Denoising of Quality Scores for Boosted Inference and Reduced Storage

Idoia Ochoa*, Mikel Hernaez*, Rachel Goldfeder†, Tsachy Weissman* and Euan Ashley†

*Department of Electrical Engineering and †Department of Medicine
Stanford University, Stanford, CA, 94305
{iochoa,mhernaez,rlg2,tsachy,euan}@stanford.edu

Abstract

Massive amounts of sequencing data are being generated thanks to advances in sequencing technology and a dramatic drop in the sequencing cost. Much of the raw data are comprised of nucleotides and the corresponding quality scores that indicate their reliability. The latter are more difficult to compress and are themselves noisy. Lossless and lossy compression of the quality scores has recently been proposed to alleviate the storage costs, but reducing the noise in the quality scores has remained largely unexplored. This raw data is processed in order to identify variants; these genetic variants are used in important applications, such as medical decision making. Thus improving the performance of the variant calling by reducing the noise contained in the quality scores is important.

We propose a denoising scheme that reduces the noise of the quality scores and we demonstrate improved inference with this denoised data. Specifically, we show that replacing the quality scores with those generated by the proposed denoiser results in more accurate variant calling in general. Moreover, a consequence of the denoising is that the entropy of the produced quality scores is smaller, and thus significant compression can be achieved with respect to lossless compression of the original quality scores. We expect our results to provide a baseline for future research in denoising of quality scores.

The code used in this work as well as a Supplement with all the results are available at http://web.stanford.edu/~iochoa/DCCdenoiser_CodeAndSupplement.zip.

1 Introduction

Recent advances in Next Generation high-throughput Sequencing (NGS) [1] have revolutionized biomedical sciences and have marked the beginning of a new era for biological research. It is now possible to identify genomic changes that predispose individuals to debilitating diseases or make them more responsive to certain therapies and emerging treatments [2].

These data sets are commonly stored in FASTQ or SAM file formats, and they are mainly composed of nucleotide sequences, called "reads", and per-base quality scores that indicate the level of confidence in the readout of these sequences. The latter are represented in the *Phred scale*, given by $Q = \lceil -10 \log_{10} P \rceil$, with P being the probability that the corresponding nucleotide is in error. In the genomic files they are stored as an ASCII character ranging from 33 to 73 (*Phred+33* scale) or from 64 to 104 (*Phred+64* scale).

The genomic data must be stored, transmitted and processed. To reduce the storage costs and facilitate its transmission, data compression techniques for both the reads and the quality scores have been proposed in recent years.

Quality scores have proven to be more difficult to compress than the reads, due in part to their higher variance and larger alphabet [3]. Moreover, there is evidence that quality scores are corrupted by some amount of noise introduced during sequencing, mainly due to the use of inaccurate models to estimate the probabilities of error [4, 5]. Thus, whereas lossless compression is preferred for the reads, lossy compression of quality scores has emerged as a natural candidate to boost compression performance (see [6, 7, 8, 9] and references therein).

The aforementioned lossy compressors address the problem of storage. However, they are not designed to simultaneously reduce the noise present in the data, and as a result the inference performed on it may be compromised. Recently, there has been some work on analyzing the effect that lossy compression has on variant calling. For example, in [10] an extensive analysis is performed which includes several lossy compressors, pipelines and datasets, and it is shown that in most cases the performance of the variant caller is comparable - and sometimes superior - to that of the original data. Qualitatively similar findings where also reported in [8]. These intriguing results seem to suggest that denoising of quality scores is possible and of potential benefit.

The data under consideration is used for biological inference, and thus reducing its noise is important, since it can improve the subsequent analysis performed on it. With this in mind, in this paper we propose a denoising scheme to reduce the noise presented in the quality scores, and demonstrate that it can potentially result in better inference. Moreover, we show that reducing the noise leads to a smaller entropy than that of the original quality scores, and thus a significant boost in compression is also achieved. Thus the angle of the present work is, in a sense, complimentary to that of [10]: while [10] focused on lossy compression and found inferential performance boosts as welcome side benefits, the starting point for the present work is denoising for improved inference, with boosted compression performance as an important benefit stemming from data processing principles. Such schemes to reduce the noise of genomic data while easing its storage and dissemination can significantly benefit the field of genomics. With this work we aim to provide a baseline for future research in this direction.

