Bioinformatics, 2018, 1–8
doi: 10.1093/bioinformatics/bty748
Advance Access Publication Date: 31 August 2018
Original Paper

Sequence analysis

Synergizing CRISPR/Cas9 off-target predictions for ensemble insights and practical applications

Shixiong Zhang [1], Xiangtao Li[1,2], Qiuzhen Lin[3] and Ka-Chun Wong [1,*]

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong, [2]Department of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China and [3]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 20, 2018; revised on August 8, 2018; editorial decision on August 23, 2018; accepted on August 30, 2018

Abstract

Motivation: The RNA-guided CRISPR/Cas9 system has been widely applied to genome editing. CRISPR/Cas9 system can effectively edit the on-target genes. Nonetheless, it has recently been demonstrated that many homologous off-target genomic sequences could be mutated, leading to unexpected gene-editing outcomes. Therefore, a plethora of tools were proposed for the prediction of off-target activities of CRISPR/Cas9. Nonetheless, each computational tool has its own advantages and drawbacks under diverse conditions. It is hardly believed that a single tool is optimal for all conditions. Hence, we would like to explore the ensemble learning potential on synergizing multiple tools with genomic annotations together to enhance its predictive abilities.
Results: We proposed an ensemble learning framework which synergizes multiple tools together to predict the off-target activities of CRISPR/Cas9 in different combinations. Interestingly, the ensemble learning using AdaBoost outperformed other individual off-target predictive tools. We also investigated the effect of evolutionary conservation (PhyloP and PhastCons) and chromatin annotations (ChromHMM and Segway) and found that only PhyloP can enhance the predictive capabilities further. Case studies are conducted to reveal ensemble insights into the off-target predictions, demonstrating how the current study can be applied in different genomic contexts. The best prediction predicted by AdaBoost is up to 0.9383 (AUC) and 0.2998 (PRC) that outperforms other classifiers. This is ascribable to the fact that AdaBoost introduces a new weak classifier (i.e. decision stump) in each iteration to learn the DNA sequences that were misclassified as off-targets until a small error rate is reached iteratively.
Availability and implementation: The source codes are freely available on GitHub at https://github.com/Alexzsx/CRISPR.
Contact: kc.w@cityu.edu.hk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The ground-breaking technology known as CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats with Cas9) system has been proved to be an efficient tool for precise gene editing (Billon *et al.*, 2017; Cho *et al.*, 2013, 2014; Jinek *et al.*, 2012; Kleinstiver *et al.*, 2016; Mali *et al.*, 2013; Scott and Zhang, 2017; Xu *et al.*, 2017). Within the system, the single guide RNA (sgRNA) is a 20-nt RNA that can guide the Cas9 endonuclease to the target double stranded DNA sequence of interest. If the sgRNA matches perfectly with one strand of the target, the Cas9 endonuclease cleaves it for precise gene editing. However, the predesigned sgRNA followed by a protospacer adjacent motif (PAM) could guide Cas9 to off-target homologous attachment and cause unexpected mutations both *in vitro* and *in vivo* (Fu *et al.*, 2013; Hsu *et al.*, 2014; Meng *et al.*, 2017; Radecke *et al.*, 2018; Rosenbluh *et al.*, 2017; Tsai *et al.*, 2015, 2017; Wang *et al.*, 2017; Wolter and Puchta, 2017; Zhou *et al.*, 2017); for instance, Cradick *et al.* (2013) found that substantial off-target cleavages lie on the target genes of the human hemoglobin β and C-C chemokine receptor type 5. Fu *et al.* (2013) also identified the off-target mutations with up to five mismatches in human cells. Hruscha *et al.* (2013) demonstrated that CRISPR/Cas9 results in multiple unwanted mutations in zebrafish cells. Those previous studies support the idea that *in vivo* off-target mutations caused by CRISPR/Cas9 can be ubiquitous and should be addressed since CRISPR/Cas9 has even been adopted to the gene editing on human embryos now (Ma *et al.*, 2017).

In most cases, those off-target genomic sequences are homologous to the on-target gene sequences with one or more mismatches. The categories of mismatches between off-targets and on-targets can be grouped into four types (Doench *et al.*, 2014; Lin *et al.*, 2014) based on sequence properties: (i) they have the same length but there are nucleotide mismatches; (ii) they have the same length and are perfectly matched but the PAM (e.g. NAG or NGG) is mismatched; (iii) they have different lengths and there are missing nucleotide bases; (iv) they have different lengths and there are some extra nucleotide bases.

Public concerns were raised on whether this technology could be adopted in the clinical setting since the off-target mutations can lead to potential risks upon the gene editing. Therefore, it is extremely critical to investigate the off-target activities and improve the fidelity of CRISPR/Cas9 application to avoid off-target mutations.

The issue of off-target mutations has recently attracted many researchers to develop genomic profiling methods to study the genome-wide DNA damage caused by CRISPR/Cas9 and those can provide useful experimental results to build computational models for predicting potential off-target sites. Fu *et al.* (2013) found that single and double mismatches are tolerated to varying degrees depending on their positions along the guide gRNA-DNA interface. Tsai *et al.* (2015) used GUIDE-seq method for global detection of DNA double-stranded breaks and revealed wide variability in off-target activities and unappreciated characteristics of off-target sequences. Kim *et al.* (2015) proposed a method called Digenome-seq for profiling genome-wide off-target effects of programmable nucleases including Cas9. Wang *et al.* (2015) proposed the IDLVs (Integrase-Defective Lentiviral Vectors) to identify the off-target sequences with at least 1% frequencies. BLESS proposed by Ran *et al.* (2015) can detect not only the off-target sequences with base mismatches (aforementioned case 1) predicted by other similarity-based methods, but also the off-targets caused by the aforementioned cases 3 and 4 (missing bases and extra bases between sgRNA and DNA) (Peng *et al.*, 2016; Ran *et al.*, 2015).

