here, we propose a framework for evaluating and comparing various perturbation strategies for MPRA experiments. Under this framework, we benchmark three different perturbation approaches from the perspectives of alteration in motif-based profiles, consistency of MPRA outputs, and robustness of models that predict the activities of putative regulatory motifs. Although our analyses show similar while significant results in multiple metrics, the method of randomly shuffling nucleotides outperform the other two methods. Thus, we still recommend designing sequences by randomly shuffling the nucleotides of the perturbed site in perturbation-MPRA. The evaluation framework, together with the benchmarking findings in our work, creates a resource of computational pipelines and illustrates the promise of perturbation-MPRA for predicting non-coding regulatory activities.

INTRODUCTION

Advances in high-throughput technologies have allowed a detailed characterization of the human genome, including regulatory elements such as enhancers which contain binding motifs for transcription factors (TFs) and play a central role in the transcriptional regulation of gene expression. Aberrations in the non-coding regions of the genome have been linked to numerous polygenic disorders such as cancer, heart, and neurological disorders (1, 2, 3), making the study of non-coding regions an important area of research.

However, linking the non-coding genome to the etiology of diseases is largely limited by the low throughput of conventional "luciferase reporter assays", especially when numerous non-coding regions are of interest. To address this challenge, massively parallel reporter assays (MPRAs) were developed to simultaneously measure the activity of thousands of regulatory elements and their variants in a single experiment (4, 5). Furthermore, a perturbation-based MPRA approach was introduced to elucidate the regulatory effects of transcription factor (TF) binding motifs, instead of single nucleotide variants (6, 7, 8). The essence of this technique is to analyze the change in the transcription activity of reporter genes after altering the DNA sequence of putative functional regulatory regions.

In our recent studies, we have utilized utilized the perturbation MPRA technique to successfully identify over 500 non-coding genomic regions that temporally regulate gene transcription during neural differentiation (9, 10). Although the potential of perturbation MPRA has been widely acknowledged, there have been limited attempts to evaluate different design strategies of the tested sequences.

Motivated by the scarcity of the gold standard for DNA sequence designing strategies for the MPRAs technique, we propose a framework for assessing and comparing perturbation strategies (Figure 1). Under this framework, we benchmark three different perturbation methods using a publicly available dataset we recently generated (9, 10).

*To whom correspondence should be addressed. Tel: +1 848-445-9809; Email: kreimer@cabm.rutgers.edu

Briefly, this dataset includes 591 wild-type (WT) sequences, 2,146 motif perturbation sequences, with each sequence perturbed using three different perturbation approaches, and 591 negative control sequences. The perturbation methods, in short, either replaced the target motifs with two different "non-motif" sequences (PERT1 and PERT2), or simply shuffled the nucleotides of target motifs (PERT3).

For benchmarking, we first define five indices to comprehensively evaluate the achievement of the perturbation goals. These indices include, for example, the perturbation rate that indicates the impact on the target motifs both *in-situ* and *ex-situ*, and the specificity index that indicates the proportion of WT motifs that survives the perturbation processes, etc. By comparing these indices, we found that the PERT3 exhibits the highest specificity while the lowest perturbation rate. Next, we compared the consistency of MPRA outputs, both in functional regulatory site (FRS) identities and numeric regulatory effects. Our analyses revealed a high correlation among the three perturbation methods, but we also found a constant bias in the results of PERT1 and PERT2. This is likely due to their insertion of fixed sequences, which may introduce systematic biases to the assayed regions. Finally, we extracted multiple genomic features for each tested sequence and used the difference in the features between the perturbation sequences and their WT equivalents as independent variables to fit predictive machine-learning models. Our results for these predictive models demonstrated the robustness of both classifiers and regressors based on PERT3 data.

To the best of our knowledge, this is the first study that assesses and compares different perturbation methods of MPRA experiments. Our study fills this gap by constructing a blueprint evaluation framework for perturbation sequence designing strategies. Additionally, our results provide guidance for establishing a gold standard of perturbation MPRAs techniques, and our prediction pipeline holds great promise for further computationally identifying functional genomic regulatory regions.

MATERIALS AND METHODS

Dataset overview

We utilized a publicly available dataset of perturbation MPRA we recently published (9). The MPRA experiment was performed in the human embryonic stem cell line across seven time points after neural differentiation induction (0, 3, 6, 12, 24, 48, and 72 hours). Specifically, it assayed three groups of genomic sequences: **a)** Wild type group: 591 wild-type sequences (denoted as "WT"): each WT sequence represents a 171-nucleotide genomic region whose regulatory activity differs over time (10), **b)** Motif perturbation group: 2,146 sequences, each containing a single-perturbed motif within the genomic region of its WT equivalent. And each sequence is perturbed using three different perturbation approaches (denoted as "motif_PERT1", "motif_PERT2" and "motif_PERT3"):

1. PERT1: a motif is replaced with the prefix of an artificially scrambled motif, while using three bp downstream and upstream of the motif in the WT sequence. Under this strategy, the sequence of the perturbed motif is original_sequence_start"scrambled motif1 prefix"original_sequence_end.
2. PERT2: similar to PERT1, a motif is replaced with the prefix of another artificially scrambled motif, while keeping the WT starting and ending sequences. Under this strategy, the sequence of the perturbed motif is original_sequence_start"scrambled motif2 prefix"original_sequence_end.
3. PERT3: the motif is scrambled by randomly shuffling its nucleotides.

and **c)** Negative control group 1: 591 scrambled sequences (denoted as "SCRAM"). Scrambled sequences are based on WT sequences with shuffled nucleotides, creating a set of negative controls, **d)** Negative control group 2: these are a set of all the 591 WT sequences where we perturbed a sub-sequence in the length of the average motif (12 bp) in a random location within the WT sequence using the same three perturbation methods (denoted as "non-motif_PERT1","non-motif_PERT2", and "non-motif_PERT3"). The non-motifs and motifs are perturbed using the same three perturbation approaches.

The experimental read-out of the perturbed sequences is then subjected to the MPRAnalyze (11) and MPRAflow (12) tools to assess the motif regulatory effect over time, which is represented by the Log2 fold changes (Log2FC) of PERT read-outs compared to the WT and SCRAM at each of the seven time points. The sequences are further classified into two according to the Log2FC values: activating (Log2FC $>$ 0) and repressing (Log2FC $<$ 0).

Additionally, to identify the functional regulatory sites (FRS), we used MPRAnalyze (11) to apply a set of four filters to the PERT sequences (9):

1. At one or more time points, the activity of a PERT sequence significantly deviates from its WT equivalent.
2. The temporal activity of a PERT sequence significantly deviates from its WT equivalent.
3. The activity of either a PERT sequence (at one or more time points) or a WT sequence (across all the time points) is significantly higher than its corresponding SCRAM negative control sequence.
4. The temporal activity of either a PERT or a WT sequence is significantly higher than its corresponding SCRAM negative control sequence.

The target motif of a sequence will be labeled as an FRS if the sequence passes all four filters and shares consistent effects (either activating or regressing) in PERT3 and either PERT1 or PERT2. In summary, the MPRA output consists of the numeric regulatory effect (Log2FC) and the multi-class FRS identities at seven time points. These two output types are used as input variables for training the prediction models.

Metrics for assessing motif-based profiles

Hit rate (HR) A "hit" sequence indicates the *in-situ* removal of its target motif (*in-situ* removal = "genomic location-specific removal"). In detail, we define "hit" as the target motif of a perturbed sequence that does not occur in the scanning results of the Find Individual Motif Occurrences (FIMO) tool (13), matched by the motif name, DNA strands, and genomic coordinates; otherwise, it's a "fail." The hit rate of $\mathrm{PERT_i}$ is denoted as $\mathrm{HR_i}$:

$$\mathrm{HR_i} = \frac{\mathrm{N_{Hit_i}}}{\mathrm{N_i}}, \tag{1}$$

where $\mathrm{N_{Hit_i}}$ is the number of "hit" sequences and $\mathrm{N_i}$ is the total number of designed sequences in $\mathrm{PERT_i}$.

Perturbation rate (PR) A "perturbed" sequence indicates that all motifs that match the target motif ID are removed within the designed genomic region. In detail, we define a sequence as "perturbed" if the name of the target motif occurs in its FIMO scanning results, regardless of its genomic position. Then, the perturbation rate of $\mathrm{PERT_i}$ is formulated as:

$$\mathrm{PR_i} = \frac{\mathrm{N_{Perturbed_i}}}{\mathrm{N_i}}, \tag{2}$$

where $\mathrm{N_{Perturbed_i}}$ is the number of "perturbed" sequences and Ni is the total number of designed sequences in $\mathrm{PERT_i}$.

Perturbation specificity (PS) To assess how many WT motifs are impacted by the perturbation, we introduce the "perturbation specificity" metric. For the designed sequence j of $\mathrm{PERT_i}$, its perturbation specificity is formulated as:

$$\mathrm{PS_{ij}} = \frac{\mathrm{M_{survived_{ij}}}}{\mathrm{M_{WT_{ij}}}}, \tag{3}$$

where $\mathrm{M_{WT_{ij}}}$ is the number of motifs that overlap with the target motif in the corresponding WT sequence of designed sequence j of $\mathrm{PERT_i}$, and $\mathrm{M_{survived_{ij}}}$ is the occurrence of wild-type motifs that are still present within the designed sequence j of $\mathrm{PERT_i}$. Both $\mathrm{M_{WT_{ij}}}$ and $\mathrm{M_{survived_{ij}}}$ are obtained from FIMO scanning results.

Newly introduced target motifs per sequence (NTM) Since the perturbation process alters the orders of nucleotides, some of the newly introduced motifs may be identical to the target motifs. To assess such impact of the perturbation methods, we calculated and compared the "number of newly introduced target motifs per sequence" among the three perturbation methods. For $\mathrm{PERT_i}$, its "newly introduced target motifs per sequence" metric is formulated as:

$$\mathrm{NTM_i} = \frac{\mathrm{q_i}}{\mathrm{N_i}}, \tag{4}$$

where $\mathrm{q_i}$ is the number of newly introduced motifs that are identical to the target motif IDs in $\mathrm{PERT_i}$, and $\mathrm{N_i}$ is the total number of designed sequences in $\mathrm{PERT_i}$.

