1 **CRISPR-GNL: an improved model for predicting CRISPR activity by machine**
2 **learning and featurization**

3 **Jun Wang**^{1,2,3,*}, **Xi Xiang**^{1,2,3}, **Lixin Cheng**⁵, **Xiuqing Zhang**^{1,2}, **Yonglun Luo**¹,
4 ^{2,3,4,*}

5 **Affiliations**

6 1 BGI Education Center, University of Chinese Academy of Sciences, Beijing, China

7 2 BGI-Shenzhen, Shenzhen, China

8 3 Lars Bolund Institute of Regenerative Medicine, BGI-Qingdao, BGI-Shenzhen,
9 Qingdao, China

10 4 Department of Biomedicine, Aarhus University, Denmark

11 5 Department of Critical Care Medicine, Shenzhen People's Hospital, The Second
12 Clinical Medicine College of Ji'nan University, Shenzhen, China

13 * To whom correspondence should be addressed

14 **ABSTRACT**

15 **Motivation:** The CRISPR/Cas9 system has been broadly used in genetic engineering.
 16 However, risks of potential off-targets and the variability of on-target activity among
 17 different targets are two limiting factors. Several bioinformatic tools have been
 18 developed for CRISPR on-target activity and off-target prediction. However, the
 19 general application of the current prediction models is hampered by the great
 20 variation among different algorithms.

21 **Results:** In this study, we thoroughly re-analyzed 13 published datasets with eight
 22 regression models. We proved that the current model gave very low cross-dataset and
 23 cross-species prediction outcome. To overcome these limitations, we have developed
 24 an improved model (a generalization score, GNL) based on normalized gene editing
 25 activity from 8,101 gRNAs and 2,488 features using Bayesian Ridge Regression
 26 model. Our results demonstrated that the GNL model is a better general algorithm for
 27 CRISPR on-target activity prediction

28 **Availability and implementation:** The prediction scorer is available on GitHub
 29 (https://github.com/TerminatorJ/GNL_Scorer).

30 **Contact:** J.W. (wangjun6@genomics.cn) or Y.L. (luoyonglun@genomics.cn)

31 **Supplementary Information:** Supplementary data are available at Bioinformatics
 32 online.

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51 **INTRODUCTION**

52 CRISPR/Cas9 is the adaptive immunity system in bacteria and most archaea
 53 (Horvath and Barrangou, 2010; Koonin and Makarova, 2009), which was first
 54 harnessed for programmable and precision gene editing in 2012. The CRISPR/Cas9
 55 gene editing system is comprised of two key components: a small guide RNA (gRNA)
 56 and a Cas9 endonucleases (Deltcheva, et al., 2011; Martin Jinek, et al., 2012). The
 57 gRNA is a chimeric RNA molecule of tracrRNA and crRNA (Ran, et al., 2013) that
 58 guides the Cas9 protein to the target site in the genome. Another important feature of
 59 the CRISPR-Cas9 gene editing system is its dependence on a high conserved DNA
 60 motif, also known as the protospacer adjacent motif (PAM). The PAM-dependent
 61 targeting of CRISPR is able to distinguish self and non-self-DNAs. During
 62 CRISPR/Cas9-mediated gene editing, a targeted double-stranded DNA break (DSB)
 63 is firstly introduced. DSBs are subsequently repaired by the DNA repair pathways in
 64 mammalian cells: nonhomologous-mediated end joining (NHEJ) and
 65 homology-directed repair (HDR). Thus, the performance of CRISPR gene editing is
 66 decided by four layers, (1) selection of the best CRISPR gRNA target site (or guide
 67 sequences); (2) an efficient delivery of the CRISPR components to cells/host; (3)
 68 introducing a precise DSB; and (4) DSB repaired by NHEJ or HDR.

69 Selecting a target site with high on-target activity and low off-target is substantially
 70 important for gene editing. Previously, we have discovered that the gene editing
 71 activity of CRISPR-Cas9 in mammalian cells is affected by several factors, such as
 72 the secondary structure and chromatin accessibility of the guide sequences (Jensen, et
 73 al., 2017). To select gRNAs with high activity, we also developed a dual-fluorescence

74 system for experimentally evaluating gRNA activity in cells (Zhou, et al., 2016).
75 Previous results from us and other research groups consistently reveal that the
76 CRISPR gRNA activities are highly variable. Thus, several *in silico* gRNA design
77 web tools and algorithms have been developed to facilitate CRISPR design and
78 applications.

79 Three types of CRISPR design software have been developed based on both
80 experimental and simulated data: (i) alignment-based, where the CRISPR guide
81 sequences were retreated based on mapping PAM sequences in the genome; (ii)
82 hypothesis-driven, in which gRNA activity score is given based on the specific
83 features such as GC content. (iii) machine learning based, where gRNA efficiency
84 score is predicted by models trained with big datasets (Chuai, et al., 2017). An
85 increasing number of CRISPR gene editing data are generated by different
86 laboratories around the world. Big data-driven learning based methods have become
87 the keys in CRISPR gRNA activity prediction. Also, a growing number of features
88 have been discovered to be of importance for gRNA activity. Thus, featurization of
89 the gRNA sequences has been widely used to improve the prediction accuracy, such
90 as SSC (Xu, et al., 2015), gRNA Designer (Doench, et al., 2016), gRNA Scorer
91 (Chari, et al., 2015), CRISPR-DT (Zhu and Liang, 2018). DeepCas9 (Xue, et al.,
92 2018), CNN_std (Lin and Wong, 2018) and CRISPRcpfl (Kim, et al., 2018) utilized
93 the convolution neural network (CNN) to predict the gRNA activity based on the
94 automatic recognition of sequence features, resulting in high performance in both
95 Cas9 and cpfl protein.

