Toward a dynamic threshold for quality-score distortion in reference-based alignment

Ana A. Hernandez-Lopez, Claudio Alberti, and Marco Mattavelli

École Polytechnique Fédérale de Lausanne
EPFL SCI-STI-MM
1015 Lausanne, Switzerland
ana.hernandezlopez@epfl.ch

Abstract. The intrinsic high entropy metadata, known as quality scores, are largely the cause of the substantial size of sequence data files. Yet, there is no consensus on a viable reduction of the resolution of the quality score scale, arguably because of collateral side effects. In this paper we leverage on the penalty functions of HISAT2 aligner to rebin the quality score scale in such a way as to avoid any impact on sequence alignment, identifying alongside a distortion threshold. We tested our findings on whole-genome sequence and RNA sequence data, and contrasted the results with three methods for lossy distortion of the quality scores.

Keywords: quality scores · reference-based alignment · quality score distortion · HISAT2 · lossy compression.

1 Introduction

In the last few years the fast-paced advancements in sequencing technology have created new challenges in the domain of genomic information. As an unprecedented amount of data is being made available [1, 2], the problem is now inclining on storing sequence data as efficiently as possible [14].

It has been noted that metadata aimed at measuring the reliability of sequence data (quality scores) take up a large chunk of the overall compressed file size. This important observation has propelled a discussion on whether these values are indeed significant for omics applications, and whether or not keeping them entirely is necessary.

2 Methodology

We investigated the penalty functions that drive the alignment score system for read sequence alignment in a well-known quality-aware aligner. We then derived a simplification in the assignment of penalty values that reduces quality score scale granularity while keeping alignment scores unaffected. Consequently, this coarser quality score scale reduces storage footprint of sequence files with the advantage of entirely preserving read mapping percentages. In other words, we distorted quality scores without collateral impact on alignment.

The aligner in question is HISAT2 [3], the modern version of the popular aligner Bowtie2, suitable for mapping genome and exome sequence data. Compared to other quality-aware aligners like Novoalign [4], HISAT2's approach to alignment score computation is straight-forward and deterministic, making it a good candidate to explore the relation and effect of quality scores and sequence alignment. HISAT2 computes an alignment score for each read sequence by adding penalty scores. There are four possible penalties: mismatches, soft-clips, gaps, and ambiguous bases. However, only the scores for mismatches and soft-clips depend on quality scores, as per their penalty functions

$$MN + \lfloor (MX - MN)\ \frac{min(Q, 40)}{40} \rfloor \tag{1}$$

For mismatches: $MX = 6, MN = 2$; for soft-clips: $MX = 2, MN = 1$.

Solving the equation for both penalties and for all possible quality score values we can group the assignment of penalty values in five bins, as shown in Fig. 1.

ASCII	QS	Penalties (mismatches, softclips)	
[73]	[40]	-6	-2
[63, 72]	[30, 39]	-5	-1
[53, 62]	[20, 29]	-4	-1
[43, 52]	[10, 19]	-3	-1
[33, 42]	[0, 9]	-2	-1

Fig. 1. Rebinning of quality score scale.

With this rebinning we can compute distortion rate baselines that represent lossy compression rates that can "at least" be applied to the quality scores of raw sequence files without compromising alignment. These baselines can be thought of as distortion thresholds, which rely on sequence files. Fig. 2 shows the setup of our experiments. An input file with undistorted quality scores (**D**) is rebinned to produce an output file with distortion rate **d**. Both undistorted and rebinned files are aligned, and produce identical alignment reports. The distortion threshold for file **D** is **d**.

To observe the effect that quality score distortion plays on alignment we ran three lossy compression tools [8–10] and set their parameters such that the output files met as close as possible the value of the distortion threshold **d**. The approximate distortion rates for each compressor are ***dA***, ***dB*** and ***dC*** (refer to Fig. 2). The distorted files were then aligned with HISAT2 to quantify mapping results.