General alteration in the number of motifs To assess the non-specific impacts of the perturbation, we obtained and compared these indices among the three perturbation methods:

1. The number of gained motifs
2. The number of lost motifs
3. The net change in the number of motifs

Consistency analysis of MPRA outputs The MPRA outputs consist of two parts: the multi-labeled FRS identities and the numerical regulatory effects. To analyze the consistency of FRS identities, we counted the number of overlapped and unique activatorsregressors that are specific to their genomic coordinates and DNA strands across three perturbation methods. And the results are visualized by an UpSet plot (14). As for the agreement in numerical regulatory effects, we tested the correlation of Log2FCs between any two of the three perturbation methods using three correlation tests: Pearson r correlation, Spearman's rank correlation, and Kendall's rank correlation test.

Features extraction for designed sequences

The features are a major determinant of the performance of predictive models (15, 16). The features used in this work can be grouped into two main categories: sequence-based features and time-specific features.

Group A: sequence-based features Since this group of features is based on the nucleotide sequences, each assayed sequence, either WT or perturbed, has its own set of features:

- DNA 5-mer frequencies: 1,024 features indicating the counts of all possible nucleotide 5-mers.
- #5-mers: a single feature summarizing the number of distinct 5-mers.
- DeepBind scores: 515 predicted scores of all pre-trained DeepBind models for transcription factor (TF) binding (17).
- #DeepBind-top: a single feature summarizing the number of models above the 90^{th} percentile across all the DeepBind models for TF binding (17).
- DeepSEA scores: 21,907 chromatin profiles (transcription factor, histone marks, and chromatin accessibility profiles across a wide range of cell types) from the underlying DeepSEA learning model (16).
- #DeepSea-top: a single feature summarizing the number of chromatin profiles above the 90^{th} percentile across all the DeepSEA profiles (16).
- DNA shape metrics: 13 predicted DNA shape features, which are: helix twist (HelT), Rise, Roll, Shift, Slide, Tilt, Buckle, Opening, propeller twist (ProT), Shear, Stagger, Stretch, and minor groove width (MGW) (18, 19).

- Max polyA/polyT lengths: two features indicating the length of the longest polyA and polyT subsequences, respectively.
- #ENCODE/CIS-BP motifs: 4,706 features, showing the number of significant DNA-binding ENCODE/CIS-BP (20, 21, 22) motifs from simple DNA-binding motif scoring using the Find Individual Motif Occurrences (FIMO) tool (13).
- ENCODE/CIS-BP motif summaries: four features indicating the number of motifs, and the maximum number of ENCODE/CIS-BP motifs within a 20 bp window in the sequence, as determined by FIMO scanning algorithm (13, 21, 22).
- #TF family: fourteen features indicating the frequency of major TF families based on the FIMO results of ENCODE/CIS-BP scannings, which are: Basic Domain Group, Beta-Scaffold Factors, Helix-turn-helix, Other Alpha-Helix Group, Unclassified Structure, and Zinc-Coordinating Group (23).

For each perturbed sequence, we subtract its sequence-specific features from that of its WT equivalent. Additionally, we calculate the Levenshtein similarity scores between the perturbed sequences and their respective correspondent WT sequences (24, 25). In total, 28,189 features are yielded from group A.

These differences in features (denoted as Δ"[feature name]", e.g., Δ#5-mers), along with the Levenshtein similarity scores, are then subject to the feature normalization process (see Section "Feature normalization").

Group B: time-specific features The time-specific features used in this study are the experimental read-outs of WT sequences (10). These features include the signals of three genomic assays at seven time points (0, 3, 6, 12, 24, 48, and 72 hours):

- ATAC-seq: the normalized number of reads using DESeq2 (26) from an overlapping ATAC-seq peaks within the designed genomic region
- H3K27ac ChIP-seq, the normalized number of reads using DESeq2 (26) from an overlapping H3K27ac peaks within the designed genomic region
- RNA-seq: mRNA expression of the nearest gene to the designed region

In total, three features are yielded from group B. For each perturbed sequence, we use the time-specific feature of its corresponding WT sequence as its feature to fit prediction models.

Feature normalization

Performing principal component analysis (PCA) is a common technique to reduce the number of features in high-dimensional data to avoid over-fitting and improve the generalization performance of machine learning models. In this case, PCA was applied to the large number of group A features (28,189) to reduce them into a smaller set of principal components (PCs) that capture the maximum amount of variability in the data. By selecting the number of PCs such that they explain at least 99% of the variance in the data, the most important information in the original features is retained while reducing their dimensionality.

In this study, we employed PCA to transform the 28,189 group A features into 1,500 PCs for each perturbation method. Together with the time-specific features of group 2, a total of 1,503 features were used as input for subsequent prediction tasks. This approach helps to prevent over-fitting and improves the accuracy of the machine learning models.

Calculation of the feature importance scores

We first defined the importance score I of feature i as the largest loading score of feature i across 1,500 PCs. In particular, from the PCA step, we obtain a matrix L to denote the loadings matrix that explains the correlations between the original features and the PCs. L is a $28{,}189 \times 1{,}500$ matrix with rows representing features and columns representing 1,500 PCs. For feature i, its loading score on the j^{th} dimension is denoted as L_{ij}. We then define the importance score I of the feature i as its largest loading score across the 1,500 PCs:

$$I_i = \max\{L_{i1}, L_{i2}, ..., L_{ij}\}, j \in \{1, ..., 1500\} \tag{5}$$

Gene ontology analysis

We conducted the Gene ontology (GO) over-representation analysis using the genes corresponding to the top 2,500 important TF binding features. The results were determined using the R package ClusterProfiler (27). The significance of GO terms was defined as an FDR-adjusted $p < 0.05$.

Model training

Classification models We utilized six classification models to predict the FRS identity of perturbed sequences:

1. SGD: linear SVM classifiers with stochastic gradient descent (SGD) training (28)
2. SVC: C-Support vector classifiers (29)
3. KNN: classifiers based on k-nearest neighbors voting (30)
4. ET: ExtraTrees classifiers (31)
5. HGB: histogram-based gradient boosting classifiers (32)
6. MLP: multilayer perceptron classifiers (33)

All classifiers were run with the default settings of the scikit-learn package (34). The 1,503 normalized feature values were used as input. To generate target values, the FRS identity labels at seven time points were concatenated and stacked into a single variable.

Regression models

1. SGD: SGD: linear regressors fitted by minimizing a regularized empirical loss with SGD training (28)
2. SVC: SVR: Epsilon-Support vector regressors (29)
3. KNN: regressors based on k-nearest neighbors voting (30)
4. ET: ExtraTrees regressors (31)
5. HGB: histogram-based gradient boosting regressors (32)
6. MLP: multilayer perceptron regressors (33)

All regressors were run with the default settings of the scikit-learn package (34). The 1,503 normalized feature values were used as input. The Log2FCs at seven time points were concatenated and stacked into a single variable, and regarded as target values.

The randomized 10-fold cross-validation test

We performed 10-fold cross-validation tests to evaluate the performances of different models. A 10-fold cross-validation test was chosen as it provides a good balance between minimizing bias and reducing variance. In detail, the dataset is randomly partitioned into ten subsets, with one subset utilized as the testing dataset and the other nine together as the training data set. This procedure was conducted 10 times, with each subset being used once as a testing dataset to generate ten models. The average performance of these ten models was used to evaluate the performance of the different models.

To ensure a fair and objective comparison among the models, we strictly implemented their algorithms and optimized parameters to build models on the same training dataset and subsequently benchmark their performance on the independent test datasets.

Model performance measures

The performance of classification models is evaluated using the area under the receiver-operating characteristic curve (AUROC). For the regression models, we evaluated their performance using three correlation tests: Pearson, Spearman, and Kendall. Specifically, we tested the correlation between the predicted Log2FC values and the observed Log2FC values for each fold.

Statistical tests

For the motif-based profile metrics, the Kruskal–Wallis one-way analysis of variance and post-hoc pairwise Dunn's multiple comparisons test were used to identify statistically significant differences in continuous variables, including the perturbation specificity and the number of gained/lost motifs. Moreover, the pairwise Fisher's exact test was conducted to compare the count data, including hit and perturbation rates. The pairwise exact binomial test was performed to compare newly introduced target motifs per sequence (NTM).

For the consistency analyses, the correlation of Log2FCs was indicated by three correlation coefficients: Pearson's r, Spearman's ρ, and Kendall's τ coefficient. The P values of correlation tests were subsequently adjusted for multiple comparisons at seven different time points by the Benjamini-Hochberg method.

For the performance evaluation of prediction models, we performed pairwise Wilcoxon rank sum tests on the AUROC and correlation coefficients. For all pairwise tests, a threshold of 0.05 was applied to the P values adjusted by the Benjamini-Hochberg method. And an α level was considered 0.05 for all statistical tests in this study.

RESULTS

To evaluate the three perturbation methods, we first defined five motif-based metrics: *1) hit rate*, representing the rate of *in-situ* motif perturbation, defined as the proportion of designed sequences that successfully eliminate the target motif at the target genomic locale, *2) perturbation rate*, which represents the rate of both *ex-situ* and *in-situ* motif perturbation and is defined as the proportion of designed sequences that eliminate all the motifs that match the target motif ID within the 171-nucleotide genomic region, *3) perturbation specificity*, indicating the global impact of perturbation on all the motifs that lie within the perturbed sequence, and is defined as the proportion of WT motifs that are still found in the perturbation sequence, *4) newly introduced target motifs per sequence*, which reflects the occurrence of gained motifs that are identical to the target motif ID, and is calculated by dividing the total number of such gained motifs by the total number of perturbation sequences, *5) non-specific changes in the number of motifs*, which include the number of gained, lost motifs, as well as the net change in the number of motifs within the perturbation sequence. We then assess the differences in these metrics across the three perturbation methods (Figure 1, part I). Next, we assess the important features representing variability among all perturbation methods (Figure 1, part II). Third, we compare the consistency of MPRA outputs (Figure 1, part III). Finally, to evaluate the generalizability in referencing the non-coding regulatory activity of the three perturbation methods, we compare how different prediction models perform across the three perturbation methods (Figure 1, part IV).

All perturbation methods achieve high *in-situ* hit rates

The basic goal of a motif perturbation is to remove the target motif at the target genomic location. To assess how well each perturbation method is in reaching this goal, we computationally identified the occurrences of the motifs in perturbed sequences, by using the FIMO tool 13 scanning results and matching the motif names, DNA strands, and genomic coordinates (Section "Methods").

