A meta-analysis of gRNA library screens enables an improved understanding of the impact of gRNA folding and structural stability on CRISPR-Cas9 activity.

Moreb, E.A.[1] , and Lynch, Michael D.[1,2,3]
[1] Department of Biomedical Engineering, Duke University
[2] To whom all correspondence should be addressed.
[3] michael.lynch@duke.edu

Abstract

CRISPR systems are known to be inhibited by unwanted secondary structures that form within the guide RNA (gRNA). The minimum free energy of predicted secondary structures has been used in prediction algorithms. However, the types of structures as well as the degree to which a predicted structure can inhibit Cas9/gRNA activity is not well characterized. Here we perform a meta-analysis of published CRISPR-Cas9 datasets to better understand the role of secondary structures in inhibiting gRNA activity. We identify two inhibitory structures and provide estimated free energy cutoffs at which they become impactful. Further, we identify the prevalence of these structures in existing datasets. The cutoffs provided help to explain conflicting impacts of free energy values in different datasets as well as providing a guideline for future gRNA designs.

Highlights:

- Clearly define two secondary structures that inhibit CRISPR-Cas9 activity
- Provide free energy calculations and cutoffs at which each structure begins to inhibit activity
- Evaluate impact of these structures in published datasets

Keywords: CRISPR, Cas9, gRNA activity, Free Energy, gRNA Secondary Structure

Introduction

CRISPR-Cas9 on-target activity is dependent on the sequence of the guide RNA (gRNA).[1–3] The mechanisms underlying this sequence dependence have been studied extensively but many unknowns remain.[3–7] One factor reported to negatively impact CRISPR-Cas9 on-target activity is the formation of unwanted secondary structures.[7,8] Unwanted RNA structures and their associated stability can be estimated computationally.[9] The stability of secondary structures in both the spacer sequence and the full gRNA sequence (including scaffold) have been estimated and have been correlated with activity. This approach is incorporated into some predictive algorithms.[5,7] However, while gRNA structure is known to inhibit activity, it has not been clear which structures impact activity, to what degree, and how prevalent or impactful this is in routine experimentation. We therefore sought to better define inhibitory gRNA structures and their impact in genome editing studies. To do so we expanded upon a meta-analysis we have recently reported, which includes 39 published datasets with gRNA libraries in various organisms.[6] We report two types of secondary structure that can inhibit Cas9 activity, provide estimated free energy cutoffs at which they begin to inhibit activity, and evaluate their prevalence within published datasets.

Methods

Datasets were compiled as described previously.[6] All calculations and generation of figures were performed in Python, using standard libraries.[10–14] Structure and minimum free energy predictions were calculated using ViennaRNA RNAfold package.[9] To calculate the average accessibility of each position along the spacer (Figure 1e), we first calculated the predicted secondary structure of the gRNA and scaffold together (Full Structure). Positions in the spacer region of the Full Structure were then assigned a 1 if predicted to be unbound or a 0 if bound and for each group of gRNA (ie, Functional gRNA), we averaged the values at each position within the gRNA. For example, the accessibility at position 1 within the Functional gRNA is the average accessibility at that site for all ~1.17 million Functional gRNA in the dataset. All code is provided as a Jupyter Notebook in Supplementary File S1.

Results & Discussion

Of the 39 datasets analyzed, a recent study in *E. coli* provides a unique opportunity to better understand the impact of gRNA structure on Cas9 activity.[8] Talas et al. 2021 measured the activity of ~1.2 million self-targeting gRNA (stgRNA) using a plasmid based screening approach in *E. coli*, which enables "the major fraction of the plasmids [to] be cleaved", resulting in a strongly binary dataset skewed towards active gRNA. Of the ~1.2 million gRNA, 86.8% had a perfect activity score of 1 while only 4.3% of gRNA are defined as inactive (activity score <0.5). The authors also noted a strong correlation between activity in this screen and the Minimum Free Energy (MFE) predictions of potential secondary structures of both the spacer sequence as well as

the full gRNA sequence.[15] We therefore hypothesized that gRNA that don't cleave in this study are "defective". In the event that unwanted hairpins (lower MFE values) inhibit the Cas9-gRNA complex, it makes sense that the activity in this study is highly correlated with MFE values.