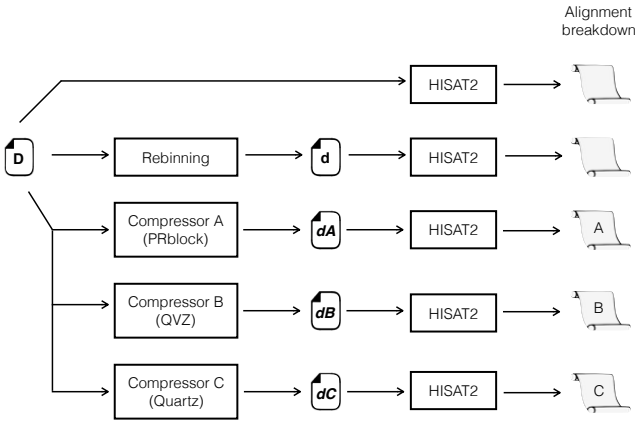

Fig. 2. Experimentation setup.

3 Results

We experimented with synthetic and natural data and are reporting results for two natural data samples: T16M Metastatic liver tumor (whole-genome sequence data) [5], and Gene expression data in skin fibroblast cells (rna-seq data) [6]. Results are reported in the tables in Fig. 3. The alignment report in the top tables is presented as the percentage of reads grouped in one of three possible sets: reads that aligned zero times (`Z`), reads that aligned exactly one time (`X`), and reads that aligned more that one time (`M`).

The bottom tables in Fig. 3 summarize alignment information as the percentage of reads whose alignment coordinate changed as a consequence of quality score distortion. We call this read relocation, and can happen between alignment sets or within alignment set `M` (see Fig. 4). For example, a read aligned before quality score distortion may be grouped in set `Z` but if that same reads is aligned after quality score distortion it may be grouped in set `X`. This type of read relocation is between sets and the percentage of reads relocated in this fashion is shown in the second column of the tables at the bottom of Fig. 3.

The second form of read relocation can occur within set `M`, when the quality scores of a read with multiple alignment locations are modified in a way such that the new alignment coordinate belongs to the set of its multiple candidate locations. The percentage of reads relocated within set `M` is shown in the third column of the bottom tables in Fig. 3. The percentages are relative to the full file (F) and to the set of multireads (M).

Note that this type of read relocation occurs even in the rebinned file. This happens when the set `M` contains reads whose set of alignment coordinates have the same alignment score. HISAT2 will select one of the candidate coordinates for each read by computing a pseudo-random number generated from the read name, the sequence string, the quality score string and an optional seed value.

wgs data

Distortion method	Parameters	Distortion rate [bits/QS]	Alignment set [% reads] Z	X	M
Rebinned	—	0.9619	49.4	40.1	10.5
PRblock	q=1, l=2	1.0281	49.3	39.9	10.8
QVZ	0.018	0.9571	49.2	39.8	11
Quartz	—	1.5583	49.7	41.1	9.2
Undistorted	—	2.3479	49.4	40.1	10.5

rna-seq data

Distortion method	Parameters	Distortion rate [bits/QS]	Alignment set [% reads] Z	X	M
Rebinned	—	0.4098	1.6	73.8	24.6
PRblock	q=2, l=20	0.4028	1.7	73.9	24.4
QVZ	0.013	0.3906	1.9	73.8	24.3
Quartz	—	0.5067	1.7	77.2	21.1
Undistorted	—	0.7715	1.6	73.8	24.6

Distortion method	Changed alignment set [% reads]	Changed alignment coordinate within set M [% reads] F	M
Rebinned	—	0.002	0.019
PRblock	0.011	0.006	0.056
QVZ	0.009	0.008	0.072
Quartz	0.026	0.002	0.021

Distortion method	Changed alignment set [% reads]	Changed alignment coordinate within set M [% reads] F	M
Rebinned	—	0.008	0.032
PRblock	0.011	0.011	0.045
QVZ	0.009	0.008	0.032
Quartz	0.026	0.020	0.094

Fig. 3. Distortion rate and alignment percentages for wgs and rna-seq samples.

Thus, modifying the quality scores will trigger HISAT2 intrinsic response toward multireads with equally likely alignment coordinates.

4 Discussion

Simplifying the representation of quality scores is arguably a natural choice in the face of the sequence data explosion, and computational methods that approach the problem introduce collateral errors that are difficult to quantify.

The assessment of quality score distortion has been attempted for some application domains [11–13] without clear consensus on the limits of "safe" lossy distortion levels. Meanwhile the increasing complexity of genomic assays, datasets and computational methods only adds to the difficulty of its potential quantification.