There are also computational tools to predict the off-targets (Bae *et al.*, 2014; Cradick *et al.*, 2014; Heigwer *et al.*, 2014; Montague *et al.*, 2014; Naito *et al.*, 2015; Sander *et al.*, 2010; Xiao *et al.*, 2014; Zhu *et al.*, 2014). The CRISPRdirect (Naito *et al.*, 2015) and Cas-OFFinder (Bae *et al.*, 2014) can report the potential off-target sites with multiple mismatches. CRISPRseek (Zhu *et al.*, 2014) calculated a cleavage score for the potential off-target candidates. CasOT (Xiao *et al.*, 2014) is another tool designed for user-friendliness. There are also tools based on different sequence alignment algorithms to predict the off-target bindings (Cradick *et al.*, 2014; Heigwer *et al.*, 2014; Montague *et al.*, 2014; Sander *et al.*, 2010). Haeussler *et al.* (2016) evaluated the above off-target scoring algorithms and integrated them into the sgRNA selection tool CRISPOR.

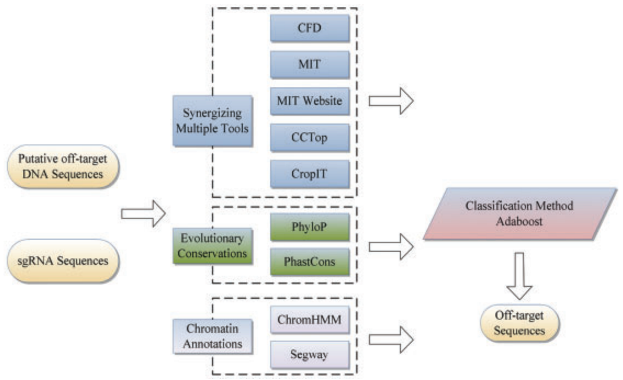
Putative off-target DNA Sequences
sgRNA Sequences
Synergizing Multiple Tools
CFD
MIT
MIT Website
CCTop
CropIT
Evolutionary Conservations
PhyloP
PhastCons
Chromatin Annotations
ChromHMM
Segway
Classification Method Adaboost
Off-target Sequences

Fig. 1. Overview of the ensemble learning framework on synergizing multiple tools, evolutionary conservation and chromatin annotations

Stemmer *et al.* (2015) and Singh *et al.* (2015) proposed off-target predictors CCTop and Cropit to rank the potential off-target sites separately. The two methods were calculated by considering the mismatch positions to PAM. Hsu *et al.* (2013) proposed a scoring algorithm that assigns a weight to each nucleotide mutation. Those three methods all consider the counts and positions of the mismatches between sgRNA and DNA. Doench *et al.* (2014) proposed a CFD scoring algorithm which not only considers the counts and positions, but also identify the related sequence context.

Given the deluge of the available tools, we would like to explore the potential of ensemble learning on synergizing multiple tools together and see if we can push our off-target prediction capabilities to the maximum. In addition, we would also like to investigate whether evolutionary conservation and chromatin annotations could be adopted to enhance the predictive capabilities further, avoiding any side effect in the near future.

2 Materials and methods

In this section, we proposed ensemble learning models to predict the off-targets activities. Different ensemble insights could be observed from the experimental results. In addition, chromatin features, including evolutionary conservation data (PhyloP and PhastCons scores) and Chromatin state segmentation data (ChromHMM and Segway) are introduced into the model to explore whether we can improve the prediction performance further.

2.1 Overview of the ensemble learning model

Figure 1 depicts the ensemble learning structure. The input of the ensemble learning model contains five scores calculated by five scoring methods (CFD, MIT, MIT Website, CCTop, and Cropit), evolutionary conservation data (PhyloP and PhastCons scores) and Chromatin state segmentation data (ChromHMM and Segway). The output of ensemble learning is a score between 0 and 1, indicating the probability of off-target activity. The value of 1 indicates perfect match off-target activity, while the value of 0 indicates the opposite.

2.2 Datasets

The dataset used in this study contains 25 332 putative off-target DNA sequences with a mismatch count of up to four (from human genome hg19) identified by CRISPOR and 26 sgRNAs with modification frequency filtering >0.01% (Haeussler *et al.*, 2016). In this dataset, there are 152 verified off-target (positive) DNA sequences

Table 1. Examples of DNA, sgRNA and the coordinates of DNA

sgRNA	Putative off-target DNA sequences	Chrom	Start	End
AAATGAGAAGAAGAGGCACA**GGG**	AAATGAGAAGAAGAGGCACA**GGG**	Chr3	46399717	46399739
AACACCAGTGAGTAGAGCGG**AGG**	TACTCCAGTGAGTAGAGAGG**CGG**	Chr16	87942783	87942805
CTTGCCCCACAGGGCAGTAA	TCAGCCCCACAGGGCAGTAA**GGG**	Chr9	104595866	104595888
GAACACAAAGCATAGACTGC**GGG**	GAACACAATGCATAGATTGC**CGG**	Chr4	90522167	90522189
GACACCGAAGCAGAGTTTTT**AGG**	GACAACAAAGTAGAGTTTTT**AGG**	Chr7	100221100	100221122

Note: In each DNA sequence, the last three nucleotides are PAM highlighted in bold.

with label 1. The others are all non-off-targets (negative). Such a class imbalance problem imposes challenges in predictive model building. It also necessitates our ensemble learning approaches for precision. Table 1 tabulates examples of putative off-target DNA sequences, sgRNAs, and the genomic coordinates of putative off-target DNA sequences.

