ACS Synthetic Biology

Technical Note

pubs.acs.org/synthbio

sgRNA Scorer 2.0: A Species-Independent Model To Predict CRISPR/Cas9 Activity

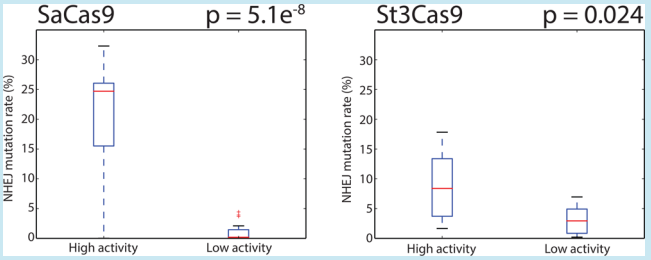
Raj Chari,*[,†] Nan Cher Yeo,[†,‡] Alejandro Chavez,[†,‡] and George M. Church*[,†,‡]

[†]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, United States

[‡]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, United States

Supporting Information

ABSTRACT: It has been possible to create tools to predict single guide RNA (sgRNA) activity in the CRISPR/Cas9 system derived from *Streptococcus pyogenes* due to the large amount of data that has been generated in sgRNA library screens. However, with the discovery of additional CRISPR systems from different bacteria, which show potent activity in eukaryotic cells, the approach of generating large data sets for each of these systems to predict their activity is not tractable. Here, we present a new guide RNA tool that can predict sgRNA activity across multiple CRISPR systems. In addition to predicting activity for Cas9 from *S. pyogenes* and *Streptococcus thermophilus* CRISPR1, we experimentally demonstrate that our algorithm can predict activity for Cas9 from *Staphylococcus aureus* and *S. thermophilus* CRISPR3. We also have made available a new version of our software, sgRNA Scorer 2.0, which will allow users to identify sgRNA sites for any PAM sequence of interest.



KEYWORDS: *CRISPR, Cas9, sgRNA activity prediction, sgRNA scorer, genome engineering*

A number of groups, including ours, have assessed hundreds to thousands of single guide RNA (sgRNAs) to identify sequence features that correlate with CRISPR/Cas9 activity and have created publically available software to aid researchers in sgRNA selection.[1−9] Specifically, our initial effort, which we called sgRNA Scorer 1.0, has had over 7000 unique users representing 87 different countries worldwide since its release in July 2015, highlighting the widespread interest in and usage of the CRISPR/Cas9 technology.

Recent concurrent efforts have also been undertaken to identify CRISPR systems from other bacterial species and to assess their potency in editing eukaryotic genomes.[10,11] However, the tractability of identifying sequence features that correlate with activity in newly discovered systems is challenging because current strategies require that a library of sgRNA sequences needs to be generated and tested for each CRISPR protein to be analyzed. Thus, a more generalized model that is able to predict sgRNA activity across a broad swath of CRISPR proteins would be of great use as it would rapidly facilitate utilization of the currently characterized CRISPR systems along with additional CRISPR systems that will undoubtedly be discovered.

We previously generated large sgRNA libraries for CRISPR/Cas9 systems from both *Streptococcus pyogenes* (SpCas9) and *Streptococcus thermophilus* CRISPR1 (St1Cas9) and demonstrated that the model for each could predict high- and low-performing sgRNAs from their sequence composition.[1] We hypothesized that by combining the models for SpCas9 and St1Cas9 we could generate a new model that could predict activity across multiple Cas9 orthologs and, potentially, across other CRISPR systems.

Here, we built a new combined support vector machine (SVM) model and assessed its predictive ability across three other Cas9 orthologs and a non-Cas9 system, Cpf1. Briefly, we took the sets of high- and low-activity sgRNAs for SpCas9 (133 high, 146 low) and St1Cas9 (82 high, 69 low) and merged them, resulting in sets of 215 high-activity and 215 low-activity sgRNAs. Subsequently, this final set of 430 sgRNA sequences was then used as the basis for the new SVM. 10-fold cross validation of our new model yielded an accuracy of 73.7%, a precision of 72.8%, and a recall of 75.8%, which are comparable to those of our previous individual models for SpCas9 and St1Cas9.

