GigaScience

GuideMaker: Software to design CRISPR-Cas guide RNA pools in non-model genomes

--Manuscript Draft--

Manuscript Number:	GIGA-D-21-00186R1
Full Title:	GuideMaker: Software to design CRISPR-Cas guide RNA pools in non-model genomes
Article Type:	Technical Note
Funding Information:	agricultural research service (6066-21310-005-D) — Dr. Adam R Rivers agricultural research service (6066-21310-005-28-S) — Dr. Christopher R. Reisch agricultural research service (0500-00093-001-00-D) — Dr. Adam R Rivers
Abstract:	Background: CRISPR-Cas systems have expanded the possibilities for gene editing in bacteria and eukaryotes. There are many excellent tools for designing CRISPR-Cas guide RNAs for model organisms with standard Cas enzymes. GuideMaker is intended as a fast and easy-to-use design tool for challenging projects with 1) non-standard Cas enzymes, 2) non-model organisms, or 3) projects that need to design a panel of guide RNAs (gRNA) for genome-wide screens. Findings: GuideMaker can rapidly design gRNAs for gene targets across the genome using a degenerate protospacer adjacent motif (PAM) and a genome. The tool applies Hierarchical Navigable Small World (HNSW) graphs to speed up the comparison of guide RNAs and optionally provides on-target and off-target scoring. This allows the user to design effective gRNAs targeting all genes in a typical bacterial genome in about 1-2 minutes. Conclusions: GuideMaker enables the rapid design of genome-wide gRNA for any CRISPR-Cas enzyme in non-model organisms. While GuideMaker is designed with prokaryotic genomes in mind, it can efficiently process eukaryotic genomes as well. GuideMaker is available as command-line software, a stand-alone web application, and a tool in the CyCverse Discovery Environment. All versions are available under a Creative Commons CC0 1.0 Universal Public Domain Dedication.
Corresponding Author:	Adam R Rivers, Ph.D. USDA Agricultural Research Service Gainesville, FL UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	USDA Agricultural Research Service
Corresponding Author's Secondary Institution:	
First Author:	Ravin Poudel, Ph.D.
First Author Secondary Information:	
Order of Authors:	Ravin Poudel, Ph.D. Lidimarie Trujillo Rodriguez, B.S. Christopher R. Reisch, Ph.D. Adam R Rivers, Ph.D.
Order of Authors Secondary Information:	
Response to Reviewers:	GIGA-D-21-00186 GuideMaker: Software to design CRISPR-Cas guide RNA pools in non-model

genomes
Ravin Poudel; Lidimarie Trujillo Rodriguez; Christopher R. Reisch; Adam R Rivers
GigaScience

October 22, 2021

Dear Dr. Edmunds:

Thank you for synthesizing the comments from the reviewers, I found them fair and constructive and have implemented or addressed all suggestions, including feature requests and requests for additional data. For ease of review, I have placed all the reviewer comments and my responses in tabular format and attached that document as well. We have also registered GuideMaker with bio.tools (https://bio.tools/guidemaker) and SciCrunch.org (SCR_021778) and included those identifiers in the paper.

Sincerely,

Adam Rivers

Point-by-point comments:

Reviewer 1

1. I tested the website and the tool, not finding any bugs and errors. Website is well made, congratulations!
Thanks.
2. Name of the tool: GuideMaker is not self-explanatory for what it is specialized for, which is pooled design. In the future consider naming your tools more distinctly as I am afraid that currently the tool will be buried under hundreds of other GuideSomething tools.
This is a good point. At this point we have a domain, website, preprint and users of the software so it would be pretty disruptive to change, but we will be more specific with future names.
Authors also claim to support Cas13 (page 3 line 65), but don't mention anything more specific about it. I mention that because design for RNA is vastly different from design for DNA and it should be explained how the tool designs for RNA.
We have removed mention of Cas13 since it was not evaluated for this application.
From my understanding the tool offers highly discriminatory settings towards off-target search for a quick resolution of the all vs all comparison problem, however authors ignore that CRISPR off-targets are not defined by the hamming distance, but levenshtein distance. This was proven already by many studies e.g. Tsai et al. 2015. I recommend that authors embrace this issue in the paper and explain why their design may be suitable, and for what kind of studies it would be alright to use hamming distance vs levenshtein distance instead of ignoring the problem.
We added an option to use Levenshtein distance to the command-line version of GuideMaker. We also evaluated the effect of using each distance metric on the guides selected there was virtually no difference in the guides selected for the bacterial genome tested. Results are reported in the manuscript on lines 127-130 and in Supplementary Table 4. We suspect this is because there were not many of indels in guides from bacterial genomes. Because hamming is much faster and gave equivalent results, we have kept it as the default distance metric.

5. Study could gain prominence by showing a couple figures and describing how the grid-optimization parameters were selected. This would be especially important for everyone that wants to use this tool for nonbacterial gnomes (page 6, lines 128-131). Although script for optimization is included, it would be good to see what are the tradeoffs.
We have added graphical outputs to the grid optimization Jupyter notebook, along with instructions on how to run it for other genomes. We explain how to use this in the paper.
I believe that Figure 4 and all other AVX2 vs nonAVX2 comparisons are not interesting enough to include multiple times. AVX2 improvements are nice, but the tool is already plenty fast, and running time of 250 vs 220 seconds does not matter for normal users.

	We have removed the redundant non-AVX figures since most processors now support AVX2. The AVX2 performance gains were larger previously but the NMSlib library improved its non-AVX performance so there is not much difference anymore. Supp fig 3 summarizes the effects of AVX2. Similarly the number of cores does not seem to influence tool speed above 8 cores and one figure should be enough to explain that. We removed additional cores above 16, but retained 16 to show the flattening out in performance between 8 and 16. Tool claims very fast running times, but does not compare to the running times of other similar tools for the design of the pooled screens, this could highlight its superiority.We now compare the tools to the command-line version of CHOPCHOP using E. coli in Supplementary Fig 5. Despite using precomputed mappings for CHOPCHOP, Guidemaker is about 100x faster and uses 60% less memory CHOPCHOP is a general tool for the design of pooled screens while here it is used as a pooled screen tool due to its configurability. Additionally, CHOPCHOP also supports all PAM and all species, but on its python version availableThat is a good point we have used the CLI version of CHOPCHOP for the added comparisons. Comparisons to CHOPCHOP focus on the guides found, but I don't understand why consensus ratio between the tools should matter. What is more important is whether GuideMaker does indeed not filter any guides that are preferable for each gene (e.g. by CHOPCHOP ranking) and whether its hamming based filter is good enough to not cause significant unknown off-target effects (levenshtein distance off-targets not found by hamming distance filter). All it takes is one bulge and the hamming distance will become large, while levenshtein distance can even be as low as 1. We used CHOPCHOP consensus because it is widely used and there is not a gold-standard ground truth data for this. Guidemaker reports about the same number of targets when these same filtering metrics are applied. We have also added Doench et al. 2016 scoring (Azimuth) and CFD scoring to evaluate on target and off target guides for Cas9, so user can sort output by these scores. We added Supp fig 6 to show the effect of selection on On-Target and Off target scores filtered with GuideMaker parameters or unfiltered like CHOPCHOP. Our lsr filtering does not affect on-target scoring but does reduce off target scoring slightly. Our testing has revealed that using Levenshtein distance does not affect guide selection (see explanation above) It is not clear to me why the tool can't be used with large genomes, filtering on the 11bp seed and hamming distance should be plenty fast for also very large genomes. It can used for larger genomes just that HNSW loses it speed advantage around at around 1E9 guides. HNSW starts out much faster than indexing and searching the whole genome conventionally but the time per query grows more slowly for the conventional methods. Eventually it becomes faster to use conventional search rather than HNSW search. Could it be that the tool should support other input, not only genbank file format?We have added support for importing sequence and annotation from GFF/GTF and Fasta files. Reviewer 2 The author developed a software, GuideMaker, for designing CRISPR-Cas guide RNA pools in non-model genomes. Three bacterial genomes, a fungal genome, and a plant genome were used in performance benchmarking, which proves that the software supports the design of gRNAs in non-standard Cas enzymes for non-model organisms at the genome-scale. However, the advantages of this software are not well estimated nor presented compared to other tools like CHOPCHOP.We have improved our explanation of the advantages of GuideMaker relative to CHOPCHOP for its intended applications, including a performance evaluation in Supplementary figure 5 and a better explanation. We have also added both on-target and off-target scoring for NGG PAMs (the only PAMs for which training data is available), from Doench et al. 2016. Also, the software was mainly evaluated in three bacteria genomes, one fungus and Arabidopsis genome. There are no tests for non-model plant or animal genomes. Therefore, the "non-model genomes" in the title are exaggerated. I list more problems as follows. We have added the genome of the 537 MG plant Phaeseolis vulgarus. Our assertion that Guidemaker can be used for non-model organisms comes from the fact that it does not require precomputed reference genomes but rather computes guide pools quickly on the fly. This feature of the software can be shown without necessarily the