96 However, limitations exist when applying these models for gRNA activity
97 prediction, such as the high variability among different species and datasets.
98 Species-specific software were developed, such as fryCRISPR for *Drosophila* (Gratz,
99 et al., 2014), CRISPR-P for plant (Lei, et al., 2014), and CRISPRscan for zebrafish
100 (Moreno-Mateos, et al., 2015). Machine learning based software, such as CRISPOR,
101 have been applied in different species (Chuai, et al., 2017). However, the prediction
102 models had a great bias and limitation for species generalization (Guo, et al., 2018). A
103 benchmark comparison has been done in CRISPOR showing that the generalization

of different scores are extremely low (Haeussler, et al., 2016). So, the heterogeneity of different species needs to be explored. In addition, batch effects and gRNA activity detection methods used different experiments have not been taken into consideration.

In this study, we applied 11 published datasets across different experiments, cells, and species, to compare the performance of each dataset using the same model and evaluation index. Eight models were used for data training to evaluate the data quality and the measurement impact of the experiments. We identified significant variations and difference by the Spearman correlation test. Our results show that the software trained by data from one species is not appropriate for the others. Additionally, an improved model, GNL Scorer (https://github.com/TerminatorJ/GNL_Scorer), was proposed based on data selection and feature collection and trained with two datasets (12340 gRNAs) (Hart, et al., 2015).

MATERIALS AND METHODS

gRNA efficiency data source

We initially retreated 14 datasets from different experiments and five species: human, mouse, zebrafish, *Drosophila*, *Ciona intestinalis* (**Table1**). The HEL dataset was used as the independent validation set together with Z_fish_GZ, Z_fish_VZ, *Ciona*, *Drosophila*, *C. elegans* which were not used to build any other model. We used human knock-out efficiency datasets from six experiments used nine cell types (HCT116 (Hart, et al., 2015), Hela (Hart, et al., 2015), HEK293T (Chari, et al., 2015), HI60 (Wang, et al., 2014), NB4 (Doench, et al., 2014), TF1 (Doench, et al., 2014), MOLMB (Doench, et al., 2014), A375 (Doench, et al., 2016), HEL (Labuhn, et al., 2017)). As for the mouse dataset, we initially used MEsc cell dataset (Koike-Yusa, et al., 2014) and EL4 dataset (Doench, et al., 2016). Three types of zebrafish datasets were used: Gagnon et al. (Gagnon, et al., 2014), Moreno-Mateos et al. (Moreno-Mateos, et al., 2015), and Varshney et al. (Varshney, et al., 2015). We also used the *Ciona* (Gandhi, et al., 2016), *Drosophila* (Ren, et al., 2014), and *C. elegans* datasets (Farboud and Meyer, 2015) as testing datasets. The MEsc and HI60 datasets were bimodal distributed, which were not included in BRR model training to

133 minimize the impact on generalization.

134 For comparison of generalizations for cross-species prediction, the datasets were
135 divided into four categories according to species. Human datasets were from all the
136 human cell type data except HEL. Mouse datasets were separated from the Doench
137 V1 datasets (Doench, et al., 2014), which was trained the gRNA Designer (Rule set I)
138 by 959 gRNA, and MEsc dataset with 949 gRNA from Koike-Yusa et al. (Koike-Yusa,
139 et al., 2014). *Ciona intestinalis* and *Drosophila* datasets were used as previously
140 described.

141 **Normalization of datasets**

142 For better comparability between different datasets, we normalized all the single
143 datasets with batch-effect to the range from -4 to +4. And the dataset of *C. elegans*
144 which could not be normalized to this standard was removed in the combining
145 normalized single dataset. The tool we used to normalize the data was SPASS 20.0.
146 All the 11 datasets (MEsc and HI60 datasets excluded) were combined into a single
147 dataset (all-in-one, AIO) for testing the influence of model performance in the same
148 standard of features and algorithms.

149 All the 10 normalized datasets were used for training when applied to select the best
150 dataset for generalized prediction. We randomly split each of them: 80% for training
151 and 20% for testing. There are six datasets that are used as the validation datasets for
152 comparing the performance of different algorithms. These six datasets were not used
153 to build the machine learning model for other software. For evaluation of the
154 generalized performance, we selected the human HEL, *Drosophila*, *C. elegans*,
155 zebrafish, and *Ciona* as the independent validation datasets.

156 **Predictive models**

157 All prediction models used in this study were based on regression. Eight regression
158 models were used: (i) Gradient-boosted regression tree (GBRT), (ii) Decision tree
159 (DT), (iii) Linear regression (Linreg), (iv) L2-regularized linear regression (L2reg), (v)
160 L1-regularized linear regression (L1reg), (vi) Bayesian Ridge Regression (BRR), (vii)
161 Random forest (RF), (viii) Neural network (NN). All these models were applied by
162 using the scikit-learn package in python. We set the learning rate of 0.01 in GBRT

163 model. The alternative max depth was set to [5, 10, 100, 1000] for grid search in the
164 DT model. We provided $\alpha = [1, 3, 5, 7, 9, 10]$ to select grid in the L2reg model
165 and $\alpha = [0.1, 0.5, 0.75, 1, 10]$ for grid search in L1reg model. As for the Lingre,
166 BBR and RF models, default settings were used. For the NN model, we used two
167 hidden layers with 100 neurons each. We use the 10-fold cross validation for selecting
168 the hyper-parameter setting of each candidate model.

