Suggested titles

ABRIDGE: a novel efficient software to compress alignment files

Sagnik Banerjee1,2\*, Carson M. Andorf1,3,4\*

1 Program in Bioinformatics and Computational Biology, Iowa State University, Ames, IA, 50011, USA

2 Department of Statistics, Iowa State University, Ames, IA, 50011, USA

3 Corn Insects and Crop Genetics Research Unit, USDA-Agricultural Research Service, Ames, IA, 50011, USA

4 Department of Computer Science, Iowa State University, Ames, IA, 50011, USA

Correspondence should be addressed to either [sagnik@iastate.edu](mailto:sagnik@iastate.edu) or [carson.andorf@usda.gov](mailto:carson.andorf@usda.gov)

**Abstract**

**Keywords:** SAM compression, quality score compression

**Introduction**

Para 1

Next generation sequencing (NGS) has opened up opportunities to study several biosystems from a quantitative viewpoint (Hunt et al., 2019; Elmore et al., 2020). Over the years, numerous sequencing protocols have been designed to probe the modus operandi of number of biological processes (Buenrostro et al., 2015). Researchers have perfected these protocols - making it more economical and effective. This made sequencing accessible to even under funded labs leading to surge in data. Short read data (generated typically on Illumina platforms) is often mapped to a reference (genomic/transcriptomic) and then used for several purposes – assembling (Kovaka et al., 2019; Song et al., 2019), annotating (Haas et al., 2003; Holt and Yandell, 2011; Bruna et al., 2020; Banerjee et al., 2021), finding differentially expressed genes (Robinson et al., 2010; Love et al., 2014) and in proteomics (Banerjee et al., 2015c, 2015a, 2015d, 2015b, 2016a, 2016b, 2020; Velásquez-Zapata et al., 2021). Most bioinformatics projects utilize a very large set of RNA-Seq or DNA-Seq samples collected from multiple tissue types and conditions. The primary step in such experiments is to align the RNA-Seq samples to a reference that generates a file in BAM (Li et al., 2009) or the CRAM (Fritz et al., 2011) format. Even though these formats offer compression to some extent, the total size of all the aligned files can often exceed the storage capacity of small labs. Hence, better compression techniques are needed that utilize the underlying structure of reference alignment files and offer a multitude of options to cater to a diverse range of user requirements.

Para 2

Short reads, generated by sequencing platforms like Illumina, need to be mapped to a reference using aligners like STAR (Dobin et al., 2013) or BWA (Abuín et al., 2015) before further processing. These aligners typically output the result in a SAM (Li et al., 2009) format which can be converted to a binary BAM format to achieve a better compression. SAM format stores the location, shape (CIGAR string) (https://genome.sph.umich.edu/wiki/SAM), nucleotide bases, quality scores and tag level information for each aligned read. Since alignments in SAM format are stored for each read, the file size grows linearly with the number of reads in the sample. Hence, there is a need to devise an algorithm that can exploit the underlying structure of SAM files and offer the best possible compression in a reasonable amount of time.

Para 3

A considerable amount of time and effort has been directed to designing algorithms to compress alignment files to reduce storage demands and facilitate file transfers (Hosseini et al., 2016; Numanagić et al., 2016). Most approaches achieve compression by eliminating redundant data by accumulating alignment information across multiple reads or alignments. SAM compressors, like NGC (Popitsch and von Haeseler, 2013), DeeZ (Hach et al., 2014) and genozip (Lan et al., 2021) are reference based while BAM, CRAM, Quip (Jones et al., 2012) and CSAM (Cánovas et al., 2016) are reference free. Reference based approaches achieve compression by representing an aligned read with a description of how it differs from the reference. This eliminates the need to store the actual read sequence thereby reducing storage demands. Quality scores do not map to any reference and hence cannot be compressed like the read string. Hence some compressors like NGC, CSAM, genozip and DeeZ offers users the option to map quality values within a range to a single value. While this can lead to a better compression, it might remove quality scores of mismatched bases which are essential for SNP detection. Quip implements Markov chains to encode read sequences and quality scores. Samcomp (Bonfield and Mahoney, 2013) compresses SAM alignments in lossless fashion by tokenizing the read identifiers and scoring the reads as a reference difference model. A very similar approach is undertaken by DeeZ where tokenized read names and read sequence is compressed with delta encoding. Boiler (Pritt and Langmead, 2016) stores data by converting the alignment space to a coverage space but ignores sequence and quality information. Table @Summary\_Of\_Compression\_SoftwareMT provides an outline of the various state-of-the art software for compressing aligned data.

