


# A statistical framework for predicting critical regions of p53-dependent enhancers

Xiaohui Niu, Kaixuan Deng, Lifan Liu, Kun Yang and Xuehai Hu 

Corresponding author: Xuehai Hu, College of Informatics, Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan, Hubei, 430070, P.R. China. Tel.: +86-18171282783; Fax: +86-27-87288509; E-mail: huxuehai@mail.hzau.edu.cn

## Abstract

P53 is the ‘guardian of the genome’ and is responsible for regulating cell cycle and apoptosis. The genomic p53 binding regions, where activating transcriptional factors and cofactors like p300 simultaneously bind, are called ‘p53-dependent enhancers’, which play an important role in tumorigenesis. Current experimental assays generally provide a broad peak of each enhancer element, leaving our knowledge about critical enhancer regions (CERs) limited. Under the inspiration of enhancer dissection by CRISPR-Cas9 screen library on genome-wide p53 binding sites, here we introduce a statistical framework called ‘Computational CRISPR Strategy’ (CCS), to predict whether a given DNA fragment will be a p53-dependent CER by employing 7-mer as feature extractions along with random forest as the regressor. When training on a p53 CRISPR enhancer dataset, CCS not only accurately fitted the top-ranked enriched single guide RNAs (sgRNAs) but also successfully reproduced two known CERs that were validated by experiments. When applying it to an independent testing dataset on a tiling of a 2K-b genomic region of CRISPR-deCDKN1A-Lib, the trained model shows great generalizability by identifying a CER containing five top-ranked sgRNAs. A feature importance analysis further indicates that top-ranked 7-mers are mapped onto informative TF motifs including POU5F1 and SOX5, which are differentially enriched in p53-dependent CERs and are potential factors to make a general p53 binding site to form a p53-dependent CER, providing the interpretability of the trained model. Our results demonstrate that CCS is an alternative way of the CRISPR experiment to screen the genome for mapping p53-dependent CERs.

**Key words:** critical enhancer regions; computational CRISPR; p53; K-mer; TF motifs

## Introduction

The development and cellular differentiation in eukaryotes require precise regulation of gene expressions, which is

governed by the orchestration of various genomic regulatory elements (GREs) [1, 2]. Enhancers are main GREs that positively regulate gene expressions in a distal manner [3–5]. On the

Xiaohui Niu is an associate professor in the College of Informatics at Huazhong Agricultural University. He is interested in research on the function of genetic variant.

Kaixuan Deng is a master in the College of Informatics at Huazhong Agricultural University. He is interested in machine learning algorithms on regulatory element predictions.

Lifan Liu is a master in the College of Informatics at Huazhong Agricultural University. She is interested in machine learning algorithms on regulatory element predictions.

Kun Yang is a master in the College of Informatics at Huazhong Agricultural University. He is interested in machine learning algorithm.

Xuehai Hu is an associate professor in the College of Informatics at Huazhong Agricultural University. He is interested in genomic regulatory element predictions and genomic prediction.

Submitted: 7 January 2020; Received (in revised form): 26 February 2020

© The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

one hand, the ENCODE project identified >500 000 putative enhancers with the evidence of accessible chromatin and histone modification markers [6, 7]. The total length of all these enhancers covers ~15% of the human genome [8], implying the enhancer element is a nonnegligible component of genome. On the other hand, genome-wide association studies (GWASs) in the past decade found that over 55% of the disease-associated SNPs are located within the noncoding regions of the human genome [9]. Some examples that noncoding GWAS SNPs are exactly located within the critical enhancer region (CER, a short but critical DNA fragment) of enhancers have been previously reported [10], implying that the disruption of a CER might cause abnormal transcriptions and thus triggers human diseases [11, 12].

In the past two decades, researchers have developed several distinct experimental strategies to detect molecular biomarkers for indirectly inferring the locations of active enhancers, such as transgenic mouse assay [13], using chromatin features from ENCODE data [14], massively parallel reporter assay (MPRA) [15–17], STARR-seq using self-transcribing transcripts [18] and cap analysis of gene expression (CAGE) [19] utilizing eRNA. Altogether, these strategies used co-occurrences of active enhancers and specific molecular signals, which include histone modification markers (such as monomethylation of lysine 4 of histone H3 (H3K4me1) and acetylation of lysine 27 of histone H3 (H3K27ac)) [6, 20], specific transcriptional factor-binding sites (TFBS) like p300 [21] and enhancer RNA (eRNA) [19, 22], to infer the locations of active enhancers. Unfortunately, the above experiments generally give a broad peak of a given enhancer. For example, FANTOM5 enhancers identified by eRNA signals are 282-bp of average length [19], and thus the critical regions within enhancers that are necessary for activating gene expressions remain unknown.

More recently, researchers employed genome-editing tools like the CRISPR-Cas9 system to screen known enhancers for identifying CERs. For the first time, Canver et al. [11] designed all possible single guide RNAs (sgRNAs) according to all possible protospacer adjacent motifs (PAMs) within the human BCL11A enhancer (a 12-kb composite enhancer) and then monitored the changes of BCL11A gene expression with deletions of every 10-bp of the cleavage positions. Three top-scoring regions were identified as CERs and further validated by additional experiments. Notably, Korkmaz et al. [12] proposed to target Cas9 to genome-wide p53 binding sites within enhancers and used a tumor-suppressive mechanism named oncogene-induced senescence (OIS) to identify p53-dependent CERs. In their results, a genomic CRISPR-Cas9 tiling screen mapped eight enriched sgRNAs, three of which target to cis-regulatory elements of CDKN1A that is a master regulator of p53-dependent OIS. Being different with the previous two studies, Klann et al. [23] designed CRISPR-Cas9-based epigenomic regulatory element screening (CERES) to identify regulatory element activity in the native genomic context. The advantage of CERES is the ability of evaluating both loss- and gain-of-function of a given regulatory element without the disruption of its DNA sequence. These notable progresses demonstrate that it is feasible to employ CRISPR-Cas9 tiling screen tools for identifying CERs.

p53 is well known as a tumor suppressor gene and the ‘guardian of the genome’. The main biological function of p53 is to regulate cell cycle and apoptosis through the p53 signaling pathway, and it is mutated in more than 50% of all human tumors [24]. As an upstream TF, it needs cooperation involving cofactors like p300 and other activating TFs including NF-Y and SP1 [24] to activate the p53 signaling pathway. The genomic regions, where p300 and activating TFs bind, are regarded as

p53-dependent enhancers, which play an important role in p53 normal function and in tumorigenesis. Therefore, accurate identification of p53-dependent enhancers, especially the CERs of them, is extremely important for prioritizing cancer-related variants and cancer pathogenesis, and thus is a whole new field in today’s biology.