	genomes of obscure organism. We have clarified this confusing part in the since Pseudomonas and Arabidopsis certainly are model organisms. The authors did not compare the computation resources and performance (running time, memory) with existing softwares like CHOPCHOP. Also, the authors need to compare the score rankings with CHOPCHOP to present the relative power of GuideMaker. Is there any score rankings concerning efficiency or off-target possibilities for the designed Guide RNAs This is a good suggestion; we have added Supp. Fig 5 that looks at the time and memory requirements for Guidemaker and CHOPCHOP CLI. We have added the same on target and off target ranking algorithms used by CHOPCHOP V3. Those algorithms are Azimuth and CDF from Doench et al. (2016). 2. It is better to add support for gff formated annotation input files since many non-model species do not have GenBank annotations.We have added support for importing sequence and annotation from GFF/GTF and Fasta files. 3. The authors mentioned GuideMaker can design gRNAs for any small to medium size genome (up to about 500 megabases). The maximum genome used in the article was Arabidopsis thaliana (114.1MB), which is obviously smaller than the described (up to about 500 megabases). We couldn't find the description whether the authors had investigated the larger genomes. Therefore, the detailed analysis or discussion of this problem is needed. We have added the 537 MB Phaseolus vulgraris genome in Supp. Fig 4 to demonstrate this claim. 4. The authors stated GuideMaker to design CRISPR-Cas guide RNA pools in non-model genomes. Arabidopsis thaliana is a model organism and test in a non-model plant genome will be highly valuable. We have added the genome of the 537 MG plant Phaeseolis vulgarus. Our assertion that Guidemaker can be used for non-model organisms comes from the fact that it does not require precomputed reference genomes but rather computes guide pools quickly on the fly. We have clarified this confusing part in the since Pseudomonas and Arabidopsis were model organisms. 5. It is also stated that GuideMaker can design gRNAs for any PAM sequence from any Cas system but the results of SaCas and StCad was described in only one sentence. This is now also shown in detail in Supplementary Figures 1-4. Guidemaker allows any PAM to be chosen and more complex PAMs run faster, Supplementary Figure 1-2. 6. The source of the genomes was missing in the manuscript. In particular, some species have multiple genome versions in the same database. Therefore, to make the results more repeatable, the specific website and version number for each species are needed.This is a good point we have added the exact Accessions to the main text of manuscript (lines 176-179) and Supplementary Table 1. Minor comments 1. Line 11, "bacteria" should be "bacterias". It appears that “bacteria” is an acceptable plural form of the singular noun “bacterium”, based on this explanation: https://www.merriam-webster.com/dictionary/bacteria 2. Line 38, delete the", including non-model organisms", prokaryotic and eukaryotic organisms include the non-model organisms. Deleted. 3. Line 111, "candidates guides" should be "candidate guides". Corrected. 4. Line154, "gRNA identify with GuideMaker" should be "gRNA identified with GuideMaker". Corrected. 5. Line 195, "The second way GuideMaker reduces…" should be "The second way that GuideMaker reduces…". This section was rewritten so the text no longer exists. 6. Line 204, "and", no need for italics. This was italicized for emphasis. I have removed the italics. 7. Line 207, "gRNA's" should be "gRNAs". Corrected 8. Lines 209-210, "we anticipate performance will…" should be "we anticipate that performance will…". Added optional that. 9. Figure. 1. It seems that the font size of the description of Control gRNAs is inconsistent with others, please check. The entire document has been reformatted to 12-point font.

10. Line 22,55,98,159,175,187,219 and 247, "Guidemaker" should be "GuideMaker".
Thanks, the format is now consistent.
11. Line 262, "CAS" should be "Cas".
Corrected
12. Supplementary Figure 4. Grammar mistake in sentence "the different number of logical cores with or without AVX2 settings are available". It should be "the different number of logical cores with or without AVX2 settings is available".
This has been rewritten for clarity.

Reviewer 3

Overall, the tool is very well documented and easy to use. In the current version of the manuscript, GuideMaker does not show a clear improvement over the state-of-the-art design tool, CHOPCHOP. The authors do not implement any existing on-target scoring methods to determine the targeting efficacy of the picked sgRNAs. This can lead to picking guides that are highly specific but not effective enough. We have improved our explanation of the advantages of GuideMaker relative to CHOPCHOP for its intended applications, including a performance evaluation in Supp. Fig. 5 and a better explanation in the text. We have also added both on-target and off-target scoring for the “NGG” PAM (the only PAM for which training data is available). Based on the model from Doench et al. 2016.
1. Implementing on-target scoring methods, at least for the Cas enzymes that have on-target efficacy information, can help improve the process of picking sgRNAs. This tool will probably be used more often with standard Cas enzymes and it will be useful to have on-target efficacy scores attached to the guide RNAs.
Good suggestion, we have implemented the Azimuth model for on-target scoring from Doench et al. 2016, specifically their “V3_nopos” model. We have also refactored the original feature calling to improve speed, updated code to Python 3.9 and transferred their original model in pickle format to a safer, reproducible, cross platform compatible model in the Onnx runtime. We have also added the off target CFD scoring from the same paper.
2. The authors do a thorough analysis of the computational performance of GuideMaker with various genomes and Cas enzymes but including a comparison of the computational performance of GuideMaker vs. CHOPCHOP will strengthen the manuscript.
We have added this comparison, in Supp. Fig 5.
3. The authors define the PAM sequence of SaCas9 to be NGRRT whereas the canonical PAM sequence of SaCas9 is NNGRRT. This should be modified throughout the manuscript and analyses involving SaCas9 should be redoneWe have fixed this issue.
A good addition to the tool would be to output a file with all the sequences that were designed targeting the region of interest with the specific PAM sequence. This gives the user a sense of the universe from which the final guides were picked.
The user can get this by filtering the current output file by the locus name.
5. Another useful input parameter would be to specify a target region that the user wants to focus on such as letting the user input genomic coordinates or a gene name or locus tag. For example, CRISPy by Blin et al., 2016 takes a GenBank file as input and allows the user to input features specific to the uploaded genome.
Minor PointsWe have added the “--filter_by_locus“ option to filter results for this application.
1. "CyVerse" is misspelled as "CyCVerse" in multiple places in the manuscript.
We have fixed this.
2. Reference Figure 2 in Line 92.
Added.
3. Line 154: "Ratios between tools were calculated by dividing the number of gRNA identified.."The sentence was rewritten for clarity.
4. In Supplementary Figure 3 "wit haVX2" should be "with aVX2".
Corrected.
5. GitHub link in Line 336 does not work.
Those links are fixed.
6. Line 225-226: "GuideMaker also creates off-target gRNAs for use as negative controls in high-throughput experiments." "Off-target gRNAs" is misleading in this context.

	We now refer to them as “off-target control RNA sequences” since they are not guides.
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.	Yes

Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	

GuideMaker: Software to design CRISPR-Cas guide RNA pools in non-model genomes

Ravin Poudel[1,2], Lidimarie Trujillo Rodriguez[2], Christopher R. Reisch[2], and Adam R. Rivers[1*]

[1]Genomics and Bioinformatics Research Unit, USDA Agricultural Research Service, Gainesville, FL, 32608, USA

[2]Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL 326011, USA

* Correspondence: adam.rivers@usda.gov

Abstract

Background:

CRISPR-Cas systems have expanded the possibilities for gene editing in bacteria and eukaryotes. There are many excellent tools for designing CRISPR-Cas guide RNAs for model organisms with standard Cas enzymes. GuideMaker is intended as a fast and easy-to-use design tool for challenging projects with 1) non-standard Cas enzymes, 2) non-model organisms, or 3) projects that need to design a panel of guide RNAs (gRNA) for genome-wide screens.

Findings:

GuideMaker can rapidly design gRNAs for gene targets across the genome using a degenerate protospacer adjacent motif (PAM) and a genome. The tool applies Hierarchical Navigable Small World (HNSW) graphs to speed up the comparison of guide RNAs and optionally provides on-target and off-target scoring. This allows the user to design effective gRNAs targeting all genes in a typical bacterial genome in about 1-2 minutes.

Conclusions:

GuideMaker enables the rapid design of genome-wide gRNA for any CRISPR-Cas enzyme in non-model organisms. While GuideMaker is designed with prokaryotic genomes in mind, it can efficiently process eukaryotic genomes as well. GuideMaker is available as command-line software, a stand-alone web application, and a tool in the CyCverse Discovery Environment. All versions are available under a Creative Commons CC0 1.0 Universal Public Domain Dedication.