Para 4

To overcome the shortcoming of previous SAM compression approaches we introduce ABRIDGE. We offer users a plethora of choices to compress SAM files. To optimize space utilization, ABRIDGE accumulates all reads that are mapped on to the same nucleotide on a reference. ABRIDGE modifies the traditional CIGAR string to store soft-clips, mismatches, insertions, deletions, and quality scores thereby removing the need to store the MD string. To further reduce space demand, ABRIDGE modifies the CIGAR information to store the strand on which the read was mapped. All features of multi-mapped reads are stored with their individual CIGAR strings. Hence reads mapping to homeologs in polyploid species will retain their alignment profile. Users can choose from three levels of compression offering varying extent of compression with the caveat of duration of compressing. ABRIDGE offers options of completely lossless compression and selectively lossy conversions. Consequently, decompressions in ABRIDGE can regenerate the entire SAM file with or without modifications depending on the choices made during compression. In this manuscript we explore the different modes in which ABRIDGE can operate and compare it with other state-of-the-art tools.

**Implementation**

Para 1

ABRIDGE accepts a single SAM file as input and returns a compressed file that occupies much less space than its BAM (or CRAM) counterpart. Users can choose to retain all the quality scores which would necessitate a lossless compression. In several applications, storing the entire quality score is often redundant. Hence, ABRIDGE can be requested to preserve only those quality values which for which the corresponding nucleotide base was a mismatch to the reference or was an insertion in the read sequence. This option considerably reduces the compressed size but stores the most relevant information which can later be used for analysis that use quality scores like variant calling. To further reduce space users can eliminate quality scores altogether. Some downstream software like transcriptome aligners do not use soft-clips or mismatches, so we designed ABRIDGE to provide options to ignore such information in the SAM file while compressing. ABRIDGE compresses SAM files in two passes – in the first pass, relevant information from the SAM file is rearranged and in the second pass, the file is compressed using generic compressors. ABRIDGE decompresses data by applying the reverse algorithm producing all the information that was requested to be stored during compression. Once the data is compressed, users have the option of retrieving alignment information from random locations making is very easy to access alignments anywhere in the genome without having to decompress the entire data. Finally, ABRIDGE also offers users the option to generate coverages directly from the compressed file.

Compressing data using abridge

ABRIDGE achieves a high compression ratio owing to the underlying strategies of eliminating redundant data. Instead of storing the entire sequence of reads, ABRIDGE stores the location of the reference to which the read mapped and relevant information about the mismatched and/or inserted base pairs. ABRIDGE also merges exact same reads originating from a particular position of the reference. Read names for uniquely mapped single-ended reads are discarded but are preserved for paired-ended reads to associate each read with the corresponding fragment. ABRIDGE offers users a multitude of choices for storing quality values. Users can request to store all the quality values without making any changes or allow ABRIDGE to modify the quality scores of some bases to facilitate better and faster compression. Instead of blindly modifying the quality scores, ABRIDGE inspects each base pair and modifies its quality value only if the base pair was aligned perfectly to the reference. Hence, the quality scores of bases which are inserts and/or mismatches are preserved. This provides the users with the opportunity to retain all the relevant information that is necessary to perform vital downstream analysis like variant calling. ABRIDGE stores a modified version of the CIGAR string by including soft clipped bases, quality scores of mismatched and inserted bases along with nucleotides that did not match with the reference. Users are also provided with the choice of achieving best compression by eliminating quality scores altogether. This option is helpful for storing alignments files for the purposing for performing transcriptome assemblies where quality scores are not typically used (Song et al., 2019).

Unlike the read sequence, quality scores cannot be “mapped” to any reference. Hence ABRDIGE stores quality values as reported and then compresses those with generic compressors. ABRIDGE can store quality values in 4 different ways – (1) discard quality values of reference matched bases and include only the mismatched and inserted bases. For this case, quality values are stored within the enhanced CIGAR, (2) Store all quality scores with altered values for reference matched bases, (3) Store all quality values without making any change in the quality values, and (4) Discard quality scores altogether .