Herein, we ask whether it is possible to identify p53-dependent CERs by a computational way only with the DNA sequences. Our motivations come from two existing facts: (a) previous studies of disrupting core TF motifs within enhancers demonstrate that specific critical regions within enhancers are necessary for enhancer activities [15], suggesting the existence of predictive sequence features of CERs; (b) the accumulation of sequence samples from CRISPR-Cas9 tiling screen experiments allows us to learn those predictive sequence features by advanced machine learning tools. In this study, a p53 CRISPR enhancer dataset [12] was constructed for testing the possibility of the above proposed computational CRISPR strategy (CCS). Importantly, when encoding primary DNA sequences with a 7-mer representation, a random forest (RF) model not only accurately fitted top-ranked enriched sgRNAs in the training dataset but also reproduced known CERs in an independent testing dataset. Furthermore, a feature importance analysis gives a good interpretability of the trained model by finding meaningful TF motifs, which are supported by biological experimental data and are important for helping a p53 binding site to form a real CER. All in all, our approach suggests the feasibility, effectiveness and substitutability of identification of CERs with a computational way. The codes and datasets for training and testing are available at [https://github.com/kaixuanDeng95/Computational\\_CRISPR\\_Strategy](https://github.com/kaixuanDeng95/Computational_CRISPR_Strategy).

## Materials and methods

### A p53 CRISPR enhancer dataset for training a prediction model

In this study, we constructed a p53 CRISPR enhancer dataset for training and testing a computational CRISPR model. Korkmarz et al. [12] designed 1286 sgRNAs to target to genome-wide p53 binding sites within enhancers and then used OIS to monitor tumorigenesis with deletion of these sites in human BJ cells (a well-characterized cell model of OIS). In their study, to measure necessity and importance of each deleted DNA fragment for OIS, they employed an indicator of Z-score, which represents the normalized enrichment of a given sgRNA in tamoxifen-inducible HRASG12V (BJ-RASG12V) relative to human BJ control cell. As a result, 1080 sgRNAs along with their Z-scores were provided in their publication, based on which they determined eight top-ranked enriched sgRNAs. And two top-scoring elements named p53<sup>enh3507</sup> and p53<sup>enh3508</sup> (located at ~10 kb upstream and proximal to the transcription start site (TSS) of CDKN1A, respectively) were validated by further experiments and were eventually considered to be two CERs.

Here we first ask whether it is possible to reproduce previous results by a computational way. It aims to construct a prediction model that will accurately identify enriched sgRNAs with high predicted Z-scores. The input of the model is those DNA fragments deleted by sgRNAs, and the output is their Z-scores. To this end, we first mapped each sgRNA to its targeting site within the human genome using a published web-server named ‘CRISPRdirect’ (<http://crispr.dbcls.jp/>) [25]. After removing those sgRNAs with multiple mappings, 980 sgRNAs with unique mapping along with their Z-scores were determined. Among them, only 20 sgRNAs are equipped with top Z-scores,

whereas the vast majority of them have very low Z-scores, suggesting that the distribution of Z-score is extremely imbalanced (Supplementary Figure 1). Most importantly, observing that our task is to identify top-ranked enriched sgRNAs with high Z-scores, we must learn informative sequence features of CERs from trustable samples with top-ranked Z-scores. To this end, we performed a data augmentation strategy by replicating the top-ranked 150 samples and obtained the final dataset containing  $980 + 150 = 1130$  samples (Supplementary Data 1). For each sample, we extracted their flanking genomic sequences with an appropriate local length using the reference human genome of 'hg19'. The optimal local length will be determined via a 10-fold-cross-validation (10-fold-CV).

In statistical or machine learning framework, a 10-fold-CV is usually required to evaluate the performance of the trained model. Generally, a 10-fold-CV adopts a random grouping strategy that randomly divides the whole dataset into 10 groups. We here must emphasize that random grouping is not appropriate for the augmented dataset, because it will take high risk of putting the same augmented sample into training group and testing group, respectively. This will lead to an over-estimated prediction result, which is analogous to a recent publication that criticized the over-estimated training accuracy of the enhancer-promoter interaction prediction problem [26]. Generally, the random grouping scheme of the 10-fold-CV chooses all samples belonging to the  $i$ th group to form the testing group in the  $i$ th fold. By contrast, we here adopt a modified grouping scheme of a modified 10-fold-CV. In the  $i$ th fold, it collects all the samples belonging to the  $i$ th group from 980 un-augmented samples, not belonging to the  $i$ th group from 1130 augmented samples, to form the testing group. It will avoid the above risk and ensure the reliability of the trained model under this modified 10-fold-CV.

### A tiling dataset of CRISPR-deCDKN1A-Lib for an independent testing

Each enhancer may contain multiple regulatory elements, whose cooperation tends to be biologically functional. Apart from the element of  $p53^{enh3507}$ , Korkmarz et al. [12] adopted a tiling strategy on a genomic region of  $\sim 2\text{Kb}$  centered at the gene of deCDKN1A and designed the CRISPR-deCDKN1A-Lib of 197 sgRNAs to search the existence of other elements. This CRISPR-deCDKN1A-Lib is an ideal dataset for an independent testing of the trained model. Similar to the training set of 980 sgRNAs, we again employed 'CRISPRdirect' to determine the cleavage sites of 197 sgRNAs. After removing multiple mapping, 195 sgRNAs with their deleted genomic sequences were obtained (Supplementary Data 2). Then we extracted their flanking genomic sequences with the optimal local length, which will be determined via the modified 10-fold-CV. We will take the extracted genomic sequences of 195 sgRNAs as the inputs for the trained model, and then measure the consistency between the outputs of the model and their experimental Z-scores.