169

170 **Featurization**

171 Three sets of features were evaluated in this study. Feature set I contains 627
172 features, and is partially derived from the Doench features (Doench, et al., 2016). The
173 feature set I includes six parts of features. (i) 604 features of “one-hot” encoding of
174 the nucleotide. There are two subsets in this category: position-dependent and
175 position-independent. And each category applies to the one nucleotide and pairwise
176 nucleotide. Such as “_nuc_pd_Order2” consisted of e.g. AA_1/AT_1/AG_1, and
177 “_nuc_pd_Order1” consisted of e.g. A_1/T_1/G_1/C_1. (ii) 3 GC features, which
178 consists of GC count, GC count < 10, GC count > 10. (iii) 16 features of the two
179 nucleotides flanking the NGG PAM in the 5’ and 3’. (iv) Seven thermodynamic
180 features. We calculated four thermodynamic features as Doench et al. (Doench, et al.,
181 2016) using the “Tm_staluc function” in Biopython package. All these features above
182 were derived from the 30mer of target sequences.

183 The feature set II is a combined set of 2488 features. In addition to those included
184 in the feature set I, we added three extra type of features. (i) Four features of a
185 consecutive of five nucleotides, such as AAAAA/GGGGG. (ii) 1856 features of three
186 nucleotides with “position-dependent” and “position-independent”, such as
187 ACG/AGG and ACG_1/ACT_2. (iii) One feature of delta G. The delta G of seed
188 spacer sequence was calculated by hybrid-ss-min program in the OligoArrayAux
189 package (Zuker, 2003). Note that, (i) and (ii) used the 30mer context sequence. And
190 (iii) used the 20nt guide sequences.

191 Feature set III contains 2701 features. On top of feature set II, we added three
192 additional feature sets: (i) 200 sequence conservation features, (ii) 10 sequence

disorder status features as described by Chen et al. (Chen, et al., 2017), and (iii) Three epigenetic features: chromosome accessibility, CTCF and H3K4me3 from the ENCODE (Consortium, 2004). Position specific iteration BLAST (PSI-BLAST) was used to search the homologous protein against the query protein. We used the NR (non-redundant) database to retrieve the homologous proteins. And the PSSM matrix was calculated by SPOT-Disorder (Heffernan, et al., 2015). The 30mer context sequences are converted to the corresponding amino acid (AA) sequences (10 aa). If the resulting length of the amino acid sequences was shorter than 10, we inserted “X” for the lacking AAs. After generating the PSSM matrix, the 10-amino acid target protein was encoded as 200 (20x10) PSSM features. The disorder status was calculated by SPOT-Disorder (Heffernan, et al., 2015). Each 10-AA sequence was given a disorder probability in this residue. In total, 10 disorder features were included.

206

207 Feature selection

The 2701 features were used to represent the factors that may influence CRISPR gene editing efficiency. However, not all of them would contribute equally. Thus, we used the mRMR method (Peng, et al., 2005) to rank and select features of importance. According to the relationship between the features and their target classes, each feature was ranked in a descending order and generated a feature list called MaxRel. The MaxRel list was generated as described below. Firstly, we calculated the mutual information (MI) to access the correlation between two variables, which are two features among the 2701 features. Using a probabilistic density function $P(x)$, $P(y)$ and $P(x,y)$, $P(x,y)$ was denoted as the joint probabilistic density of $P(x)$ and $P(y)$:

$$I(x; y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy. (1)$$

Next, the highest value of MI to the target class c was calculated. $\{x_i, i=1, \dots, m\}$ was the feature set of m features in S , required to have the largest value of MI to the target class. The MI was denoted as D here.

$$\max D(S, c), D = I(\{x_i, i = 1, \dots, m\}; c). (2)$$

220 Third, the minimal redundancy between each feature pair was calculated. x_j was
 221 selected feature in the whole feature set S , and x_j is the feature need to be selected.
 222 When we calculate in the constrain of (2) and (3) is called
 223 “minimal-redundancy-maximal-relevance” (mRMR) (Ding and Peng, 2005).

224

$$\min R(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j). \quad (3)$$

225 For each x_j in the reset of feature in the S , D-R is calculated, Then the x_j is
 226 transferred to x_i . When all the x_j is turn to x_i , the whole procedures will be stopped.

$$\max \phi(D, R), \phi = D - R. \quad (4)$$

227 After completing the calculation as described above, we generate list as in (5). And
 228 we will also generate a serial of feature sets like S_1, S_2, S_3 in the S . Then, each sub
 229 feature set was used to train the BBR model and measured by a corresponding
 230 evaluation value (Chen, et al., 2017). This method was called IFS (incremental
 231 feature selection).

$$S = [f_1, f_2, \dots, f_N] \quad (5)$$

232 To select the best combination of features, we use the value of Spearman
 233 Correlation when training the model in the way of IFS. The best BBR was trained and
 234 got the best Spearman value of 0.49 when use 1839 features.

235

236 **Model evaluation**

237 We used Spearman Correlation to assess the regression model, the Spearman
 238 Correlation could make use of the real-valued scores better than classification
 239 evaluation (Fusi, et al., 2015). The Classification-based algorithm discards the
 240 real-valued assay measurement, and converts them to binary active/inactive. Thus, it
 241 is unable to capture the more nuanced information available in the data. The
 242 Spearman Correlation conducted using the *Scipy* package used in Python.