2.3 Features

For the putative off-target sequences, there are five scoring methods to give scores to each off-target sequence based on the counts, locations, and identities of mismatches between DNA and sgRNA. These scores are considered as the features of classifiers. As the designed structure of CRISPR/Cas9, the Cas9 endonuclease is guided by sgRNA to the target sites and then cleaved for precise gene editing. In addition, several recent works indicate that the chromatin organization can influence Cas9 endonuclease binding by limiting the accessibility of the target site (Chen *et al.*, 2016; Knight *et al.*, 2015). In particular, Uusi-Mäkelä *et al.* (2018) found a correlation between chromatin accessibility and the efficiency of CRISPR/Cas9 mutagenesis; its experimental results indicate that CRISPR/Cas9 mutagenesis is influenced by chromatin accessibility in zebrafish embryos. Therefore, in this study, we introduce chromatin features, including evolutionary conservation data (PhyloP and PhastCons scores) and chromatin state segmentation data (ChromHMM and Segway), into the model. The following Table 2 demonstrates examples of those scores and chromatin features for each putative off-target DNA sequence in Table 1.

2.3.1 Off-target scoring methods

CFD score. The CFD scores stand for cutting frequency determination, and is proposed by Doench *et al.*, (2014) to calculate the off-target potential of sgRNA-DNA intersections. In CFD scoring method, the position, number, and identity of mismatches between the sgRNA and target DNA sequences play the major roles in determining activity (Doench *et al.*, 2014).

CCTop score. Stemmer *et al.* (2015) proposed a CRISPR/Cas9 off-target online predictor (CCTop, http://crispr.cos.uni-heidelberg.de) and it provides a user interface to tune the parameters for condition adjustment. In CCTop scoring method, the position and number of mismatches between the sgRNA and target DNA sequences are considered.

Cropit score. A web-based CRISPR/Cas9 Off-target Prediction and Identification Tool (Cropit) is introduced to perform improved off-target binding and cleavage site predictions. Cropit not only utilized the position and number of mismatches, but also used whole-genome chromatin state information (DNase I HS data) as the features.

MIT Website. Hsu *et al.* (2013) proposed a scoring method to calculate and evaluate the potential off-target sequences by assigning a weight per position of mismatch between sgRNA and target DNA. However, it can only consider the cases up to four mismatches.

MIT score. MIT score (Haeussler *et al.*, 2016) reused the rules from CRISPOR but MIT score can deal with over 4 mismatches (Hsu *et al.*, 2013).

2.3.2 Evolutionary conservation

PhyloP scores. PhyloP scores measure evolutionary pattern of conservation or acceleration at individual alignment site. The positive PhyloP score measures the evolutionary conservation which is slower than the evolution expected under neutral drift, at sites that are predicted to be conserved. The negative PhyloP score measures the evolutionary conservation which is faster than the evolution expected under neutral drift (i.e. fast-evolving sites).

PhastCons scores. The first difference between PhastCons and PhyloP is that PhyloP only considers the current sequence positions of interest, while PhastCons also consider the adjacent sequence context. The second difference is that PhastCons scores have been normalized between 0 and 1.

2.3.3 Chromatin annotations

ChromHMM. ChromHMM is implemented as a multivariate Hidden Markov Model (HMM) to learn and characterize chromatin states (Ernst and Kellis, 2012, 2017). ChromHMM could model the presence or absence of each chromatin mark for genomic annotations (Ernst *et al.*, 2011).

Segway. Segway relies on a dynamic Bayesian network (DBN) model to automatically segment the genome using ChIP-seq, DNase-seq, and FAIRE-seq data from ENCODE (Hoffman *et al.*, 2012). The segmentation task can be divided into two parts: (i) finding genomic segment boundaries; (ii) and assigning labels to those genomic segments (Hoffman *et al.*, 2012). There are multiple labels and we focus on the regulatory regions which are labeled as ‘active promoter’ and ‘active enhancer’ in Segway.

2.4 Classification methods

In this study, we designed and implemented five machine learning methods to evaluate the ensemble performance of the above five scoring features and chromatin features. The classification methods are listed as follows: AdaBoost (Freund and Schapire, 1996), Random Forest (Breiman, 2001), Multilayer Perceptron (MLP) (Bishop, 1995), Support Vector Machine (SVM) (Burges, 1998), and Decision Tree (Quinlan, 1999).

2.5 Evaluation criteria

We measured two criteria to evaluate the prediction performance of the aforementioned methods and select the best method with the best feature combination. We evaluated the models by using 5-fold stratified cross-validation, because the dataset utilized is imbalanced with only 152 validated off-targets. The first one is the area under the receiver operating characteristic (ROC) curves (AUC_{ROC}) which are based on two metrics: the true positive rate (TPR) and the false

Table 2. Example of off-target scores and Chromatin features

CFD	MIT website	MIT	Cropit	CCTop	PhyloP	PhastCons	ChromHMM		Segway	
0.023	0	0.061	455	189.753	2	0.757	0	0	0	0
0	0.030	0.031	425	184.402	−0.043	0	0	0	0	0
0.028	0	1.275	530	207.452	0.543	0.300	1	1	0	0
0.010	0	0.372	552.5	193.381	3	0	0	0	0	0
0.001	0.085	0.082	327.5	188.390	−0.109	0	0	0	0	1

Note: ChromHMM and Segway contain two columns where the left column denotes the presence of the genomic annotation "active promoter" and the right column denotes the presence of the genomic annotation active enhancer.

positive rate (FPR). The second one is the area under the precision-recall curves (PRC; AUC_{PRC}) which is constituted by precision and recall. Two criteria are adopted because the off-target prediction problem has the class imbalance issues which ROCs and PRCs can complement each other to address.

3 Results

In this section, we first evaluate different ensemble combinations of five scoring methods using the AdaBoost model and compare those to individual scoring methods. After that, we explore to aggregate chromatin features (including evolutionary conservation and chromatin annotations) into the performance of the best scoring methods' combinations to test whether it can improve the prediction of CRISPR off-target activities further. Finally, we compare the proposed ensemble learning model with other classification models, demonstrating that the proposed model can yield robust and competitive performance.

