Dear Editor,

We are writing this letter to request that our manuscript entitled “**ABRIDGE: An ultra-compression software for SAM alignment files**” by Sagnik Banerjee et al. be reviewed and considered for publication in the journal Nucleic Acids Research Genomics and Bioinformatics.

Technological advancements in the field of sequencing have made genome-wide studies affordable. Reduced costs and increased access to sequencing facilities have led to a surge in genetic data. Sequencing RNA and/or DNA has become crucial to answering many high-impact problems ranging from treating emerging diseases to producing resilient crops. For example, detecting genes participating in immune response towards novel microbial infections. Host-pathogen interactions result in the birth of new strains with novel infection mechanism. This emergence of new strains necessitates the understanding of novel gene structures and decipher the host genes that respond to the new variant which can be achieved by high depth sequencing. The accuracy of sequencing technologies to pin-point a locus of interest down to a single nucleotide precision has gained much appreciation and will hence continue as an established platform to study a wide variety of biosystems. Sequencing technologies produce readouts in the form of short sequences – ranging from 50 to 300 nucleotides. One of the very first steps is to align these data to a selected reference using software like STAR or HiSAT2. These software report the alignments in SAM format that contains important information about the location and the nature of alignment. Newer sequencing platforms produce billions of reads that can help study lowly expressed genes but also raises space demands after alignment. Current approaches to compress SAM file focus more on offering quicker random access than on providing a high compression ratio. Also, most software seldom allows users the choice to discard specific quality scores. Therefore, software is needed which can compress SAM files and enable users a multitude of options for lossy compression.

We present ABRIDGE, a state-of-the-art compressor for SAM files that achieves ultra-high compression while offering a wide variety of options for lossy compression. ABRIDGE aggregates reads mapping to the same location on the reference and retains only those nucleotides that deviate from the reference sequence resulting in a much smaller file. The software also employs techniques to store only the most relevant quality scores thereby further reducing file sizes. We demonstrate the superiority of ABRIDGE by comparing it with other compressors. Our results prove that ABRIDGE can compress SAM alignment files achieving greater compression (more than 30% in some cases) than other contemporary software. The USDA-ARS model organism databases MaizeGDB will provide long-term support to maintain both the source code and data sets from this study.

We hope you share our excitement for this study and find it suitable for review by the editors of Nucleic Acids Research Genomics and Bioinformatics. We strongly believe this is an important paper that will be utilized by the community to reduce space demands and facilitate data transfer.

Sincerely,

Sagnik Banerjee

Ph.D. Candidate,

Iowa State University,

437 Bessey Hall,

Ames, IA 50011

Phone: (725) 333-0696

Email: sagnikbanerjee15@gmail.com

Carson Andorf

USDA-ARS, Computational Biologist

1027 Crop Genome Informatics Laboratory

Ames, IA 50011

Phone: (515) 294-2019

Email: carson.andorf@usda.gov