We first evaluated how well our new model could predict sgRNA activity for SpCas9 and St1Cas9. Using the prediction data from targeted loci sequencing that we had previously generated,[1] we rescored these sgRNAs and found that the correlation of prediction scores between our old and new models was 0.997 for SpCas9 and 0.940 for St1Cas9, suggesting that combining the data sets from the two different Cas9 orthologs did not adversely impact the predictive ability for each Cas9 (Table S1).

We then assessed whether the new model could predict activity across other Cas9 orthologs. Indeed, we found that our new combined model not only accurately predicts activity for SpCas9 and St1Cas9 but also has strong predictive ability for

Received: November 12, 2016
Published: February 1, 2017

DOI: 10.1021/acssynbio.6b00343
ACS Synth. Biol. XXXX, XXX, XXX−XXX

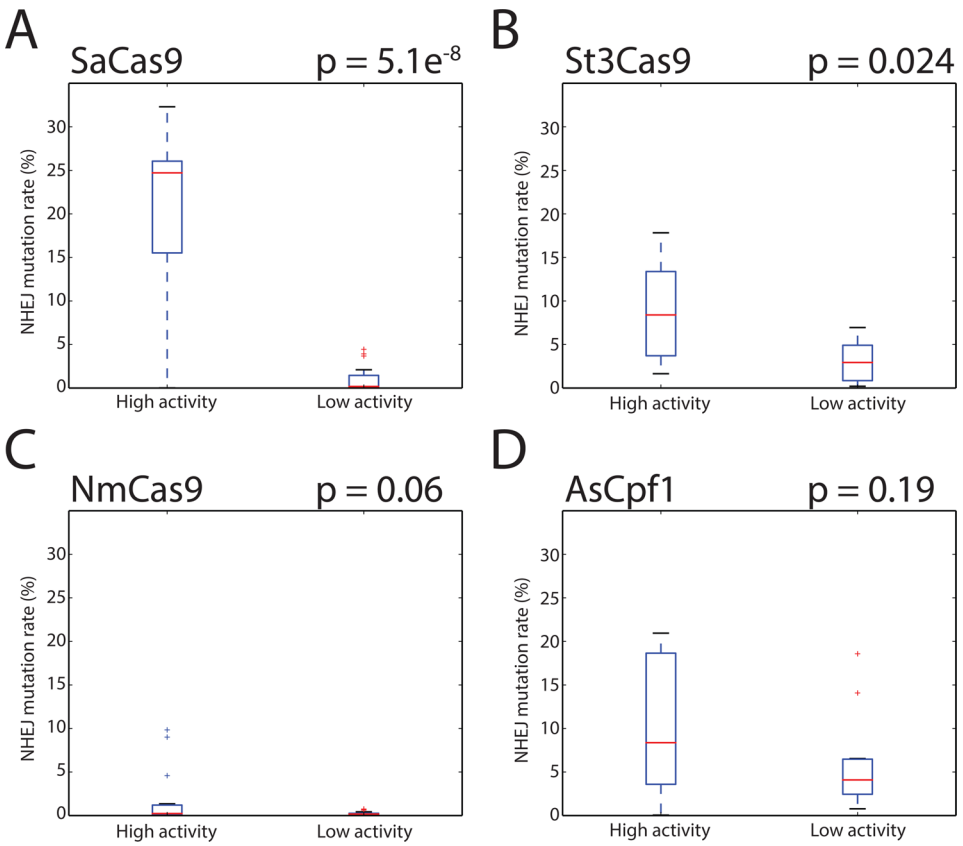


Figure 1. Activity prediction across Cas9 orthologs. For each Cas9 ortholog and Cpf1, a set of sgRNAs predicted to have high activity and a set predicted to have low activity were cloned and transfected into 293T cells in an individual manner. At 72 h post transfection, DNA was harvested and PCR amplicons were generated for targeted loci and sequenced using an Illumina MiSeq. (A) Cas9 from *S. aureus* (PAM: 3′ NNGRRT), (B) Cas9 from *S. thermophilus* CRISPR3 (PAM: 3′ NGGNG), (C) Cas9 from *N. meningitidis* (PAM: 3′ NNNNGATT), and (D) Cpf1 from *Acidaminococcus* sp. (PAM: 5′ TTTN). For Cpf1, the Cpf1_R model was used to predict high- and low-activity sgRNAs. *P*-values were calculated using Student's *t* test with unequal variance.