3.1 Ensemble of scoring methods exhibits better predictive power on CRISPR off-target activities

We started by training and testing different ensemble combinations of the five aforementioned tools including CFD (Doench *et al.*, 2014), the current state-of-the-art according to (Haeussler *et al.*, 2016), MIT Website (Hsu *et al.*, 2013), MIT (Haeussler *et al.*, 2016), Cropit (Singh *et al.*, 2015), CCTop (Stemmer *et al.*, 2015) to explore the synergy among the off-target prediction capabilities. First, we aggregated all five scoring tools to a summary score using AdaBoost and compared it with the individual scoring tools as benchmarks under 5-fold stratified cross-validation. As depicted in Figure 2, we can observe that the ensemble learning of all five individual scoring tools synergized the off-target predictive performance to a very impressive performance of $AUC_{ROC} = 0.938$ which outperforms the individual tools. The AUC_{PRC} of ensemble learning on all five scoring tools also outperforms other individual tools. Second, we tested different possible combinations of 4 individual scoring tools. The results are depicted in Figure 3, interestingly, the combination of CFD, MIT Website, MIT, and Cropit performs better than other combinations in ROC ($AUC_{ROC} = 0.9374$), while its $AUC_{PRC} = 0.299$ is lower than $AUC_{PRC} = 0.312$ of the combination of CFD, MIT Website, MIT, and CCTop. It indicates that there is a performance trade-off between Cropit and CCTop if we already have the scores from CFD, MIT Website, and MIT. We also observe that the ensemble without CFD performed notably worse than the other combinations. It suggests that CFD plays a driver role in the predictive performance. To investigate it further, we tested all possible combinations of two individual scoring tools with CFD. The results can be observed from Figure 4. Consistently, the combination of CFD, MIT Website, Cropit performs better than other combinations in ROC ($AUC_{ROC} = 0.936$), while its $AUC_{PRC} = 0.273$ is lower than $AUC_{PRC} = 0.286$ of the combination of CFD, MIT, and CCTop. It suggests that MIT Website and Cropit can help increase the sensitivity of CFD at the expense of precision loss which may not be acceptable in the clinical settings. To conclude the ensemble learning study, we tested different pair-wise combinations. The results are plotted in Figure 5 where the combination of CFD and MIT performs better than other combinations both in ROC ($AUC_{ROC} = 0.931$) and PRC ($AUC_{PRC} = 0.271$), demonstrating that such simple pair-wise combination could result in the predictive performance, compared to the complete ensemble performance of the tools tested. It hints us that we should build on top of CFD and MIT to explore how we can improve the performance further.

3.2 Combinatorial contribution of evolutionary conservation data to predict CRISPR off-target activities

In this study, we introduced PhyloP and PhastCons to the above ensemble learning models to explore if the evolutionary conservation can enhance its predictive capabilities further. We firstly retrieved the PhyloP and PhastCons data according to the genomic coordinates of the putative off-target sequences using the annotation package GenomicScores (v1.3.3) of the R software (v3.4.3 - 2017-11-30). After that, we aggregated the two evolutionary conservation data into the four ensemble learning models with the best performance in each case of multiple scoring tools combinations. The comparison results are visualized in Supplementary Table S1 according to the mean AUCs across 5-fold stratified cross-validation of ROC and PRC curves. In addition, the standard deviations are provided in Supplementary Figures S1–S3. We can see that PhyloP can enhance the predictive capabilities of off-target activities for all four ensemble cases and the most obvious performance improvement is in the fourth case of the ensemble of CFD and MIT. However, we found that PhastCons didn't increase the performance improvement at all. Therefore, we decide to forgo including PhastCons into our deployed model on GitHub.

3.3 Combinatorial contribution of chromatin annotations data to predict CRISPR off-target activities

In this part, we tested two labels ('active promoter' and 'active enhancer') of ChromHMM and Segway to the above ensemble learning models to see if the chromatin annotations can enhance its predictive capabilities further. The ChromHMM and Segway data can be accessed according to the genomic coordinate of each putative off-target sequence using the annotation package Genomation (v1.10.0) of the R software (v3.4.3 - 2017-11-30). After that, we augmented the ensemble learning models with the best performance in each case of multiple scoring tool combinations to aggregate the two chromatin annotations data separately. Here we evaluated the results using the AUCs across 5-fold stratified cross-validation of

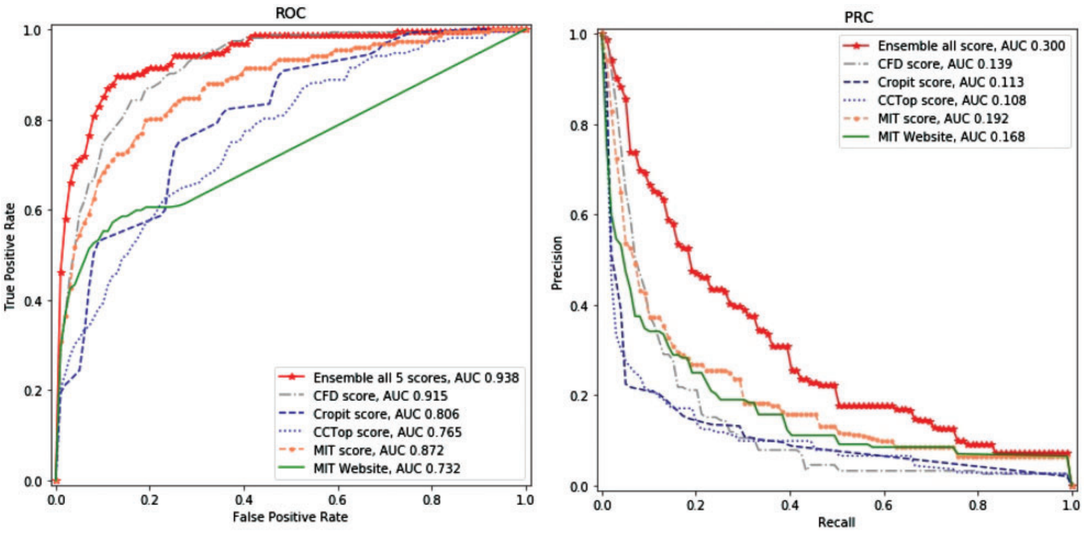

Fig. 2. Performance comparison of the ensemble all 5 scoring methods and single scoring method using AdaBoost under 5-fold stratified cross-validation