Keywords PAM, CRISPR-Cas, gRNA, HNSW

Introduction

CRISPR-Cas technology enables rapid and efficient genome editing in both prokaryotic and eukaryotic cells [1,2]. CRISPR-based systems are set apart from other genome editing tools by the ease with which they can be programmed to target specific sequences. Almost any DNA sequence in the cell can be targeted if it possesses a compatible protospacer adjacent motif (PAM). The PAM is a sequence that flanks the DNA target site, known as the protospacer, and must be present for target recognition [3]. The target specifying guide-RNA (gRNA) can be supplied as RNA, or encoded in DNA, depending on the organism under investigation. Although CRISPR-Cas is often used to edit single genes in eukaryotes, it is increasingly used for other purposes in prokaryotic and eukaryotic organisms [4].

The *Streptococcus pyogenes* Cas9 (SpCas9) was the first Cas described [5] and it is still the most widely used enzyme in CRISPR gene editing. Other Cas enzymes described early in the CRISPR revolution, such as the *Staphylococcus aureus* Cas9 and the *Acidaminococcus* Cas12a, are also commonly used [6,7]. Accordingly, the parameters for these enzymes are often included in computational tools to identify CRISPR target sites [8–11]. Cas9 enzymes from other organisms and other Cas-associated proteins that can cleave dsDNA, ssDNA, ssRNA, and insert transposon elements have also been described and have their place in molecular toolkits [12–18]. Each of these enzymes generally has specific requirements, such as PAM sequence constraints, PAM

orientation, and protospacer length. Many of these CRISPR-Cas systems have been repurposed to enable molecular genetics techniques like gene deletions, gene insertions, transcriptional depletion and activation, and translational repression [12,19–22]. Some of these techniques can be scaled to the genome level with chip-synthesized oligonucleotides and pooled approaches to screening [23]. In pooled screens, high-throughput DNA sequencing is used to identify how the pool has changed over time to elucidate genes that affect cells' fitness in specific conditions. Given the diversity of the CRISPR systems and their uses, identifying appropriate target sites is not trivial, especially for the number of targets needed for genome-scale experiments.

Here we introduce GuideMaker, a computational tool to identify target sites and design gRNA sequences that is not limited to any specific CRISPR system or organism. GuideMaker is most useful for a few kinds of CRISPR experiments. The first use case is designing pools of gRNAs for genome-wide screening experiments like Perturb-seq and CRISPR pool [23,24]. GuideMaker is optimized for making the all-versus-all comparisons necessary to design a genome-wide screen and return candidate gRNAs for every gene locus. The tool allows the user to filter targets based on their proximity to features of interest, like the start codon for any coding sequence. The second major use case is for researchers working with non-model organisms. Online gRNA design tools often have a limited number of preselected genomes available for analysis because most methods require PAM site positions to be precomputed. GuideMaker rapidly computes all guide positions on demand from user-provided GenBank files or a set of GFF/GTF (general feature format/general transfer format) files and fasta files from any organism. The third use case is for researchers working with Cas enzymes other than the canonical versions of Cas9, Cas12a (Cpf1), or Cas13 with different PAM and target site requirements. GuideMaker allows the user to specify a custom PAM with variable length, including degenerate nucleotides and allows the PAM to be on either the 3' or 5' side of the protospacer. These features allow GuideMaker to support any current or future CRISPR-Cas system. Since the determination of which CRISPR-Cas system functions best in any given organism is not predictable, this tool is highly relevant to researchers developing CRISPR tools in new species. For SgCas9 GuideMaker also implements on-target and off-target scoring from Doench et al. (2016). Because there is limited experimental

data on most Cas/organism combinations, cannot calculate target scoring for other Cas enzymes but instead uses design heuristics that prioritize uniqueness in the seed region of the guide.

Methods

Main features, input parameters, and workflow

GuideMaker is designed to be easy to use as either a web application (Figure 2) or a command-line utility. The key features of GuideMaker are:

1. All the potential guides in a genome can be quickly designed in one run.
2. It can design gRNAs for any PAM sequence from any Cas system.
3. Search is customizable through user-defined guide parameters (as highlighted in Figure 1). These features are specific to organisms, CRISPR-Cas systems, and experiments. Tuning these parameters can improve the sensitivity and specificity of gRNA.
4. Users can exclude specific restriction sites from guides to preserve those sites for downstream experiments.
5. It creates control sequences based on the input genome. In CRISPR experiments it is often desirable to create negative control sequences to evaluate off-target binding. GuideMaker provides the user with realistic control gRNAs that are highly divergent from sequences adjacent to PAM sites.
6. It provides an option to select the subset of results by locus tags of interest.
7. It rovides off-target Cutting Frequency Determination (CFD) scores for gRNAs [8].
8. Provides on-target efficacy score for canonical "NGG" PAM. These efficiency scores are based on Azimuth algorithm[8].
9. Provides tabular result files which can be used for the design and ordering of gRNA pools.
10. Provides an interactive visualization and exploratory tool to evaluate the guides.
11. The software can be run as a web application [25], a CyVerse application, or a command-line application [26]. Server code is included for running local instances of the web application as well.

A typical workflow of GuideMaker involves three major steps (Figure 2). In the first step, the user uploads the input genome in one or more GenBank or GFF/GTF and fasta files (gzipped or uncompressed) and defines the PAM and gRNA parameters (as highlighted in Figure 1). GuideMaker identifies and filters target sites, then returns summary data to the graphical environment (Figure 2). Users can inspect the interactive plots to learn more about the identified gRNAs and sort them by genome coordinates or locus tag. In the final step, GuideMaker provides the results as downloadable files under the results section. These files are used for synthesizing the guides. The command-line version of GuideMaker has similar input parameters as the web application, with the flexibility to generate plots, configure the underlying hyper-parameters for the Hierarchical Navigable Small World (HNSW) graph, filter the results by specific locus tag, select Hamming or Levinschein as the edit distance, predict on-target scores for "NGG" PAM, off-target CFD scores, or to run the web application locally. To make the application easier to install we distribute the application as a Bioconda environment [27], Docker container [28], Python package on Github [26], through the CyVerse discovery environment [29] or as an online web application [25]. Detailed information on accessing the software through various methods is available on the project homepage [30].

Search method

GuideMaker initially scans the genome, recording all candidate guide sequences adjacent to the specified PAM sequence on both DNA strands (Figure 3). Candidate guides are then optionally checked for the restriction sites. Next, the candidate guides are searched for a unique "seed region" closest to the PAM site and candidate gRNAs that are not unique in their "seed region" are removed. Then, approximate nearest neighbor search is used to remove candidate guides too similar to PAM adjacent sequences in the genome, based on Hamming distance by default (the number of substitutions required to turn one DNA sequence into another equal-length sequence). Levenshtein distance is optionally available on the command line and CyVerse versions of GuideMaker, but it is substantially slower and returns guides that are ~99.9% similar for tests in *E. coli*. (Supplementary Table 4). The approximate nearest neighbor search is performed using the Hierarchical Navigable Small World (HNSW) graph method in the Non-Metric Space Library (NMSLIB) [31,32]. An index of all the initial candidate guides is created using the bitwise Hamming distance metric.

Each guide with a unique "seed region" is compared to all candidate guides and any guides with edit distances (user can select either the Hamming or Levenshtein distance, with Hamming being the default) below the user-set threshold are removed. This differs from the standard procedure of indexing the genome and mapping each candidate guide against the whole genome then parsing each result. HNSW has a search complexity of $O(\log N)$ and index complexity of $O(N \cdot \log N)$ [31]. Finally, user-defined criteria are applied to specify the proximity and orientation of guides relative to genomic features like genes. A list of guides is then returned to the user with relevant information about the guide and its target genomic features.

GuideMaker provides an option to predict off-target CFD scores (flag *–cfd_score*) on the predicted guides, and on-target scores (flag *–doench_efficiency_score*) for the canonical NGG PAM using the the "version3 no position" gradient boosted regression treed model from Doench et al (2016). These models have been converted into json format and Open Neural Network exchange format respectively for reproducibility, and featurization and scoring has been refactored to run in Python 3 using the Onnx model runtime [33].

The core of GuideMaker's search method is the HNSW method in NMSLIB [32]. The method builds a multilayer graph index of the input data and has several parameters that can be optimized for index building and search to trade-off speed and accuracy. Graph construction is the most time-consuming step in our tests, and thus grid optimization was run to minimize run time while keeping recall above 99% relative to the ground truth exact nearest-neighbor search. The grid-optimization parameters: [M, efc, ef, and post] used in the HNSW graph for approximate nearest neighbor search have been optimized for bacterial genomes. A Jupyter notebook [34] script for re-optimization and visualization of these hyper-parameters is included in the test directory of the command-line version of the software and optimized parameters can be passed to GuideMaker with the *--config* flag.