243

244

245 **RESULTS**

246 **Data normalization can alleviate batch-effect among different CRISPR activity** 247 **datasets**

248 We used 13 broadly used datasets (summarized in **Table 1**) with the objective of
249 developing a new model with good performance and generalization. The Doench
250 dataset (Doench, et al., 2016) was divided into 2 versions (V1 and V2) based on how
251 CRISPR gRNA activity was evaluated (resistance assay or flow cytometer). Since all
252 the 13 different datasets are from different studies and species, we first evaluated the
253 data normality and found that not one was normally distributed (**Figure S1**). Thus a
254 data normalization step was applied (see method) and after normalization, the *C.*
255 *elegans* dataset (Farboud and Meyer, 2015) was removed due to failure in data
256 normalization. We normalized the remaining 12 datasets to the gRNA activity range
257 from -4 to 4 (**Figure S2**). Then we combine them as a new dataset called “normalized
258 all-in-one (AIO) dataset”.

259

260 **Table1:** Summary of CRISPR activity datasets used in this study

Dataset	Nr. of gRNAs	Species	Cell type	Software	Model	Detection method	Ref
Ciona	72	Ciona	1-cell embryos	/	/	HTS ^a -% Indel	Gandhi, et al., 2016
Drosophila	39	<i>Drosophila</i>	syncitial embryos	/	/	%heritable mutation rate	Ren, et al., 2014
C.Elegans	50	<i>C.elegans</i>	hermaphrodite worms	/	/	%mutant F1 progeny	Farboud and Meyer, 2015
Hct116	4239	human	Hct116	DeepCRISPR	CNN ^c	HTS ^a -quantification	Hart, et al., 2015
Hek293t	1234	human	Hek293T	SgRNA-Scorer/ DeepCas9/ DeepCRISPR	SVM/CNN ^c	HTS ^a -% delta Indel	Chari, et al., 2015
Hela	8101	human	Hela	DeepCRISPR	CNN ^c	HTS ^a -quantification	Hart, et al., 2015
Hi60	2076	human	Hi60	SSC/DeepCas9/ DeepCRISPR	Elastic Net/ CNN ^c	HTS ^a	Wang, et al., 2014
Z_fish_GZ	111	zebrafish	1-cell embryos	/	/	PCR	Gagnon, et al., 2014
Z_fish_MM	1020	zebrafish	1-cell embryos	CRISPR_Scan	Linear regression	PCR	Moreno-Mateos, et al., 2015
Z_fish_VZ	102	zebrafish	1-cell embryos	/	/	CE ^b	Varshney, et al., 2015
Doench_V1	1841	human	NB4/TF1/ /mouse	sgRNA designer (rs I)/ WU-CRISPR	Logistic regression/SVM	FCM	Doench, et al., 2014
Doench_V2	3473	human	A375	sgRNA designer (rs II)	GBRT	resistance assay	Doench, et al., 2016
MEsc	949	mouse	MEsc	SSC	Elastic Net	FCM	Koike-Yusa, et al., 2014
HEL	430	human	HEL	CRISPRater	L1 Linear-regression	FCM	Labuhn, et al., 2017

^a High-Throughput Sequencing.

^b Capillary Electrophoresis.

^c Convolution Neural Network.

To evaluate whether data normalization can alleviate batch effect, we train the normalized and non-normalized Doench (V1 + V2) and AIO datasets using the GBRT (Gradient Boosted Regression Tree) model, a state-of-art model used in gRNA Designer (rule set II) in the feature set I. Our results showed that the non-normalized AIO dataset could not improve the Spearman Correlation by just adding the data without considering the batch-effect of different datasets from different experiments (**Figure 1**). The Spearman Correlations for normalized AIO dataset were significantly improved (p value = 0.045) compared to the non-normalized AIO data, while no difference was found for the normalized and non-normalized the Doench datasets (**Figure 1**). However, even with the combined normalized AIO dataset, the GBRT Spearman Correlation value is significantly lower (p value = 0.044) compared to the Doench datasets. This results collectively suggested that there is a strong batch effect among the different datasets. Most importantly, it indicates that algorithms developed with a specific training dataset may have a strong bias when used to predict the other datasets.

277

288 with a 10-fold approach. Doench dataset outperforms than AIO from left to right, Significance
 289 results are from paired t-test, Asterisk(*) indicates a p value less than 0.05.

290

291 Selection of datasets with the best prediction outcome

292 To select good datasets that give the better general prediction of gRNA activity, we
 293 used 10 normalized datasets and randomly divided each dataset: 80% for training and
 294 20% for testing. And we trained 8 models (see methods) with the 80% training dataset.
 295 The MESC and H160 were not included due to bimodule distribution of the dataset
 296 (Figure S1). We next chose the model with the highest Spearman Correlation for each
 297 specific dataset (Figure S3), and evaluated the generalization of the remaining 20%
 298 testing datasets in all 10 groups. We found that HeLa, HCT116 and Doench V1
 299 datasets had the best performance in overall gRNA activity prediction (Figure 2a, b).
 300 The activity of the gRNAs in the HeLa and HCT116 datasets were measured by a
 301 quantized sequencing detected method (Hart, et al., 2015). In addition, the gRNA
 302 activity datasets generated by sequencing give the best generalization performance
 303 compared to the rest of the datasets generated by other methods. This indicates that
 304 the dataset detected by sequencing-based method could have great generalization
 305 ability. Consistent with our finding, the HeLa and HCT116 datasets had a “cleaner”
 306 supervisory signal for machine learning the same as FC to RES dataset (Fusi, et al.,
 307 2015).

