Author's Response To Reviewer Comments

Dear Dr. Edmunds,

We have completed minor revisions to the manuscript requested by the reviewers.

Specifically, we have:

Included the bio.tools identifier in the manuscript
Included the Scicrunch.org identifier in the manuscript

We have addressed these remaining comments:

"Maybe you intended to remove mentions of Cas13, but in the current version it still stands out. Page 3/line 66."

Sorry, that last reference to Cas13 has now been removed.

"It is hard for me to believe that edit distance search for off-targets is equal to the hamming distance. This might be true small bacterial genomes, but for larger genomes (eg. human/mouse) this probably can't hold. It could also be that your implementation of the edit distance calculation for the guides could be flawed and therefore not reflecting the actuality. Co adding tests for that "leven" option."

We have addressed this in two ways:

We added a unit test (test_levin_dist) to the test code verifying that both Levenshtein and Hamming distance are being ca as expected. This test code can be found here https://github.com/USDA-ARS-
GBRU/GuideMaker/blob/main/tests/test_core.py#L319-L347

In that unit test we created a test sequence:

CGTAGCTAGTCACTAGCTGACAGCAAGGTTTTTCGTAGCTAGACACTAGCTGACAGCAAGGTTTTTCGTAGCTAGTCACTAGCTGACTA

That test sequence had three guide areas embedded in it (changes are shown with brackets and underscores):
1. CGTAGCTAG[T]CACTAGCTGACA_GCA|AGG
2. CGTAGCTAG[A]CACTAGCTGACA_GCA|AGG
3. CGTAGCTAG[T]CACTAGCTGACTAGCA|AGG

Guide 2 has 1 substitution (in brackets) and guide 3 has 1 insertion (underscore) relative to guide 1.

The Levenshtein distances for sequence 1 vs. [2, 3] are [1, 2], while the Hamming distances for sequence 1 vs. [2, 3] are

The test code verifies that these edit distances are calculated correctly by the functions in Guidemaker. These edit distanc calculations come directly from the highly-used NMSLIB library.

To address the concern that the guides designed with Leven and Hamming distance would diverge more for longer genom tested the effect of using Levin and Hamming on the 537 MB genome of Phaseolus vulgaris (NC_023759). That data has b added to Supplementary Table 4.

Indeed, fewer guides were identical when Levin distance was used for the longer genomes, but the guides designed with L Hamming were still 98% similar (versus 99.9% similar for E coli. MG 1655). For the larger Phaseolus vulgaris genome usin Distance with the "NGG" PAM took about twice as long, while. for E coli it took about 15x as long. This is likely because ind

not distance computation, makes up a larger part of the compute time for larger genomes.

We agree that Levin distance the more biologically relevant measure of efficiency but think that for most users designing guides per gene and working on smaller genomes the data supports the conclusion that Hamming is an appropriate distan approximation.

In the last revision we added Levin distance an an option for users who need it. We discuss the results in lines 233-242.

We have also added Supplementary Table 2 which summarizes the runtime to compute all guides for the PAMs "NGG", "NI and "NNAGAAW" in the Homo sapiens (GRCh38.p13) genome. We added this benchmark for the large community of huma researchers.

We have made additional improvements to the bibliography and abbreviation sections.

Sincerely,

Adam Rivers

Close