The denoising scheme described in this paper is based on the one outlined in [11], which in brief consists of applying lossy compression and decompression to the noisy signal followed by a post-processing step. This scheme achieves denoising under certain conditions, and it has been successfully tested in practice in both simulated and real data [12].

One of the challenges with the proposed denoiser is that the lossy compression should be performed under a distortion measure and distortion level determined by the statistics of the noise, which are not known in the case of quality scores. Thus, instead, we use the lossy compressors for quality scores - targeting different distortion levels - proposed in the literature. We will show that the denoiser scheme improves the inference performed on the data in this case, for some distortion levels.

To assess the performance of the aforementioned denoising scheme, we focus on one of the most widely used downstream applications in practice: variant calling, or detection of single nucleotide polymorphisms (SNPs) and insertions or deletions (INDELs). Specifically, we evaluate the accuracy of variant calls when the raw quality scores are replaced with denoised scores. For the analysis, we use the methodology described in [10], which includes several pipelines and datasets for which a consensus of SNPs and INDELs exists.

2 Denoising of Quality Scores

We first formalize the problem of denoising of quality scores, and then describe the proposed denoising scheme in detail. We conclude this section with a description of the evaluation criteria.

2.1 Problem Setting

Let $\boldsymbol{X}_i = [X_i(1), X_i(2), \ldots, X_i(n)]$ be a sequence of true quality scores of length n, and $\boldsymbol{X} = \{\boldsymbol{X}_i\}_{i=1}^N$ a set of quality score sequences. We further let $\boldsymbol{Q} = \{\boldsymbol{Q}_i\}_{i=1}^N$ be the set of noisy quality score sequences that we observe and want to denoise[1], where $Q_i(j) = X_i(j) + Z_{i,j}$ and $\boldsymbol{Q}_i = [Q_i(1), Q_i(2), \ldots, Q_i(n)]$. Note that $\{Z_{i,j} : 1 \leq i \leq N, 1 \leq j \leq n\}$ represents the noise added during the sequencing process. This noise comes from different sources of generally unknown statistics, some of which are not reflected in the mathematical models used to estimate the quality scores [4, 5].

Our goal is to denoise the noisy quality scores $\boldsymbol{Q}$ to obtain a version closer to the true underlying quality score sequences $\boldsymbol{X}$. We further denote the output of the denoiser by $\widehat{\boldsymbol{X}} = \{\widehat{\boldsymbol{X}}_i\}_{i=1}^N$, with $\widehat{\boldsymbol{X}}_i = [\widehat{X}_i(1), \widehat{X}_i(2), \ldots \widehat{X}_i(n)]$.

2.2 Denoising Scheme

The suggested denoising scheme is depicted in Fig. 1. It consists of a lossy compressor applied to the noisy quality scores $\boldsymbol{Q}$, the corresponding decompressor, and a post-processing operation that uses both the reconstructed quality scores $\widehat{\boldsymbol{Q}}$ and the original ones. The output of the denoiser is the sequence of noiseless quality scores $\widehat{\boldsymbol{X}}$. In order to compute the final storage size, a lossless compressor for quality scores is applied to the denoised signal $\widehat{\boldsymbol{X}}$. Note that we can not simply store the output of the lossy compressor and use that as the final size, since the post-processing operation also needs access to the original quality scores. That is, the denoiser needs to perform both the lossy compression and the decompression, and incorporate the original (uncompressed) noisy data for computing its final output.

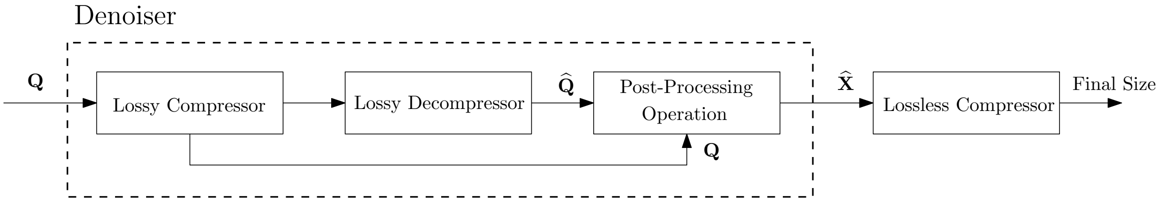

Figure 1: Outline of the proposed denoising scheme.