Computational performance

Genomes of different sizes, GC content, and chromosome numbers were used to test the speed and scalability of GuideMaker (Supplementary Table 1). For benchmarking the performance, the same parameters were used unless a specific parameter was being tested: a PAM motif of 'NGG', 3' pam orientation, target

length of 20, lsr (length of seed region) of 11, before and after parameters of 500, knum of 10, controls of 10, dist of 3 and threads of 16. We profiled the performance of GuideMaker with different threads [1, 2, 4, 8, 16] in processors with and without the AVX2 processor instruction set. All tests were run on a single compute node with 2 x 24 core Intel® X®(R) Platinum 8260 CPU @ 2.40 GHz with Cascade Lake microarchitecture. Three bacterial genomes, a fungal genome, and a plant genome were used in performance benchmarking: *Escherichia coli* K12 (NC_000913), *Pseudomonas aeruginosa* PAO1 (NC_002516), *Burkholderia thailandensis* E264 (NC_007651), *Arabidopsis thaliana* (NC_003070)*, Aspergillus fumigatus* (NC_007194)*,* and *Phaseolus vulgaris* (NC_023759). For the gene or locus-specific comparisons, only the guides within the locus coordinates (i.e., zero feature distance) were considered.

Comparison to existing design method

We compared the results of GuideMaker with the results of the online and command-line versions of CHOPCHOP[35]. GuideMaker and CHOPCHOP parameters were set to approximate the same search. The length of the target sequence was set to 20 and zero mismatches were allowed in the seed region (11bp) of the target. The *Escherichia coli* (str. K-12/MG1655) genome was used with the online version of CHOPCHOP since it has a limited number of genomes. Targets were searched in 40 Kbp increments to account for CHOPCHOP's size limitations. Target sequences were searched across multiple 40 Kbp segments *of E.coli* genome (NC_000913.3:2001-42000, NC_000913.3:80001-120000, NC_000913.3:160001-200000, NC_000913.3:240001-280000, and NC_000913.3:320001-360000). We also searched for target sequences and genes/locus_tags within 40Kbp of (NC_000913.3:2001-42000) to compare identifications at the locus level. The ratio between the tools was calculated by dividing the number of gRNA identified with GuideMaker by the number of guides identified by CHOPCHOP to represent the proportion of guides identified by both GuideMaker and CHOPCHOP.

The command-line version of CHOPCHOP was used to compare the memory usage and computation time of CHOPCHOP and GuideMaker over an entire genome. The *E. coli* K-12 genome was chosen for comparison because the precomputed 2bit genome files and Bowtie indexes were provided with

CHOPCHOP v 3. The matching GenBank file was downloaded for Guidemaker and both programs were run 5 times on the same machine using different numbers of processor cores [1, 2, 4, 8, 16].

Results

The time for GuideMaker to complete a typical run identifying all SpCas9 gRNAs (PAM 'NGG') in a bacterial genome using 8 compute cores was 75 seconds for *E. coli* and 130 seconds for *P. aeruginosa* (Figure 4). For SaCas9 and StCas9, which have a longer PAM sequence ('NGRRT' and 'NNAGAAW' respectively, with 3' PAM orientation) and thereby fewer potential targets, the same genomes ran in 19 or 5 seconds (Supplementary Figures 1). The fungus *Aspergillus fumigatus* (28MB) and the plants *Arabidopsis thaliana* (114 MB) and *Phaseolus vulgaris* (537MB) have larger genomes but are still processed quickly. *A. fumigatus* processed between 23-304 seconds, while *A. thaliana* processed in 250-921 and *P. vulgaris* processed in 333-4162 seconds depending on the number of cores, AVX2 instructions, and PAM sequence (Supplementary Figure 2). GuideMaker can take advantage of Advanced Vector Extensions (AVX2) on newer x86 processors, which improves the search speed because HNSW search is accelerated with AVX2 (Supplementary Figure 3). The acceleration was larger when fewer processors were available (Supplementary Figure 3). The HNSW algorithms are parallelized, and indexing-and-search takes most of the compute time in GuideMaker so the software scales well when additional cores are added up to 8 cores (Supplementary Figure 3). In practice it scaled up sub-linearly with genome size, globally estimating Cas9 guides for *E. coli* MG1655 (4.6MB) in 75 seconds and *Phaseolus vulgaris* (537MB) in 1549 seconds, both on 8 cores (Memory usage: 1.9GB for *E. coli* and 46.9GB for *P. vulgaris,* Supplementary Figure 4).

The results of GuideMaker were compared with the popular guide design software CHOPCHOP version 3 [35]. When GuideMaker's filtering settings are set to match CHOPCHOP, the results are very similar and 99.9% of the targets identified by GuideMaker fall within 2bp of target coordinates returned by CHOPCHOP. When GuideMaker's unique seed region criterion was not applied at the loci level, the average number of guides identified by the two approaches was similar per locus (Mean GuideMaker = 116.8, Mean

CHOPCHOP = 113.6, p-value = 0.86, Supplementary Table 2). Although the number of guides identified per gene locus differed, none of the genes were missed by either tool. GuideMaker's default requirement of a unique seed region is more stringent than CHOPCHOP, and with it enabled, GuideMaker returns (count=1787) 38.4% (for 2Kbp-42Kbp regions) of the targets compared to CHOPCHOP (count=4651) *E. coli* K12. At the sequence level, 96.7% of the identified gRNA (1729/1787) from both tools had identical sequences. The more stringent seed region filtering used by default in GuideMaker reduced the CFD scores of guides suggesting that it would reduce off target binding (Supplementary Figure 6), but that would need to be experimentally validated in a range of organisms. The ratio of gRNA found by both the tools across the multiple 40Kbp regions was 39.2% (sd= 1.9%, Supplementary Table 3) when using GuideMaker's more stringent default settings. This ratio was calculated by dividing the number of gRNA from GuideMaker by the number from CHOPCHOP for each 40Kb region. GuideMaker processed an entire *E. coli* genome about 60 times faster than the command line version of CHOPCHOP v3. It also used almost 50% less memory across all the compared processor cores (Supplementary Figure 5).

Discussion

Designing gRNAs is a two-step process where GuideMaker first identifies potential guides adjacent to PAM sequences and then filters the potential guides based on multiple criteria. The most important criterion is that each guide has a minimum edit distance from any other sequence adjacent to a PAM site in the genome; this decreases the likelihood of off-target binding. The second way GuideMaker reduces off-target binding is by requiring that a set number of bases near the PAM site are unique from any other candidate guide. The 8 bases nearest the PAM are the most important for target specificity, and any mismatch is sufficient to prevent binding [36,37]. The length of the unique region should be set with consideration for the size of the genome since requiring short unique regions will limit the number of total guides that can be found. For example, requiring that every gRNA be unique in the first 3 bp would only allow for $4^3 = 64$ possible guides to be designed. For normal *--lsr* values of 9-12 this is only limiting for human-sized genomes and can be disabled by

setting *--lsr* to 0. All guides designed by GuideMaker are perfect matches to a single site in the genome. Additional specificity is obtained by requiring all similar PAM-adjacent sequences to be unique in the critical "seed region" and have a total number of mismatches that exceed the user-defined threshold. This double criterion is expected to increase specificity.

The primary goal of the current version of our software is to support the design of gRNAs for non-standard Cas enzymes or non-model organisms at the genome scale. Guide RNAs do not perform equally, thus empirical experiments will be needed to fully validate the functionality and efficacy of gRNA predictions. Given the similarity in targets identified by GuideMaker and CHOPCHOP, we anticipate that performance is similar to the current state of the art but applicable to more design use cases. When a unique seed region and edit distance-based filters were applied, GuideMaker created guides more conservatively, generating only about 40% of the guides created by CHOPCHOP. While CHOPCHOP has an option to specify the maximum number of mismatches in the first 9 bp or the whole guide, it does not allow the application of both criteria. While there are small differences in the number and position of guides generated by GuideMaker, with GuideMaker being more conservative by default, both programs create enough guides to target nearly all gene loci in the genome of *E. coli*. The current version of the GuideMaker provides options to predict off-target CFD scores and on-target scores for the canonical NGG PAM. Both scoring approaches are based on the publicly available models trained on empirical data with SpCas9. If experimentally validated data become available from genome-wide screens with different Cas enzymes, future versions of GuideMaker could potentially incorporate new scoring models to help rank candidate guides.