Para 3

ABRIDGE will generate the compressed file in `.abridge` format which essentially compresses several files using one of the compressors as discussed in Table XXX. During decompression, a SAM is produced from the compressed files. The decompression step might require substituting dummy quality scores for some cases, depending on how the quality scores were stored during compression. The decompressed file will be sorted, and dummy read names will be produced where they were discarded to save space. Some applications, like genome-guided assembling, do not require the nucleotide sequence. Hence, ABRIDGE allows the user to decompress without generating the actual read sequence. This option is much faster to execute since it does not require the reference to be loaded and read from.

Para 4

Para 5

Para 6

Finally, users have the option to generate nucleotide coverage from the compressed files. ABRIDGE will partially decompress the files and generate coverage information in BED format. Nucleotide coverage can be generated in two modes - (1) Overlapping mode - reports the total number of reads that span a certain nucleotide, and (2) Non-overlapping mode - reports the number of reads that start at a particular nucleotide. The results produced are exactly the same as bedtools (Quinlan, 2014).

**Results and Discussion**

Selecting data for analysis

A total of 6 RNA-Seq, 6 DNA-Seq, 3 ATAC-Seq and 3 Bisulfite-Seq samples were chosen from NCBI SRA for testing and comparing all the compression software. All the samples chosen were paired-ended. Single-ended samples were generated by merging the two mate pairs together. Almost all the samples were 150 bp (except the 3 Bisulfite-Seq which were 125 bp). We chose samples from different sequencing assays to demonstrate the superiority of ABRIDGE over other compression software across the entire spectrum.

Alignment to reference

We used STAR to align the short reads with a threshold of 75% of the length of reads mapping to the reference. Even-though STAR is designed to align RNA-Seq reads we modified the settings to enforce STAR to map DNA-Seq reads without any splices. For more details please see .

Fidelity issue with downstream analysis

Abridge can decompress data faster than other softwares

Retrieve data randomly

Obtaining coverage data from Abridge files

Wheat homeologs

**Conclusion**

Future works: Extend to other types of files like BED, VCF etc.

Reference:

Abuín, J. M., Pichel, J. C., Pena, T. F., and Amigo, J. (2015). BigBWA: approaching the Burrows–Wheeler aligner to Big Data technologies. *Bioinformatics*, btv506.

Banerjee, S., Basu, S., Ghosh, D., and Nasipuri, M. (2015a). PhospredRF: Prediction of protein phosphorylation sites using a consensus of random forest classifiers. in *2015 International Conference and Workshop on Computing and Communication, IEMCON 2015* (Institute of Electrical and Electronics Engineers Inc.). doi:10.1109/IEMCON.2015.7344514.

Banerjee, S., Bhandary, P., Woodhouse, M., Sen, T. Z., Wise, R. P., and Andorf, C. M. (2021). FINDER: An automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinformatics*, 2021.02.04.429837. doi:10.1186/s12859-021-04120-9.

Banerjee, S., Ghosh, D., Basu, S., and Nasipuri, M. (2016a). *JUPred\_MLP: Prediction of phosphorylation sites using a consensus of MLP classifiers*. doi:10.1007/978-81-322-2695-6\_4.

Banerjee, S., Ghosh, D., Basu, S., and Nasipuri, M. (2016b). JUPred\_SVM : Prediction of Phosphorylation Sites using a consensus of SVM classifiers. in *Proceedings of Fifth International Conference on Soft Computing for Problem Solving* (Springer), 1–8. doi:10.1007/978-981-10-0448-3\_45.

Banerjee, S., Guha, S., Dutta, A., and Dutta, S. (2015b). Improvement of protein disorder prediction by brainstorming consensus. in *Computing and Communication (IEMCON), 2015 International Conference and Workshop on* (IEEE), 1–7. doi:10.1109/IEMCON.2015.7344428.

Banerjee, S., Mitra, B., Chatterjee, A., Santra, A., and Chatterjee, B. (2015c). Identification of relevant physico chemical properties of amino acids with respect to protein glycosylation prediction. in *Computing and Communication (IEMCON), 2015 International Conference and Workshop on* (IEEE), 1–7. doi:10.1109/IEMCON.2015.7344520.