### A K-mer feature representation

For a computational approach, each DNA sequence needs to be encoded with a feature representation. K-mer is a simple and valid way for coding DNA sequence as well as a widely used feature representation for a variety of prediction tasks including enhancer predictions [27] and regulatory element predictions [28]. A K-mer feature representation counts occurrence frequencies of  $4^K$  distinct K-continuous nucleotide combinations.

More precisely, a DNA sequence  $P$  with  $L$  bases can be expressed as:

$$P = B_1 B_2 \cdots B_L, \text{ where } B_i \in \{A, C, G, T\}, i = 1, \cdots, L.$$

The K-mer feature representation of the DNA sequence  $P$  is defined as the normalized frequency vector of all possible substrings of length  $K$  in that DNA sequence, i.e.

$$K\text{-mer} = [f_1, f_2, \cdots, f_{4^K}]^T,$$

where  $f_i = \frac{n_i}{L-K+1}$ , and  $n_i$  is the occurrence number of the  $i$ -th K-mer in the DNA sequence  $P$  for each  $i$  ( $i = 1, 2, \cdots, 4^K$ ). In this study, we tried three choices of  $K$ :  $K=6$ ,  $K=7$  and  $K=8$ , and the corresponding feature dimensions are  $4^6 = 4096$ ,  $4^7 = 16384$  and  $4^8 = 65536$ .

### Machine learning prediction methods

In current study, due to our regression task, five widely used statistical or machine learning models, i.e. least absolute shrinkage and selection operator (LASSO), elastic net (EN), support vector regression (SVR), random forest (RF) and gaussian process regression (GPR), were adopted to perform prediction comparisons (for details of these five conventional regression methods, please refer to Supplementary Materials). An optimal model will be chosen based on a comprehensive comparison among these five regression methods.

### Evaluation of the prediction performance

In statistical prediction, a 10-fold-CV test is usually used to examine a predictor for its effectiveness in practical applications. In this study, the main concern is the degree of consistency between the predicted Z-scores and the experimental Z-scores, with the emphasis on those top-ranked enriched samples. Specifically, we chose a modified 10-fold-cross-validation to evaluate the performance of the trained model. Accordingly, we employed two indexes, Pearson correlation coefficient (PCC<sub>-all</sub>) between the experimental Z-scores and the predicted Z-scores of all the samples, and PCC<sub>-20</sub> on the top-20-ranked samples.

The reason why we choose top 20 ranked samples is that Z-score is the normalized score of the degree of enrichments of all sgRNAs and it is assumed to follow standard normal distribution. The 0.95 quantile of standard normal distribution is about 1.65. Therefore, we select all the samples with their Z-scores larger than 1.65, which are exactly the top-20-ranked samples (Supplementary Figure 1).

## Results

### A trained RF model accurately fits top-ranked enriched sgRNAs and reproduces known CERs

We first ask whether the results of p53 enhancer CRISPR experiments can be reproduced by a computational way. To this end, we try to construct a prediction model that will accurately fit the existing p53 enhancer CRISPR data.

It is widely accepted that the disruption of a core motif is an important factor for influencing the activity of a cis-regulatory element. Actually, recent studies demonstrated that local sequences flanking the core motif contain useful information for discriminating functional regions from nonfunctional ones

**Table 1.** PCC<sub>-all</sub> of triad combinations of K-mer, local sequence length and regression method via a modified 10-fold-CV test

Regression method	6-mer				7-mer				8-mer			
	50-bp	100-bp	200-bp	500-bp	50-bp	100-bp	200-bp	500-bp	50-bp	100-bp	200-bp	500-bp
GPR	0.1288	0.1127	0.1205	0.1249	0.0160	0.0327	0.0838	0.1050	0.0153	0.0308	0.0837	0.1140
SVR	0.0835	0.0643	0.0688	0.0791	0.1040	0.0688	0.0648	0.0770	0.1000	0.0708	0.0631	0.0669
LASSO	0.0679	0.0458	0.1249	0.1363	0.0363	0.0896	0.1185	0.1429	0.0297	0.1256	0.0818	0.1180
EN	0.0748	0.1004	0.1376	0.0993	0.1066	0.1021	0.1230	0.1117	0.0833	0.1018	0.1121	0.1114
RF	0.0954	0.1305	0.1365	0.1442	0.1346	0.1465	0.1420	0.1340	0.1145	0.1408	0.1247	0.1222

**Table 2.** PCC<sub>-20</sub> of triad combinations of K-mer, local sequence length and regression method via a modified 10-fold-CV test

Regression method	6-mer				7-mer				8-mer			
	50-bp	100-bp	200-bp	500-bp	50-bp	100-bp	200-bp	500-bp	50-bp	100-bp	200-bp	500-bp
GPR	0.5975	0.6241	0.6253	0.5186	0.5975	0.6248	0.6251	0.5839	0.5975	0.6246	0.0651	0.4856
SVR	0.4713	0.5401	0.5377	0.5388	0.4807	0.5406	0.5375	0.5473	0.4914	0.5411	0.5378	0.5944
LASSO	0.5160	0.5544	0.3724	0.4037	0.4351	0.4117	0.4562	0.4155	0.1973	0.3848	0.4138	0.4426
EN	0.4650	0.4714	0.4169	0.4695	0.4997	0.4499	0.4339	0.4478	0.4401	0.4481	0.4307	0.4294
RF	0.5651	0.5473	0.5083	0.5366	0.5563	0.6722	0.5526	0.5192	0.5295	0.5178	0.5346	0.5240

[29–32]. However, an optimal length near the core motif needs to be determined. We tested triad combinations of three different choices of K-mer (6-mer, 7-mer and 8-mer), four local sequence lengths (50-bp, 100-bp, 200-bp and 500-bp) and five regression methods (LASSO, EN, SVR, RF and GPR) to search an optimal combination.