GuideMaker is a fast and flexible tool for designing guide RNA across the entire genome in non-model organisms or with non-canonical Cas enzymes. It takes advantage of fast HNSW search to quickly index and search new genomes. Several parameters can be tuned to ensure compatibility with the specific application of the user. For example, GuideMaker checks the designed gRNA for a given restriction enzyme site to prevent incompatibility with the cloning strategy. Second, the maximum distance from a target sequence from the start of an annotated feature can be chosen to disrupt promoters or the beginning of the coding sequence, since these sites are preferred for CRISPRi experiments. GuideMaker also creates off-target

control RNA sequences for use as negative controls in high-throughput experiments. Lastly, the program plots the results for visual exploration of the targets and exports the data as .csv files. The software is available as a command-line application, a web application, and is integrated into the CyVerse Discovery Environment to provide users with a range of usage options. Guidemaker is a fast, flexible design tool for the creation of challenging guide RNA pools.

Availability and Requirements

Project name: GuideMaker

Project home page: https://guidemaker.org

Operating system(s): Linux or macOS

Programming language: Python >=3.6

Other requirements: pybedtools>=0.8.2, nmslib>=2.0.6, streamlit>1.2.6, biopython >=1.79, pandas >= 1.0, onnx>= 1.8.1

License: CC0 1.0 Public Domain Dedication

Competing Interests

Authors declare no competing interests

Data Availability

The source code and command-line executables for GuideMaker are available at the Zenodo [38] and can be installed directly from Github [26], Bioconda [27], or as a Docker container [28]. Data and code to reproduce the analysis in the paper are available at Zenodo [38]. As a work of the United States Department of Agriculture, GuideMaker is released to the public domain under a Creative Commons (CC0) public domain attribution. The program is also available as a web application through the CyVerse discovery environment [29], and as a stand-alone web application [25].

Additional Files

Supplementary Figure 1. Performance of GuideMaker for SaCas9 and StCas9.

Supplementary Figure 2. Performance of GuideMaker for SpCas9, SaCas9, and StCas9.

Supplementary Figure 3. Performance of GuideMaker with AVX2 settings.

Supplementary Figure 4. Memory usage of GuideMaker for SpCas9, SaCas9, and StCas9.

Supplementary Figure 5. Performance and memory usage comparisons between CHOPCHOP (CLI version) and GuideMaker.

Supplementary Figure 5. Comparison of efficiency and CFD scores with or without GuideMaker based filters

Supplementary Table 1: Organisms features

Supplementary Table 2: Comparison of the average number of gRNA identified by GuideMaker and CHOPCHOP.

Supplementary Table 3: Comparison of consensus ratio between GuideMaker and CHOPCHOP.

Supplementary Table 4: Comparison of processing times and results for Levenshtein and Hamming distances

List of abbreviations

BWA: Burrows-Wheeler Aligner; Cas: CRISPR-associated protein; CDS: CoDing Sequences; CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats; gRNA: Guide RNA; HMSW: Hierarchical Navigable Small World; NMSLIB: Non-Metric Space Library; PAM: Protospacer Adjacent Motif

Funding

The research was supported by the United States Department of Agriculture (USDA), Agricultural Research Service (ARS) project number 6066-21310-005-D, and ARS cooperative agreement 6066-21310-005-28-S to the University of Florida. This research used resources provided by the SCINet scientific computing initiative of the USDA-ARS, ARS project number 0500-00093-001-00-D.

Author Contributions

R.P., L.T.R., C.R.R., and A.R.R. conceived and designed the study. R.P. and A.R.R developed and optimized the software and performed the experiments. R.P., L.T.R., C.R.R., and A.R.R, tested the software, wrote, and revised the manuscripts. All authors read and approved the final manuscript.

References

1. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol. 2013; doi: 10.1038/nbt.2508.

2. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nat Protoc. 2013; doi: 10.1038/nprot.2013.143.

3. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology. 2009; doi: 10.1099/mic.0.023960-0.

4. Pickar-Oliver A, Gersbach CA. The next generation of CRISPR–Cas technologies and applications. Nat Rev Mol Cell Biol. 2019; doi: 10.1038/s41580-019-0131-5.

5. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, et al.. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011; doi: 10.1038/nature09886.

6. Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, et al.. In vivo genome editing using Staphylococcus aureus Cas9. Nature. 2015; doi: 10.1038/nature14299.

7. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, et al.. Cpf1 Is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015; doi: 10.1016/j.cell.2015.09.038.

8. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al.. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016; doi: 10.1038/nbt.3437.

9. Hiranniramol K, Chen Y, Liu W, Wang X. Generalizable sgRNA design for improved CRISPR/Cas9 editing efficiency. Luigi Martelli P, editor. Bioinformatics. 2020; doi: 10.1093/bioinformatics/btaa041.

10. Xu H, Xiao T, Chen CH, Li W, Meyer CA, Wu Q, et al.. Sequence determinants of improved CRISPR sgRNA design. Genome Res. 2015; doi: 10.1101/gr.191452.115.

11. Perez AR, Pritykin Y, Vidigal JA, Chhangawala S, Zamparo L, Leslie CS, et al.. GuideScan software for improved single and paired CRISPR guide RNA design. Nat Biotechnol. 2017; doi: 10.1038/nbt.3804.

12. Anzalone A V., Koblan LW, Liu DR. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. Nat Biotechnol. 2020; doi: 10.1038/s41587-020-0561-9.

13. Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, et al.. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell. 2009; doi: 10.1016/j.cell.2009.07.040.

14. Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DBT, et al.. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science. 2016; doi: 10.1126/science.aaf5573.

15. Ma E, Harrington LB, O'Connell MR, Zhou K, Doudna JA. Single-stranded DNA cleavage by divergent CRISPR-Cas9 enzymes. Mol Cell. 2015; doi: 10.1016/j.molcel.2015.10.030.

16. Klompe SE, Vo PLH, Halpin-Healy TS, Sternberg SH. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. Nature. 2019; doi: 10.1038/s41586-019-1323-z.

17. Strecker J, Ladha A, Gardner Z, Schmid-Burgk JL, Makarova KS, Koonin E V, et al.. RNA-guided DNA insertion with CRISPR-associated transposases. Science. 2019; doi: 10.1126/science.aax9181.

18. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012; doi: 10.1126/science.1225829.

19. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014; doi: 10.1016/j.cell.2014.05.010.

20. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al.. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013; doi: 10.1016/j.cell.2013.02.022.

21. Cox DBT, Gootenberg JS, Abudayyeh OO, Franklin B, Kellner MJ, Joung J, et al.. RNA editing with CRISPR-Cas13. Science. 2017; doi: 10.1126/science.aaq0180.

22. Yan WX, Chong S, Zhang H, Makarova KS, Koonin E V., Cheng DR, et al.. Cas13d Is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein. Mol Cell. 2018; doi: 10.1016/j.molcel.2018.02.028.

23. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al.. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell. 2016; doi: 10.1016/j.cell.2016.11.038.

24. Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, et al.. A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. Cell. 2016; doi: 10.1016/j.cell.2016.05.003.

25. Zhu LJ. Overview of guide RNA design tools for CRISPR-Cas9 genome editing technology. Front Biol. 2015; doi: 10.1007/s11515-015-1366-y.

26. Cui Y, Xu J, Cheng M, Liao X, Peng S. Review of CRISPR/Cas9 sgRNA design tools. Interdiscip Sci Comput Life Sci. 2018; doi: 10.1007/s12539-018-0298-z.

27. GuideMaker. The GuideMaker web app. https://guidemaker.app.scinet.usda.gov. Accessed 2021 May 27.

28. GuideMaker 2021. GuideMaker (Version 0.2.0). https://github.com/USDA-ARS-GBRU/GuideMaker/releases/tag/v0.2.0.

29. GuideMaker. The GuideMaker bioconda installation. https://anaconda.org/bioconda/guidemaker.

30. GuideMaker. The GuideMaker docker container. https://github.com/orgs/USDA-ARS-GBRU/packages?repo_name=GuideMaker.

31. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, et al.. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. PLOS Biol. 2016; doi: 10.1371/journal.pbio.1002342.

32. GuideMaker. The GuideMaker project homepage. https://guidemaker.org. Accessed 2021 June 18.

33. Malkov YA, Yashunin DA. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Trans Pattern Anal Mach Intell. 2020; doi: 10.1109/TPAMI.2018.2889473.

34. Naidan B, Boytsov L, Malkov Y, Novak D. Non-metric space library manual. 2015; doi: http://arxiv.org/abs/1508.05470.

35. Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. CHOPCHOP v3: Expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res. 2019; doi: 10.1093/nar/gkz365.

36. Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, et al.. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A. 2011; doi: 10.1073/pnas.1104144108.

37. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al.. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013; doi: 10.1038/nbt.2647.

38. Poudel R, Rodriguez LT, Reisch CR, Rivers AR. Source code and command-line executables for GuideMaker: Software to design CRISPR-Cas guide RNA pools in non-model genomes. doi: 10.5281/zenodo.4849258.

39. Poudel R, Rodriguez LT, Reisch CR, Rivers AR. Supporting material and analysis code for GuideMaker: Software to design CRISPR-Cas guide RNA pools in non-model genomes. doi: 10.5281/zenodo.4898253.

