Author's Response To Reviewer Comments

Close

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 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 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 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.

Close