Compared to the desired gRNA structure (Figure 1a, structure i), several structures could inhibit activity: hairpins within the spacer (structure ii), binding between the spacer and the non-structured sequence of the scaffold (structure iii), or binding between the spacer and the scaffold such that the natural hairpins of the scaffold are disrupted (structure iv). Structures ii and iii have been reported previously to impact gRNA activity.[7] To derive inhibitory gRNA structures from these data, we first used the ViennaRNA RNAFold package[9] to predict the self-folding structures of just the spacer (referred to as Spacer Structure) as well as the full length gRNA (including both the spacer and the scaffold, referred to as Full Structure). We also calculated the MFE for the Spacer Structure (Spacer MFE) and the Full Structure (Full MFE). We saw a strong sigmoidal impact on activity in the dataset from Talas et al., for both the Spacer MFE and Full MFE (Figure 1b and 1c, respectively).[8] However, the Full MFE also contains the spacer sequence and therefore strong hairpins in the spacer would also impact the Full MFE results. To better separate these two potential effects, we binned the gRNA activities into two groups, Functional gRNA with an activity score >0.5 and Nonfunctional gRNA with activity <0.5, and compared the relationship between Spacer MFE and Full MFE (Figure 1d). In the Nonfunctional gRNA, two unique populations can be identified, confirming that while there is overlap between structures generating low Spacer and Full MFEs, they collectively capture two unique gRNA structures.

We next turned to better computationally define these two distinct groups of gRNA. In contrast to the Full MFE, the Spacer MFE represents only the stability of structure ii. As a result, we first divided the Nonfunctional gRNA (activity <0.5) into two groups with Spacer MFE values above or below -5 kcal/mol. We then looked at the 20bp spacer region of the Full Structure and assigned either a 1 or 0 to each base position, depending on whether the base was predicted to be bound (0, in a stem loop) or free (1) (Figure 1e). For Functional gRNA, all positions were, on average, accessible. In contrast, the two Nonfunctional gRNA groups both showed reduced accessibility (Figure 1e). For gRNA with a Spacer MFE < -5 kcal/mol, we identified a pattern of inaccessible nucleotides separated by a group of accessible nucleotides, which is consistent with structure ii and expected for strongly negative Spacer MFE values. For the other Nonfunctional gRNA, there is a drop in accessibility in the seed region of the gRNA. This is consistent with either structure iii or iv but suggests that a defining feature of these inhibitory structures is obstruction of the seed region. To help delineate between structure iii and iv, we next grouped all gRNA by the PAM proximal 5bp of the spacer, calculated the average activity for each group, and rank ordered the sequences by average activity (Figure 1f). The least active fifteen 5bp sequences all have strong homology to the non-structured sequence of the gRNA scaffold, consistent with structure iii (Figure 1g-h). Notably, this explains a previously reported but unexplained inhibitory GCC motif in the PAM proximal 5bp sequence.[8,16] Furthermore, we used RNAFold[9] to calculate the duplex stability between the spacer and non-structured sequence of the scaffold (hereafter

referred to as Duplex Stability, Figure 1i). We see a sigmoidal relationship between activity and Duplex Stability (Figure 1j). This both confirms structure iii as inhibitory of Cas9 activity and provides a method for predicting this structure that is not confounded by predicted structures within the spacer.

We next wanted to assess the impact of these structures on gRNA activity in a larger group of datasets. As mentioned, we have previously compiled 39 datasets from various species.[6] We filtered data from these 39 studies and included only libraries with 10,000 or more gRNA and then compared Spacer MFE and gRNA activity across these datasets, including datasets with different Cas9 variants (Figure 2a-f).[5,17–20] Despite differences between datasets, there is a clear relationship between Spacer MFE values and activity when the Spacer MFE is below -5 kcal/mol but not when it is greater. We repeated this for Duplex Stability, setting a similar cutoff, in this case at -15 kcal/mol (Figure 2g-l). We noted that in some datasets less stable Duplex Stability values (closer to 0) appeared to negatively impact activity, contrary to expectations. This trend appeared to be species dependent, with human datasets most impacted. To better understand this relationship, we compared the GC content of a gRNA to its Duplex Stability and noted a strong correlation, particularly in the range of low GC (Figure 3a). Extreme GC content has previously been reported to negatively impact Cas9 activity in certain datasets.[1,3] In the human data set from Wang et al 2019[5], if we remove gRNA with GC content less than 30%, we see improved on-target activity within the same range of less stable Duplex Stability values (Figure 3b). Interestingly, the observation that low GC content reduces gRNA activity is not shared among different species and therefore does not impact the relationship between Duplex Stability and activity in the same manner (Supplemental Figure S1). This result, suggests that this observation is due to other context-dependent factors impacted by sequence composition rather than Duplex Stability.[6] This supports a model in which Duplex Stability below the cutoff we established is representative of formation of structure iii, while Duplex Stability above the cutoff does not indicate formation of inhibitory structures but rather may be correlated with other sequence-dependent features (Figure 3c).