40. Hirano S, Abudayyeh OO, Gootenberg JS, Horii T, Ishitani R, Hatada I, et al.. Structural basis for the promiscuous PAM recognition by *Corynebacterium diphtheriae* Cas9. Nat Commun. 2019; doi: 10.1038/s41467-019-09741-6.

41. Gleditzsch D, Pausch P, Müller-Esparza H, Özcan A, Guo X, Bange G, et al.. PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. RNA Biol. 2019; doi: 10.1080/15476286.2018.1504546.

42. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat Biotechnol. 2014; doi: 10.1038/nbt.2808.

403 43. Wu X, Kriz AJ, Sharp PA. Target specificity of the CRISPR-Cas9 system. Sander JD, Joung JK, editors.
404 Quant Biol. 2014; doi: 10.1007/s40484-014-0030-x.

405

406

407 **Figure 1. Input parameters for GuideMaker**

408

409 **Figure 2. A typical workflow of GuideMaker:** 1) A user uploads the input genome (single or multiple) as
410 GenBank file, then defines the PAM sequence along with all the associated parameters and submits them to
411 run the program. 2) GuideMaker processes the input files and generates the interactive plots. Users can use
412 these interactive plots to explore the results and sort them by locus tag and genome coordinates. 3)
413 GuideMaker provides all the results and log files as downloads under the "Results" section.

414

415 **Figure 3. Entity Relationship Diagram showing the operation of the GuideMaker core program.**

416

417 **Figure 4. Performance of GuideMaker for SpCas9.** Evaluating the performance of GuideMaker across
418 three bacterial genomes using the **"NGG**" PAM motif with a target length of 20, unique zone of 11, 3prime
419 PAM orientation, before and into parameters of 500, knum of 10, controls of 10, and dist of 3. The mean of
420 10 runs was used for the evaluation, where dot and bar represent the mean and standard error, respectively.

Figure 1

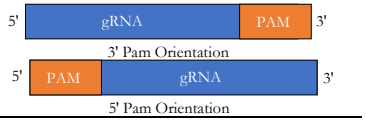
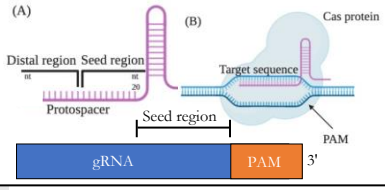
Inputs	Descriptions	Notes/Examples
Genome File	GuideMaker accepts one or more Genbank (.gbk or gzipped .gbk.gz) files with sequence data from a single genome as an input. GuideMaker extracts all the required information from the Genbank file to identify gRNAs and genomic features, allowing users to globally create gRNAs without preprocessed mapping files. Option: --genbank	E.g. *Carsonella_ruddii*.**gbk.gz**, *Carsonella_ruddii*.**gbk**
PAM	The Protospacer Adjacent Motif (PAM) is the short, generally 2-8 bp, sequence essential for binding by the Cas protein[3,40,41]. GuideMaker provides users the flexibility to define the PAM sequence for any Cas protein, enabling usage of new CRISPR-Cas systems. Degenerate PAM sequences are allowed. Option: --pamseq	E.g. NGG (SpCas9) NGRRT (SaCas9)
Restriction Enzymes	It can be useful to avoid sequences with restriction endonuclease recognition sites for used cloning guide library. GuideMaker allows users to provide a list of defined or degenerate restriction site sequences to avoid targeting. Option: --restriction_enzyme_list.	E.g. NGRT; Default: None
PAM Orientation	The PAM orientation parameter defines PAM position relative to the protospacer. Depending on the CRISPR-Cas system, the orientation of PAM could be 5' or 3' to the guide sequence. For instance, SpCas9 recognizes 'NGG' PAM on the 3' end of the guide (i.e. 5'-[guide][pam]-3'), whereas the Cpf1 PAM is on the 5' end of the guide sequence (i.e. 5'-[pam][guide]-3'). To accommodate such differences, GuideMaker offers flexibility to define the PAM orientation. Option: --pam_orientation.	
Guidelength	Guidelength defines the length of gRNA. Changing the guide length allows the user to adjust the gRNA efficacy and specificity [42]. GuideMaker allows users to select the length of gRNA within 10-27 bp. Option: --guidelength.	
Length of seed region	The seed region is the guide sequence closest to the PAM recognition site, and the distal region is the region furthest from the PAM. GuideMaker divides each guide into the seed and distal regions (Figure A and B). For instance, if the guide length is 22bp, and the length of the seed region is 10, then the size of the seed and the distal regions is 10 and 12, respectively. It has been shown that the region close to PAM is sensitive [36,43], and non-uniqueness in this region can lead to off-target matches; however, the importance of the seed region is specific to the CRISPR-Cas system and the organism. Thus, GuideMaker allows the user to define the seed region with the maximum length of 27 bp; although, the length of the seed region must be less than or equal to the Guidelength. Additionally, the length of the seed region should not be too small because the total number of possible guides is limited to 4 raised to the power of the seed length. Option: --lsr.	
Edit Distance	Edit distance defines the number of substitutions required to turn one DNA sequence into another sequence. GuideMaker calculates the pairwise edit distance between all the candidate gRNAs and all sequences adjacent to a PAM site. gRNAs with a distance less than or equal to the user-defined value are considered too similar and removed to minimize off-targeting. Option: --dist	Options: [0 – 5]; Default: 2
Distance type	Defines the edit distance type. GuideMaker provides two edit distance type: hamming ; and leven. Option: -- dtype	Options:[hamming, leven]; Default hamming
Before	Before parameter allows user to select gRNAs that are upstream of a feature's start site. For example, if "before" is set to 100, each gRNA within 100 bp upstream of a feature will be retrieved. Option: --before	Options: [1 – 500]; Default: 100
Into	The into parameter allows the user to select gRNAs that are downstream of a feature's start. For example, if "into" is set to 100, each gRNA within 100 bp downstream of a feature will be retrieved. Option: --into.	Options: [1 – 500]; Default: 200
Locus tag	List of locus tag for subsetting the final output so the gRNA specific to the listed locus tag are retrieved. Option: --filter_by_locus	Default: None
CFD score	Cutting Frequency Determination (CFD) score for accessing off-target activity of gRNAs. Option: --cfd_score	Default: None
Efficiency score	On-target efficiency score predicted based on Azimuth 2.0.– only for NGG PAM. Option: --doench_efficiency_score	Default: None
Similar guides	Retrieves the number of sequences similar to the gRNA. Option: --knum	Options: [2 – 20]; Default: 3
Control gRNAs	Provides the set number of random control gRNAs. Option: --controls	Default: 1000