If a perturbed sequence yields a "non-occurrent" result, it is defined as a "hit" indicating a successful perturbation, otherwise a "fail" (Section "Methods"; Figure 2A). We then calculated and compared the proportion of hit and fail sequences for each perturbation method (equation **1**). The hit rates of PERT1 and PERT2 are similar ($\mathrm{HR}_1 = 98\%$, $\mathrm{HR}_2 = 99\%$), and both are significantly higher than that of PERT3 ($\mathrm{HR}_3 = 98\%$, pairwise Fisher's exact test, PERT1 vs. PERT2, $P = 1.00$; PERT1 vs. PERT3, $P = 1.42 \times 10^{-3}$;

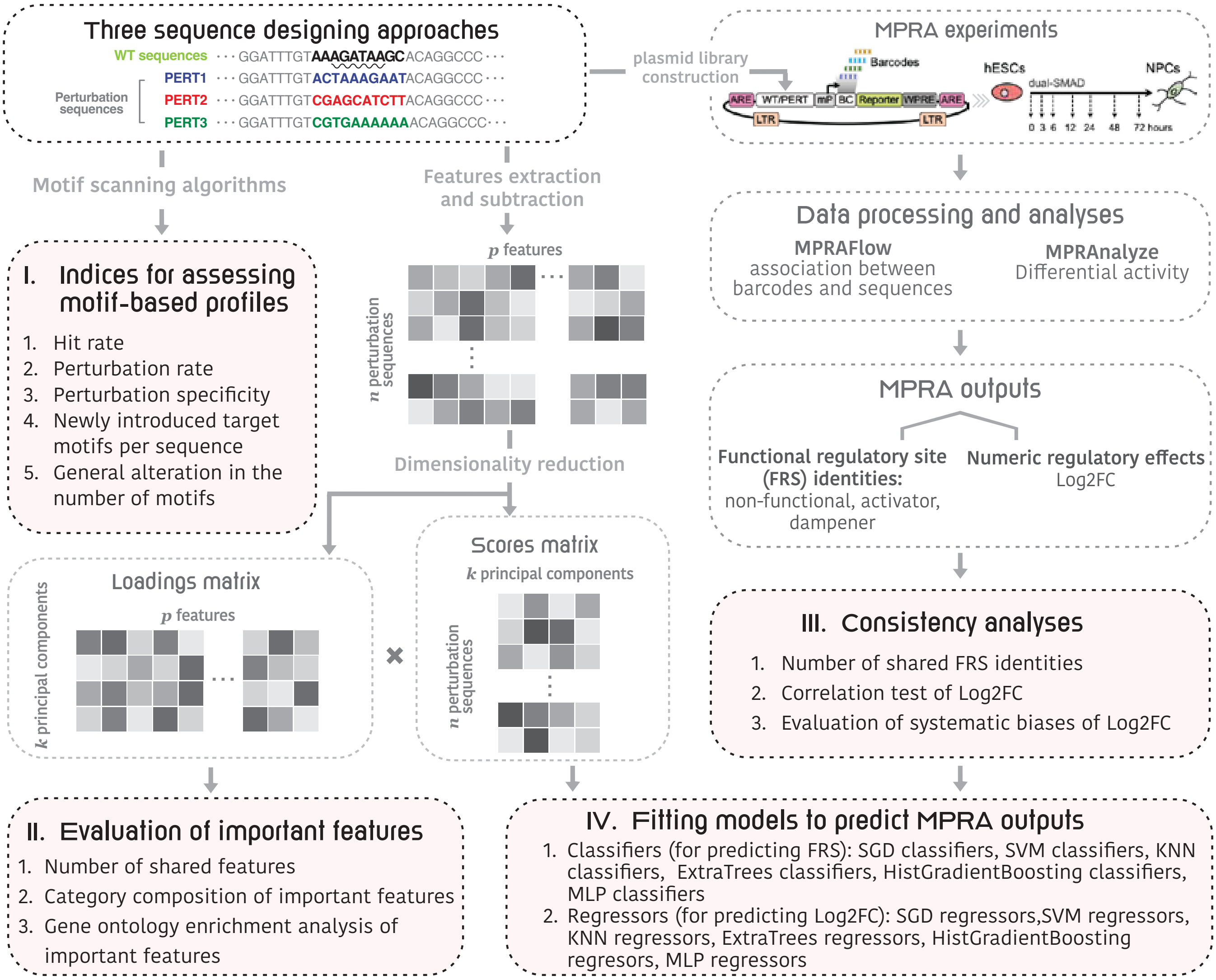

Three sequence designing approaches
WT sequences ···GGATTTGTAAAGATAAGCACAGGCCC···
Perturbation sequences
PERT1 ···GGATTTGTACTAAAGAATACAGGCCC···
PERT2 ···GGATTTGTCGAGCATCTTACAGGCCC···
PERT3 ···GGATTTGTCGTGAAAAAAACAGGCCC···
plasmid library construction
MPRA experiments
Barcodes
ARE WT/PERT mP BC Reporter WPRE ARE
LTR LTR
hESCs
dual-SMAD
NPCs
0 3 6 12 24 48 72 hours
Motif scanning algorithms
Features extraction and subtraction
p features
n perturbation sequences
Dimensionality reduction
I. Indices for assessing motif-based profiles
1. Hit rate
2. Perturbation rate
3. Perturbation specificity
4. Newly introduced target motifs per sequence
5. General alteration in the number of motifs
Data processing and analyses
MPRAFlow
association between barcodes and sequences
MPRAnalyze
Differential activity
MPRA outputs
Functional regulatory site (FRS) identities:
non-functional, activator, dampener
Numeric regulatory effects
Log2FC
Loadings matrix
p features
k principal components
×
Scores matrix
k principal components
n perturbation sequences
III. Consistency analyses
1. Number of shared FRS identities
2. Correlation test of Log2FC
3. Evaluation of systematic biases of Log2FC
II. Evaluation of important features
1. Number of shared features
2. Category composition of important features
3. Gene ontology enrichment analysis of important features
IV. Fitting models to predict MPRA outputs
1. Classifiers (for predicting FRS): SGD classifiers, SVM classifiers, KNN classifiers, ExtraTrees classifiers, HistGradientBoosting classifiers, MLP classifiers
2. Regressors (for predicting Log2FC): SGD regressors,SVM regressors, KNN regressors, ExtraTrees regressors, HistGradientBoosting regresors, MLP regressors

Figure 1. An outline of the framework for evaluation of perturbation-bases massively parallel assays technique. In the "Three sequence designing approaches" box, we used the "GATA_known9" motif as an example. In detail, the GATA motifs are a group of sequences conforming to the consensus WGATAR (W = A or T and R = A or G) (marked by the wavy underline), that can be recognized and bound by GATAbinding transcription factors (35).

PERT2 vs. PERT3, $P = 1.42 \times 10^{-3}$). Still, all three PERTs exhibit high hit rates of over 98% (Figure 2B).

The non-location-specific perturbation rate of PERT3 is the lowest

Apart from the basic goal, one of the advanced goals of motif perturbation is to reduce the regulatory activity of the target motif to the baseline, that is, to eliminate all the motifs that are identical to the target motif ID within the 171-nucleotide genomic region of perturbation sequence. Hence, we further quantified the occurrence of the target motif in each "hit" sequence using the FIMO scanning results, by matching only the motif name and not its location. Sequences were defined as "perturbed" if no designed target motif was found within their genomic region, and the perturbation rate was then calculated as the proportion of 'perturbed' sequences (Figure 2C). In simple words, this metric indicates the rate of both *ex-situ* and *in-situ* motif perturbation, that is not specific to the target genomic location (Section "Methods", equation **2**).

Comparing the perturbation rate of the three PERTs, we found that PERT1 and PERT2 possess similar perturbation rates of over 80%. Although the perturbation rate of PERT3 is significantly lower than those of the other two , it is still as high as 79% (Figure 2D, $PR_1 = 84\%$, $PR_2 = 83\%$, $PR_3 = 79\%$; pairwise Fisher's exact test, PERT1 vs. PERT2, $P = 0.649$; PERT1 vs. PERT3, $P = 8.79 \times 10^{-5}$; PERT2 vs. PERT3, $P = 4.58 \times 10^{-4}$). These results indicate that the strategic design of perturbation sequences (PERT1 and PERT2), instead of simply shuffling the nucleotide sequences (PERT3), leads to a higher chance of perturbing non-location-specific target motifs within genomic regions.

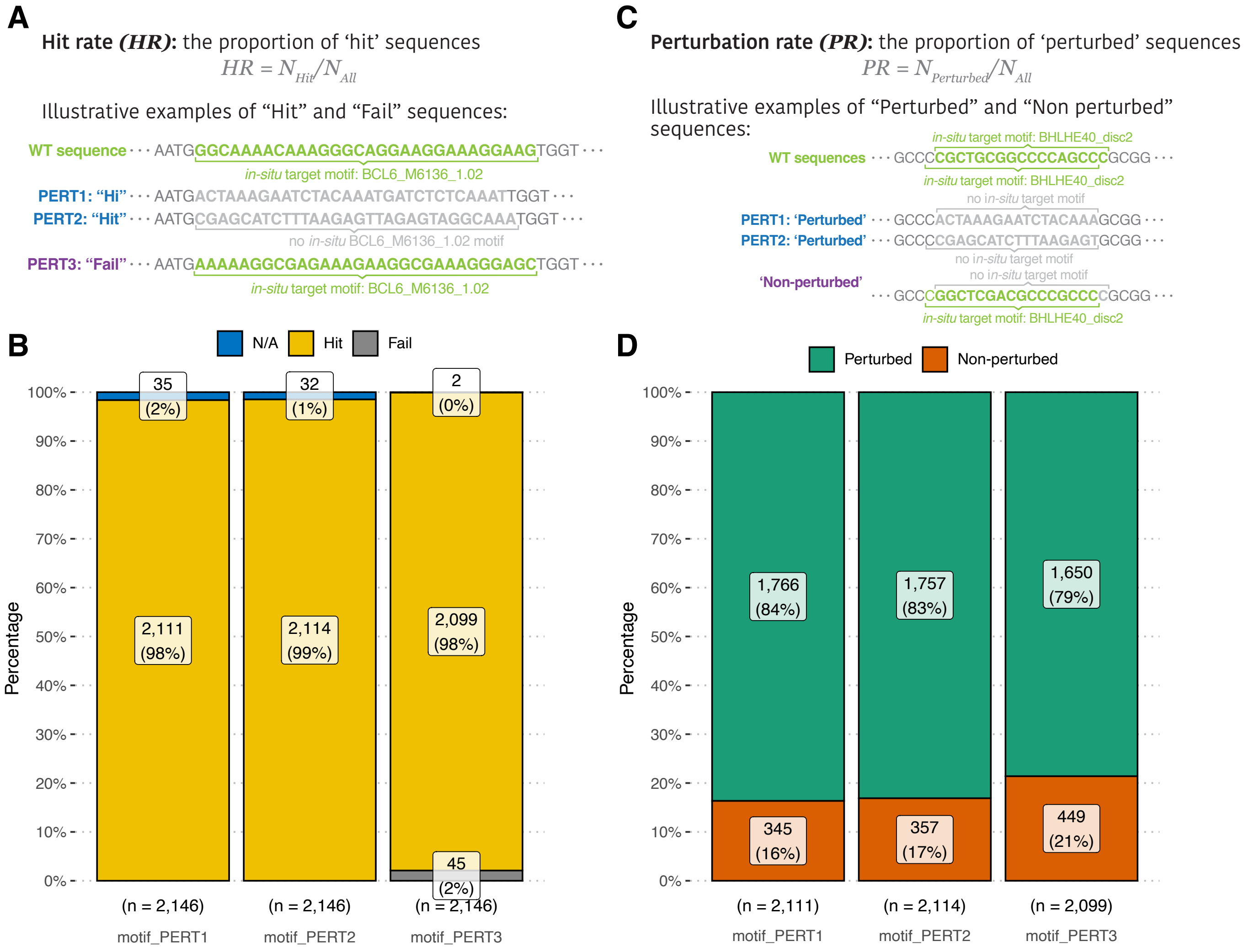


Figure 2. Evaluations of perturbation-wise metrics. **(A)** Toy examples of "hit" and "fail" sequences. **(B)** A comparison of hit rates among three perturbation methods. **(C)** Toy examples of "perturbed" and "non-perturbed" sequences. **(D)** A comparison of perturbation rates among three perturbation methods.