Staphylococcus aureus Cas9 (SaCas9) (Figure 1A) as well as modest predictive ability for *S. thermophilus* CRISPR3 Cas9 (St3Cas9) (Figure 1B). However, we also found that the overall gene editing activity of Cas9 from *Neisseria meningitidis* (NmCas9) was relatively weak compared to that for any of the other Cas9 systems that we tested, affecting our ability to accurately assess our algorithm for this ortholog (Figure 1C). Comparing the editing rates of the sgRNAs predicted to have high activity across different orthologs revealed that SaCas9 showed the closest potency to SpCas9, as has been observed previously,[11,12] whereas St3Cas9 demonstrated one-third of the median editing rate of SaCas9.

Finally, we attempted to predict guide RNA activity for a completely different CRISPR system from the Cpf1 endonuclease family.[10] Briefly, this system, unlike Cas9, has a PAM sequence (TTTN) at its 5′ end, compared to the 3′ end for Cas9. Since our SVM was based on the PAM sequence residing on the 3′ end, we applied our model in two different orientations. The first orientation applied the model directly to the spacer sequence as is (Cpf1_F), and the second orientation applied the model based on the distance of each spacer nucleotide relative to the PAM (Cpf1_R) (Figure S1).

Using *Acidaminococcus* sp. Cpf1 (AsCpf1), we generated a series of sgRNAs predicted to be high versus low performing and quantified their gene editing efficiency for both orientations, Cpf1_F and Cpf1_R. Although the Cpf1_F based model exhibited virtually no difference between predicted high- and low-activity sgRNAs (Figure S1), the Cpf1_R based model showed modest, but statistically insignificant, predictive ability (Figure 1D). Thus, it is likely a separate data set examining a large number of sgRNA sequences specifically for Cpf1 would be needed to generate a specific model for this CRISPR system; indeed, such a study has recently been published.[13]

To this point, the majority of efforts to develop algorithms to predict sgRNA activity have focused on SpCas9.[1−9] This has enabled not only the development of interactive web tools to extract and score sgRNAs from a specific sequence but also the production of genome-wide libraries of theoretically highly efficient sgRNAs targeting all genes.[3,14−18] As new orthologous CRISPR systems have been identified and characterized, these software tools have been extended accordingly by enabling the selection of sgRNA sequences based on the length of the spacer and the identity of the PAM recognition site. However, to our knowledge, the predictions of activity made by SpCas9 sgRNA design tools with respect to sgRNAs for orthologous CRISPR systems have not been experimentally verified. The work reported here represents the development of the first generalized sgRNA scorer that has also had its predictions validated in human cells across multiple CRISPR/Cas9 systems. Furthermore, to enable the scientific community at large to take advantage of our improved design metrics, we have implemented a new version of our software, sgRNA Scorer 2.0 (http://crispr.med.harvard.edu/sgRNAScorerV2), which

DOI: 10.1021/acssynbio.6b00343
ACS Synth. Biol. XXXX, XXX, XXX−XXX

will allow users to identify target sites for any CRISPR system for which the PAM sequence and spacer length can be specified.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.6b00343.

Comparison of SpCas9 and St1Cas9 (XLSX)
Observed editing rates for all loci assessed and primer sequences used to amplify each locus (XLSX)
Methods, endogenous gene editing rates, and weights for the SVM applied in two orientations (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: rchari@genetics.med.harvard.edu (R.C.).
*E-mail: gchurch@genetics.med.harvard.edu (G.M.C.).

ORCID

Raj Chari: 0000-0002-2216-313X

Author Contributions

R.C. designed and performed experiments and wrote the manuscript. A.C. and N.C.Y. performed experiments and edited the manuscript. G.M.C. is the principal investigator of the laboratory in which the work was performed.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by NIH grants RM1 HG008525 and P50 HG005550.

■ REFERENCES

(1) Chari, R., Mali, P., Moosburner, M., and Church, G. M. (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods 12*, 823−826.