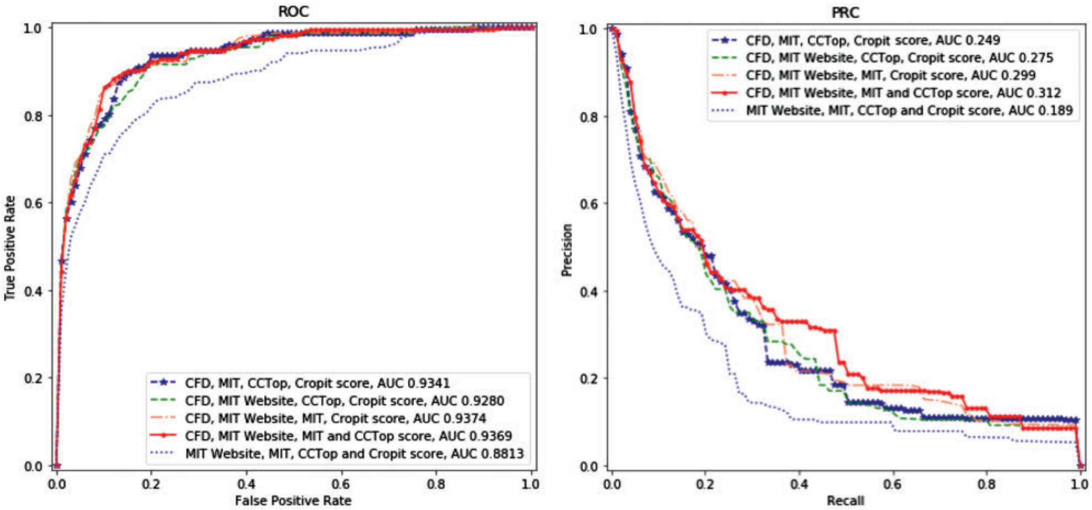

Fig. 3. Performance comparison of the combination of 4 scoring methods using AdaBoost under 5-fold stratified cross-validation

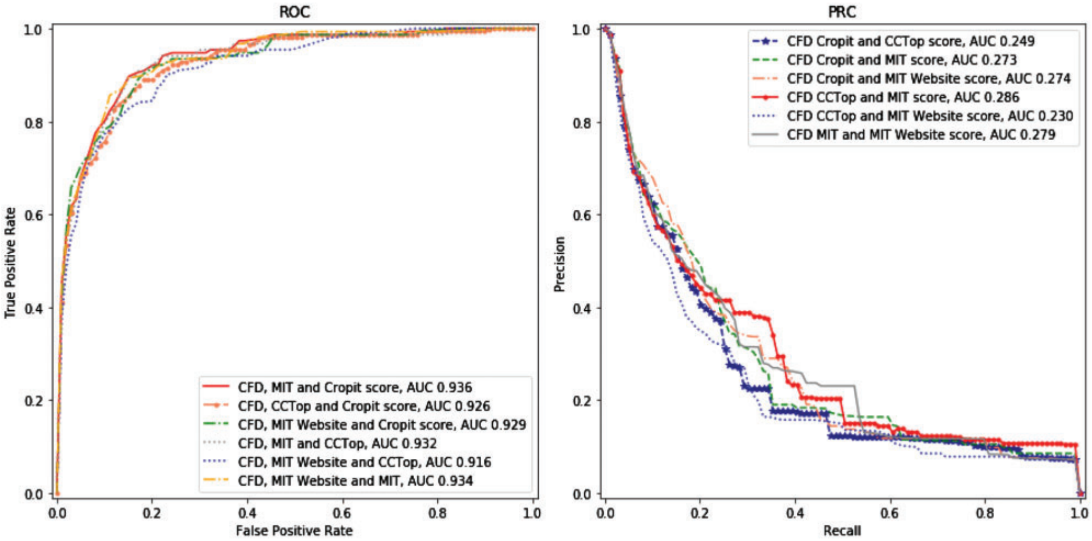

Fig. 4. Performance comparison of the combination of 3 scoring methods using AdaBoost under 5-fold stratified cross-validation

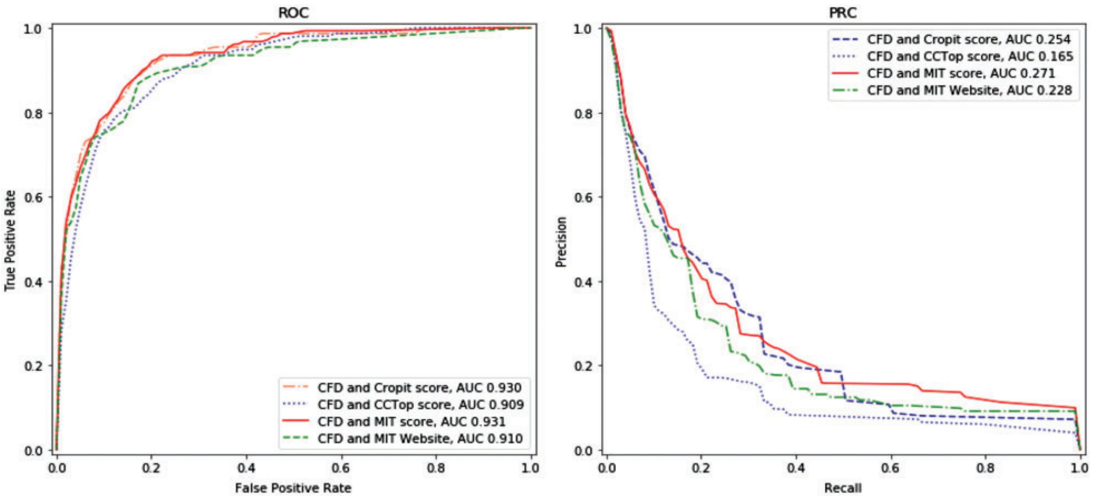

Fig. 5. Performance comparison of the combination of 2 scoring methods using AdaBoost under 5-fold stratified cross-validation