Perturbation specificity are similar among three methods

Another advanced goal of motif perturbation is to keep the impact on the overall motifs as low as possible: since the perturbation process essentially alters the DNA sequence within a certain range of the genome, the motifs that overlap with the target motifs are likely to be affected. To assess such a global impact of the perturbation on all the motifs that lie within the perturbation sequence, we introduced the perturbation specificity metric. It is defined as "the proportion of WT motifs that are still present within the genomic region after perturbation" (Section "Methods", equation **3**, Figure 3).

Comparing the perturbation specificity among three PERTs, we found that all three perturbation methods vastly affect the WT motifs. Namely, only 10% of the overlapping WT motifs "survived" the perturbation processes. Specifically, PERT3 has the highest perturbation specificity, which implies that randomly shuffling nucleotides exerts the least overall impact within the genomic regions of perturbed sequences (Figure 3B, $PS_1=7\%$, $PS_2=7\%$, $PS_3=11\%$; pairwise Dunn's test, PERT1 vs. PERT2, $P=5.73\times10^{-3}$; PERT1 vs. PERT3, $P=8.95\times10^{-27}$; PERT2 vs. PERT3, $P=1.14\times10^{-15}$).

On the other hand, another advanced goal is to avoid "creating" target motifs in the perturbation sequences. To this end, we sought to investigate which perturbation approach introduces the highest number of new motifs that are identical to the target motif ID. We defined the newly introduced target motifs per sequence metric, which is calculated by dividing the total number of "newly introduced target motifs" by the total number of sequences for each perturbation method (Section "Methods", equation **4**, Figure 3C). The highest metric is produced by PERT3, indicating that shuffling the nucleotides increases the probability of generating the same motifs as the target ones (Figure 3D, $NTM_1=0.0043$, $NTM_2=0.0085$, $NTM_2=0.17$; pairwise exact binomial test, PERT1 vs. PERT2, $P=0.122$; PERT1 vs. PERT3, $P=2.35\times10^{-92}$; PERT2 vs. PERT3, $P=1.73\times10^{-82}$).

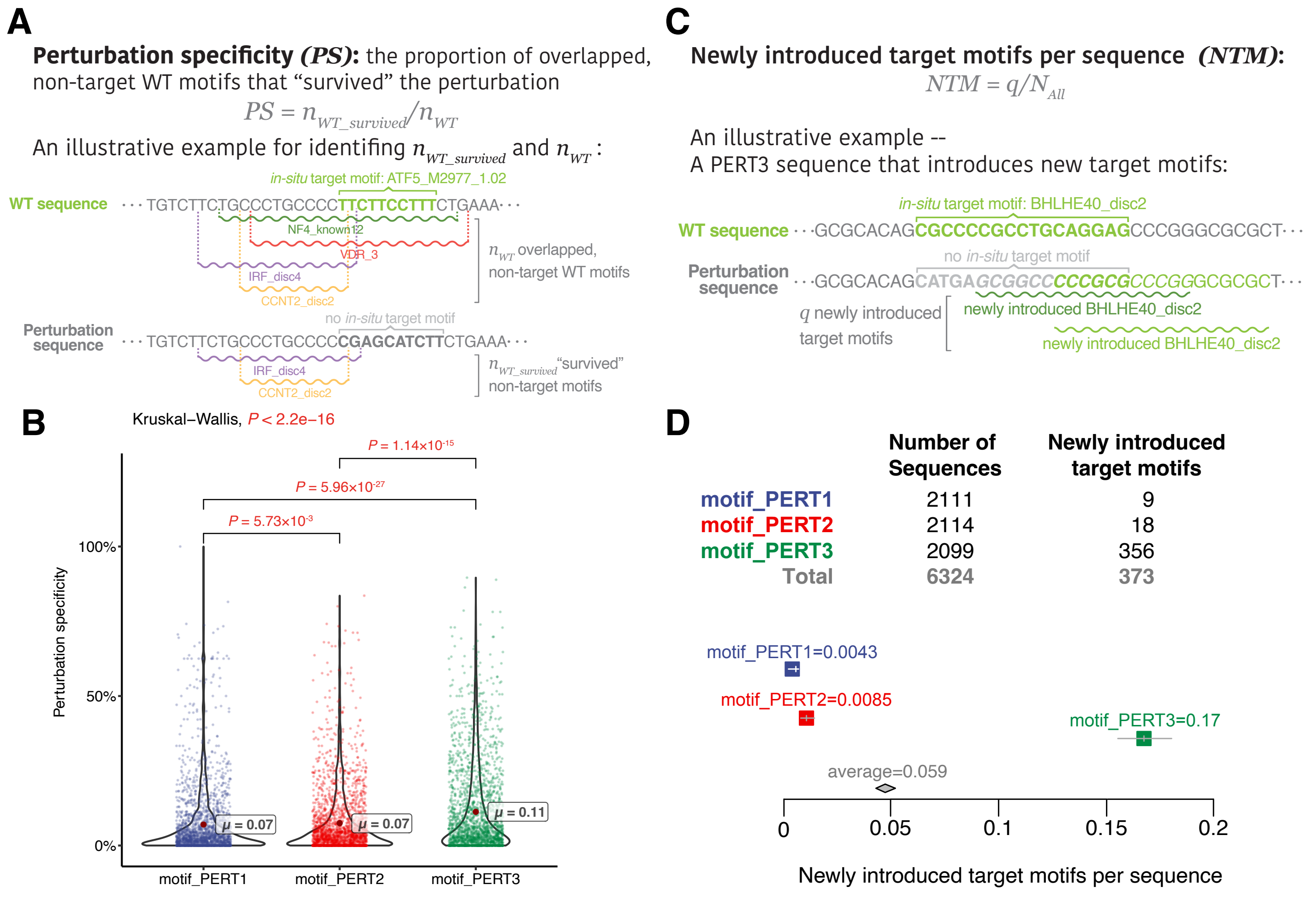


Figure 3. Evaluations of motif-based metrics. **(A)** An toy example of calculating perturbation specificity. **(B)** A comparison of perturbation specificity among three perturbation methods. Significant P values ($P < 0.05$) are shown in red. **(C)** An toy of calculating "newly introduced target motifs per sequence". **(D)** A comparison of "newly introduced target motifs per sequence" among three perturbation methods.

All three perturbation approaches vary in motif gain/loss

To gain a better perturbation effect, the impacts that are non-specific to the target motifs should also be minimized as much as possible. To address such impacts, we evaluated the overall motifs gained or lost across motif perturbation approaches (Figure 4A), and found that PERT3 gains significantly over 30 more motifs than PERT1 and PERT2 (Figure 4B, $PERT1 \sim= 8.45$, $PERT2 \sim= 10.63$, $PERT3 \sim= 43{,}26$; pairwise Dunn's test, PERT1 vs. PERT2, $P = 1.18 \times 10^{-5}$; PERT1 vs. PERT3, $P = 1.55 \times 10^{-206}$; PERT2 vs. PERT3, $P = 2.18 \times 10^{-199}$). However, the number of motifs lost was similar among the three methods (Figure 4C, $PERT1 \sim= 101.12$, $PERT2 \sim= 99.95$, $PERT3 \sim= 90.73$; pairwise Dunn's test, PERT1 vs. PERT2, $P = 0.771$; PERT1 vs. PERT3, $P = 0.0811$; PERT2 vs. PERT3, $P = 0.109$).

We also compared the net change in the number of motifs for each perturbation approach. We observed that PERT3 resulted in a significantly greater net change compared to the other two approaches, whereas there was no significant difference between PERT1 and PERT2 (Figure 4D, $PERT1 \sim= -92.72$, $PERT2 \sim= -89.38$, $PERT3 \sim= -47.50$; pairwise Dunn's test, PERT1 vs. PERT2, $P = 0.232$; PERT1 vs. PERT3, $P = 3.63 \times 10^{-69}$; PERT2 vs. PERT3, $P = 1.58 \times 10^{-60}$).

We then compared these non-specific metrics for the non-motif perturbation sequences. We found similar results to the motif perturbation group: PERT3 resulted in the most motif gains (Figure 4E, $PERT1 \sim= 8.02, PERT2 \sim= 9.98, PERT3 \sim= 29.35$; pairwise Dunn's test, PERT1 vs. PERT2, $P = 0.451$; PERT1 vs. PERT3, $P = 8.03 \times 10^{-55}$; PERT2 vs. PERT3, $P = 5.74 \times 10^{-50}$), with no significant difference in the number of lost motifs (Figure 4F, $PERT1 \sim= 44.07, PERT2 \sim= 43.56, PERT3 \sim= 38.67$). In addition, the net change in the number of motifs of PERT3 is negative but the highest (Figure 4G, $PERT1 \sim= -38.63, PERT2 \sim= -36.12, PERT3 \sim= -12.02$; pairwise Dunn's test, PERT1 vs. PERT2, $P = 0.44$; PERT1 vs. PERT3, $P = 7.03 \times 10^{-23}$; PERT2 vs. PERT3, $P = 8.26 \times 10^{-20}$). These findings further support that the differences in the non-specific impacts are due to the perturbation method used.