Finally, we turned to evaluate the impact of these structures across the remaining smaller datasets. We first calculated the percent of gRNA that are below our proposed cutoffs in each dataset (Figure 4a).[3,4,17,18,20–31] Overall, structure ii is much more prevalent than structure iii, ranging from 2.7-23.4% of gRNA compared to 0-1.6%, respectively. This variability in relative prevalence of these structures within datasets may have confounding effects on activity predictions. For example, one of the datasets, reported in Doench et al 2014[3] and used to train the WU-CRISPR algorithm[7], has 23.4% of gRNA below our Spacer MFE cutoff. In contrast, the second highest percentage is 12.6%. This may lead the WU-CRISPR algorithm to more heavily weight the Spacer MFE. Surprisingly, we still see a large number of datasets published after these studies with a relatively high percent of gRNA predicted to contain structure ii or iii, indicating that a more clear consensus on the impact of structure on activity is needed.

We next calculated Pearson correlations between Spacer MFE above and below the cutoffs and gRNA activity (Figure 4b). We see broad agreement across datasets that activity is correlated with Spacer MFE below -5 kcal/mol while the correlations above -5 kcal/mol are either reduced or even negative, further highlighting that gRNA with an MFE above -5 kcal/mol are not likely forming structures that inhibit activity. In this regime, other factors are likely contributing to sequence dependent activity. Three datasets show negative correlations with Spacer MFE below -5 kcal/mol but all three of these datasets have fewer than 205 gRNA and at most 25 gRNA with Spacer MFE below the cutoff. These types of correlations on gRNA libraries this small are unlikely to capture meaningful features. We also calculated Pearson correlations between the Duplex Stability above and below -15 kcal/mol and gRNA activity (Figure 4c). Again, we see several datasets with negative or no correlation between Duplex Stability below -15 kcal/mol and activity. These datasets have relatively few gRNA predicted to have structure iii or are datasets where activity is low overall for other reasons (such as the evoCas9 data from Kim et al 2020a).[26] Taken together, these results strongly support the use of these new proposed cutoffs for eliminating improperly folded gRNAs in future designs.

In this analysis, we have confirmed two types of secondary structure that can inhibit on-target activity. While previous algorithms for gRNA design have relied on weighted MFE values, this analysis more strongly supports the use of free energy cutoffs to identify "improperly functioning" gRNA. Interestingly, this analysis did not identify structures wherein the natural stem loops of the gRNA scaffold are disrupted. Free energy values above our proposed cutoffs are not meaningfully connected to inhibitory structures but rather may be correlated with other sequence-dependent factors. Furthermore, we have confirmed these cutoffs are practically meaningful across a variety of contexts. These results provide a clear cutoff that can be immediately deployed in new gRNA designs as well as help towards improving algorithms to predict gRNA activity.

Acknowledgements

We would like to acknowledge the following support: ONR YIP #12043956, and DOE EERE grant #EE0007563. We would also like to acknowledge financial support from Duke Innovation & Entrepreneurship Initiative.

Author contributions

E.A. Moreb performed computational analyses. E.A. Moreb and M.D. Lynch designed analyses, analyzed results, wrote, revised and edited the manuscript.

Conflicts of Interest

M.D. Lynch has a financial interest in DMC Biotechnologies, Inc., M.D. Lynch and E.A. Moreb have a financial interest in Roke Biotechnologies, Inc.

References

(1) Wang, T.; Wei, J. J.; Sabatini, D. M.; Lander, E. S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **2014**, *343* (6166), 80–84.

(2) Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A.; Charpentier, E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **2012**, *337* (6096), 816–821.

(3) Doench, J. G.; Hartenian, E.; Graham, D. B.; Tothova, Z.; Hegde, M.; Smith, I.; Sullender, M.; Ebert, B. L.; Xavier, R. J.; Root, D. E. Rational Design of Highly Active sgRNAs for CRISPR-Cas9-Mediated Gene Inactivation. *Nat. Biotechnol.* **2014**, *32* (12), 1262–1267.

(4) Moreno-Mateos, M. A.; Vejnar, C. E.; Beaudoin, J.-D.; Fernandez, J. P.; Mis, E. K.; Khokha, M. K.; Giraldez, A. J. CRISPRscan: Designing Highly Efficient sgRNAs for CRISPR-Cas9 Targeting in Vivo. *Nat. Methods* **2015**, *12* (10), 982–988.

(5) Wang, D.; Zhang, C.; Wang, B.; Li, B.; Wang, Q.; Liu, D.; Wang, H.; Zhou, Y.; Shi, L.; Lan, F.; Wang, Y. Optimized CRISPR Guide RNA Design for Two High-Fidelity Cas9 Variants by Deep Learning. *Nat. Commun.* **2019**, *10* (1), 4284.

(6) Moreb, E. A.; Lynch, M. D. An Analysis of gRNA Sequence Dependent Cleavage Highlights the Importance of Genomic Context on CRISPR-Cas Activity. *bioRxiv*, 2021, 2021.05.06.442929. https://doi.org/10.1101/2021.05.06.442929.

(7) Wong, N.; Liu, W.; Wang, X. WU-CRISPR: Characteristics of Functional Guide RNAs for the CRISPR/Cas9 System. *Genome Biol.* **2015**, *16*, 218.