The proposed denoiser is based on the one outlined in [11], which is universally optimal in the limit of large amounts of data when applied to a stationary ergodic

[1] For example, the quality score sequences found in a FASTQ or SAM file.

source corrupted by additive white noise. Specifically, consider a stationary ergodic source X^n and its noisy corrupted version Y^n given by

$$Y_i = X_i + Z_i,$$

for $1 \leq i \leq n$, where Z^n is an additive white noise process. Then, the first step towards recovering the noiseless signal X^n consists of applying a lossy compressor under the distortion measure $\rho : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}^+$ given by

$$\rho(y, \hat{y}) \triangleq \log \frac{1}{P_Z(y - \hat{y})}, \tag{1}$$

where $P_Z(\cdot)$ is the probability mass function of the random variable Z. Moreover, the lossy compressor should be tuned to distortion level $H(Z)$, that is, the entropy of the noise.

In the case of the quality scores, the statistics of the noise are unknown, and thus we can not set the right distortion measure and level at the lossy compressor. In the presence of such uncertainty, one could make the worst case assumption that the noise is Gaussian [13] of unknown variance, which would translate into a distortion measure given by the square of the error (based on Eq. (1)). However, even with this assumption, the correct distortion level depends on the unknown variance. Thus, instead, we take advantage of the extensive work performed on lossy compressors for quality scores in the past, and use them for the lossy compression step. Since we do not know the right distortion level to set at the lossy compressor, we apply each of them with different distortion levels. This decision, although lacking theoretical guarantees, works in practice, as demonstrated in the following section. Moreover, it makes use of the lossy compressors that have already been proposed and tested.

The second step consists of performing a post-processing operation based on the noisy signal Y^n and the reconstructed sequence $\widehat{Y}^n$. For a given integer $m = 2m_0+1 > 0$, $y^m \in \mathcal{Y}^m$ and $\hat{y} \in \widehat{\mathcal{Y}}$, define the joint empirical distribution as

$$\hat{p}^{(m)}_{Y^n \widehat{Y}^n}(y^m, \hat{y}) \triangleq \frac{|\{m_0 + 1 \leq i \leq n - m_0 : (Y_{i-m_0}^{i+m_0}, \widehat{Y}_i) = (y^m, \hat{y})\}|}{n - m + 1}. \tag{2}$$

Thus, Eq. (2) represents the fraction of times $Y_{i-m_0}^{i+m_0} = y^m$ while $Y_i = \hat{y}$, for all i. Once the joint empirical distribution is computed, the denoiser generates its output as

$$\widehat{X}_i = \underset{\hat{x} \in \widehat{\mathcal{X}}}{\operatorname{argmin}} \sum_{x \in \widehat{\mathcal{Y}}} \hat{p}^{(m)}_{Y^n \widehat{Y}^n}(Y_{i-m_0}^{i+m_0}, x) d(\hat{x}, x), \tag{3}$$

for $1 \leq i \leq n$. Note that $d : \widehat{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}^+$ is the original loss function under which the performance of the denoiser is to be measured, and $\widehat{\mathcal{X}}$ is the alphabet of the denoised sequence $\widehat{X}^n$.

For the case of the quality scores, the joint empirical distribution can be computed mostly as described in Eq. (2). However, since now we have a set of quality score sequences, we redefine it as

$$\hat{p}^{(m)}_{\boldsymbol{Q}, \widehat{\boldsymbol{Q}}}(q^m, \hat{q}) \triangleq \frac{|\{(i,j) : (Q_i(j - m_0, \dots, j + m_0), \widehat{Q}_i(j)) = (q^m, \hat{q})\}|}{nN}, \tag{4}$$

where $Q_i(j-m_0,\ldots,j+m_0)$ is a short hand of $(Q_i(j-m_0),\ldots,Q_i(j+m_0))$, and $Q_i(j)=0$ for $j<1$ and $j>n$. Finally, the output of the denoiser is given by

$$\widehat{X}_i(j)=\operatorname*{argmin}_{\hat{x}\in\widehat{\mathcal{X}}}\sum_{\hat{q}\in\widehat{\mathcal{Q}}}\hat{p}^{(m)}_{\boldsymbol{Q},\widehat{\boldsymbol{Q}}}(Q_i(j-m_0,\ldots,j+m_0),\hat{q})d(\hat{x},\hat{q}), \tag{5}$$

for $1\leq i\leq N$ and $1\leq j\leq n$, with d being squared distortion. Note also that the alphabets of the original, reconstructed and denoised quality scores are the same, i.e., $\mathcal{Q}=\widehat{\mathcal{Q}}=\widehat{\mathcal{X}}$.