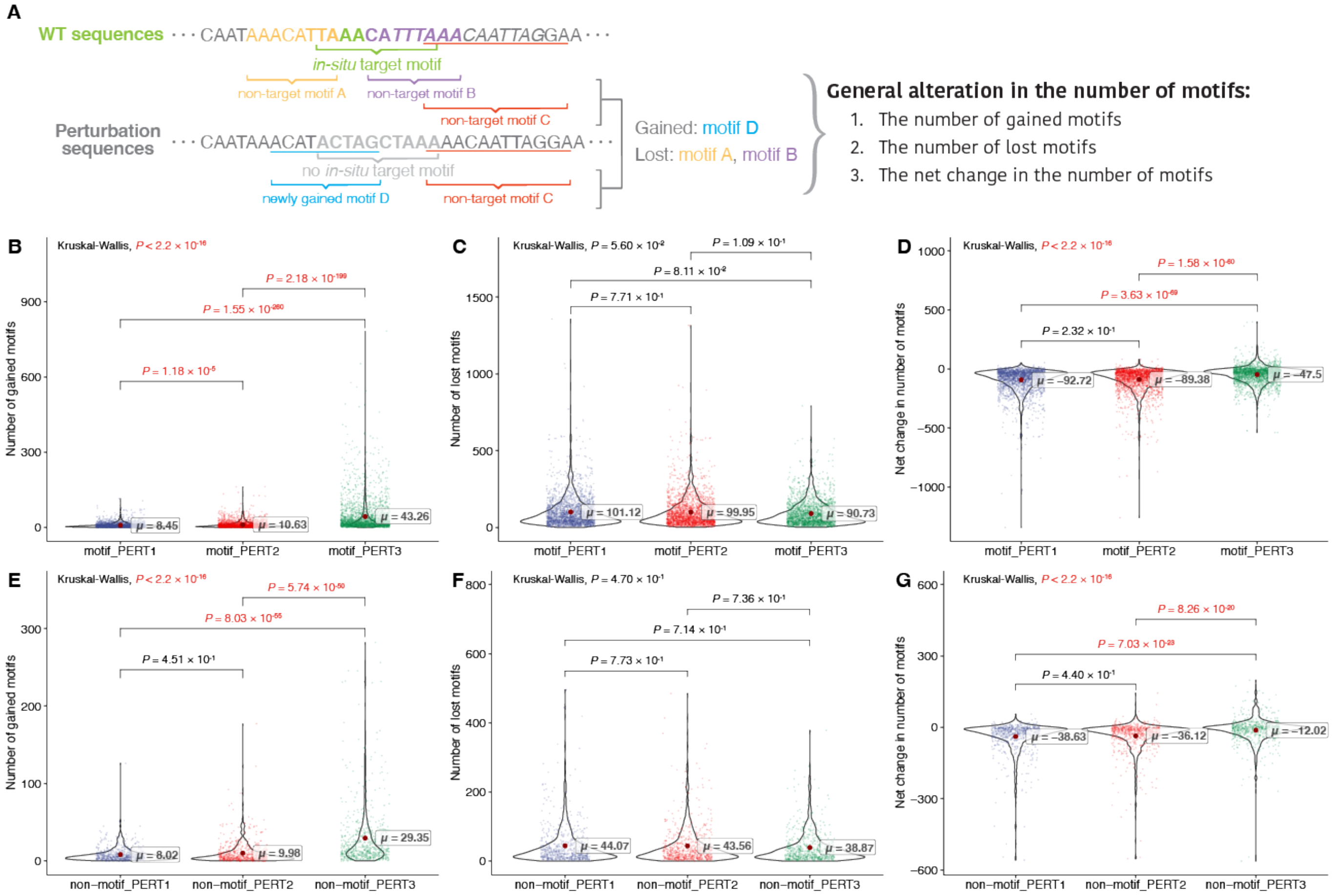


Figure 4. Evaluation of general alteration in the number of motifs. **(A)** Toy examples of calculating general alteration in the number of motifs. **(B-D)** The results for motif perturbations: **(B)** the number of gained motifs, **(B)** the number of lost motifs, and **(B)** the net change in the number of motifs. Significant P values ($P < 0.05$) are shown in red. **(E-G)** The results for non-motif perturbations: **(E)** number of gained motifs, **(F)** number of lost motifs, and **(G)** net change in the number of motifs. Significant P values ($P < 0.05$) are shown in red.

The three perturbation approaches share similar important features, specifically neural developmental features

We then set out to investigate which innate features represent the variances among perturbation sequences, and whether these features differ using different perturbation methods. First, we queried the top 10% of the features (2,500) that explain the variability among perturbed sequences (Section "Methods"), and found that a majority of these features (1,601) are shared by at least two perturbation methods (Figure 5A). Notably, these features mainly fall into "the change in the number of ENCODECIS-BP motifs" and "5-mers frequencies" categories.

Further scrutiny of the top 30 features revealed a substantially large overlap among the three perturbation methods (Figure 5B). Since a majority of the shared features are transcription factor (TF) binding motifs, we conducted gene ontology analysis on the TFs corresponding to the top 2,500 binding motifs. The analysis revealed consistent enrichment of early embryonic development ontologies, including neural development pathways among three perturbation approaches (Figure 5C). These findings suggest that the three perturbation approaches share important features related to neural development.

The MPRA outputs are largely consistent across different perturbations

After assessing the basic and advanced goals of perturbation methods, we next evaluated the consistency of MPRA outputs among three perturbation methods. The MPRA output consists of two parts: the multi-class FRS identities, and the numeric regulatory effect (Log2FC) at seven time points of neural differentiation (Section "Methods").

For the FRS identities, the activities of 419 functional regulatory sites are consistent across three perturbation methods, and 95% (399) of them are activators (Figure 6A). Additionally, 262 sites are consistent in any of the two approaches but not the remaining one (Figure 6A). In terms of the Log2FC, we found a high correlation among all three perturbations across all the time points (Figure 6B-D). However, we found that PERT2 yielded higher Log2FC than the other two approaches (Supplementary Figure 1). This indicates that using a perturbation approach where the same sequence is being introduced, can cause a constant bias in the

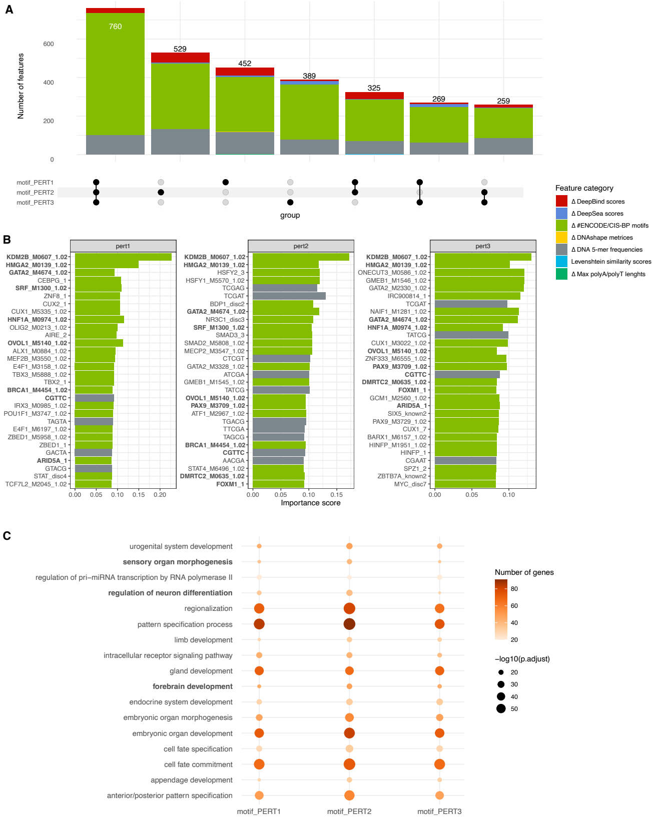


Figure 5. Assessment of the important features representing perturbation sequences. **(A)** The number of important features shared by three perturbation methods. **(B)** Top 30 important features of each perturbation method. The names of features that are shared by at least two perturbation methods are marked in bold. **(C)** Gene ontology enrichment analysis of the top 2500 genes represented by the TF binding factors.

results (e.g., higher or lower Log2FC for PERT2 or PERT1 respectively).

Predictive models of MPRA activity perform the best in PERT3

The perturbation MPRA technique, if designed appropriately, has the potential to predict the activity of non-coding regulatory genomic regions 15. Namely, it is feasible to predict the regulatory activity of a motif by fitting predictive models using the difference in the features between its WT sequence and perturbation sequence. Consequently, this leads to a critical question: which sequence design method for motif perturbation could yield the best performance of such prediction models? This suggests that by designing the perturbation sequences, we may expand the applicability of perturbation MPRA from experimentally identifying regulatory motifs only within designed genomic regions to computationally predicting regulatory elements throughout the non-coding genome. In light of this, we further compared the performances of three perturbation methods using the supervised models as described in the Methods section.

Briefly, we use the difference of features between perturbation sequences and their equivalent WT sequence as the independent variables to fit both classification and regression models. Next, we perform a 10-fold cross-validation for each perturbation data. To benchmark the performance of the models, we statistically compared the AUROC for classifiers and the Pearson correlation coefficient for regressors on the independent test data sets in each fold.

For the classification models that predict the measure of motif FRS identities, we report the receiver-operating characteristic curve (AUROC) of three perturbation approaches (Figure 7). We found that three non-linear models (ET, HGB, and MLP) exhibit high robustness in predicting the FRS identities in the three perturbations. Furthermore, using the results from ET models, we found that PERT3 significantly outperforms PERT2 and PERT1, and PERT1 significantly outperforms PERT2 (pairwise Wilcoxon rank sum test, PERT1 vs. PERT2, $P = 5.58 \times 10^{-5}$; PERT1 vs. PERT3, $P = 3.24 \times 10^{-5}$; PERT2 vs. PERT3, $P = 3.24 \times 10^{-5}$).

For the regression models that predict the quantitative measure of motif regulatory effect, we report the Pearson correlation coefficients for the three perturbation approaches (Figure 8, Supplementary Figure 2, Supplementary Figure 3). Similarly, the model-wise comparison shows the robustness of the ET and HGB model, and PERT3 significantly outperforms the other two methods, while PERT2 outperforms PERT1 (pairwise Wilcoxon rank sum tests, PERT1 vs. PERT2, $P = 2.57 \times 10^{-3}$; PERT1 vs. PERT3, $P = 3.89 \times 10^{-5}$; PERT2 vs. PERT3, $P = 3.89 \times 10^{-5}$).