(8) Tálas, A.; Huszár, K.; Kulcsár, P. I.; Varga, J. K.; Varga, É.; Tóth, E.; Welker, Z.; Erdős, G.; Pach, P. F.; Welker, Á.; Györgypál, Z.; Tusnády, G. E.; Welker, E. A Method for Characterizing Cas9 Variants via a One-Million Target Sequence Library of Self-Targeting sgRNAs. *Nucleic Acids Res.* **2021**. https://doi.org/10.1093/nar/gkaa1220.

(9) Lorenz, R.; Bernhart, S. H.; Höner Zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **2011**, *6*, 26.

(10) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **2007**, *9* (3), 90–95.

(11) Waskom, M. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6* (60), 3021.

(12) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* **2011**, *13* (2), 22–30.

(13) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for

Scientific Computing in Python. *Nat. Methods* **2020**, *17* (3), 261–272.
(14) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423.
(15) Zuker, M.; Stiegler, P. Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information. *Nucleic Acids Res.* **1981**, *9* (1), 133–148.
(16) Graf, R.; Li, X.; Chu, V. T.; Rajewsky, K. sgRNA Sequence Motifs Blocking Efficient CRISPR/Cas9-Mediated Gene Editing. *Cell Rep.* **2019**, *26* (5), 1098–1103.e3.
(17) Schwartz, C.; Cheng, J.-F.; Evans, R.; Schwartz, C. A.; Wagner, J. M.; Anglin, S.; Beitz, A.; Pan, W.; Lonardi, S.; Blenner, M.; Alper, H. S.; Yoshikuni, Y.; Wheeldon, I. Validating Genome-Wide CRISPR-Cas9 Function Improves Screening in the Oleaginous Yeast Yarrowia Lipolytica. *Metab. Eng.* **2019**, *55*, 102–110.
(18) Guo, J.; Wang, T.; Guan, C.; Liu, B.; Luo, C.; Xie, Z.; Zhang, C.; Xing, X.-H. Improved sgRNA Design in Bacteria via Genome-Wide Activity Profiling. *Nucleic Acids Res.* **2018**, *46* (14), 7052–7069.
(19) Kim, H. K.; Kim, Y.; Lee, S.; Min, S.; Bae, J. Y.; Choi, J. W.; Park, J.; Jung, D.; Yoon, S.; Kim, H. H. SpCas9 Activity Prediction by DeepSpCas9, a Deep Learning-Based Model with High Generalization Performance. *Sci Adv* **2019**, *5* (11).
(20) Moreb, E. A.; Hoover, B.; Yaseen, A.; Valyasevi, N.; Roecker, Z.; Menacho-Melgar, R.; Lynch, M. D. Managing the SOS Response for Enhanced CRISPR-Cas-Based Recombineering in E. Coli through Transient Inhibition of Host RecA Activity. *ACS Synth. Biol.* **2017**, *6* (12), 2209–2218.
(21) Moreb, E. A.; Hutmacher, M.; Lynch, M. D. CRISPR/Cas "non-Target" Sites Inhibit on-Target Cutting Rates. *bioRxiv*, 2020, 2020.06.12.147827. https://doi.org/10.1101/2020.06.12.147827.
(22) Chari, R.; Mali, P.; Moosburner, M.; Church, G. M. Unraveling CRISPR-Cas9 Genome Engineering Parameters via a Library-on-Library Approach. *Nat. Methods* **2015**, *12* (9), 823–826.
(23) 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.; Liu, X. S. Sequence Determinants of Improved CRISPR sgRNA Design. *Genome Res.* **2015**, *25* (8), 1147–1157.
(24) 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.; Moffat, J. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **2015**, *163* (6), 1515–1526.
(25) 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.; Root, D. E. Optimized sgRNA Design to Maximize Activity and Minimize off-Target Effects of CRISPR-Cas9. *Nat. Biotechnol.* **2016**, *34* (2), 184–191.
(26) Kim, N.; Kim, H. K.; Lee, S.; Seo, J. H.; Choi, J. W.; Park, J.; Min, S.; Yoon, S.; Cho, S.-R.; Kim, H. H. Prediction of the Sequence-Specific Cleavage Activity of Cas9 Variants. *Nat. Biotechnol.* **2020**, *38* (11), 1328–1336.
(27) Kim, H. K.; Lee, S.; Kim, Y.; Park, J.; Min, S.; Choi, J. W.; Huang, T. P.; Yoon, S.; Liu, D. R.; Kim, H. H. High-Throughput Analysis of the Activities of xCas9, SpCas9-NG and

SpCas9 at Matched and Mismatched Target Sequences in Human Cells. *Nat Biomed Eng* **2020**, *4* (1), 111–124.

