How the other compression softwares were executed?

How was the NCBI-SRA data downloaded and mapped? Provide full commands

Best is to create a separate GitHub repo/Code Ocean to reproduce the entire workflow Things to mention:

How was STAR used for mapping?

Code to select only those mappings that have no soft-clips, mismatches, indels, etc.

How single ended data was generated. How time and memory was calculated

\section{Supplemental Materials}

\subsection{SAM file format}

We illustrate the various fields of SAM format below. The first 11 fields of the SAM format is mandatory while the rest are optional.

\subsubsection{Required fields}

1. \textbf{QNAME} (Type: String) - Name of the query template \\

2. \textbf{FLAG} (Type: Integer) - A format flag to denote the type of alignment\\

3. \textbf{RNAME} (Type: String) - The name of the reference\\

4. \textbf{POS} (Type: Integer) - The position where the short read is mapped\\

5. \textbf{MAPQ} (Type: Integer) - A score to denote the quality of alignment\\

6. \textbf{CIGAR} (Type: String) - Stands for Compressed Idiosyncratic Gapped Alignment Report. A string that outlines the matches, inserts and deletions is the alignment\\

7. \textbf{RNEXT} (Type: String) - Reference name to which the read in the other pair has mapped\\

8. \textbf{PNEXT} (Type: Integer) - Position to which the read in the other pair has mapped\\

9. \textbf{TLEN} (Type: Integer) - Length of the template\\

10. \textbf{SEQ} (Type: String) - The nucleotide sequence of the read\\

11. \textbf{QUAL} (Type: String) - The quality scores as reported by the sequencing machine\\

\subsubsection{Optional fields}

A number of optional fields can be included in a SAM file. Among them, the most important are NH, MD and XS. All three tags are required by ABRIDGE.

\subsection{Selection of samples for testing}

A total of five RNA-Seq and five DNA-Seq samples were chosen from NCBI SRA for testing and for comparing all the compression software. All the samples chosen were paired-ended and sequenced to 150 bp. Single-ended samples were generated by merging the two mate pairs together. We chose samples from different sequencing assays to demonstrate the superiority of ABRIDGE over other compression software across the entire spectrum. To demonstrate the linear increase of compressed size with increase in file size, samples were merged together to mimic deeply sequenced samples \textbf{Supplementary Table \ref{abridge\_ListOfNCBI-SRASamplesForExperimentST}}.

\subsection{Alignment to reference}

We used STAR (\cite{Dobin2013STAR:Aligner}) to align the short reads with a minimum mapping threshold of 75\%. 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 DNA-Seq reads mapping was carried out with the following parameters - \texttt{--scoreGap -100, --scoreGapNoncan -100, --scoreGapGCAG -100, --scoreGapATAC -100 and --alignSJoverhangMin 500}. The intentional high penalty for splice generation forced all the alignments to be unspliced. All the alignments of DNA-Seq samples were inspected to ensure that there were no spliced alignments present.

\subsection{ABRIDGE Compression}

ABRIDGE performs compression in two steps. During the first step, ABRIDGE restructures the data to discard any redundant information. Since the alignment file is always accepted in sorted format, ABRIDGE stores only the difference between consecutive mapped positions. For a deeply sequenced sample, this helps save a lot of space. Nucleotide sequences are entirely eliminated except for mismatches and indels. The sequences can be later reconstructed from the reference and the position stored in the compressed file. Unlike nucleotide sequence, quality scores cannot be "mapped" to any reference. Hence, all quality scores need to be stored if and when the user requests for it. This can lead to a rise in the space demand. Hence, we employed a different method to compress quality scores.

\subsubsection{Run Length encoding of quality scores}

To compress quality scores, we implement run length encoding. Instead of encoding quality scores of each read, we apply run-length encoding simultaneously for all quality scores in a particular nucleotide position of the read. If the user allows ABRIDGE to modify the quality scores of matched bases, then the software achieves even better compression.

\subsection{Commands to run other software}

In this section we discuss the commands used to generate compressed files by DEEZ, SAMCOMP, GENOZIP and CSAM. DEEZ was executed with lossy values of \texttt{0, 50 and 99} and both modes to encode quality. SAMCOMP was executed only with the sorted alignment file. Compression using GENOZIP was performed by both modes of compression and was launched with the `optimize` parameter. CSAM was also launched in both lossy and lossless modes.

Future Work

ABRIDGE will be updated in the future to compress other types of files that store biological information like BED, VCF, etc. To enhance compression, we will further explore other techniques to compress quality scores. Currently, ABRIDGE stores the read names for paired ended reads and also for multi-mapped single ended reads. We will modify our algorithm to retain all relevant information without having to store read names.