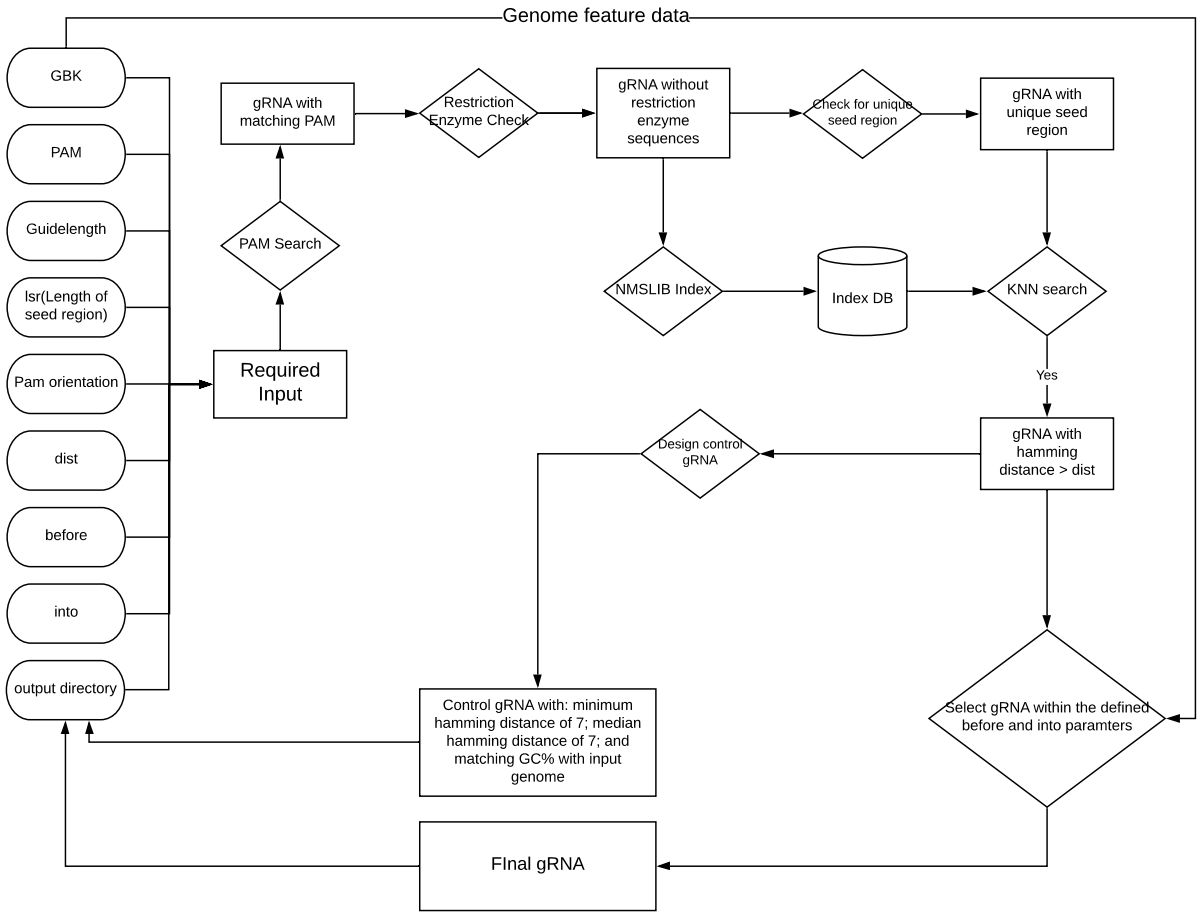
Figure 2

Click here to access/download;Figure;Figure 2.docx

Update fig

Figure 3

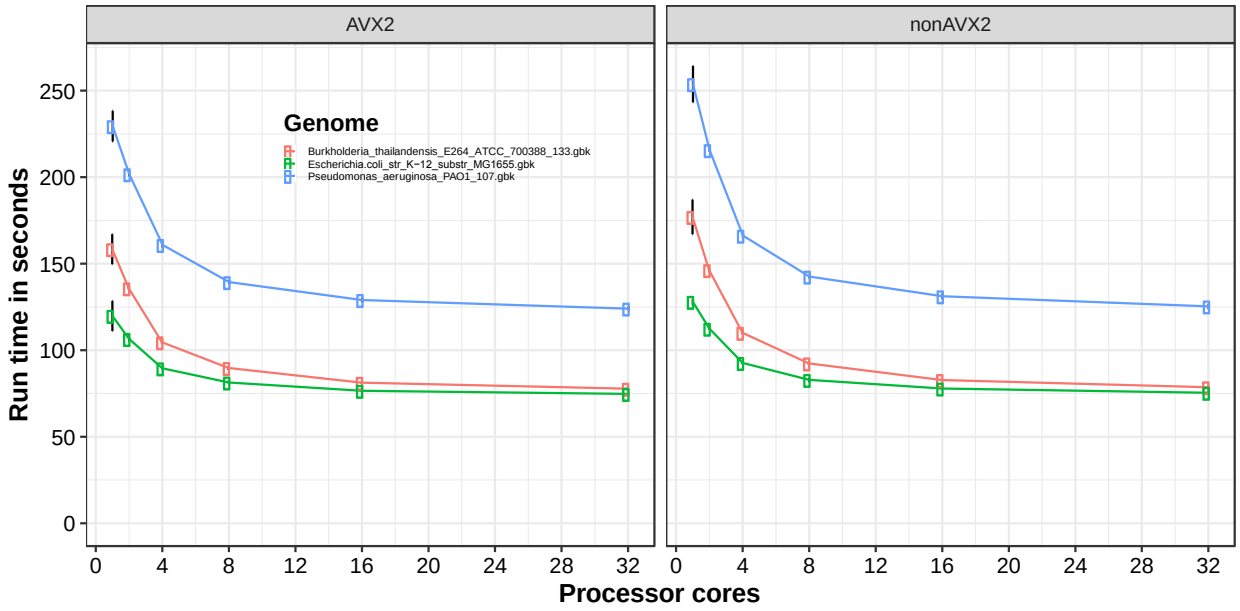
Click here to access/download;Figure;Figure 3.docx



Update fig with additional option – may be not –as its very specific to NGG

Figure 4

Click here to access/download;Figure;Figure 4.docx



Remove Non-avx2 to supplement

GIGA-D-21-00186
GuideMaker: Software to design CRISPR-Cas guide RNA pools in non-model genomes
Ravin Poudel; Lidimarie Trujillo Rodriguez; Christopher R. Reisch; Adam R Rivers
GigaScience

October 22, 2021

Dear Dr. Edmunds:

Thank you for synthesizing the comments from the reviewers, I found them fair and constructive and have implemented or addressed all suggestions, including feature requests and requests for additional data. For ease of review, I have placed all the reviewer comments and my responses in tabular format. We have also registered GuideMaker with bio.tools (https://bio.tools/guidemaker) and SciCrunch.org (SCR_021778) and included those identifiers in the paper.

Sincerely,

Adam Rivers

Reviewer 1

1. I tested the website and the tool, not finding any bugs and errors. Website is well made, congratulations!	Thanks.
2. Name of the tool: GuideMaker is not self-explanatory for what it is specialized for, which is pooled design. In the future consider naming your tools more distinctly as I am afraid that currently the tool will be buried under hundreds of other GuideSomething tools.	This is a good point. At this point we have a domain, website, preprint and users of the software so it would be pretty disruptive to change, but we will be more specific with future names.
Authors also claim to support Cas13 (page 3 line 65), but don't mention anything more specific about it. I mention that because design for RNA is vastly different from design for DNA and it should be explained how the tool designs for RNA.	We have removed mention of Cas13 since it was not evaluated for this application.
From my understanding the tool offers highly discriminatory settings towards off-target search for a quick resolution of the all vs all comparison problem, however authors ignore that CRISPR off-targets are not defined by the hamming distance, but levenshtein distance. This was proven already by many studies e.g. Tsai et al. 2015. I recommend that authors embrace this issue in the paper and explain why their design may be suitable, and for what kind of studies it	We added an option to use Levenshtein distance to the command-line version of GuideMaker. We also evaluated the effect of using each distance metric on the guides selected there was virtually no difference in the guides selected for the bacterial genome tested. Results are reported in the manuscript on lines 127-130 and in Supplemental Table 4. We suspect this is because there were not many of indels in guides from bacterial genomes. Because hamming is much

would be alright to use hamming distance vs levenshtein distance instead of ignoring the problem.	faster and gave equivalent results, we have kept it as the default distance metric.
5. Study could gain prominence by showing a couple figures and describing how the grid-optimization parameters were selected. This would be especially important for everyone that wants to use this tool for nonbacterial gnomes (page 6, lines 128-131). Although script for optimization is included, it would be good to see what are the tradeoffs.	We have added graphical outputs to the grid optimization Jupyter notebook, along with instructions on how to run it for other genomes. We explain how to use this in the paper.
I believe that Figure 4 and all other AVX2 vs nonAVX2 comparisons are not interesting enough to include multiple times. AVX2 improvements are nice, but the tool is already plenty fast, and running time of 250 vs 220 seconds does not matter for normal users.	We have removed the redundant non-AVX figures since most processors now support AVX2. The AVX2 performance gains were larger previously but the NMSlib library improved its non-AVX performance so there is not much difference anymore. Supp fig 3 summarizes the effects of AVX2.
Similarly the number of cores does not seem to influence tool speed above 8 cores and one figure should be enough to explain that.	We removed additional cores above 16, but retained 16 to show the flattening out in performance between 8 and 16.
Tool claims very fast running times, but does not compare to the running times of other similar tools for the design of the pooled screens, this could highlight its superiority.	We now compare the tools to the command-line version of CHOPCHOP using E. coli in Supplemental Fig 5. Despite using precomputed mappings for CHOPCHOP, Guidemaker is about 100x faster and uses 60% less memory
CHOPCHOP is a general tool for the design of pooled screens while here it is used as a pooled screen tool due to its configurability. Additionally, CHOPCHOP also supports all PAM and all species, but on its python version available	That is a good point we have used the CLI version of CHOPCHOP for the added comparisons.
Comparisons to CHOPCHOP focus on the guides found, but I don't understand why consensus ratio between the tools should matter. What is more important is whether GuideMaker does indeed not filter any guides that are preferable for each gene (e.g. by CHOPCHOP ranking) and whether its hamming based filter is good enough to not cause significant unknown off-target effects (levenshtein distance off-targets not found by hamming distance filter). All it takes is one bulge and the hamming distance will become large, while levenshtein distance can even be as low as 1.	We used CHOPCHOP consensus because it is widely used and there is not a gold-standard ground truth data for this. Guidemaker reports about the same number of targets when these same filtering metrics are applied. We have also added Doench et al. 2016 scoring (Azimuth) and CFD scoring to evaluate on target and off target guides for Cas9, so user can sort output by these scores. We added Supp fig 6 to show the effect of selection on On-Target and Off target scores filtered with GuideMaker parameters or unfiltered like CHOPCHOP. Our lsr filtering does not affect on-target scoring but does reduce off target scoring slightly. Our testing has revealed that using Levenshtein distance does not affect guide selection (see explanation above)

It is not clear to me why the tool can't be used with large genomes, filtering on the 11bp seed and hamming distance should be plenty fast for also very large genomes.	It can used for larger genomes just that HNSW loses it speed advantage around at around 1E10 guides. HNSW starts out much faster than indexing and searching the whole genome conventionally but the time per query grows more slowly for the conventional methods. Eventually it becomes faster to use conventional search rather than HNSW search.
Could it be that the tool should support other input, not only genbank file format?	We have added support for importing sequence and annotation from GFF/GTF and Fasta files.