Nevertheless, even uniform requantization of the quality scores is a suitable approximation for high accuracy applications [7], and we have shown that this approach can be extended further to rebin coarsely quality scores without impact in sequence alignment.

In the light of the fast-paced sequencing technology progress, the utility of quality scores is at stake, as they are arguably unnecessary for many omics applications. We must therefore advocate for a feasible and pertinent granularity that suits each host application.

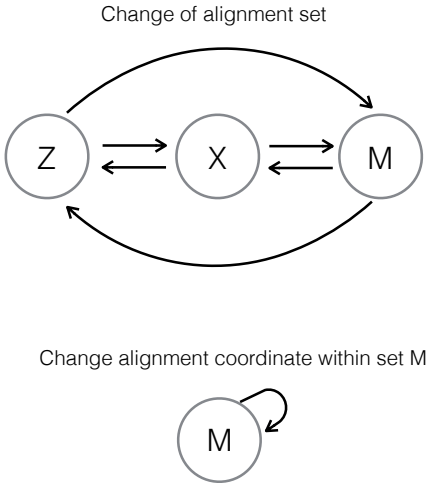

Fig. 4. Read relocation between sets (top), and within set `M` (bottom).

References

1. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLOS Biology 13(7): e1002195. https://doi.org/10.1371/journal.pbio.1002195
2. He KY, Ge D, He MM. Big Data Analytics for Genomic Medicine. Int J Mol Sci. 2017;18(2):412. Published 2017 Feb 15. doi:10.3390/ijms18020412
3. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357-60.
4. http://www.novocraft.com/
5. T16M Metastatic liver tumor; SRR089705.fastq.gz; https://www.ebi.ac.uk/ena/data/view/SRR089705
6. Gene expression data in skin fibroblast cells; SRR7093809.fastq.gz; https://www.ebi.ac.uk/ena/data/view/PRJNA454681
7. Reducing Whole-Genome Data Storage Footprint. Illumina white paper, April 2014.
8. Rodrigo Cánovas, Alistair Moffat, Andrew Turpin, Lossy compression of quality scores in genomic data, Bioinformatics, Volume 30, Issue 15, 1 August 2014, Pages 21302136, https://doi.org/10.1093/bioinformatics/btu183
9. Greg Malysa, Mikel Hernaez, Idoia Ochoa, Milind Rao, Karthik Ganesan, Tsachy Weissman, QVZ: lossy compression of quality values, Bioinformatics, Volume 31, Issue 19, 1 October 2015, Pages 31223129, https://doi.org/10.1093/bioinformatics/btv330
10. Yu YW, Yorukoglu D, Peng J, Berger B. Quality score compression improves genotyping accuracy. Nat Biotechnol. 2015;33(3):240-3.
11. Ochoa, Idoia Hernaez, Mikel Goldfeder, Rachel Weissman, Tsachy Ashley, Euan. (2016). Effect of lossy compression of quality scores on variant calling. Briefings in Bioinformatics. 18. bbw011. 10.1093/bib/bbw011.

12. A. Hernandez-Lopez, Ana Voges, Jan Alberti, Claudio Mattavelli, Marco Ostermann, Jorn. (2018). Lossy Compression of Quality Scores in Differential Gene Expression: A First Assessment and Impact Analysis. 167-176. 10.1109/DCC.2018.00025.
13. Alberti, Claudio Daniels, Noah Hernaez, Mikel Voges, Jan Goldfeder, Rachel A Hernandez-Lopez, Ana Mattavelli, Marco Berger, Bonnie. (2016). An Evaluation Framework for Lossy Compression of Genome Sequencing Quality Values. 2016 Data Compression Conference (DCC). 2016. 221-230. 10.1109/DCC.2016.39.
14. Alberti, Claudio Paridaens, Tom Voges, Jan Naro, Daniel Ahmad, Junaid Jameel Ravasi, Massimo Renzi, Daniele Zoia, Giorgio Ochoa, Idoia Mattavelli, Marco Delgado, Jaime Hernaez, Mikel. (2018). An introduction to MPEG-G, the new ISO standard for genomic information representation. 10.1101/426353.