(2) Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S., Brown, M., and Liu, X. S. (2015) Sequence determinants of improved CRISPR sgRNA design. *Genome Res. 25*, 1147−1157.

(3) Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., and Root, D. E. (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol. 34*, 184−191.

(4) Moreno-Mateos, M. A., Vejnar, C. E., Beaudoin, J.-D., Fernandez, J. P., Mis, E. K., Khokha, M. K., and Giraldez, A. J. (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods 12*, 982−988.

(5) Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B., and Valen, E. (2016) CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res. 44*, W272−276.

(6) Horlbeck, M. A., Gilbert, L. A., Villalta, J. E., Adamson, B., Pak, R. A., Chen, Y., Fields, A. P., Park, C. Y., Corn, J. E., Kampmann, M., and Weissman, J. S. (2016) Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife 5*, e19760.

(7) Park, J., Bae, S., and Kim, J.-S. (2015) Cas-Designer: a web-based tool for choice of CRISPR-Cas9 target sites. *Bioinformatics 31*, 4014−4016.

(8) Heigwer, F., Kerr, G., and Boutros, M. (2014) E-CRISP: fast CRISPR target site identification. *Nat. Methods 11*, 122−123.

(9) Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., Joly, J.-S., and Concordet, J.-P. (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol. 17*, 148.

(10) Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., van der Oost, J., Regev, A., Koonin, E. V., and Zhang, F. (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell 163*, 759−771.

(11) Ran, F. A., Cong, L., Yan, W. X., Scott, D. A., Gootenberg, J. S., Kriz, A. J., Zetsche, B., Shalem, O., Wu, X., Makarova, K. S., Koonin, E. V., Sharp, P. A., and Zhang, F. (2015) In vivo genome editing using Staphylococcus aureus Cas9. *Nature 520*, 186−191.

(12) Friedland, A. E., Baral, R., Singhal, P., Loveluck, K., Shen, S., Sanchez, M., Marco, E., Gotta, G. M., Maeder, M. L., Kennedy, E. M., Kornepati, A. V. R., Sousa, A., Collins, M. A., Jayaram, H., Cullen, B. R., and Bumcrot, D. (2015) Characterization of Staphylococcus aureus Cas9: a smaller Cas9 for all-in-one adeno-associated virus delivery and paired nickase applications. *Genome Biol. 16*, 257.

(13) Kim, H. K., Song, M., Lee, J., Menon, A. V., Jung, S., Kang, Y.-M., Choi, J. W., Woo, E., Koh, H. C., Nam, J.-W., and Kim, H. (2016) In vivo high-throughput profiling of CRISPR-Cpf1 activity. *Nat. Methods 14*, 153.

(14) Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S., and Moffat, J. (2015) High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell 163*, 1515−1526.

(15) Sanjana, N. E., Shalem, O., and Zhang, F. (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods 11*, 783−784.

(16) Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., and Sabatini, D. M. (2015) Identification and characterization of essential genes in the human genome. *Science 350*, 1096−1101.

(17) Tzelepis, K., Koike-Yusa, H., De Braekeleer, E., Li, Y., Metzakopian, E., Dovey, O. M., Mupo, A., Grinkevich, V., Li, M., Mazan, M., Gozdecka, M., Ohnishi, S., Cooper, J., Patel, M., McKerrell, T., Chen, B., Domingues, A. F., Gallipoli, P., Teichmann, S., Ponstingl, H., McDermott, U., Saez-Rodriguez, J., Huntly, B. J. P., Iorio, F., Pina, C., Vassiliou, G. S., and Yusa, K. (2016) A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep. 17*, 1193−1205.

(18) Ma, H., Dang, Y., Wu, Y., Jia, G., Anaya, E., Zhang, J., Abraham, S., Choi, J.-G., Shi, G., Qi, L., Manjunath, N., and Wu, H. (2015) A CRISPR-Based Screen Identifies Genes Essential for West-Nile-Virus-Induced Cell Death. *Cell Rep. 12*, 673−683.

DOI: 10.1021/acssynbio.6b00343
ACS Synth. Biol. XXXX, XXX, XXX−XXX