308

Figure 2. The performance of each best model in specific dataset. a) The x axis is the best model for specific 80% train dataset, the y axis was the 20% test datasets. The heatmap shows from the color of yellow to crimson. The darker color means the higher Spearman correlation. b) The mean value of Spearman correlation for each model trained by 80% of specific dataset and predicted by the remained 20% datasets.

Next, we investigated whether the combination of best-performing datasets could further improve model performance compared to the best performing Hela dataset, which we shown that has a great generalization gRNA activity prediction. All three datasets used the BRR as their best performance model (**Figure S3**). We next trained BRR model with four sets of data: (1) Hela dataset; (2) Hela+Doench_V1 datasets, (3) Hela + HCT116 datasets, (4) Hela + HCT116 + Doench_V1 datasets. The result show that the combined HCT116 and Hela dataset has a higher generalized prediction power than the other datasets combination (**Figure 3**). and similarly well as compared to the Hela dataset along.

Figure 3. Combination of the normalized datasets with the highest Spearman correlation. All the datasets are trained by BRR (Bayesian Ridge Regression), which perform the best among each dataset as mentioned before. Here we just split the best three datasets that have the best generalization. Due to the other datasets were less likely to perform well than these. x-axis shows the 20% testing datasets for each one, the Spearman correlation value is the average score of 10-fold cross validation. The thumbnail shows the mean value across all these test datasets for four combined models

Datasets from different species have pool generalization of gRNA activity prediction

As observed above, we next investigated whether models trained by one specific species could be used for gRNA activity prediction in other species. Currently, although all *in silico* CRISPR gRNA design web tools are defined for each species mainly due to their genome specificity. However, the prediction of gRNA activity is not defined as species-specific. To validate that, we grouped the normalized datasets into five species: human, mouse, zebrafish, *C. elegans*, *Drosophila*. Firstly, we trained each regression models using the combined data (the training dataset, 80%) and selected the best performing model for each combined dataset (**Figure S4**). Next, we evaluated the accuracy of gRNA activity prediction of species-specific datasets (the testing dataset, 20%) using the best performing model. Our results showed that each best performing model trained with a species-specific dataset can well achieve

self-prediction (**Figure 4**). However, all models have a poor generalization of cross-species gRNA activity prediction. It was noted that the models trained with human or mouse datasets are more effective in cross-species prediction.

Figure 4: The generalization of the models trained by different species. The rows represent the best model of specific species trained by 80% of each dataset, the optimized model is tested by 20% dataset remained from each species dataset as shown in the columns.

361 Evaluate features importance for gRNA activity prediction

362 Sequence specific, thermal dynamic, and epigenetic features affect CRISPR gRNA
 363 activity. However, one limitation of including epigenetic features is that epigenetic
 364 data is not available for all cell types or species. This might cause decision challenges
 365 when trying to add these epigenetic features into learning model. To evaluate which
 366 combination of features is optimal for gRNA activity prediction, we generated a
 367 combined set of 2701 features including epigenetic markers (see method, feature set
 368 III). These features have been validated or hypothetically suggested to be important
 369 for CRISPR gRNA activity.

370 Next, we used minimal redundancy maximal relevance (mRMR) method to
 371 evaluate which features are important for gRNA activity prediction in our model. we
 372 choose a subset of the human Doench V1 data (NB4) as test data to do the feature
 373 selection (Table 1), because epigenetic features are available for the NB4 cell and can
 374 test the generalized performance of the selected features of whether IFS or not. In
 375 addition, performing feature selection and training in the same dataset can lead to
 376 overfitting. We applied the incremental feature selection (IFS) method to choose the
 377 optimum features for the dataset and model. First, we assumed that $S_1, S_2 \dots S_N$ was

in the set of $S_i = [S_1, S_2 \dots S_N]$. Then, each feature set was used to train the model, and the feature set yielding the best performance was extracted. Based on the Bayesian Ridge Regression (BBR), we tested the optimal combination of features and dataset for gRNA activity prediction. Two best performing datasets were used: the normalized Hela dataset and the combined normalized Hela and HCT116 dataset. The model with the highest Spearman Correlation Coefficient (0.413) was achieved when 1839 features were used after IFS in feature set III of hct116 combine Hela dataset. Moreover, the Spearman Correlation Coefficient is 0.385 when 2029 features were used through IFS using the Hela dataset in feature set III. With the pre-feature set II (see method), we obtained 1163 and 777 features by IFS with the best Spearman correlation for the combined and Hela data, respectively (Figure 5a).

Figure 5: The optimized features selection result. a) The feature selection is conducted by IFS in feature set III and feature set II respectively for searching the best combination of feature in each optimized selected dataset. The upper two figures are the feature selection applied in feature set III, which results in 2,701 features in total, including pssm, disorder, epigenetic features; the two figures below are the feature selection used in the 2,488 features from feature set II. b) Evaluating the optimized models using the NB4 dataset. Two models are trained by feature set II

is shown on the left side, while the ones trained by feature set III is shown on the right side. Each feature set is processed by incremental feature selection (IFS) compared with that of not feature selection. The number of features selected by IFS in feature set II are 777 and 1,163 in the model of Hela and Hct116+Hela, respectively. In feature set III, the features number are 2,029 and 1,839, respectively. c) The statistic of the top 200 optimized features selected by the NB4 cell type dataset after ranking by the algorithm of mRMR in the dataset of Hela using the whole 2,488 features in feature set II. Statistical significance is derived from t-test. *Asterisk* (*) indicates a *p* value less than 0.05, (**) less than 0.01, (***) less than 0.0001, and (****) less than 0.00001.