To control overfitting, we adopted a testing strategy via a modified 10-fold-CV (Materials and Methods) to objectively report the performance under each triad combination. We reported the results of PCC<sub>-all</sub> and PCC<sub>-20</sub> in Tables 1 and 2, respectively. PCC<sub>-all</sub> between the predicted Z-score and the experimental Z-score of the whole samples gets its maximum of 0.1465 at the combination of 7-mer, 100-bp and the RF method (Table 1). Although this result is not very good, we must reaffirm that our main task is to identify potential CERs with top-ranked enriched sgRNAs. Therefore, we are more concerned with PCC<sub>-20</sub> between the predicted Z-score and the experimental Z-score among top-20-ranked samples. Importantly, a maximum PCC<sub>-20</sub> of 0.6722 was achieved at the same triad combination (Table 2), implying that the corresponding trained RF model will accurately fit top-20-ranked samples. Notably, the choice of 100-bp implies a biological implication of the optimal flanking sequence length of 50-bp. Actually, this finding is quite consistent with a recent study, which claimed that a flanking length of 50-bp contains the majority of necessary sequence features for activating cis-regulatory element [29]. Based on the published paper and the result of PCC<sub>-20</sub> of top-20-ranked sgRNAs, we next fix the optimal triad combination of 7-mer, 100-bp and the RF method in the rest of this study.

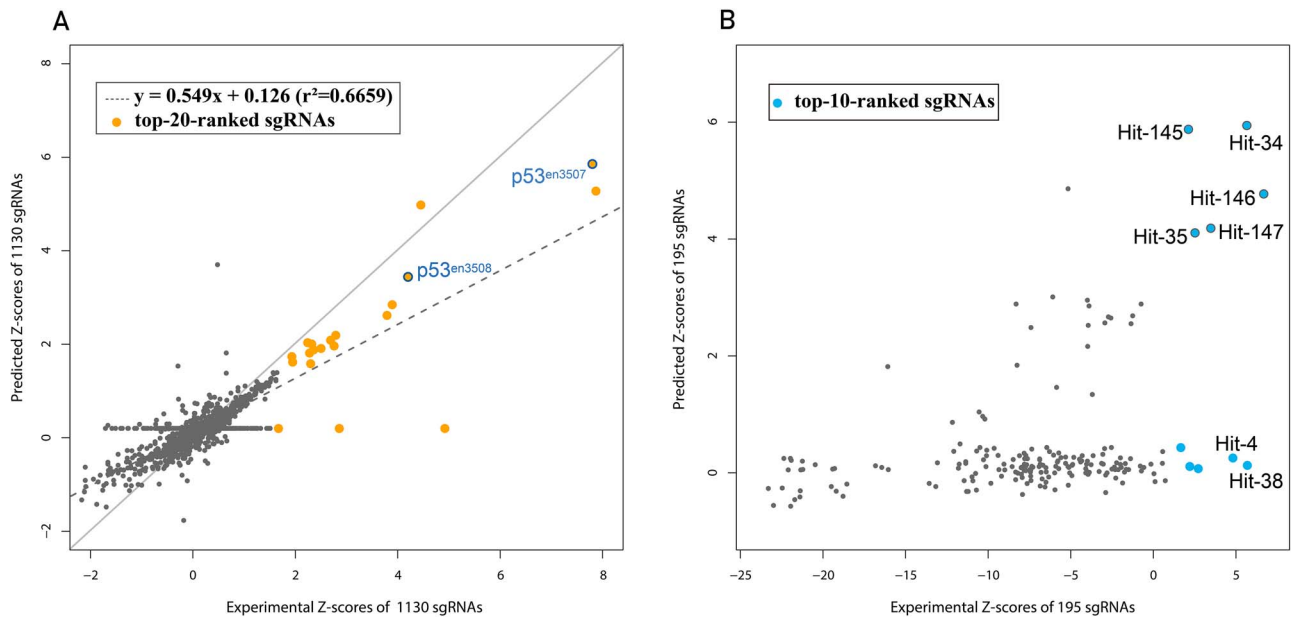
We next test the reproducibility of the RF model and focus on whether the trained RF model can successfully reproduce two known CERs: p53<sup>enh3507</sup> and p53<sup>enh3508</sup>. To this end, we trained a RF model under the optimal triad combination with 1130 augmented samples and obtained the predicted Z-scores of all 980 sgRNAs. When comparing with their experimental Z-scores, they fit a regression line  $y = 0.549x + 0.126$  with  $r^2 = 0.6659$  (Figure 1A), suggesting the trained RF model can fit the p53 enhancer CRISPR dataset to a certain extent. Notably, when we focused on the top-20-ranked sgRNAs, the experimental

top-20-ranked sgRNAs (points marked with the orange color in Figure 1A) had gotten their predicted values with top ranks as well. Most importantly, two experimentally validated CERs of p53<sup>enh3507</sup> and p53<sup>enh3508</sup> achieved the first and the fifth ranks on their predicted Z-scores (two points with the blue circle in Figure 1: p53<sup>enh3507</sup>: 7.80 versus 5.857; p53<sup>enh3508</sup>: 4.2 versus 3.439), suggesting that the trained RF model can be an alternative computational strategy for reproducing known CERs.