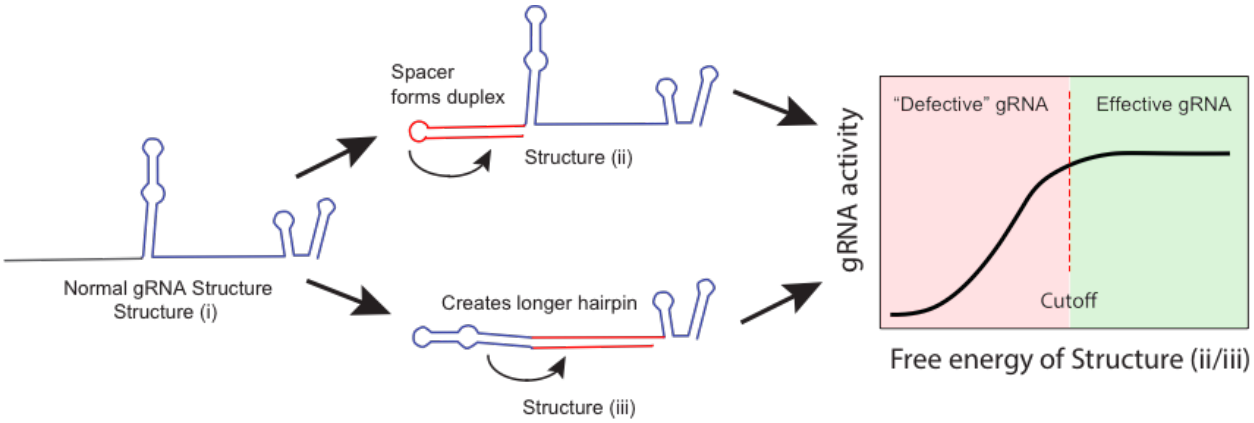
(28) Park, J.; Lim, J. M.; Jung, I.; Heo, S.-J.; Park, J.; Chang, Y.; Kim, H. K.; Jung, D.; Yu, J. H.; Min, S.; Yoon, S.; Cho, S.-R.; Park, T.; Kim, H. H. Recording of Elapsed Time and Temporal Information about Biological Events Using Cas9. *Cell* **2021**. https://doi.org/10.1016/j.cell.2021.01.014.

(29) Liu, X.; Homma, A.; Sayadi, J.; Yang, S.; Ohashi, J.; Takumi, T. Sequence Features Associated with the Cleavage Efficiency of CRISPR/Cas9 System. *Sci. Rep.* **2016**, *6*, 19675.

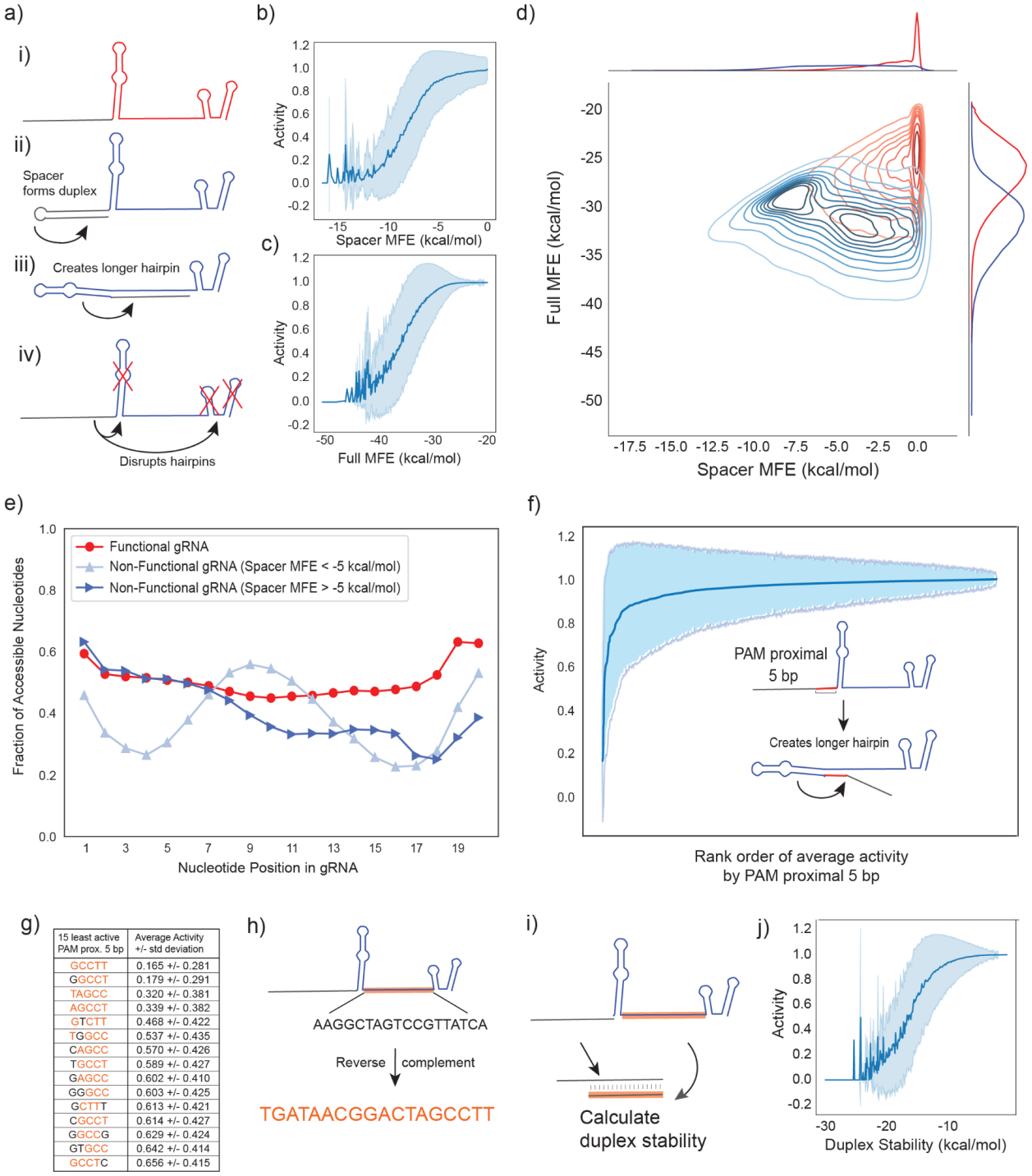
(30) Gagnon, J. A.; Valen, E.; Thyme, S. B.; Huang, P.; Akhmetova, L.; Pauli, A.; Montague, T. G.; Zimmerman, S.; Richter, C.; Schier, A. F. Efficient Mutagenesis by Cas9 Protein-Mediated Oligonucleotide Insertion and Large-Scale Assessment of Single-Guide RNAs. *PLoS One* **2014**, *9* (5), e98186.