Finally, as mentioned above, we apply a lossless compressor to the output of the decoder to compute the final size.

As outlined in [12], the intuition behind the proposed scheme is as follows. First, note that adding noise to a signal always increases its entropy, since

$$I(X^n+Z^n;Z^n)=H(X^n+Z^n)-H(X^n+Z^n|Z^n)=H(X^n+Z^n)-H(X^n)\geq 0, \tag{6}$$

which implies $H(Y^n)\geq H(X^n)$, with $Y^n=X^n+Z^n$. Also, lossy compression of Y^n at distortion level D can be done by searching among all reconstruction sequences within radius D of Y^n, and choosing the most compressible one. Thus, if the distortion level is set appropriately, a reasonable candidate for the reconstruction sequence can be the noiseless sequence X^n. The role of the lossy compressor is to partially remove the noise and to learn the source statistics in the process, such that the post-processing operation can be though of as performing Bayesian denoising. Therefore, we also expect the denoised quality scores to be more compressible than the original ones, due to the reduced entropy.

2.3 Evaluation Criteria

To measure the quality of the denoiser we cannot compare the set of denoised sequences $\widehat{\boldsymbol{X}}$ to the true sequences $\boldsymbol{X}$, as the latter are unavailable. Instead, we analyze the effect on variant calling when the original quality scores are replaced by the denoised ones. For the analysis, we follow the methodology described in [10], which consists of several pipelines and datasets specific for SNP calling and INDEL detection. In brief, the considered pipelines for SNP calling are GATK-HC (Haplotype Caller) [14, 15, 16], htslib.org [17] and Platypus [18], and for INDEL detection we used Dindel [19], GATK-UG (Unified Genotyper), GATK-HC and Freebayes [20].

All datasets in this study have a consensus sequence, making it possible to analyze the accuracy of the variant calls. That is, given a set of called SNPs (or INDELs) as output of the variant calling and the consensus set, we can determine how many of the called variants are True Positives (T.P.) and False Positives (F.P.), and how many variants were not called, that is, the number of False Negatives (F.N.). Once these values are estimated, we compute the following performance metrics: *sensitivity*, which measures the proportion of all the positives that are correctly called ($\frac{T.P.}{T.P.+F.N.}$), *precision*, which measures the proportion of called positives that are true ($\frac{T.P.}{T.P.+F.P.}$), and *f-score*, which is the harmonic mean of the sensitivity and precision.

We expect that using the denoised in lieu of the original data would yield higher sensitivity, precision and f-score.

3 Results and Discussion

We analyze the performance of the proposed denoiser for both SNP calling and INDEL detection. For the lossy compressor block we used the algorithms RBlock [7], PBlock [7], QVZ [9] and Illumina's proposed binning (http://goo.gl/d5TYDk). Since the right distortion level at which they should operate is unknown, we run each of them with different parameters (i.e., different distortion levels)[2]. Specifically, we employ QVZ with MSE distortion criteria, one and three clusters and rate ranging from 0 to 1, PBlock with values of p ranging from 1 to 32, and RBlock with values of r ranging from 3 to 30. Regarding the post-processing operation, we set m in Eq. (4) to be equal to three in all the simulations. This choice was made to reduce the running time and complexity, because of the large alphabet of the quality scores. As the entropy encoder we applied QVZ in lossless mode, which offers competitive performance [9].

Due to the extensive amount of simulations and the space constraint, we focus on the most representative results, and refer the reader to the Supplement (in the form of Excel files) for a complete analysis of all the considered datasets and pipelines. For ease of visualization of the results presented in the Excel files, we colored in red those that improve with respect to the original data, in yellow those that improve with respect to applying solely lossy compression, and in green those that improve upon the previous two. For completeness, we also added the results obtained when only lossy compression is applied to the quality scores (i.e., without the post-processing operation). The post-processing operation improves the performance beyond that achieved by applying only lossy compression in most cases. Finally, the size of the data for each case is also stated. The denoised data occupies less than the original one, corroborating our expectation that the denoiser reduces the noise of the quality scores and thus the entropy, and consistent with the data processing principle. As an example, Fig. 2 compares the size of the chr. 20 of the data ERR262997 (used for the analysis on SNP calling), with that generated by the denoiser with different lossy compressors targeting different distortion levels (x-axis). As can be observed, for all distortion levels above 4, the reduction in size is between 30% and 44%. Interestingly, similar results were obtained with all the tested datasets, which suggests that more than 30% of the entropy (of the original data) is due to noise.