### An independent testing on the CRISPR-deCDKN1A-Lib shows the generalizability of the trained RF model

We next show the generalizability of the trained RF model by an independent testing on the CRISPR-deCDKN1A-Lib, which comprises a total of 195 sgRNAs tilling on a genomic region of chr6: 36,634,056-36,636,070. By putting their 7-mer features of 100-bp local genomic sequences into the trained RF model, we can collect their predicted Z-scores. By plotting the predicted Z-scores against the experimental Z-scores in Figure 1B, we found that the majority of 195 sgRNAs (171 out of 195, 87.69%) got very low predicted Z-scores (less than 1), implying most sgRNAs were predicted to be nonfunctional. As the same criterion, we only concern about top-10-ranked enriched sgRNAs with their Z-score larger than 1.65. When focusing on these top-10-ranked sgRNAs, we found a mixed result. On one hand, half sgRNAs (5 out of 10, Hit-146, Hit-34, Hit-147, Hit-35 and Hit-145) achieved consistent predicted Z-scores with experimental Z-scores. On the other hand, the remainders (Hit-38, Hit-4, Hit-41, Hit-134 and Hit-7) have not gotten high predicted Z-scores (around zero). A direct inference might lead us to conclude that the trained RF model can only reproduce a half of experimental results.

We next try to investigate the reason why the five sgRNAs of Hit-38, Hit-4, Hit-41, Hit-134 and Hit-7 were not successfully identified with the trained RF model. Let us first focus on the correctly identified five sgRNAs: they cluster together within a continuous 100-bp genomic region of chr6: 36 635 000–36 635 100 (Figure 2) and more molecular markers including p53<sub>ChIP-seq</sub>, H3K4Me1 and H3K27Ac all support that this genomic region is the p53-dependent critical region of an enhancer element.



**Figure 1.** The prediction performances of the trained random forest model on the p53 CRISPR enhancer training dataset (A) and on the independent testing dataset of CRISPR-deCDKN1A-Lib (B).

This implies that the trained RF model can accurately identify this CER. On the contrary, the five sgRNAs of Hit-38, Hit-4, Hit-41, Hit-134 and Hit-7 scatter around the whole region (refer to [Supplementary Data 2](#) for their detailed genomic positions). For studying the biological functions of Hit-38 and Hit-4, Korkmarz et al. [12] performed a series of following experiments: a reporter assay proves that they are not enhancers; a p53- ChIP-qPCR shows that p53 does not bind to these two sites; a CEBPB- ChIP-qPCR confirms that Hit-38 is a CEBPB-dependent element. The authors deduced that Hit-38 contributes to CEBPB recruitment to deCDKN1A and to its function in OIS, while the regulatory mechanism of the element of Hit-4 is still poorly understood. Recall that the trained RF model is trained on a p53 CRISPR enhancer dataset, and it is originally supposed to act as a predictor to predict CERs of p53-dependent enhancer elements. It fulfils the original task by accurately identifying a 100 bp-CER comprising five sgRNAs. It cannot identify other sgRNAs because the remaining elements perform their regulatory functions through different or downstream pathways (For example, Hit-38 is CEBPB-dependent). In a word, we conclude that the trained RF model can accurately identify p53-dependent CERs and it cannot be used for identifying other types of regulatory elements.

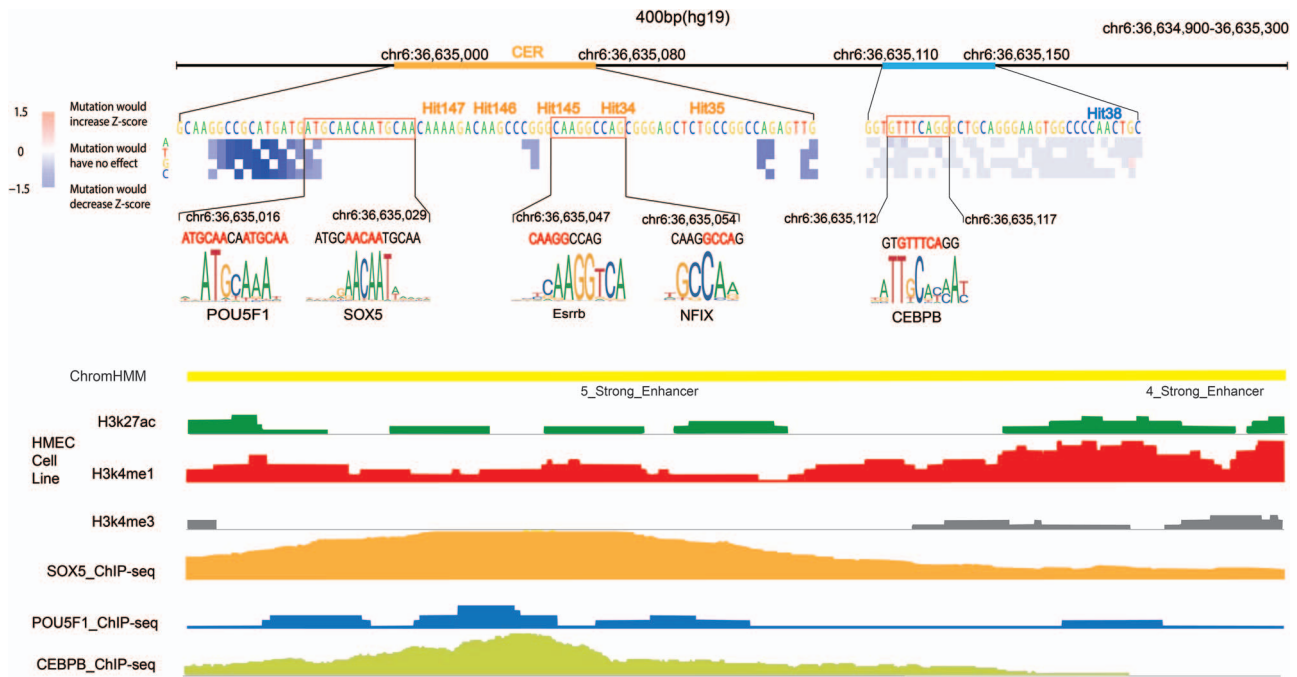
### The trained RF model has learned informative TF motif features

We next try to give the interpretability of the trained RF model equipped with a 7-mer feature representation and then focus on their further biological implications. One of the advantages of RF is the interpretability of each input feature, which is quantified with the index of ‘importance’. Higher value of importance of a given feature implies its more significant role in constructing the RF model. As a result, we listed the importance of top-10-ranked 7-mers in [Table 3](#). Furthermore, to give the interpretability, we mapped each top-10-ranked 7-mer to a comprehensive and latest database of ‘JASPAR 2020’ (<http://jaspar.genereg.net/>), which integrated 245 Position Weight Matrices (PWMs) in their

CORE collection [33]. Specifically, each top-ranked 7-mer was put into the webserver to search the most similar DNA-binding TF motif, which was displayed along with each 7-mer in [Table 3](#).

