17-Jan-2022  
  
NARGAB-2021-216  
ABRIDGE: An ultra-compression software for SAM alignment files  
  
Dear Mr Banerjee,  
  
Thank you for giving us the opportunity to consider your manuscript.  
  
Neither reviewer can recommend that it be published in NAR Genomics and Bioinformatics. Their comments are below and are for your information should you wish to consider resubmission elsewhere.  
  
The reviewers' reports are confidential and should not be published without the express permission of the Editors.  
  
Once again thank you for considering NAR Genomics and Bioinformatics for the publication of your manuscript.  
  
Sincerely,  
##########################################  
Dr Cedric Notredame, PhD  
Group Leader  
Notredame's lab - Comparative Bioinformatics Group Bioinformatics and Genomics Programme Room 440.03  
  
Centre de Regulació Genòmica (CRG)  
Dr. Aiguader, 88  
08003 Barcelona  
Spain  
  
Ph#     + 34 93 316 02 71  
Fax#    + 34 93 316 00 99  
Mobile# + 34 66 250 47 82  
  
email   [cedric.notredame@crg.eu](mailto:cedric.notredame@crg.eu)  
url     [www.tcoffee.org](http://www.tcoffee.org/)  
blog    [cedricnotredame.blogspot.com](http://cedricnotredame.blogspot.com/)  
ORC-ID: 0000-0003-1461-0988  
###########################################  
  
  
Reviewer: 1  
  
Comments to the Author  
With the advances in sequence technology, more data are generated. Consequently, it brings a problem for data storage. To save the storage and the budget to buy the disk, an efficient data compression method is needed. Here, Banerjee and Andorf introduced a new tool ABRIDGE to help compress and decompress SAM files in a lossless or lossy fashion. Overall, the study is interesting. However, after reading the manuscript and testing the tool, I have some concerns which the authors may need to address.  
  
Major:  
1. Although the authors claim that ABRIDGE can handle SAM files from DNA-seq and RNA-seq alignments, it is weird that the authors only used STAR to perform the alignment. STAR was specifically designed for RNA-seq alignment, it might not be accurate in DNA-seq alignment. I noticed that ABRIDGE requires a SAM file with an ‘NH’ tag. However, for the most used DNA-seq aligners, such as BWA and Bowtie2, they do not produce such a tag, which means the adoption of this tool would be limited.  
2. Based on my Question 1, the authors may need to change the title to avoid misleading.  
3. In ABRIDGE, the authors offer different options to use ABRIDGE in either a lossless or a lossy manner. However, the authors did not state clearly in what kind of situation different options can be selected. In the updated version, the authors may want to give a guide to do so.  
4. According to the specs (<https://samtools.github.io/hts-specs/SAMv1.pdf>), in CIGAR, it seems that sequence mismatches can be documented. Not sure why the authors claim that CIGAR is not designed to store mismatched nucleotides.  
5. The authors state that ABRIDGE would help small labs to save storage spaces with a limited budget to buy disks. However, when using ABRIDGE, extra RAM and CPU are needed, would this be paradoxical? As the expense of RAM and CPUs are also high.  
6. During random access, the file needs to be entirely decompressed. Extra storage and time are also needed. I guess this would be one of the concerns for the adoption of this tool, as such as BAM/CRAM formats are good enough to quickly do so without extra storage needed.  
7. The authors claim that ‘ABRIDGE produced a file which is 164 Mb smaller than the next best compressor’. So, for how large a file, ABRIDGE can have such an improvement? Is this improvement for all files or it has some association with the number of reads, types of reads and divergence of the alignment?  
8. In what kind of system did the authors test the tool? Were there some beta tests before the release? I downloaded the tool and had difficulties in running it which prevented me from assessing the performance. I guess the README file in the GitHub repo needs to document how to set the tool and why is docker or singularity need (I have singularity installed on my platform)?  
9. In some cases, a BAM file is directly generated from software to save space. However, ABRIDGE needs to convert BAM to SAM first and then perform SAM compression. Would it be possible to avoid such a step to directly compress a BAM file to save the time and storage used in getting a SAM file?  
  