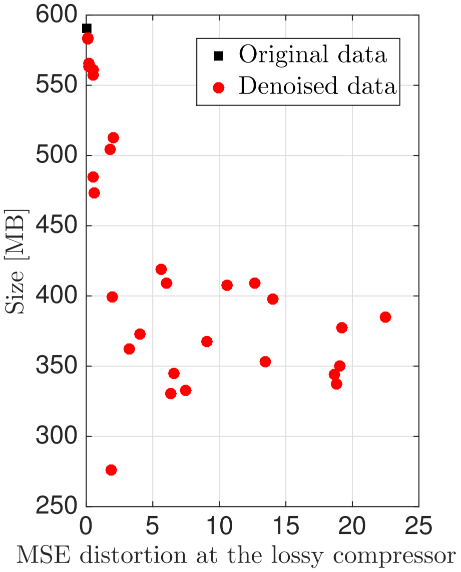

Figure 2: Reduction in size achieved by the denoiser when compared to the original data (when losslessly compressed).

In the following we focus on the performance of the denoiser in terms of its effect on SNP calling and INDEL detection.

[2] Except for Illumina's proposed binning which generates only one point in the rate-distortion plane.

3.1 SNP calling

We use data that belongs to the human individual NA12878, which has been thoroughly characterized in the past and for which two consensus of SNPs are available. One was released by the GIAB (Genome In A Bottle) consortium [21], and the other was released by Illumina. In particular, we consider the chromosomes 11 and 20 of the pair-end whole genome sequencing datasets ERR174324 (15x) and ERR262997 (30x). We observe that the results for the chromosomes 11 and 20 of the 30x coverage dataset are very similar for all the considered pipelines, and thus due to space constraints, we restrict our attention to chromosome 20. Regarding the 15x coverage dataset, we focus on chromosome 11 and the SNP consensus produced by Illumina (similar results where obtained with the GIAB consensus).

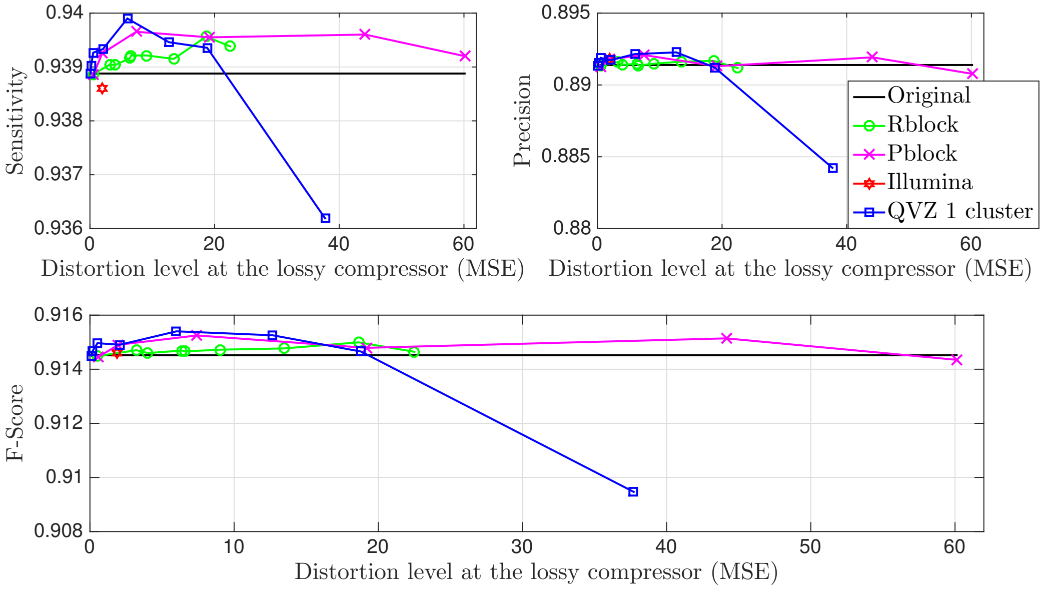

Figure 3: Denoiser performance on the GATK-HC pipeline (30x dataset, chr. 20). Different points of the same color correspond to running the lossy compressor with different parameters.