(31) Varshney, G. K.; Pei, W.; LaFave, M. C.; Idol, J.; Xu, L.; Gallardo, V.; Carrington, B.; Bishop, K.; Jones, M.; Li, M.; Harper, U.; Huang, S. C.; Prakash, A.; Chen, W.; Sood, R.; Ledin, J.; Burgess, S. M. High-Throughput Gene Targeting and Phenotyping in Zebrafish Using CRISPR/Cas9. *Genome Res.* **2015**, *25* (7), 1030–1042.

Figures & Captions



Graphical Abstract



15 least active PAM prox. 5 bp	Average Activity +/- std deviation
GCCTT	0.165 +/- 0.281
GGCCT	0.179 +/- 0.291
TAGCC	0.320 +/- 0.381
AGCCT	0.339 +/- 0.382
GTCTT	0.468 +/- 0.422
TGGCC	0.537 +/- 0.435
CAGCC	0.570 +/- 0.426
TGCCT	0.589 +/- 0.427
GAGCC	0.602 +/- 0.410
GGGCC	0.603 +/- 0.425
GCTTT	0.613 +/- 0.421
CGCCT	0.614 +/- 0.427
GGCCG	0.629 +/- 0.424
GTGCC	0.642 +/- 0.414
GCCTC	0.656 +/- 0.415

Figure 1: a) Possible inhibitory gRNA structures. Predicted minimum free energy (MFE) of folding of the b) spacer alone (Spacer MFE) and c) the spacer plus scaffold (Full MFE) both show sigmoidal impact on activity. d) Spacer MFE plotted against Full MFE after splitting the dataset into Functional (red) and Non-Functional (blue) gRNA with activity above or below 0.5, respectively. e) Accessibility of each position in the spacer for Functional gRNA (red), Nonfunctional gRNA with Spacer MFE below -5 kcal/mol (light blue), and Nonfunctional gRNA with Spacer MFE above -5 kcal/mol (dark blue). Accessible nucleotides are assigned a 1, bound nucleotides are assigned a 0, and the averaged value for each position is shown. f) All gRNA were grouped by the PAM proximal 5bp of the gRNA and average activity was calculated.