ROC and PRC curves shown in Supplementary Table S2. In addition, the standard deviations are provided in Supplementary Figures S4 and S5. We found that the promoter and enhancer annotations of ChromHMM and Segway included in the ensemble learning models did not obviously enhance the predictive capabilities of off-target activities in any ensemble case unfortunately. One of the reasons could be that only very few of our putative off-target sequences have the labels of promoter and enhancer. Nonetheless, we do believe such a limitation will be alleviated in the near future when more CRISPR/Cas9 datasets are released.

3.4 Comparison of different classification methods using the best feature combination to predict CRISPR off-target activities

After we tested different ensemble combinations of five scoring methods and investigated the contributions of evolutionary conservation and chromatin annotations for improving the predictive performance, we have arrived at the ensemble learning model with the best combination of scoring tools and the effective chromatin features. Therefore, in this section, we explore to train and test different classification models (AdaBoost, Random Forest, MLP, SVM, and Decision Tree) from scikit-learn (0.19.1) for ensemble learning. Figure 6 compares the ROC and PRC curves of the five methods. It can be observed that AdaBoost outperforms all other methods on both AUC_{ROC} and AUC_{PRC}, whereas MLP has the second performance. SVM has the worst predictive performance. It should be noted that AdaBoost is an iterative ensemble learning method. In each iteration, AdaBoost introduces a new weak classifier (i.e. decision stump) to learn the DNA sequences that were misclassified to off-targets until a small error rate is reached. Therefore, AdaBoost can learn from the imbalanced dataset in a regularized manner, achieving the best predictive performance of CRISPR off-targets.

Below is the parameter setting of the used classification methods: For AdaBoost, the decision stumps are used as the base classifiers, the learning rate is set to 1.0, and the SAMME is used as the discrete boosting algorithm. For Random Forest, the number of trees is set to 200 and the max depth of tree is set to 1. For MLP, the activation is set to relu, the learning rate is 0.001. For SVM, the RBF kernel is used. Decision Tree's parameters are set to be default.

Supplementary Figures S6 and S7 illustrate the misclassification rate and weight of each base classifier (decision stump) for 5-fold