We also compare the IFS features to the non-IFS model by using 2488 of feature set II and 2701 of feature set III. Our results showed that the model trained by Hct116 + Hela performed the best in the NB4 test data without IFS used the feature set II (2488 features in total) (0.442) (**Figure 5b**). Also, we found that the hela+hct116 model outperforms the Hela model in the human dataset, so the feature selection is not necessary for the best performing model. For example, sgRNA Designer (Rule set II) picks up 13 types of features out of 627 features to train the GBRT regression model (Doench, et al., 2016) and sgRNA scorer utilizes 92 features without any selection to develop a classify model for stCas9 and spCas9 (Chari, et al., 2015).

Next, we evaluated which type of features (among these 2488 optimal feature sets trained with the combined hct116 and Hela datasets) plays the most decisive role in CRISPR gRNA activity. The GC count, nuc_pi_Order1, thermodynamic temperature, the secondary structure of the guide (deltaG) were regarded as the optimal features, because they all showed in the top 200 important features in the mrmr result (**Figure 5c**). The secondary structure of the guide (deltaG) affect CRISPR activity is consistent with our previous findings (Jensen, et al., 2017). Our featurization results suggest that the sequence composition, physicochemical characteristics greatly influence CRISPR gene editing activity. Our analysis also demonstrates that prediction software based on CNN such as DeepCRISPR (Chuai, et al., 2018) and CRISPRCpf1 (Kim, et al., 2018) can soundly predict the efficiency by extracting the feature using the convolution. When ranking was considered, the sequence thermodynamic temperature (TM) is ranked as the most important and decisive feature on gRNA activity. Previously studies have proven that unwinding was the

primary determinant of CRISPR activity (Gong, et al., 2018) and Hui Peng et al. also observed the same outcome (Peng, et al., 2018). Among the top 20 most important features, the composition of continuous bases in specific position in the 20nt of sgRNA take the major importance in the sgRNA activity decision (80%) and the TM was rank as the most important feature, which mean position dependent and the TM the major role in the efficiency in the model trained by combined hct116+hela datasets.

435

436 **Develop new algorithms with improved general gRNA activity prediction**

437 Current algorithms were trained with datasets from a specific cell type and
438 frequently overfitting. This makes the algorithms incompatible with other cell type
439 and species (Yan, et al., 2018). Labuhn et al. had previously proved that the
440 state-of-art algorithm cannot be effectively generalized in small datasets (and) as the
441 Spearman Correlation was just 0.2011 (Labuhn, et al., 2017). To develop an improved
442 CRISPR gRNA activity prediction algorithm with better compatibility among
443 different cell types and species (or generalization), we used the two best performing
444 datasets (Hela and HCT116), the feature set II (see method without IFS), and BRR
445 model to develop two generalization scores (GNL1.0 and GNL2.0). GNL1.0 was
446 based on BBR trained with the normalized Hela dataset and feature set II. GNL2.0
447 was based on BBR trained with the combine Hela + HCT116 datasets and feature set
448 II, whose datasets was seen as the best performance in the human dataset (**Figure 5**).
449 The outcome of our study supports the result of Wilson et al. training on the dataset
450 measured by sequencing results in more accurate activity predictions (Wilson, et al.,
451 2018).

452 To evaluate the generalization, we compared GNL scores to seven algorithms
453 (sgRNA Scorer, DeepCas9, sgRNA designer (rule set I), sgRNA designer (rule set II),
454 SSC, CRISPR_Scan and Wu_CRISPR). Six datasets were used for evaluation (HEL
455 (Labuhn, et al., 2017), Ciona (Gandhi, et al., 2016), *C. elegans* (Farboud and Meyer,
456 2015), *Drosophila* (Ren, et al., 2014), *Z_fish_VZ* (Varshney, et al., 2015), *Z_fish_GZ*
457 (Gagnon, et al., 2014)). The GNL Scores showed the best generalized prediction

outcome (**Figure 6**). The CRISPR-scan algorithm performed well in the zebrafish dataset, but gave a poor prediction outcome in human CRISPR gRNA activity. Previously, using the Hela dataset, it was found that the gRNA Designer (rule set II) performs better than the gRNA designer (rule set I), gRNA Scorer, SSC, and CRISPR-Scan (Chuai, et al., 2018). Consistent with that, our analysis showed that the gRNA Designer (rule set II) performs better than most of the algorithms evaluated in the human HEL dataset. Most importantly, considered all six testing datasets, the GNL1.0 and GNL2.0 have the best generalized prediction. The GNL2.0 is ranked as the best algorithms in predicting the human HEL dataset. Taken together, we have developed two CRISPR gRNA activity prediction algorithms with improved generalization and human gRNA activity prediction. The GNL scores can be freely accessed in GitHub (https://github.com/TerminatorJ/GNL_Scorer).

Figure 6: Comparison of model performance. GNL1.0 (general scorer 1.0) and GNL2.0 (general scorer 1.0) are models we training in this paper, using the best performance datasets (hela or hela+hct116) and the optimized feature set. These two models outperform the seven other algorithms in different species datasets include human.