Minor:  
1. L23, Column 1, Page 1: ‘Most genomic software utilizes read alignments for several purposes’ -- This is not accurate, may reword  
2. L20, Column 2, Page 1: ‘several purposes – assembling, annotating’ -- May change to ‘several purposes, such as assembly, annotation’  
3. L22, Column 2, Page 1: ‘Most bioinformatics projects utilize …’ -- It depends on what kind of species the researchers working on. 'Most' is not an appropriate word here. May reword  
4. L25, Column2, Page 1: ‘The primary step’ -- May change to ‘The routine step’  
5. L36, Column 2, Page 1: ‘need to be mapped’ -- May change to ‘are usually’    
6. As there are different options to select the compression level, the authors may need to make this clear when talking about ABRIDGE. For instance, at L33, Column 1, Page 2, the authors say, ‘ABRIDGE modifies the traditional CIGAR’. Is this for all compression or only for lossy compression? If this is for lossy compression, would the modification be for all conditions or some of them?  
7. L15, Column2, Page 2: ‘ABRIDGE compresses SAM files in two passes – in the first pass, relevant information …’ -- Make this clear that what the relevant information is.  
8. L17, Column 2, Page 2: ‘using generic compressor’ -- Please clearly list the compressors used in ABRIDGE  
9. L23-24, Column 2, Page 2: ‘ABRIDGE achieves a high compression ratio … redundant data’ -- Is this for all modules in ABRIDGE or some of it? Please make this clear  
10. L25-26, Column 1, Page 3: ‘an index file … in the future’ -- Is this for current usage or only for future?  
11. L34, Column 1, Page 3: ‘compressed file in .abridge format’ -- It seems the compression relies on third-party compressors and the actual format is not '.abridge'. It's a rename of the original compression format, right?  
12. L38, Column 1, Page 3: ‘The decompression step might … during compression’ -- May clearly state in what kind of situation a dummy quality score is used and this would affect the accuracy of some downstream analyses, such as variant calling.  
13. L5-7, Column 2, Page 6: ‘Although the decompression … to the reference’ -- Not sure what kind of information the authors want to deliver here.  
14. L16, Column 2, Page 6: ‘without depressing the entire file’ -- Previously, the authors mentioned that during a random search, the entire file needs to be decompressed, but here they stated that there is no need to decompress, which confuses me. Please check.  
15. Figure 1: There are two integrated CIGARs from ‘Construct the final Integrated CIGAR’ to ‘Exact same mapping of adjoining sequence with different SAM format Flag’. What’s the difference between the two in each section?  
16. Figure 2: Does the comparison in the same level, for instance, was the file size calculated after a lossless compression or a lossy compression? If the compression is lossy, did they discard the same information? I guess all relevant figures need to state this clearly.  
17. Figure 3: From this figure, it seems ABRIDGE can only produce a modified SAM file and a subset of SAM files, is this true? Can the users get the original SAM file after decompression?  
18. GitHub README ‘samtools calmd -bAr aln.bam > aln\_baq.bam’ -- a reference file is missed here, right?  
19. For the usage of ‘abridge’, the ‘-aq’ option says ‘Adjust quality scores for matched bases to achieve better encoding. For more details please check ...’. Please indicate what to ‘check’ here.  
  
Reviewer: 2  
  
Comments to the Author  
The authors present a novel compressor for the information stored in a SAM file. The manuscript is well-written, but I do not think it adds valuable contributions to the field.  
- My main concern is that the algorithm itself does not add new ideas, as most of the presented methodological steps have already been applied in other algorithms, except maybe for the extended CIGAR idea. Note however that this steps assumes the availability of the MD field, which is not mandatory and hence most of the available SAM files may not contain it. This should be clearly stated, as it is an important drawback of the algorithm. Moreover, the fields NH and XS are also not compulsory, and expected by ABRIDGE.  
- The performance assessment only uses datasets of one species, and all datasets are very small. The method should be tested in large human datasets.  
- The choice of methods for comparison is very limited, note that NGC is from 2013, and DeeZ from 2014.  
- Following the previous point, important references to recent work are missing, such as CALQ, QVZ, GeneComp, SPRING... Note that although Spring compresses FASTQ files instead of SAM, it has been shown to compare favorable to methods that compress SAM files.  
- Following previous point, a throughful comparison should include comparison on compression performance for only reads, only QS, only identifiers, only additional information. This way one can really assess where the gain from ABRIDGE comes. currently this info is provided only for ABRIDGE.  
- Supplementary material contains only figures and tables, but the main document says that details about data acquisition are specified in the supplementary.  
- Regarding the datasets, the link used for download should be provided. I was not able to find the exact files online.  
  
Associate Editor: Himmelbauer, Heinz  
Comments to the Author:  
(There are no comments.)