It is not surprising that the motif of p53 itself is not ranked high because all training samples have p53 binding motif, implying p53 binding motif is not the determinant for making a specific p53 binding site to be a CER. Among top-10 7-mers, three of them are mapped into the binding motif of the TF of POU5F1 with 6 perfect matches. Another example of 6 perfect matches happens on the TF of NFATC2, and five 7-mers are mapped into SOX5, TWIST1, NF1X, HNF1B and Esrrb with five perfect matches.

Let us again focus on the identified CER of a 100-bp genomic region (chr6: 36 635 000-36 635 100) by analyzing its TF motif composition. Among the top-10-ranked 7-mer features, four TF motifs of POU5F1, SOX5, Esrrb and NF1X were found within this CER ([Figure 2](#)). We further downloaded three widely used epigenomic markers of H3K27ac, H3K4me1 and H3K4me3 of Human Mammary Epithelial Cells (HMEC Cell Lines) from ENCODE data (<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmHmecHMM.bed.gz>). Meanwhile, we searched the ChIP-seq data of the four mentioned TFs from ENCODE data and found two of them of POU5F1 and SOX5. We demonstrate all the molecular marker information as the evidence in [Figure 2](#), from which we can conclude that the CER of chr6: 36 635 000–36 635 100 is a p53-dependent CER with binding sites of POU5F1 and SOX5. Interestingly, POU5F1 contains a POU homeodomain that plays a key role in embryonic development and stem cell pluripotency. Aberrant expression of this gene in adult tissues was reported to be associated with tumorigenesis [34]. Moreover, SOX5 is a homologous gene with SOX2, and it was reported that the bindings of Nanog, POU5F1 and SOX2 tend to co-occur, thus forming Nanog-POU5F1-SOX2 cluster, which is believed to promote p300 recruitment in embryonic stem cells [35]. Observing the fact that EP300 is a broad molecular marker of active enhancers, these findings together imply that the core element of ‘ATGCAACAATGCAACA’ (chr6: 36 635 ,016–36 635 031) is a determinant to make this p53 binding site to form a real CER.



**Figure 2.** Top panel, a snapshot of the genomic region of chr6: 36,634,900-36,635,300 with two DNA regulatory elements. One is the identified p53-dependent CER with five top-ranked sgRNAs colored orange. The other is a CEBPB-dependent element with one top-ranked sgRNA of Hit-38 colored blue. At the middle, the detailed DNA sequences along with two mutation maps and specific TF motifs are shown. At the bottom, chromatin annotation and TF ChIP-seq peaks are shown. HMEC, human mammary epithelial cell; ChromHMM, chromatin state segmentation by HMM; H3K4Me1, histone H3 lysine 4 monomethylation; H3K4Me3, histone H3 lysine 4 trimethylation; H3K27Ac, histone H3 lysine 27 acetylation.

**Table 3.** Top-10-ranked important 7-mer features and their similar TF motifs

Rank	7-mer	Importance	Similar to	Perfect matches	TF motif	Rank	7-mer	Importance	Similar to	Perfect matches	TF motif
1	TGATGCG	11.02	NF-YA	4/7	FX 1	6	CTTTCCT	8.16	NFATC2	6/7	FX 6
2	CCAGAGT	9.62	TWIST1	5/7	FX 2	7	CGGCCAG	6.86	NF1X	5/7	FX 7
3	ATGCCGAG	8.73	POU5F1	4/7	FX 3	8	AGTTAAG	6.80	HNF1B	5/7	FX 8
4	GCAACAA	8.45	SOX5	5/7	FX 4	9	GCAAGGC	6.64	Esrrb	5/7	FX 9
5	ATGCAAC	8.37	POU5F1	6/7	FX 5	10	GATGCAA	6.45	POU5F1	6/7	FX 10

We next focus on another genomic region containing Hit-38 (chr6: 36 635 110–36 635 150). We want to use the information of epigenomic markers and ChIP-seq data of binding TFs to explain why it is not a p53-dependent CER. From [Figure 2](#), it is clear that this region is not an enhancer with the evidence of deficiency of H3K4ac. Unsurprisingly, it has none of binding sites of POU5F1, SOX5 and other top-ranked TFs, which gives an explanation of its low predicted Z-score by the trained RF model. Interestingly, it has a binding site of the TF of CEBPB with the evidence of CEBPB\_ChIP-seq experimental data, supporting the published conclusion that the genomic region targeted by sgRNA-Hit-38 is a CEBPB-dependent DNA element for regulating the expression of CDKN1A and thus for its function in OIS [12].

These pieces of evidence all support the result that the trained RF model has learnt informative TF motif features, which are potential determinants to make a p53 binding site to form a real p53-dependent CER, thus enhancing the biological interpretability of the trained RF model.