Averages are presented by rank order of the PAM proximal 5bp. g) The 15 least active 5bp sequences all have homology to h) the non-structured portion of the gRNA scaffold. i) We calculated Duplex Stability of the spacer sequence binding to the non-structured sequence of the scaffold and k) show a sigmoidal impact on activity. All data presented here is from wild-type Cas9 in the Talas et al dataset.[8]

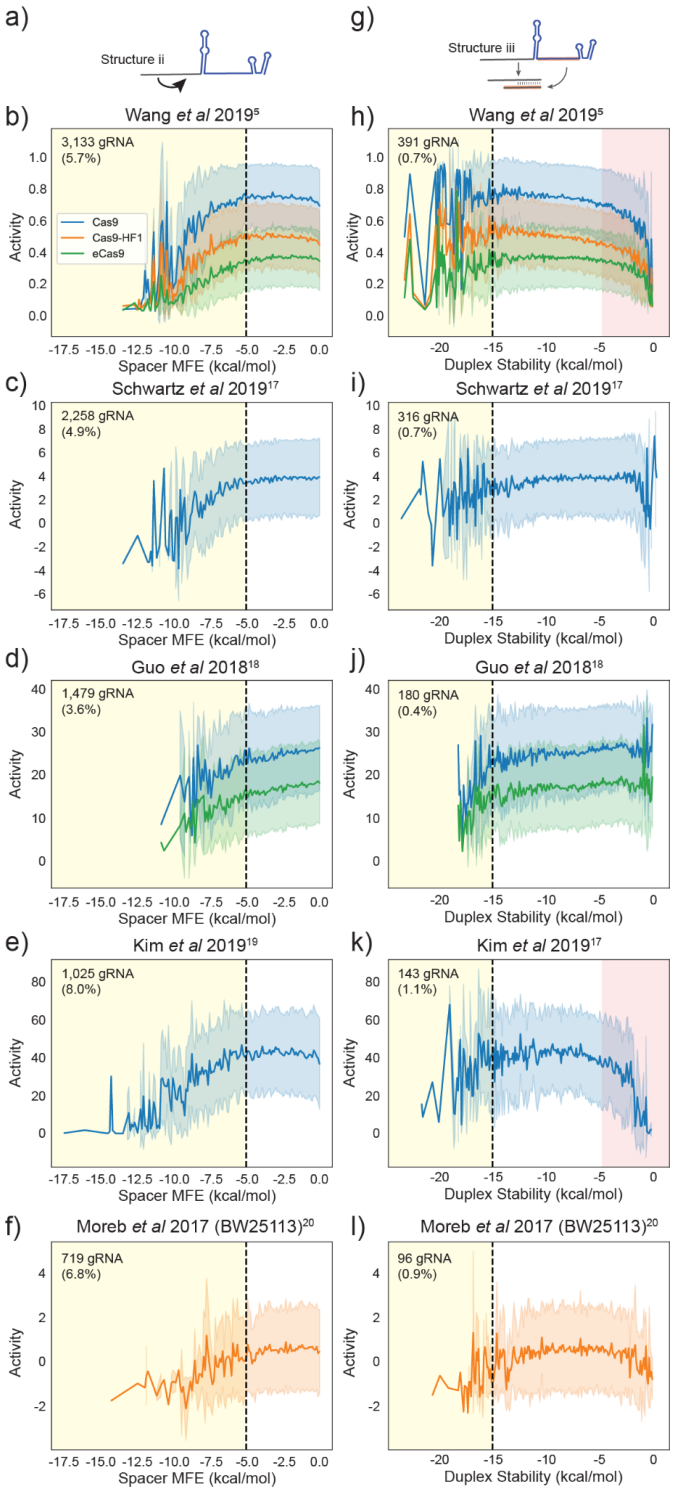


Figure 2: The impact of a) structure ii on gRNA activity is seen in the relationship between Spacer MFE and gRNA activity is highlighted in the five largest datasets with over 10,000 gRNA each (b-f) and across different Cas9 variants (wild-type Cas9: blue, Cas9-HF1: orange, eCas9: green). The yellow shaded region shows gRNA with a Spacer MFE below -5 kcal/mol. Similarly, the impact of g) structure iii, calculated as Duplex Stability, is shown for the same five datasets (h-l) and Cas9 variants. The yellow shaded region highlights gRNA below -15 kcal/mol. The red shaded regions in human datasets highlight an observed negative impact on activity for predicted unstable (low GC or high AT) duplexes, which is contrary to expectations based on structure ii) and iii) alone (see Figure 3).

