To Do list:

1. Write up a python script to detect SAM files that do not have any indication of multi-mapped reads. Typical multi-mapping TAGs for different aligners are listed below:

|  |  |  |
| --- | --- | --- |
| Name of the aligner | TAG name | Only multi-mapped or number of multimaps specified |
| STAR | NH | Yes |
| BWA | None | None |
| Bowtie1 | XM | Yes |
| Bowtie2 | None | None |
| Hisat2 | NH | Yes |
| MUMmer4 |  |  |

Use sort | uniq -c to find number of multi-mapped reads. The script will append NH tags and XS tags to each alignment.

1. Rewrite the C code to extract multi-mapping information from non-NH tags
2. Make a list of RNA-Seq and DNA-Seq datasets from several different organisms to test the software
3. Write up a code base in a way that the analysis is fully reproducible
4. BWA produces sam files with ‘\*’ in the read sequence – deal with it
5. Configure abridge to accept bam files as well
6. Bowtie1 XS filed has a different meaning. Attend to it.
7. Add checks to ensure MD field is present. If not then generate it.
8. Fidelity checks with salmon quant, repeat with most compressed version as well
9. Remove Brotli, 7z from the pipeline. Only zpaq should be there
10. Incorporate FCLQC to compress quality scores
11. Add more references to other compressors in the manuscript
12. Work on the coverage generation program
13. Think about doing away with random access
14. Mark out places in the manuscript where the language needs to be improved
15. Shift the focus of the manuscript towards lossy compression techniques