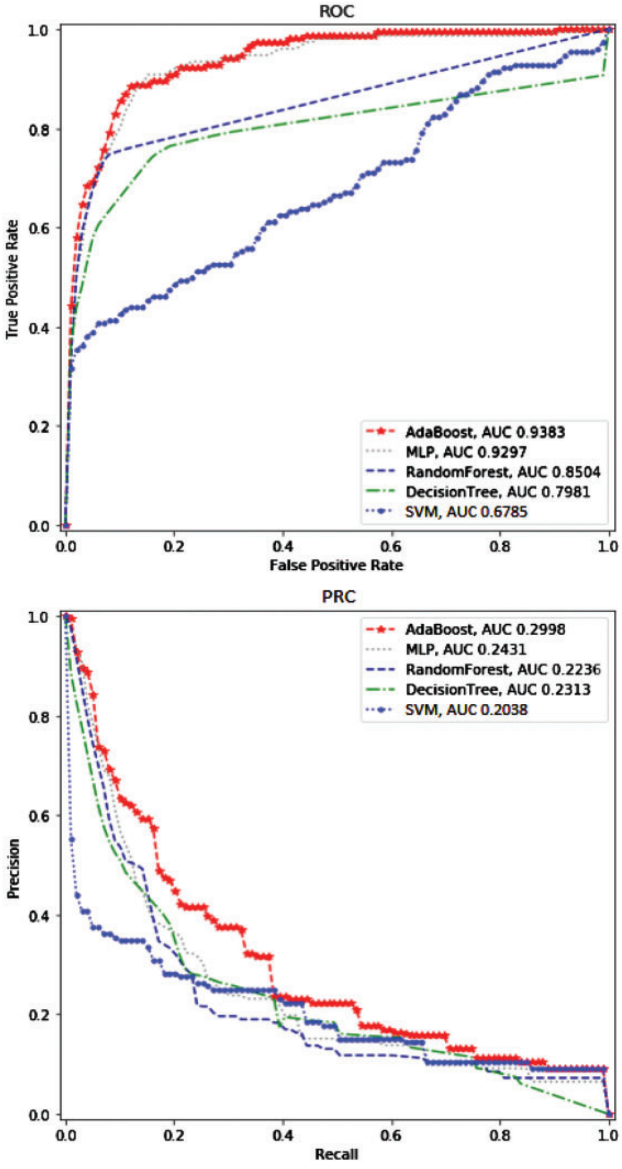

Fig. 6. Performance comparison of AdaBoost and other classifiers

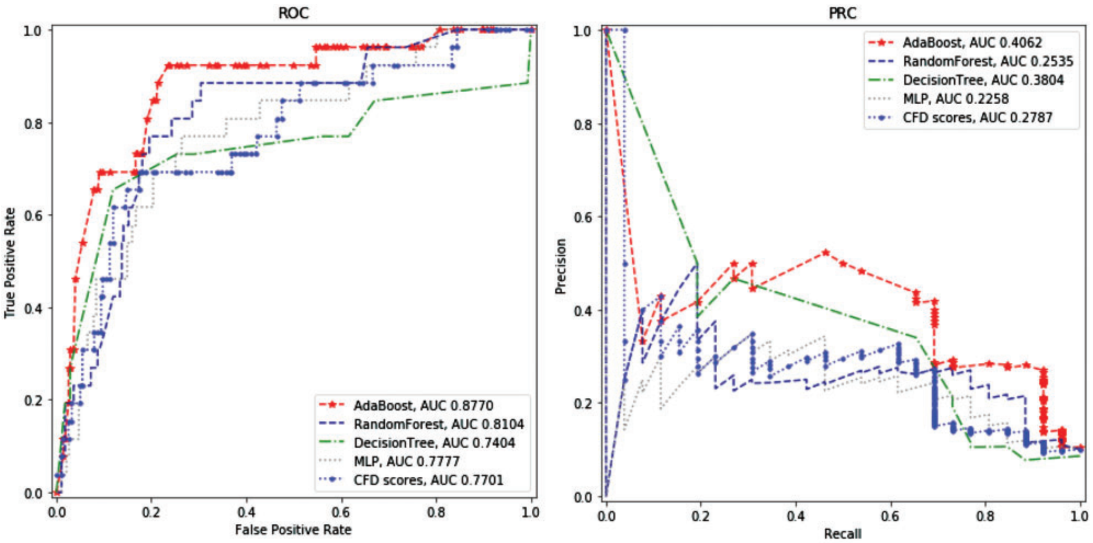

Fig. 7. Performance comparison of AdaBoost, other classifiers and CFD scores

cross-validation respectively. Supplementary Table S3 tabulates the feature importance of 5-fold cross-validation. Supplementary Table S4 tabulates the data examples of six features used in the base classifiers. The misclassification rate of each base classifier is only a little better than random guess (0.5). In initialization, the weight of each training data is $1/N$ (N is the number of training data). In each iteration, the weight distribution of the training data is constantly modified according to the error rate e_m (m-th base classifier) since it is necessary to ensure that the misclassification rate e of the weak classifier is less than 0.5. The data weight updating rule is to increase the resampling weight of misclassified data stumpsfor retraining base classifiers (i.e. decision stumps) in each iteration. According to the weight α_m of each base classifier, AdaBoost aggregates all the weak base classifiers to obtain a strong classifier $G(X) = \sum_{m=1}^{M} \alpha_m G_m(x)$ where M denotes the number of iterations (base classifier). Supplementary Table S5 tabulates the decision results of the first ten base classifiers for the five data examples shown in Supplementary Table S4. For example, $\sum_{m=1}^{200} \alpha_m G_m(Data1) = 0.3194 < 0.5$, then $G(Data1) = 0$; and $\sum_{m=1}^{200} \alpha_m G_m(Data2) = 0.5364 > 0.5$, then $G(Data2) = 1$.

3.5 Case studies

The first case study was conducted to reveal the ensemble insights into the off-target predictions. From the above investigations, we have found that the final deployed model by AdaBoost (the combination of all five scoring tools included PhyloP) showed the best off-target predictive performance ($\text{AUC}_{\text{ROC}} = 0.9383$ and $\text{AUC}_{\text{PRC}} = 0.2998$) than other methods. We picked the corresponding best predictions predicted by the deployed model and other individual scoring methods (CFD, MIT, MIT Website, CCTop, and Cropit). Because the predict results of individual tools are just scores and have different statistical schemes, it is unreasonable to select a cutoff score to all methods. For the whole dataset used in this study, it has 152 verified off-target DNA sequences. We selected the top 152 predictions from the above methods on whole dataset. We found that there are 13 DNA sequences predicted as off-targets by the deployed model but cannot be predicted by other five scoring methods. This is ascribable to the two facts that (i) AdaBoost is an iterative algorithm that focus on the DNA sequences that were misclassified. In addition, AdaBoost relies on the weak classifier to boost the predictive performance; (ii) the genomic feature (PhyloP) plays an important role in picking up the 13 DNA sequences that cannot be recognized by individual scoring methods.

The second case study is conducted on the GUIDE-Seq dataset (Tsai *et al.*, 2015) to test the ensemble learning model (trained on the CRISPOR dataset). The GUIDE-Seq dataset contains 403 putative off-targets and 28 off-targets validated by GUIDE-Seq. We compared the AdaBoost model with other classification methods and CFD scores. The results are depicted in Figure 7; the AdaBoost model still outperforms other classifiers and CFD scores both in ROC and PRC.

4 Discussion

Last year, CRISPR/Cas9 has been adopted to the gene editing on human embryos (Ma *et al.*, 2017). It raised numerous concerns and public debates. Even worse, the whole-genome sequencing study on a CRISPR/Cas9-edited human and zebrafish cells have revealed an unexpectedly a number of off-target mutations incurred by CRISPR/Cas9 (Fu *et al.*, 2013; Hruscha *et al.*, 2013;). It is extremely necessary for us, human, to manage and control the off-target activities before it becomes too late.

To practically address it, we have explored different ensemble learning approaches to predict the off-targets effects of the CRISPR/Cas9 system. The proposed ensemble learning method is based on synergizing multiple CRISPR off-target prediction tools together to push our off-target prediction capabilities to the maximum. Through systematic testing on different combinations of those tools, we included the best predictive one into the final deployed model which can perform better than the individual prediction tools. In addition, we also investigated the effect of evolutionary conservation and chromatin annotations on improving the predictive capabilities. Unfortunately, we found that only PhyloP could be adopted to enhance the predictive capabilities further.

Throughout the benchmark studies on the experimentally verified CRISPR/Cas9 data, we also observed different mechanistic and functional insights into the advantages and disadvantages of different tools through ensemble learning. We do hope that those insights could help paving the foundation of precision medicine, mitigating unexpectedly incurred side-effect in living organisms to the best of our efforts in the near future.