476

477

478

479

480

481

482 **DISCUSSION**

483 In this study, using combined normalized datasets, we have developed an improved
 484 model for predicting (or an improved prediction model) for CRISPR gRNA activity.
 485 One advantage of our model is that it provides a better generalized prediction
 486 compared to the existing algorithms. This provides a solution to the biggest
 487 disadvantage that most of the machine-learning based software is encountering.
 488 However, as shown in our analysis of the 13 datasets, there exists a great batch effect
 489 and variability of the CRISPR gRNA activity. This might be caused by many factors,
 490 such as the experimental condition, the method of gRNA activity measurement, the
 491 endogenous DNA repair machinery. The most basic principle of CRISPR gene editing
 492 is the precision introduction of a DNA double-strand break (DSB) to the target site,
 493 and repair of the DSB by the endogenous DNA repair machinery (Jiang and Doudna,
 494 2017). In mammalian cells, the DSBs are predominantly repaired by the error-prone
 495 non-homologous end joining (NHEJ) and the microhomology-mediated end joining
 496 (MMEJ) pathways (Yao, et al., 2017), which will cause small deletions or insertions
 497 to the target site (also known as indels).

498 However, due to the impracticability of measuring the indel frequency caused by
 499 thousands of individual gRNA, all current approaches used to quantify CRISPR
 500 gRNA activity are based on functional readouts such as quantification of fluorescent
 501 inactivation by FACS (Doench, et al., 2014) , number of resistant clones (Doench, et
 502 al., 2016), or quantification based on expression (Hart, et al., 2015¹). This partially
 503 explains the poor cross-dataset validation and correction of the different gRNA

activity prediction algorithms, as well as the poor generalized cross-species prediction. To partially overcome this limitation, we in this study applied data normalization, dataset filtering, and dataset combination to develop an algorithm with improved generalized CRISPR gRNA activity prediction. Using the most representative detect datasets, we developed our own model based on the Bayesian Ridge Regression by machine learning. However, the current datasets are still not sufficient enough to build a strong prediction model with high correlations. The batch-effect contributes greatly to the generalization of different models. Thus, there is still a great need to build a benchmark and large-scale approach to experimentally measure the CRISPR activity from a large number of gRNAs.

Both the Spearman Correlation and ROC-AUC curve are typical measurements for evaluating the performance of CRISPR gRNA gene editing activity prediction models. We only used the Spearman Correlation Coefficient in this study, because the ROC-AUC curve was previously shown to be less informative for those RES and FC datasets generated by flowcytometry and resistant assay for gene KO (Fusi, et al., 2015).

Another improvement of our model is featurization. Factors, also known as features, systematically and coherently decides why CRISPR-Cas9 can achieve high targeting efficiency at one locus but not the other. The majority of these decisive features are sequence dependent, while there are still many other features such as chromatin accessibility, DNA repair machinery, epigenetic markers, expression states that are not related to sequences. We had previously experimentally identified some of these important features (Jensen, et al., 2017). However, there might be still many unknown features, sequence-dependent and sequence-independent features that need to be identified in the future. Nevertheless, using machine learning in this study we identified and proved that physicochemical characteristics, sequence composition, the epigenetic features and gRNA secondary structure greatly affect CRISPR gRNA gene editing efficiency. Taken together, we have developed an improved CRISPR gRNA activity prediction algorithms which will contribute to the current in silico design of CRISPR gRNA and will promote the appropriate applications of the CRISPR gene

534 editing technology.

535 **ACKNOWLEDGEMENTS**

536 We thank the Dr. Shan and Dr. Bai for many instructive advices in the statistics and
537 algorithms respectively. And Haeussler who shared the data they collected which save
538 the time of us for initially data processing.

539 **FUNDING**

540 This work was supported by the Lundbeck Foundation (R219–2016-1375,
541 R173–2014-1105), the Danish Research Council for Independent Research
542 (DFF–1337–00128), the Sapere Aude Young Research Talent Prize
543 (DFF-1335–00763A), the Innovation Fund Denmark (BrainStem), and Aarhus
544 University Strategic Grant (AU-iCRISPR). Y.L. is also supported by Guangdong
545 Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011).

546 **REFERENCE**

- 547 Chari, R., *et al.* Unraveling CRISPR-Cas9 genome engineering parameters via a
548 library-on-library approach. *Nat Methods* 2015;12(9):823-826.
- 549 Chen, L., *et al.* Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE*
550 *Access* 2017;5:26582-26590.
- 551 Chuai, G.-h., Wang, Q.-L. and Liu, Q. In Silico Meets In Vivo : Towards Computational
552 CRISPR-Based sgRNA Design. *Trends in Biotechnology* 2017;35(1):12-21.
- 553 Chuai, G., *et al.* DeepCRISPR: optimized CRISPR guide RNA design by deep learning.
554 *Genome Biol* 2018;19(1):80.
- 555 Consortium, E.P. The ENCODE (ENCyclopedia of DNA elements) project. *Science*
556 2004;306(5696):636-640.
- 557 Deltcheva, E., *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor
558 RNase III. *Nature* 2011;471(7340):602-607.

559 Ding, C. and Peng, H. Minimum redundancy feature selection from microarray gene
560 expression data. *Journal of bioinformatics and computational biology* 2005;3(02):185-205.

561 Doench, J.G., *et al.* Optimized sgRNA design to maximize activity and minimize off-target
562 effects of CRISPR-Cas9. *Nat Biotechnol* 2016;34(2):184-191.

563 Doench, J.G., *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene
564 inactivation. *Nat Biotechnol* 2014;32(12):1262-1267.

565 Farboud, B. and Meyer, B.J. Dramatic enhancement of genome editing by CRISPR/Cas9
566 through improved guide RNA design. *Genetics* 2015;genetics. 115.175166.