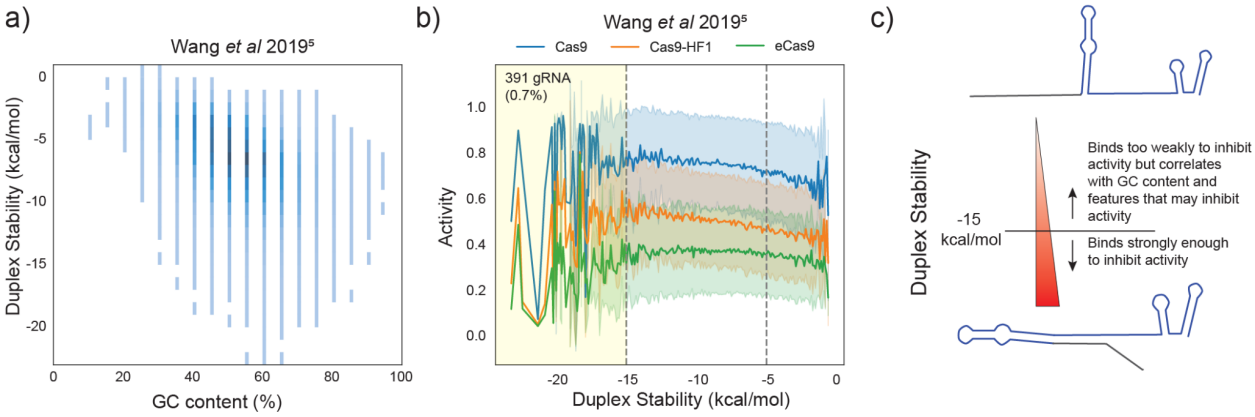


Figure 3: a) Duplex Stability is correlated with gRNA GC content. b) To demonstrate the impact of this correlation on the high range of Duplex Stability, data from Wang et al 2019[5] were re-plotted without gRNA containing GC content below 30%. c) This supports a model in which Duplex Stability below the cutoff we established is inhibitory based on formation of inhibitory structures, while Duplex Stability above the cutoff does not indicate formation of inhibitory structures but may correlate with other sequence-dependent influences on activity.

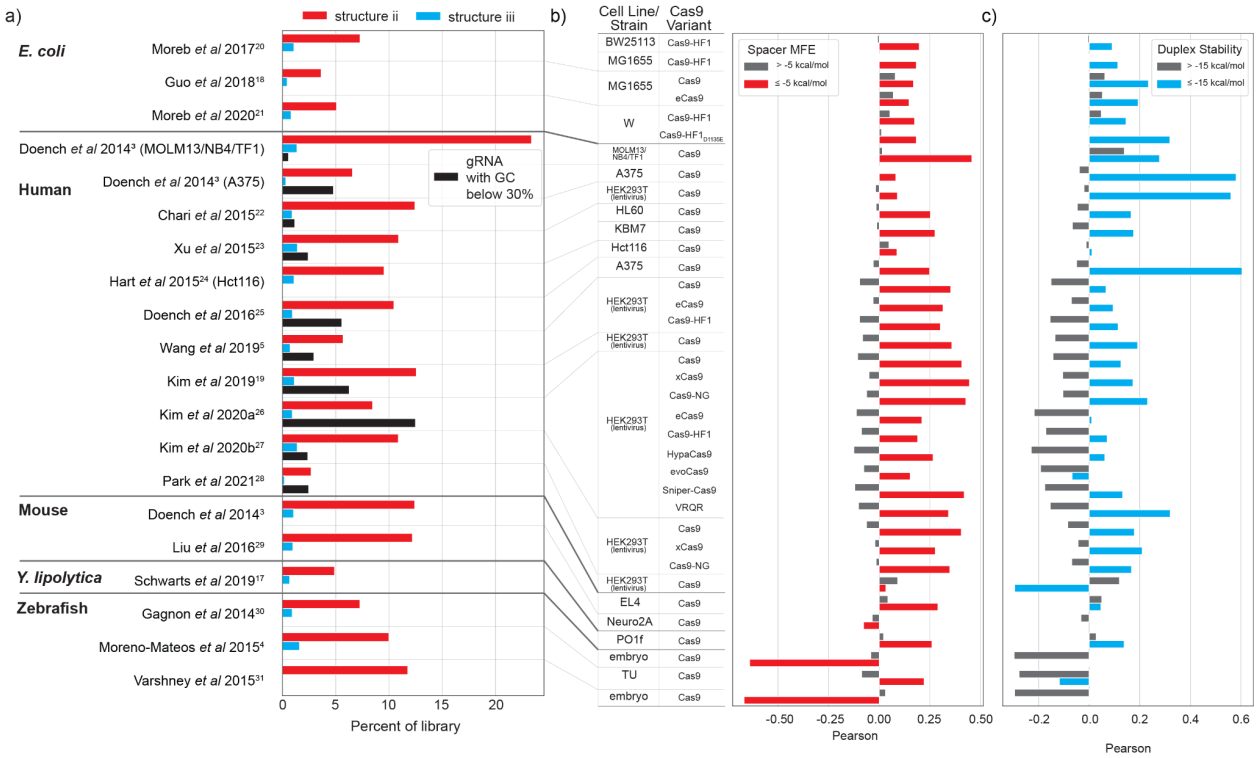


Figure 4: The impact of gRNA structure is highlighted in different datasets. a) The percent of gRNA in each library that contain a predicted structure ii or iii below the relevant cutoffs. For human datasets, we also calculate the percent of gRNA with less than 30% GC. b) Correlation between gRNA activity and Spacer MFE above (grey) and below (red) the -5 kcal/mol cutoff. c) Correlation between gRNA activity and Duplex Stability above (grey) and below (blue) the -15 kcal/mol cutoff.