Acknowledgements

The authors would like to thank the anonymous reviewers for their time in reading our manuscript.

Funding

The work described in this paper was substantially supported by three grants from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 21200816], [CityU 11203217], and [CityU 11200218].

Conflict of Interest: none declared.

References

Bae,S. *et al.* (2014) Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics*, **30**, 1473–1475.

Billon,P. *et al.* (2017) CRISPR-Mediated base editing enables efficient disruption of eukaryotic genes through induction of STOP codons. *Mol. Cell*, **67**, 1068–1079.e4.

Bishop,C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford UniversityPress, Inc., New York, NY, USA.

Breiman,L. (2001) Random forests. *Mach. Learn*, **45**, 5–32.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov*, **2**, 121–167.

Chen,X. *et al.* (2016) Probing the impact of chromatin conformation on genome editing tools. *Nucleic Acids Research*, **44**, 6482–6492.

Cho,S.W. *et al.* (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.*, **31**, 230–232.

Cho,S.W. *et al.* (2014) Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.*, **24**, 132–141.

Cradick,T.J. *et al.* (2013) CRISPR/Cas9 systems targeting *β*-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.*, **41**, 9584–9592.

Cradick,T.J. *et al.* (2014) COSMID: a web-based tool for identifying and validating CRISPR/Cas off-target sites. *Mol. Ther. Nucleic Acids*, **3**, e214.

Doench,J.G. *et al.* (2014) Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1262–1267.

Ernst,J., and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Ernst,J., and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.

Freund,Y., and Schapire,R.E. (1996) Experiments with a new boosting algorithm. *Mach. Learn*, **96**, 148–156.

Fu,Y. *et al.* (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.

Haeussler,M. *et al.* (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, **17**, 1–12.

Heigwer,F. *et al.* (2014) E-CRISP: fast CRISPR target site identification. *Nat. Methods*, **11**, 122–123.

Hoffman,M.M. *et al.* (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.

Hruscha,A. *et al.* (2013) Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. *Development*, **140**, 4982–4987.

Hsu,P.D. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.

Hsu,P.D. *et al.* (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262–1278.

Jinek,M. *et al.* (2012) A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, **2012**, 1225829.

Kim,D. *et al.* (2015) Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods*, **12**, 237–243.

Kleinstiver,B.P. *et al.* (2016) High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**, 490–495.

Knight,S.C. *et al.* (2015) Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science*, **350**, 823–826.

Lin,Y. *et al.* (2014) CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.*, **42**, 7473–7485.

Ma,H. *et al.* (2017) Correction of a pathogenic gene mutation in human embryos. *Nature*, **548**, 413–419.

Mali,P. *et al.* (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.

Meng,X. *et al.* (2017) Construction of a genome-wide mutant library in rice using CRISPR/Cas9. *Mol. Plant*, **10**, 1238–1241.

Montague,T.G. *et al.* (2014) CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.*, **42**, W401–W407.

Naito,Y. *et al.* (2015) CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. *Bioinformatics*, **31**, 1120–1123.

Quinlan,J. (1999) Simplifying decision trees. *Int. J. Hum. Comput. Stud*, **51**, 497–510.

Peng,R. *et al.* (2016) Potential pitfalls of CRISPR/Cas9-mediated genome editing. *Febs J.*, **283**, 1218–1231.

Radecke,S. *et al.* (2018) Genome-wide mapping of off-target events in single-stranded oligodeoxynucleotide-mediated gene repair experiments. *Mol. Ther.*, **26**, 115–131.

Ran,F.A. *et al.* (2015) In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature*, **520**, 186–191.

Rosenbluh,J. *et al.* (2017) Complementary information derived from CRISPR Cas9 mediated gene deletion and suppression. *Nat. Commun.*, **8**, 15403.

Sander,J.D. *et al.* (2010) ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. *Nucleic Acids Res.*, **38**, W462–W468.

Scott,D.A., and Zhang,F. (2017) Implications of human genetic variation in CRISPR-based therapeutic genome editing. *Nat. Med.*, **23**, 1095–1101.

Singh,R. *et al.* (2015) Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Res*, **43**, e118–e118.

Stemmer,M. *et al.* (2015) CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction Tool. *PLoS One*, **10**, e0124633.

Tsai,S.Q. *et al.* (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.*, **33**, 187–197.

Tsai,S.Q. *et al.* (2017) CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat. Methods*, **14**, 607–614.

Uusi-Mäkelä,M.I. *et al.* (2018) Chromatin accessibility is associated with CRISPR-Cas9 efficiency in the zebrafish (*Danio rerio*). *PLoS One*, **13**, e0196238.

Wang,M. *et al.* (2017) Multiplex gene editing in rice using the CRISPR-Cpf1 system. *Mol. Plant*, **10**, 1011–1013.

Wang,X. *et al.* (2015) Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat. Biotechnol.*, **33**, 175–178.

Wolter,F., and Puchta,H. (2017) Knocking out consumer concerns and regulator's rules: efficient use of CRISPR/Cas ribonucleoprotein complexes for genome editing in cereals. *Genome Biol.*, **18**, 43.

Xiao,A. *et al.* (2014) CasOT: a genome-wide Cas9/gRNA off-target searching tool. *Bioinformatics*, **30**, 1180–1182.

Xu,X. *et al.* (2017) CRISPR-Cas9 cleavage efficiency correlates strongly with target-sgRNA folding stability: from physical mechanism to off-target assessment. *Sci. Rep.*, **7**, 143.

Zhou,M. *et al.* (2017) Mathematical and computational analysis of CRISPR Cas9 sgRNA off-target homologies. *Proc. 2016 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2016*, 449–454.

Zhu,L.J. *et al.* (2014) CRISPRseek: a bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. *PLoS One*, **9**, e108424.