DISCUSSION

Comprehensively deciphering the regulatory activity of non-coding loci is crucial to the understanding of gene expression dynamics. Shedding light on this, the perturbation-based MPRA technique has enabled the identification of regulatory elements such as enhancers, promoters, and silencers (6, 9, 10). However, insufficient attention has been given to the comprehensive evaluation of various perturbation approaches. As a result, a gold standard of perturbation sequence design strategies remains scant.

Motivated by this scarcity, we proposed a framework for assessing different perturbation approaches, with the aim of better identifying regulatory elements using the perturbation-based MPRA technique. Further, we took advantage of a publicly available data set, which contains the MPRA results acquired from three perturbation approaches (PERT1, PERT2, and PERT3), to conduct an all-inclusive characterization and comparison of these approaches. In short, PERT1 and PERT2 replaced the target motifs with two different "non-motif" sequences, and PERT3 simply shuffled the nucleotides of target motifs.

Starting from the essential ideas of perturbation, which is to eliminate the regulatory effects from target motif(s) within a certain genomic region, we first defined five metrics for assessing the impact from different perturbation approaches (hit rate, perturbation rate, perturbation specificity, newly introduced target motifs per sequence, and general alteration in the number of motifs, see Section "Methods"). These metrics allowed us to scrutinize the overall modification of motif-based profiles within perturbation sequences from different perspectives. Based on our findings, the three approaches exhibit consistently high rates of removing the

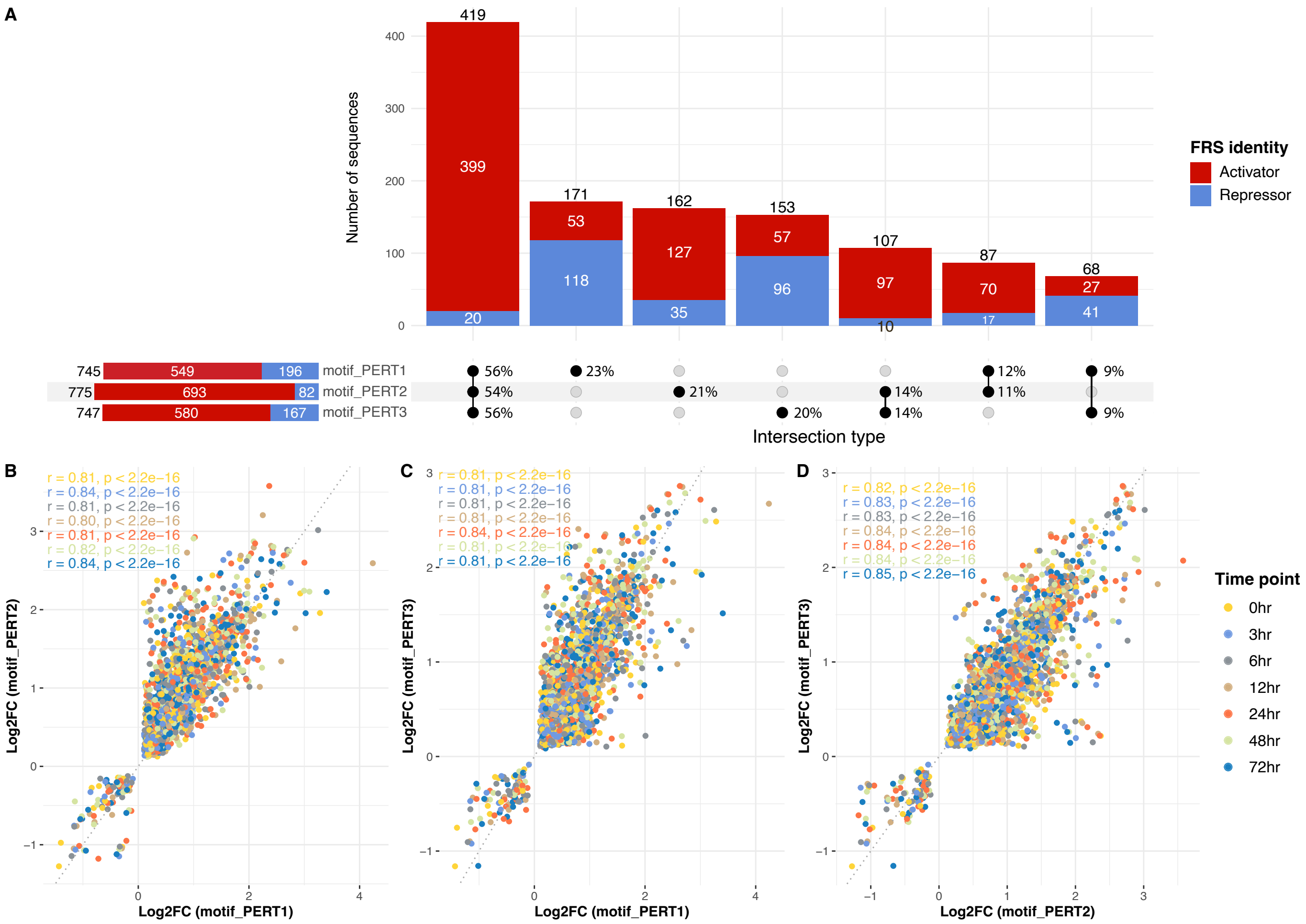
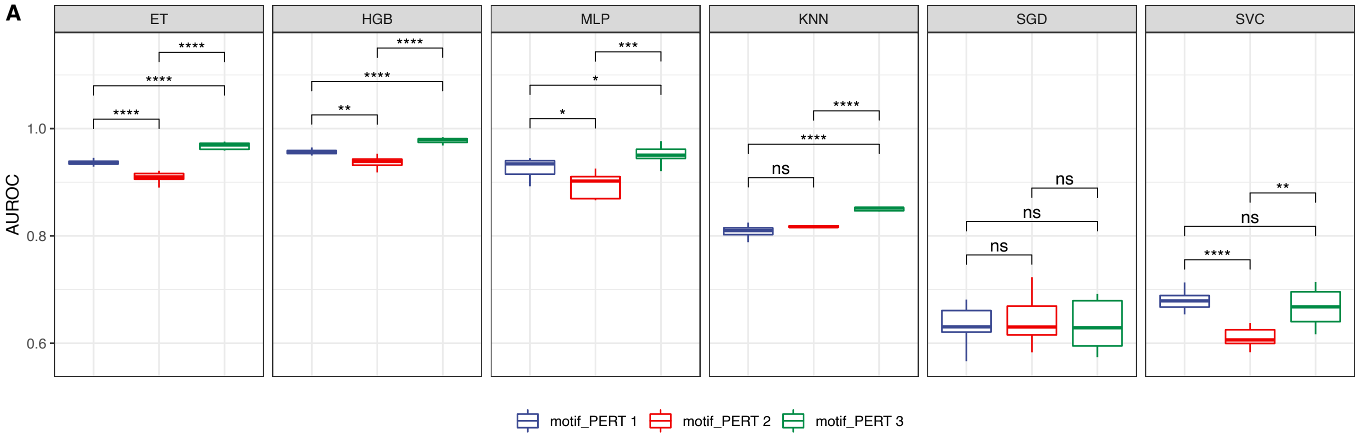


Figure 6. Consistency of MPRA outputs among three perturbations. **(A)** Number of sequences that share the same FRS identities. The bars are colored by activators (red) and repressors (blue). In the "intersection type" matrix. The percentages are row-normalized, indicating the proportion of sequences belonging to different intersection types within each perturbation method. **(B)** The correlation of Log2FC between motif_PERT1 and motif_PERT2. Each dot is a perturbation sequence and is colored by the time point. **(C)** The correlation of Log2FC between motif_PERT1 and motif_PERT3. **(D)** The correlation of Log2FC between motif_PERT2 and motif_PERT3.

target motifs at their targeted locations, which indicates success in in-situ motif perturbation. Additionally, the perturbation rate is kept high across the three perturbation methods (80%), with PERT3 being the lowest (79%), while not significantly different. This implies a further achievement in both in-situ and ex-situ removal of target motifs of the three methods. We note that PERT3 shows a higher probability of introducing target-identical motifs. Despite these, PERT3 brings minimal alterations to the WT motifs within the sequence region, implying that the perturbation specificity of PERT3 is the highest. Moreover, PERT3 leads to the least non-specific motif changes. So far, our observation suggests that the selection of perturbation approaches is a trade-off: for the researchers, it becomes a question of whether to sacrifice the perturbation specificity to achieve a high perturbation rate, or whether to pursue a higher specificity at the cost of a lower perturbation rate.

The next part of our framework is the comparison of MPRA outputs since they are crucial for inferring the activity of target motifs. Particularly, MPRA outputs consist of two parts: 1) the functional regulatory site (FRS) identities that indicate whether the target motif is a non-functional, repressing, or activating element, 2,) the numeric regulatory effects (Log2FC) that quantify the FRS motifs. According to our results, the FRS identities are largely consistent and the Log2FC are highly correlated among all three perturbations. Yet, we also observed a constant skew in the results of PERT1 and PERT2, which indicates that inserting repeated/fixed sequences across the assayed regions is likely to introduce systematic biases in downstream results. The results of this part demonstrated that PERT3 is less likely to introduce systematic biases in MPRA outputs, albeit the high-consistency and high-accuracy profiling for the regulatory activity across all three perturbation methods.