Reviewer 2

The author developed a software, GuideMaker, for designing CRISPR-Cas guide RNA pools in non-model genomes. Three bacterial genomes, a fungal genome, and a plant genome were used in performance benchmarking, which proves that the software supports the design of gRNAs in non-standard Cas enzymes for non-model organisms at the genome-scale. However, the advantages of this software are not well estimated nor presented compared to other tools like CHOPCHOP.	We have improved our explanation of the advantages of GuideMaker relative to ChopChop for its intended applications, including a performance evaluation in Supplemental figure 5 and a better explanation. We have also added both on-target and off-target scoring for NGG PAMs (the only PAMs for which training data is available), from Doench et al. 2016.
Also, the software was mainly evaluated in three bacteria genomes, one fungus and Arabidopsis genome. There are no tests for non-model plant or animal genomes. Therefore, the "non-model genomes" in the title are exaggerated. I list more problems as follows.	We have added the genome of the 537 MG plant *Phaeseolis vulgarus.* Our assertion that Guidemaker can be used for non-model organisms comes from the fact that it does not require precomputed reference genomes but rather computes guide pools quickly on the fly. This feature of the software can be shown without necessarily the genomes of obscure organism. We have clarified this confusing part in the since Pseudomonas and Arabidopsis certainly are model organisms.
The authors did not compare the computation resources and performance (running time, memory) with existing softwares like CHOPCHOP. Also, the authors need to compare the score rankings with CHOPCHOP to present the relative power of GuideMaker. Is there any score rankings concerning efficiency or off-target possibilities for the designed Guide RNAs	This is a good suggestion; we have added Supp. Fig 5 that looks at the time and memory requirements for Guidemaker and CHOPCHOP CLI. We have added the same on target and off target ranking algorithms used by CHOPCHOP V3. Those algorithms are Azimuth and CDF from Doench et al. (2016).
2. It is better to add support for gff formated annotation input files since many non-model species do not have GenBank annotations.	We have added support for importing sequence and annotation from GFF/GTF and Fasta files.
3. The authors mentioned GuideMaker can design gRNAs for any small to medium size genome (up to about 500 megabases). The maximum genome used in the article was Arabidopsis thaliana	We have added the 537 MB *Phaseolus vulgraris* genome in Supp. Fig 4 to demonstrate this claim.

(114.1MB), which is obviously smaller than the described (up to about 500 megabases). We couldn't find the description whether the authors had investigated the larger genomes. Therefore, the detailed analysis or discussion of this problem is needed.	
4. The authors stated GuideMaker to design CRISPR-Cas guide RNA pools in non-model genomes. Arabidopsis thaliana is a model organism and test in a non-model plant genome will be highly valuable.	We have added the genome of the 537 MG plant *Phaeseolis vulgarus.* Our assertion that Guidemaker can be used for non-model organisms comes from the fact that it does not require precomputed reference genomes but rather computes guide pools quickly on the fly. We have clarified this confusing part in the since Pseudomonas and Arabidopsis were model organisms.
5. It is also stated that GuideMaker can design gRNAs for any PAM sequence from any Cas system but the results of SaCas and StCad was described in only one sentence.	This is now also shown in detail in Supp. Fig 4. Guidemaker allows any PAM to be chosen and more complex PAMs actually run faster.
6. The source of the genomes was missing in the manuscript. In particular, some species have multiple genome versions in the same database. Therefore, to make the results more repeatable, the specific website and version number for each species are needed.	This is a good point we have added the exact Accessions to the manuscript.
Minor comments	
1. Line 11, "bacteria" should be "bacterias".	It appears that “bacteria” is an acceptable plural form of the singular noun “bacterium”, based on this explanation: https://www.merriam-webster.com/dictionary/bacteria
2. Line 38, delete the", including non-model organisms", prokaryotic and eukaryotic organisms include the non-model organisms.	Deleted.
3. Line 111, "candidates guides" should be "candidate guides".	Corrected.
4. Line154, "gRNA identify with GuideMaker" should be "gRNA identified with GuideMaker".	Corrected.
5. Line 195, "The second way GuideMaker reduces…" should be "The second way that GuideMaker reduces…".	This section was rewritten so the text no longer exists.
6. Line 204, "and", no need for italics.	This was italicized for emphasis. I have removed the italics.
7. Line 207, "gRNA's" should be "gRNAs".	Corrected

8. Lines 209-210, "we anticipate performance will…" should be "we anticipate that performance will…".	Added optional that.
9. Figure. 1. It seems that the font size of the description of Control gRNAs is inconsistent with others, please check.	The entire document has been reformatted to 12-point font.
10. Line 22,55,98,159,175,187,219 and 247, "Guidemaker" should be "GuideMaker".	Thanks, the format is now consistent.
11. Line 262, "CAS" should be "Cas".	Corrected
12. Supplementary Figure 4. Grammar mistake in sentence "the different number of logical cores with or without AVX2 settings are available". It should be "the different number of logical cores with or without AVX2 settings is available".	This has been rewritten for clarity.

Reviewer 3

Overall, the tool is very well documented and easy to use. In the current version of the manuscript, GuideMaker does not show a clear improvement over the state-of-the-art design tool, CHOPCHOP. The authors do not implement any existing on-target scoring methods to determine the targeting efficacy of the picked sgRNAs. This can lead to picking guides that are highly specific but not effective enough.	We have improved our explanation of the advantages of GuideMaker relative to ChopChop for it's intended applications, including a performance evaluation in Supp. Fig. 5 and a better explanation. We have also added both on-target and off-target scoring for NGG PAMs (the only PAMs for which training data is available). Based on the model from Doench et al. 2016.
1. Implementing on-target scoring methods, at least for the Cas enzymes that have on-target efficacy information, can help improve the process of picking sgRNAs. This tool will probably be used more often with standard Cas enzymes and it will be useful to have on-target efficacy scores attached to the guide RNAs.	Good suggestion, we have implemented the Azimuth model for on-target scoring from Doench et al. 2016, specifically their "V3_nopos" model. We have also refactored the original feature calling to improve speed, updated code to Python 3.9 and transferred their original model in pickle format to a safer, reproducible, cross platform compatible model in the Onnx runtime. We have also added the off target CFD scoring from the same paper.
2. The authors do a thorough analysis of the computational performance of GuideMaker with various genomes and Cas enzymes but including a comparison of the computational performance of GuideMaker vs. CHOPCHOP will strengthen the manuscript.	We have added this comparison, in Supp. Fig 5.

3. The authors define the PAM sequence of SaCas9 to be NGRRT whereas the canonical PAM sequence of SaCas9 is NNGRRT. This should be modified throughout the manuscript and analyses involving SaCas9 should be redone	We have fixed this issue.
A good addition to the tool would be to output a file with all the sequences that were designed targeting the region of interest with the specific PAM sequence. This gives the user a sense of the universe from which the final guides were picked.	The user can get this by filtering the current output file by the locus name.
5. Another useful input parameter would be to specify a target region that the user wants to focus on such as letting the user input genomic coordinates or a gene name or locus tag. For example, CRISPy by Blin et al., 2016 takes a GenBank file as input and allows the user to input features specific to the uploaded genome. Minor Points	We have added the "--filter_by_locus" option to filter results for this application.
1. "CyVerse" is misspelled as "CyCVerse" in multiple places in the manuscript.	We have fixed this.
2. Reference Figure 2 in Line 92.	Added.
3. Line 154: "Ratios between tools were calculated by dividing the number of gRNA identified.."	The sentence was rewritten for clarity
4. In Supplementary Figure 3 "wit haVX2" should be "with aVX2".	Corrected
5. GitHub link in Line 336 does not work.	Those links are fixed
6. Line 225-226: "GuideMaker also creates off-target gRNAs for use as negative controls in high-throughput experiments." "Off-target gRNAs" is misleading in this context.	We now refer to them as "off-target control RNA sequences" since they are not guides.