### The motifs of TWIST1, POU5F1 and SOX5 are differentially enriched in p53-dependent enhancers

We next perform a TF motif enrichment analysis to check whether the important TF motifs are enriched in p53-dependent enhancers with top-ranked experimental Z-scores. To this end, we selected top-20-ranked sgRNAs as top-ranked p53-dependent enhancers (the study group), and meanwhile, we selected the bottom-100-ranked sgRNAs with lowest Z-scores (5 times of top-20-ranked sgRNAs) as the negative control group. For each TF motif from top-10 important TF motifs, we counted its appearance numbers both in the study group and in the negative control group by employing a motif detection tool called ‘Clover’ with default threshold of matching score  $\geq 6$  [36]. A fisher exact test was then performed with the above two appearance numbers. We illustrated enrichment results as p.value and odd.ratio of each TF motif in [Supplementary Figure 2](#) and [Supplementary Table 1](#), from which we found that five (TWIST1, POU5F1, SOX5, etc.) out of top-10 TF motifs are

significantly enriched ( $P$  value  $<0.05$  and  $\text{odds.ratio} >1$ ) in p53-dependent enhancers, implying that some important TF motifs including TWIST1, POU5F1 and SOX5 are differentially enriched in p53-dependent enhancers.

### Genetic variants adjacent to the core element have large effects on its predicted score

To investigate the functional effects of genetic variants, we next employ the trained RF model to identify which mutation at which position within a CER would increase or decrease predicted Z-scores. As similar to 'DeepBind' [37], we here use a 'mutation map' (the middle panel of Figure 2) to illustrate the effect of every possible point mutation within the identified CER of chr6: 36 635 000–36 635 080 may have on its predicted Z-score. In greater detail, we mutated each single base to one of three possible directions and calculated the difference between the predicted Z-scores of after and before mutation. A heat map of difference scores was then drawn to demonstrate how important each base is for Z-score of the whole CER.

From left of the middle panel of Figure 2, we found two facts: one is that every point mutation within the CER would not increase predicted Z-score, implying high conservation of this CER; another fact is that the majority of point mutation would not greatly change predicted Z-score except a 10-bp fragment of chr6: 36 635 004–36 635 014. Interestingly, this 10-bp fragment is adjacent to the core element of 'ATGCAACAATGCAACA' (chr6: 36 635 016–36 635 031), implying that a single SNP adjacent to the core element might lead to a sample being wrongly assigned a low Z-score. In contrast, we performed the same job on a CEBPB-dependent DNA element containing Hit-38 (chr6: 36 635 110–36 635 150) and found that every point mutation within this region would lead to no significant effect because it is not a p53-dependent enhancer element (right of the middle panel of Figure 2). These results demonstrate that genetic variants adjacent to the core element have large effects on its predicted score, and these also imply that the trained RF model has the potential to be a convenient tool to help people to search for potential causal SNPs of p53-related cancers.

## Discussion

Current epigenomic markers, p300 binding sites or eRNA signals only give a broad peak of an enhancer element. Recently, genome-editing tools including CRISPR-Cas9 were proven effective and powerful for fine mapping of CERs in base resolution [11, 12, 38]. In this paper, we investigated the possibility of a computational strategy for accurately identifying CERs by usage of the above experimental data as training samples. For this reason, we designed a statistical framework called 'computational CRISPR strategy' (CCS) to predict whether a given DNA fragment will be a CER. By testing triad combinations of four local sequence lengths, five regular regression methods and three choices of K-mer, we constructed a prediction model with the optimal combination of the local sequence length of 100-bp, the regression method of RF and the feature representation of 7-mer. The trained RF prediction model showed its reproducibility not only by accurately fitting the top-ranked enriched sgRNAs, but also by reproducing two known CERs. An independent testing on the CRISPR-deCDKN1A-Lib of 195 tiling sgRNAs further demonstrates the generalizability of the trained RF model. A feature importance analysis indicates that top-ranked 7-mer features are mapped as motifs of important enhancer TFs including POU5F1 and SOX5, providing the interpretability of the trained RF

model. We will next discuss the applicability to other TFs, main contributions and some limitations of the statistical framework.

### The statistical framework is valid to ER $\alpha$ with available training dataset

To discuss how applicable the statistical framework obtained in this paper is to other TFs, we directly applied the optimal triad combination of 7-mer, 100-bp and the RF method to another TF of ER $\alpha$ . In the same publication, Korkmarz et al. performed a similar process on a dropout screen for identifying critical regions of ER $\alpha$ -bound enhancers. A total of 97 sgRNAs were designed to target ER $\alpha$ -bound enhancers and three candidate sgRNAs of ER $\alpha^{\text{enh588}}$ , ER $\alpha^{\text{enh1830}}$  and ER $\alpha^{\text{enh1986}}$  were experimentally validated to be CERs. Similarly, we applied our statistical framework on this dataset of ER $\alpha$ . In greater detail, 94 out of 97 sgRNAs (three sgRNAs with multiple mappings were excluded) were determined as statistical samples, and the top-5-ranked samples of ER $\alpha^{\text{enh1986}}$ , ER $\alpha^{\text{enh1830}}$ , ER $\alpha^{\text{enh812}}$ , ER $\alpha^{\text{enh588}}$  and ER $\alpha^{\text{enh723}}$  were determined as top-ranked samples (Z-score  $< -1.38$ , which is the 0.1 quantile of standard normal distribution). When we successively expanded their lengths to 100-bp, extracted corresponding 7-mer features and trained them via RF and a 10-fold-CV, a comparable PCC $_{\text{all}}$  of 0.2552 and an acceptable PCC $_{.5}$  of 0.3203 were found. Although the result of ER $\alpha$ -dependent enhancers is not good as that of p53 (PCC $_{.20}$  of 0.6722), the statistical framework can fit top-5-ranked sgRNAs of ER $\alpha$ -bound enhancers to a certain extent. This demonstrates that the statistical framework is valid to ER $\alpha$  with available training dataset.