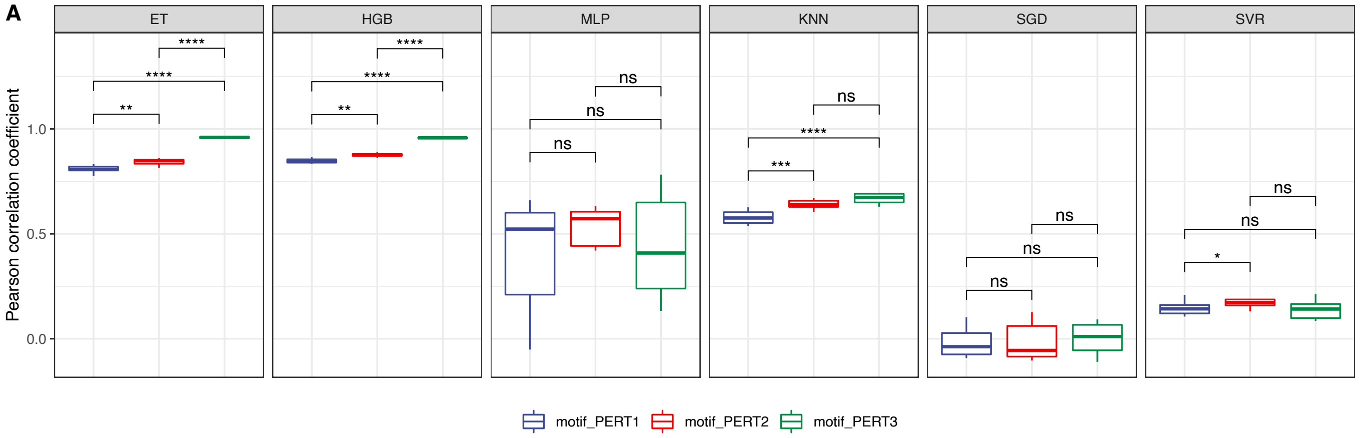
The final part of the framework is to evaluate the potential of perturbation-MPRA in predicting the regulatory activity of non-coding motifs, since our previous works have shown robustness in predicting the activity of putative regulatory elements (15, 36). Specifically, by adequately designing perturbation sequences, the MPRA outputs could be computationally predicted by machine-learning models using the biological features of designed sequences as predictor



B

	ET	HGB	MLP	KNN	SGD	SVR
motif_PERT1	0.94 ± 0.01	0.96 ± 0.01	0.93 ± 0.02	0.81 ± 0.01	0.63 ± 0.04	0.68 ± 0.02
motif_PERT2	0.91 ± 0.01	0.94 ± 0.01	0.90 ± 0.04	0.82 ± 0.01	0.63 ± 0.04	0.61 ± 0.02
motif_PERT3	0.97 ± 0.01	0.98 ± 0.01	0.95 ± 0.02	0.85 ± 0.01	0.63 ± 0.04	0.67 ± 0.03

Figure 7. Performance of classification models. **(A)** The area under the receiver-operating characteristic curve (AUROC) of different classification 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 AUROCs of classification models.



B

	ET	HGB	MLP	KNN	SGD	SVR
motif_PERT1	0.81 ± 0.02	0.85 ± 0.01	0.39 ± 0.26	0.58 ± 0.03	-0.02 ± 0.07	0.14 ± 0.03
motif_PERT2	0.84 ± 0.02	0.87 ± 0.02	0.46 ± 0.22	0.64 ± 0.02	-0.02 ± 0.09	0.18 ± 0.04
motif_PERT3	0.96 ± 0.00	0.96 ± 0.01	0.44 ± 0.24	0.67 ± 0.02	0.00 ± 0.07	0.14 ± 0.05

Figure 8. Performance of regression models. **(A)** The Pearson 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 Pearson correlation coefficients of regression models.

variables. This approach, in some cases, can efficiently identify functional regulatory regions so as to reduce the time and cost of wet lab experiments. Therefore, we developed data-driven models to predict the regulatory activity of target motifs by using the difference in over 28,000 predictive features between perturbation and wild-type sequences. Comparing the performance of models that are built upon the three perturbation methods, we found that PERT3 significantly outperforms the other two in both classification and regression tasks. These findings further support the notion that using a perturbation approach where the nucleotides are being shuffled randomly, works generally better than a replacement with a constant "non-motif" sequence approach.

In summary, we proposed a framework for the evaluation of perturbation sequence design strategies for MPRA experiments, and we utilized this framework to compare three perturbation-based MPRA approaches. From a computational perspective, this study is the first to comprehensively evaluate the library design of the MPRA technique. From an experimental perspective, our results provide deep insights into understanding the impacts of motif perturbation in MPRA experiments. Although it is challenging to offer strict guidance in the absence of *in-vivo* ground truth, we recommend designing sequences by randomly shuffling the nucleotides of the perturbed site when possible.

We anticipate that our findings, together with the proposed framework, will instill a new momentum for the non-coding genomic studies using MPRA techniques, as well as inspire the development of novel comprehensive computational methods. Such efforts and studies will continually contribute to improving our understanding of the functional effects of non-coding regulatory elements.

FUNDING

This work was supported by the the National Institute of Mental Health [R00MH117393] (to A.K.)

ACKNOWLEDGEMENTS

We would like to acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here. URL: https://oarc.rutgers.edu

DATA AND CODE AVAILABILITY

The datasets are available at the NCBI Gene Expression Omnibus (GEO) as accession number GEO: GSE115046.

Conflict of interest statement. None declared.

REFERENCES

1. Esther Rheinbay, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A. Wala, Ofer Shapira, Grace Tiao, Henrik Hornshøj, Julian M. Hess, Randi Istrup Juul, Ziao Lin, Lars Feuerbach, Radhakrishnan Sabarinathan, Tobias Madsen, Jaegil Kim, Loris Mularoni, Shimin Shuai, Andrés Lanzós, Carl Herrmann, Yosef E. Maruvka, Ciyue Shen, Samirkumar B. Amin, Pratiti Bandopadhayay, Johanna Bertl, Keith A. Boroevich, John Busanovich, Joana Carlevaro-Fita, Dimple Chakravarty, Calvin Wing Yiu Chan, David Craft, Priyanka Dhingra, Klev Diamanti, Nuno A. Fonseca, Abel Gonzalez-Perez, Qianyun Guo, Mark P. Hamilton, Nicholas J. Haradhvala, Chen Hong, Keren Isaev, Todd A. Johnson, Malene Juul, Andre Kahles, Abdullah Kahraman, Youngwook Kim, Jan Komorowski, Kiran Kumar, Sushant Kumar, Donghoon Lee, Kjong-Van Lehmann, Yilong Li, Eric Minwei Liu, Lucas Lochovsky, Keunchil Park, Oriol Pich, Nicola D. Roberts, Gordon Saksena, Steven E. Schumacher, Nikos Sidiropoulos, Lina Sieverling, Nasa Sinnott-Armstrong, Chip Stewart, David Tamborero, Jose M. C. Tubio, Husen M. Umer, Liis Uusküla-Reimand, Claes Wadelius, Lina Wadi, Xiaotong Yao, Cheng-Zhong Zhang, Jing Zhang, James E. Haber, Asger Hobolth, Marcin Imielinski, Manolis Kellis, Michael S. Lawrence, Christian von Mering, Hidewaki Nakagawa, Benjamin J. Raphael, Mark A. Rubin, Chris Sander, Lincoln D. Stein, Joshua M. Stuart, Tatsuhiko Tsunoda, David A. Wheeler, Rory Johnson, Jüri Reimand, Mark Gerstein, Ekta Khurana, Peter J. Campbell, Núria López-Bigas, Joachim Weischenfeldt, Rameen Beroukhim, Iñigo Martincorena, Jakob Skou Pedersen, and Gad Getz. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 578(7793):102–111, February 2020. Number: 7793 Publisher: Nature Publishing Group.
2. Vikram Agarwal, Fumitaka Inoue, Max Schubach, Beth K. Martin, Pyaree Mohan Dash, Zicong Zhang, Ajuni Sohota, William Stafford Noble, Galip Gürkan Yardimci, Martin Kircher, Jay Shendure, and Nadav Ahituv. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. *bioRxiv*, March 2023. Pages: 2023.03.05.531189 Section: New Results.
3. Justin Koesterich, Joon-Yong An, Fumitaka Inoue, Ajuni Sohota, Nadav Ahituv, Stephan J. Sanders, and Anat Kreimer. Characterization of De Novo Promoter Variants in Autism Spectrum Disorder with Massively Parallel Reporter Assays. *International Journal of Molecular Sciences*, 24(4):3509, January 2023. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.
4. Chengyu Deng, Sean Whalen, Marilyn Steyert, Ryan Ziffra, Pawel F. Przytycki, Fumitaka Inoue, Daniela A. Pereira, Davide Capauto, Scott Norton, Flora M. Vaccarino, Alex Pollen, Tomasz J. Nowakowski, Nadav Ahituv, and Katherine S. Pollard. Massively parallel characterization of psychiatric disorder-associated and cell-type-specific regulatory elements in the developing human cortex, February 2023. Pages: 2023.02.15.528663 Section: New Results.
5. Kyung Duk Koh, Luke R. Bonser, Walter L. Eckalbar, Ofer Yizhar-Barnea, Jiangshan Shen, Xiaoning Zeng, Kirsten L. Hargett, Dingyuan I. Sun, Lorna T. Zlock, Walter E. Finkbeiner, Nadav Ahituv, and David J. Erle. Genomic characterization and therapeutic utilization of IL-13-responsive sequences in asthma. *Cell Genomics*, 3(1):100229, December 2022.
6. 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.
7. Michael A. White, Connie A. Myers, Joseph C. Corbo, and Barak A. Cohen. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences*, 110(29):11952–11957, July 2013. Publisher: Proceedings of the National Academy of Sciences.
8. Xinchen Wang, Liang He, Sarah M. Goggin, Alham Saadat, Li Wang, Nasa Sinnott-Armstrong, Melina Claussnitzer, and Manolis Kellis. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nature Communications*, 9(1):5380, December 2018. Number: 1 Publisher: Nature Publishing Group.
9. Anat Kreimer, Tal Ashuach, Fumitaka Inoue, Alex Khodaverdian, Chengyu Deng, Nir Yosef, and Nadav Ahituv. Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. *Nature Communications*, 13(1):1504,

March 2022. Number: 1 Publisher: Nature Publishing Group.