567 Fusi, N., *et al.* In silico predictive modeling of CRISPR/Cas9 guide efficiency. *BioRxiv*
568 2015:021568.

569 Gagnon, J.A., *et al.* Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion
570 and large-scale assessment of single-guide RNAs. *PloS one* 2014;9(5):e98186.

571 Gandhi, S., Christiaen, L. and Stolfi, A. Rational design and whole-genome predictions of
572 single guide RNAs for efficient CRISPR/Cas9-mediated genome editing in Ciona. *bioRxiv*
573 2016:041632.

574 Gong, S., *et al.* DNA Unwinding Is the Primary Determinant of CRISPR-Cas9 Activity. *Cell*
575 *Rep* 2018;22(2):359-371.

576 Gratz, S.J., *et al.* Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed
577 repair in Drosophila. *Genetics* 2014;genetics. 113.160713.

578 Guo, J., *et al.* Improved sgRNA design in bacteria via genome-wide activity profiling. *Nucleic*
579 *Acids Res* 2018;46(14):7052-7069.

580 Haeussler, M., *et al.* Evaluation of off-target and on-target scoring algorithms and integration

581 into the guide RNA selection tool CRISPOR. *Genome Biol* 2016;17(1):148.

582 Hart, T., *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and

583 Genotype-Specific Cancer Liabilities. *Cell* 2015;163(6):1515-1526.

584 Heffernan, R., *et al.* Highly accurate sequence-based prediction of half-sphere exposures of

585 amino acid residues in proteins. *Bioinformatics* 2015;32(6):843-849.

586 Horvath, P. and Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea.

587 *Science* 2010;327(5962):167-170.

588 Jensen, K.T., *et al.* Chromatin accessibility and guide sequence secondary structure affect

589 CRISPR-Cas9 gene editing efficiency. *FEBS Lett* 2017;591(13):1892-1901.

590 Jiang, F. and Doudna, J.A. CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys*

591 2017;46:505-529.

592 Kim, H.K., *et al.* Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat*

593 *Biotechnol* 2018;36(3):239-241.

594 Koike-Yusa, H., *et al.* Genome-wide recessive genetic screening in mammalian cells with a

595 lentiviral CRISPR-guide RNA library. *Nat Biotechnol* 2014;32(3):267-273.

596 Koonin, E.V. and Makarova, K.S. CRISPR-Cas: an adaptive immunity system in prokaryotes.

597 *F1000 biology reports* 2009;1.

598 Labuhn, M., *et al.* Refined sgRNA efficacy prediction improves large-and small-scale

599 CRISPR-Cas9 applications. *Nucleic acids research* 2017;46(3):1375-1385.

600 Lei, Y., *et al.* CRISPR-P: a web tool for synthetic single-guide RNA design of CRISPR-system

601 in plants. *Mol Plant* 2014;7(9):1494-1496.

602 Lin, J. and Wong, K.-C. Off-target predictions in CRISPR-Cas9 gene editing using deep

603 learning. *Bioinformatics* 2018;34(17):i656-i663.

604 Martin Jinek, *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive
605 Bacterial Immunity. *Science* 2012;337(6096)(816-21).

606 Moreno-Mateos, M.A., *et al.* CRISPRscan: designing highly efficient sgRNAs for
607 CRISPR-Cas9 targeting in vivo. *Nat Methods* 2015;12(10):982-988.

608 Peng, H., Long, F. and Ding, C. Feature selection based on mutual information criteria of
609 max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern
610 analysis and machine intelligence* 2005;27(8):1226-1238.

611 Peng, H., *et al.* CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and
612 Markov sequence profiling. *Bioinformatics* 2018;34(18):3069-3077.

613 Ran, F.A., *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*
614 2013;8(11):2281-2308.

615 Ren, X., *et al.* Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized
616 sgRNA parameters in Drosophila. *Cell reports* 2014;9(3):1151-1162.

617 Varshney, G.K., *et al.* High-throughput gene targeting and phenotyping in zebrafish using
618 CRISPR/Cas9. *Genome research* 2015.

619 Wang, T., *et al.* Genetic screens in human cells using the CRISPR-Cas9 system. *Science*
620 2014;343(6166):80-84.

621 Wilson, L.O.W., *et al.* High Activity Target-Site Identification Using Phenotypic Independent
622 CRISPR-Cas9 Core Functionality. *The CRISPR Journal* 2018;1(2):182-190.

623 Xu, H., *et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res*
624 2015;25(8):1147-1157.

625 Xue, L., *et al.* Prediction of CRISPR sgRNA activity using a deep convolutional neural network.
626 *J Chem Inf Model* 2018.

627 Yan, J., *et al.* Benchmarking CRISPR on-target sgRNA design. *Briefings in bioinformatics*
628 2018;19(4):721-724.

629 Yao, X., *et al.* Homology-mediated end joining-based targeted integration using CRISPR/Cas9.
630 *Cell Res* 2017;27(6):801-814.

631 Zhou, Y., *et al.* Enhanced genome editing in mammalian cells with a modified dual-fluorescent
632 surrogate system. *Cell Mol Life Sci* 2016;73(13):2543-2563.

633 Zhu, H. and Liang, C. CRISPR-DT: designing gRNAs for the CRISPR-Cpf1 system with
634 improved target efficiency and specificity. *bioRxiv* 2018:269910.

635 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids*
636 *research* 2003;31(13):3406-3415.

637