Banerjee, S., Nag, S., Tapadar, S., Ghosh, S., Guha, S., and Bakshi, S. (2015d). Improving protein protein interaction prediction by choosing appropriate physiochemical properties of amino acids. in *Computing and Communication (IEMCON), 2015 International Conference and Workshop on* (IEEE), 1–8. doi:10.1109/IEMCON.2015.7344458.

Banerjee, S., Velásquez-Zapata, V., Fuerst, G., Elmore, J. M., and Wise, R. P. (2020). NGPINT: a next-generation protein–protein interaction software. *Brief. Bioinform.* 2020, 1–14. doi:10.1093/bib/bbaa351.

Bonfield, J. K., and Mahoney, M. V (2013). Compression of FASTQ and SAM format sequencing data. *PLoS One* 8, e59190.

Bruna, T., Hoff, K., Stanke, M., Lomsadze, A., and Borodovsky, M. (2020). BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. *bioRxiv*.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC‐seq: a method for assaying chromatin accessibility genome‐wide. *Curr. Protoc. Mol. Biol.* 109, 21–29.

Campagne, F., Dorff, K. C., Chambwe, N., Robinson, J. T., and Mesirov, J. P. (2013). Compression of structured high-throughput sequencing data. *PLoS One* 8, e79871.

Cánovas, R., Moffat, A., and Turpin, A. (2016). Csam: Compressed sam format. *Bioinformatics* 32, 3709–3716.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.

Elmore, M. G., Banerjee, S., Pedley, K. F., Ruck, A., and Whitham, S. A. (2020). De novo transcriptome of Phakopsora pachyrhizi uncovers putative effector repertoire during infection. *Physiol. Mol. Plant Pathol.* 110. doi:10.1016/j.pmpp.2020.101464.

Fritz, M. H.-Y., Leinonen, R., Cochrane, G., and Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 21, 734–740.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666.

Hach, F., Numanagic, I., and Sahinalp, S. C. (2014). DeeZ: reference-based compression by local assembly. *Nat. Methods* 11, 1082–1084.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491. doi:10.1186/1471-2105-12-491.

Hosseini, M., Pratas, D., and Pinho, A. J. (2016). A survey on data compression methods for biological sequences. *Information* 7, 56.

https://genome.sph.umich.edu/wiki/SAM.

Hunt, M., Banerjee, S., Surana, P., Liu, M., Fuerst, G., Mathioni, S., et al. (2019). Small RNA discovery in the interaction between barley and the powdery mildew pathogen. *BMC Genomics* 20, 610.

Jones, D. C., Ruzzo, W. L., Peng, X., and Katze, M. G. (2012). Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *Nucleic Acids Res.* 40, e171–e171.

Kovaka, S., Zimin, A. V, Pertea, G. M., Razaghi, R., Salzberg, S. L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 20, 1–13.

Lan, D., Tobler, R., Souilmi, Y., and Llamas, B. (2021). Genozip-A Universal Extensible Genomic Data Compressor. *Bioinformatics*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Love, M., Anders, S., and Huber, W. (2014). Differential analysis of count data–the DESeq2 package. *Genome Biol.* 15, 550.

Numanagić, I., Bonfield, J. K., Hach, F., Voges, J., Ostermann, J., Alberti, C., et al. (2016). Comparison of high-throughput sequencing data compression tools. *Nat. Methods* 13, 1005–1008.

Popitsch, N., and von Haeseler, A. (2013). NGC: lossless and lossy compression of aligned high-throughput sequencing data. *Nucleic Acids Res.* 41, e27–e27.

Pritt, J., and Langmead, B. (2016). Boiler: lossy compression of RNA-seq alignments using coverage vectors. *Nucleic Acids Res.* 44, e133–e133.

Quinlan, A. R. (2014). BEDTools: the Swiss‐army tool for genome feature analysis. *Curr. Protoc. Bioinforma.* 47, 11–12.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140.

Song, L., Sabunciyan, S., Yang, G., and Florea, L. (2019). A multi-sample approach increases the accuracy of transcript assembly. *Nat. Commun.* 10, 5000. doi:10.1038/s41467-019-12990-0.

Velásquez-Zapata, V., Elmore, J. M., Banerjee, S., Dorman, K. S., and Wise, R. P. (2021). Next-generation yeast-two-hybrid analysis with Y2H-SCORES identifies novel interactors of the MLA immune receptor. *PLoS Comput. Biol. (In Press.* doi:10.1371/journal.pcbi.1008890.