10. 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.
11. Tal Ashuach, David S. Fischer, Anat Kreimer, Nadav Ahituv, Fabian J. Theis, and Nir Yosef. MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biology*, 20(1):183, September 2019.
12. M. Grace Gordon, Fumitaka Inoue, Beth Martin, Max Schubach, Vikram Agarwal, Sean Whalen, Shiyun Feng, Jingjing Zhao, Tal Ashuach, Ryan Ziffra, Anat Kreimer, Ilias Georgakopoulos-Soares, Nir Yosef, Chun Jimmie Ye, Katherine S. Pollard, Jay Shendure, Martin Kircher, and Nadav Ahituv. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nature Protocols*, 15(8):2387–2412, August 2020. Number: 8 Publisher: Nature Publishing Group.
13. 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.
14. Jake R Conway, Alexander Lex, and Nils Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, September 2017.
15. Anat Kreimer, Zhongxia Yan, Nadav Ahituv, and Nir Yosef. Meta-analysis of massively parallel reporter assays enables prediction of regulatory function across cell types. *Human Mutation*, 40(9):1299–1313, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.23820.
16. Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, August 2015. Number: 8 Publisher: Nature Publishing Group.
17. Kathleen M. Chen, Aaron K. Wong, Olga G. Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature Genetics*, 54(7):940–949, July 2022. Number: 7 Publisher: Nature Publishing Group.
18. Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, April 2016.
19. Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Research*, 41(Web Server issue):W56–W62, July 2013.
20. Jamie C. Kwasnieski, Christopher Fiore, Hemangi G. Chaudhari, and Barak A. Cohen. High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, 24(10):1595–1602, October 2014.
21. 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.
22. Pouya Kheradpour and Manolis Kellis. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, 42(5):2976–2987, March 2014.
23. Hui Hu, Ya-Ru Miao, Long-Hao Jia, Qing-Yang Yu, Qiong Zhang, and An-Yuan Guo. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Research*, 47(D1):D33–D38, January 2019.
24. William E. Winkler. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. Proceedings of the Section on Survey Research, 1990.
25. Murat Sariyar and Andreas Borg. The RecordLinkage Package: Detecting Errors in Data. *The R Journal*, 2(2):61, 2010.
26. Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014.
27. Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, August 2021.
28. Léon Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD.
29. Nello Cristianini and Elisa Ricci. Support Vector Machines. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*, pages 928–932. Springer US, Boston, MA, 2008.
30. Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, 4(11):218, June 2016.
31. Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, April 2006.
32. Moshe Sipper and Jason H. Moore. AddGBoost: A gradient boosting-style algorithm based on strong learners. *Machine Learning with Applications*, 7:100243, March 2022.
33. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
34. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.
35. M Merika and S H Orkin. DNA-binding specificity of GATA family transcription factors. *Molecular and Cellular Biology*, 13(7):3999–4010, July 1993.
36. Anat Kreimer, Haoyang Zeng, Matthew D. Edwards, Yuchun Guo, Kevin Tian, Sunyoung Shin, Rene Welch, Michael Wainberg, Rahul Mohan, Nicholas A. Sinnott-Armstrong, Yue Li, Gökcen Eraslan, Talal Bin Amin, Ryan Tewhey, Pardis C. Sabeti, Jonathan Goke, Nikola S. Mueller, Manolis Kellis, Anshul Kundaje, Michael A Beer, Sunduz Keles, David K. Gifford, and Nir Yosef. Predicting gene expression in massively parallel reporter assays: A comparative study. *Human Mutation*, 38(9):1240–1250, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.23197.

SUPPLEMENTARY FIGURES AND TABLES

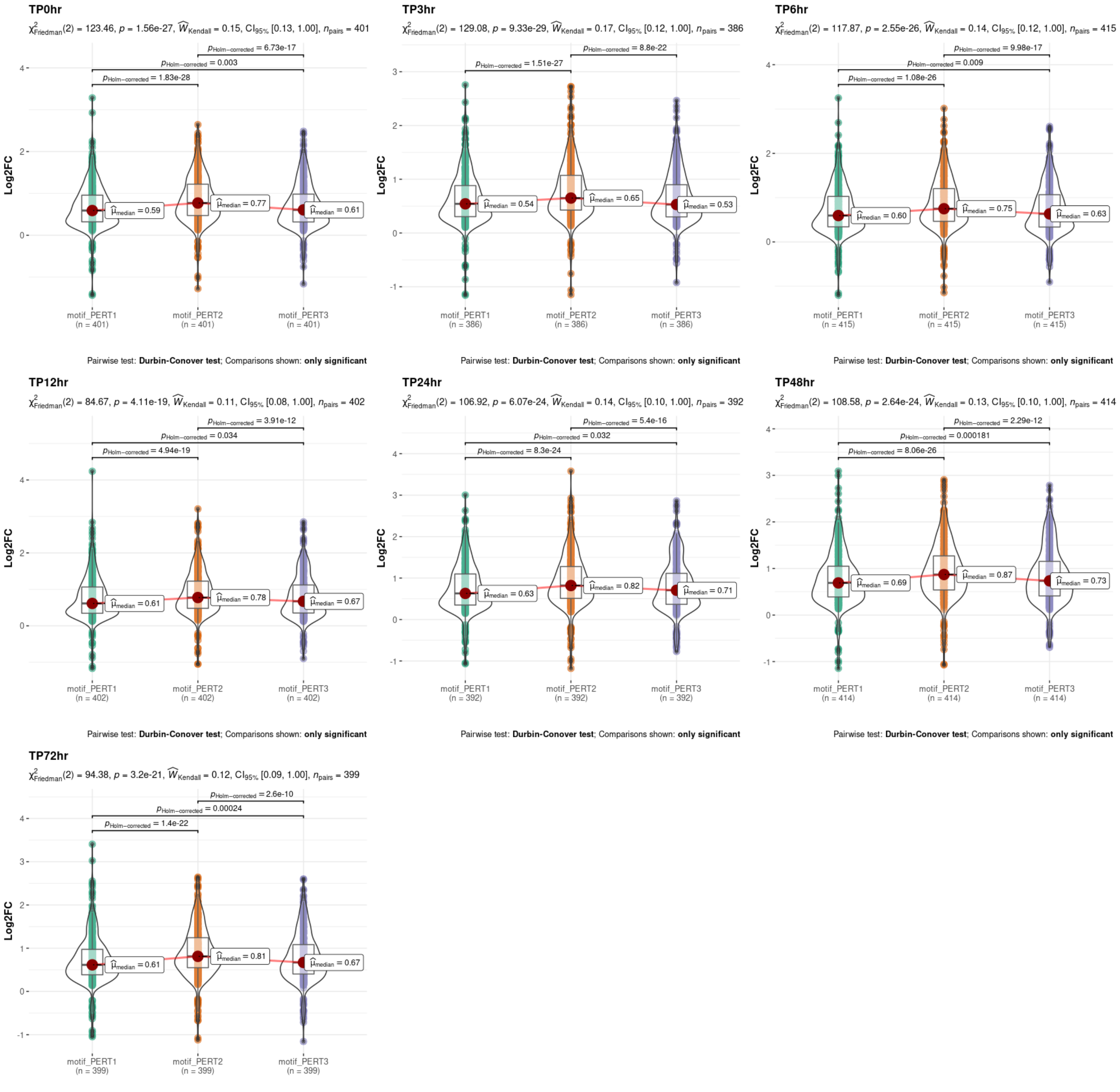
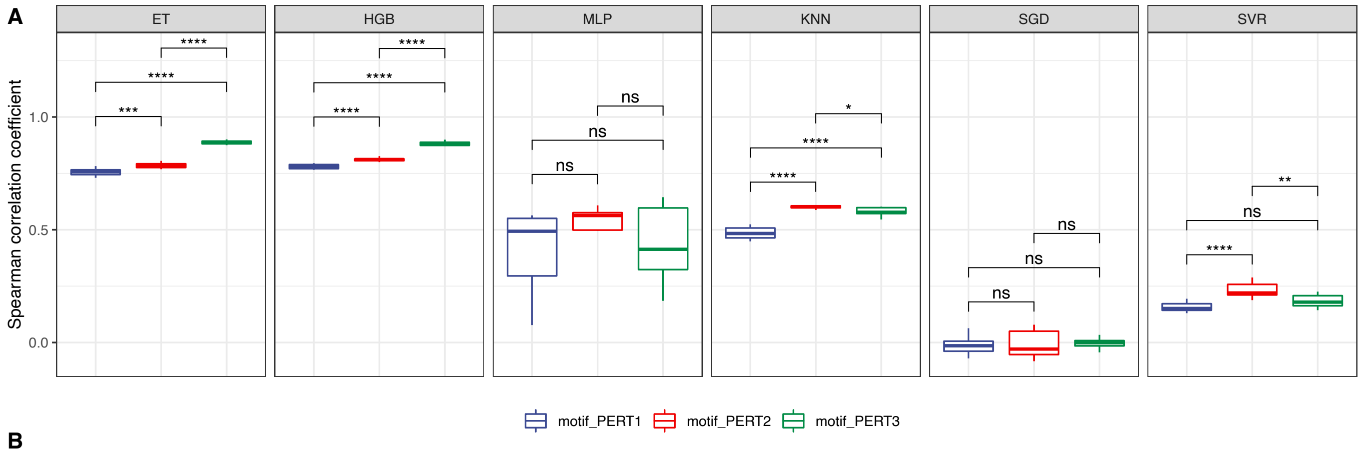


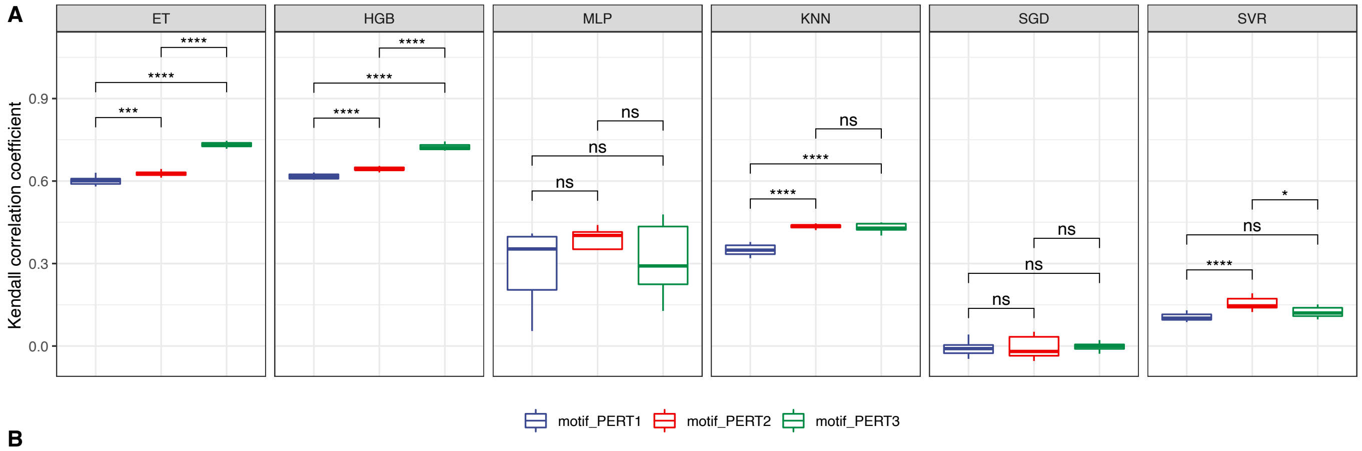
Figure Supplementary Figure 1. Comparison of Log2FC among three perturbation methods.



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

Figure Supplementary Figure 2. 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 ± standard deviation values for Spearman correlation coefficients of regression models.



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

Figure Supplementary Figure 3. 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 ± standard deviation values for Kendall correlation coefficients of regression models.