### Main contributions

On the basis of all the above efforts, we conclude that our current work brings some new contributions into the area of identification of CERs:

- (1) Possibility of CCS: CCS is the first attempt for identifying CERs with CRISPR enhancer data by a computational way. By employing an experimental dataset of p53 enhancers as training samples, we conclude that the trained RF model can reproduce the majority of experimental data, suggesting the possibility of CCS.
- (2) Local sequence length of 100-bp: by testing triad combinations, we confirm that a local length of 100-bp around CER is sufficient for constructing an accurate prediction model, which is quite consistent with a recent publication [29].
- (3) Development of practical training and testing strategy for CRISPR screen experiments: CRISPR screen library experiments all aim to identify a small number of target genomic regions among massive genome-wide regions. Their tasks are to focus on top-enriched samples. If we want to develop machine learning models to successfully predict top-enriched samples, we actually meet a problem of extremely imbalanced dataset. The current study brings a novel contribution by providing a practical strategy, which first augments top-enriched samples and then adopts the modified 10-fold-CV to evaluate the prediction model. We hope that CCS is becoming a standard framework of such related studies.

### Some limitations

Finally, although CCS achieved great progresses described above, it has some limitations that need more biological experiments and investigations in the future study. The first limitation is

that our trained RF model is valid only for identification of p53-dependent CERs. CERs depending on other regulatory pathways might not be accurately identified by our model. For solving this problem, we need more experimental data concerning CRISPR screen libraries of various types of enhancers. Only with massive data, we will have a chance for training a comprehensive model to identify various types of CERs. Another problem is that CCS has a certain risk of false positive rate, which can be found in the independent testing result from Figure 1B. For this problem, we think that those regions with high predicted Z-scores have top-ranked TF motifs in sequence but might not locate within open chromatin at epigenomic views. How to control false positive rate by introducing more other information, including open chromatin and co-occurrence of TFBS, is the main task of future works.

### Key Points

- CCS is a novel computational framework that employs 7-mer as feature extractions along with random forest as the regressor to identify critical enhancer regions (CERs) of p53-dependent enhancers.
- CCS is the first attempt to identify CERs at base-resolution by a computational way learning from the data of CRISPR-Cas9 screen library.
- CCS successively shows reproducibility, generalizability and substitutability of identification of CERs with a computational way.
- CCS has learnt informative TF motifs that might be determinants to make p53 binding sites to form real CERs.

### Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

### Acknowledgements

We acknowledge Prof. Weibo Xie for helpful discussions.

### Funding

This work was supported by the National Natural Science Foundation of China (NSFC, 11671003 to Xuehai Hu) and the Fundamental Research Funds for the Central University HZAU (Grant No. 2662017JC048 to Xiaohui Niu).

### Conflict of interest

The authors declare no conflict of interest.

### References

1. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**(7414):57.
2. Klefogiannis D, Kalnis P, Bajic VB. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform* 2016;**17**(6):196–200.
3. Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat Rev Genet* 2016;**17**(4):207–23.
4. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* 2014;**15**(4):272.
5. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* 2011;**144**(3):327–39.
6. Heintzman ND, Hon GC, Hawkins RD, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 2009;**459**(7243):108.
7. Hoffman MM, Ernst J, Wilder SP, et al. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 2012;**41**(2):827–41.
8. Fishilevich S, Nudel R, Rappaport N, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* 2017;**2017**(1):1–17.
9. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012;**337**(6099):1190–1195.
10. Sur IK, Hallikas O, Vähärautio A, et al. Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* 2012;**338**(6112):1360–3.
11. Canver MC, Smith EC, Sher F, et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* 2015;**527**(7577):192.
12. Korkmaz G, Lopes R, Ugalde AP, et al. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* 2016;**34**(2):192.
13. Visel A, Minovitsky S, Dubchak I, et al. VISTA enhancer browser—a database of tissue-specific human enhancers. *Nucleic Acids Res* 2007;**35**(Database issue):D88–92.
14. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;**9**(3):215.
15. Kwasnieski JC, Fiore C, Chaudhari HG, et al. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* 2014;**24**(10):1595–602.
16. Melnikov A, Murugan A, Zhang X, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 2012;**30**(3):271–7.
17. Shen SQ, Myers CA, Hughes AE, et al. Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res* 2016;**26**(2):238–55.
18. Arnold CD, Gerlach D, Stelzer C, et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 2013;**339**(6123):1074–7.
19. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature* 2014;**507**(7493):455.
20. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**(7539):317.
21. Visel A, Taher L, Girgis H, et al. A high-resolution enhancer atlas of the developing telencephalon. *Cell* 2013;**152**(4):895–908.
22. Liu F. Enhancer-derived RNA: a primer. *Genomics Proteomics Bioinformatics* 2017;**15**(3):196–200.
23. Klann TS, Black JB, Chellappan M, et al. CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat Biotechnol* 2017;**35**(6):561.
24. Muller PA, Vousden KH. p53 mutations in cancer. *Nat Cell Biol* 2013;**15**(1):2–8.



25. Naito Y, Hino K, Bono H, et al. CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. *Bioinformatics* 2015;**31**(7):1120–3.
26. Xi W, Beer MA. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Comput Biol* 2018;**14**(12):e1006625.
27. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* 2011;**21**(12):2167–80.
28. Ghandi M, Lee D, Mohammad-Noori M, et al. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* 2014;**10**(7):e1003711.
29. Chaudhari HG, Cohen BA. Local sequence features that influence AP-1 cis-regulatory activity. *Genome Res* 2018;**28**(2):171–81.
30. Farley EK, Olson KM, Zhang W, et al. Suboptimization of developmental enhancers. *Science* 2015;**350**(6258):325.
31. Farley EK, Olson KM, Zhang W, et al. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc Natl Acad Sci U S A* 2016;**113**(23):6508.
32. Levo M, Zalckvar E, Sharon E, et al. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res* 2015;**25**(7):1018–29.
33. Fornes O, Castro-Mondragon JA, Khan A, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2019;**48**(D1):D87–D92.
34. Boiani M, Schöler HR. Developmental cell biology: regulatory networks in embryo-derived pluripotent stem cells. *Nat Rev Mol Cell Biol* 2005;**6**(11):872.
35. Chen X, Xu H, Yuan P, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 2008;**133**(6):1106–17.
36. Frith MC, Fu Y, Yu L, et al. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 2004;**32**(4):1372–81.
37. Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.
38. Diao Y, Fang R, Li B, et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* 2017;**14**(6):629.