Fig. 3 shows the results for the 30x coverage dataset on the GATK-HC pipeline. As can be observed, for MSE distortion levels between 0 and 20 approximately, and any lossy compressor, the denoiser improves all three metrics; f-score, sensitivity and precision. Among the analyzed pipelines, GATK-HC is the most consistent and the one offering the best results. This suggests that the GATK-HC pipeline uses the quality scores in the most informative way.

For htslib.org and Platypus, we also observe that the points that improve upon the original one exhibit an MSE distortion less than 20 in general. However, in this case the lossy compressors perform differently. For example, QVZ improves the precision and fscore with the htslib.org pipeline and the sensitivity with the platypus one. On the other hand, Pblock and Rblock achieve best results in terms of precision and fscore with the platypus pipeline and sensitivity with htslib.org.

With the 15x dataset, the denoiser achieves in general better performance when

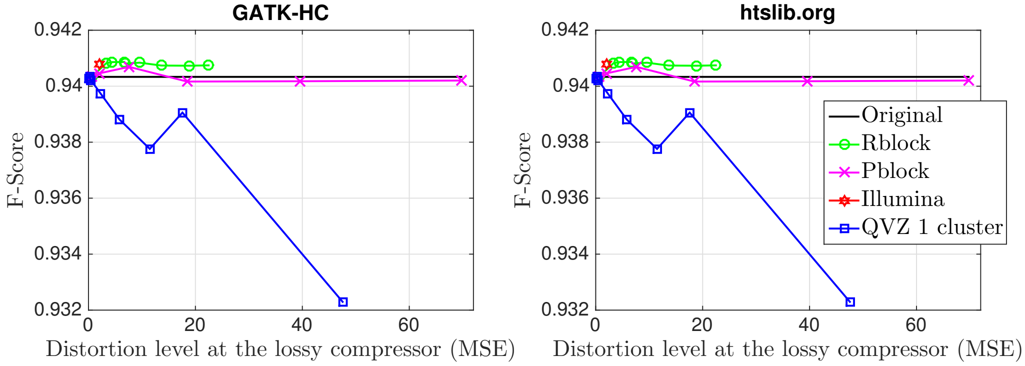

Figure 4: Denoiser performance on the GATK-HC and hstlib.org pipelines (15x dataset, chr. 11).

using the lossy compressors Rblock and Pblock. For example, Fig. 4 shows the fscore for the GATK-HC and htslib.org pipelines. As can be observed, the denoised data improves upon the uncompressed in both cases with Rblock and Illumina's proposed binning, and with Pblock when the distortion level is below 20. With QVZ the denoiser achieves better precision with the GATK-HC and htslib.org pipelines, and better sensitivity with Platypus.

Finally, it is worth noticing the potential of the post-processing operation to improve upon the performance when applying only lossy compression. As can be observed from the detailed results shown in the Supplement, this is true for all the four considered datasets (chromosomes 11 and 20 and coverages 15x and 30x), and the three pipelines (GATK-HC, htslib.org and platypus). To give some concrete examples, with the platypus pipeline the post-processing operation boosts the performance of the sensitivity when applying any lossy compressor, for all datasets. The general improvement is more noticeable for the 15x coverage datasets, where all metrics improve in most of the cases. See for example the performance with the chromosome 11, for all pipelines and both sets of SNPs consensus (in the Supplement). Recall that the results colored in yellow and/or green correspond to improvement with respect to merely applying lossy compression.

3.2 INDEL detection

We use the data employed in [10] for INDEL detection. That is, a chromosome containing 3000 heterozygous INDELs from which 100bp paired-end sequencing reads were generated with ART [22].

Among the analyzed pipelines, the denoiser exhibits the best performance on the GATK-HC pipeline. For example, in terms of f-score, we observe that the proposed scheme with Illumina's binning and Rblock as the lossy compressor achieves better performance than the original data. QVZ and Pblock also improve for the points with smaller distortions. Similar results are obtained for the sensitivity and precision. Moreover, in this case the potential of applying the post-processing operation after any of the considered lossy compressors becomes particularly apparent, as the performance always improves (see Fig. 5).

We also observe improved performance using Freebayes with QVZ in terms of

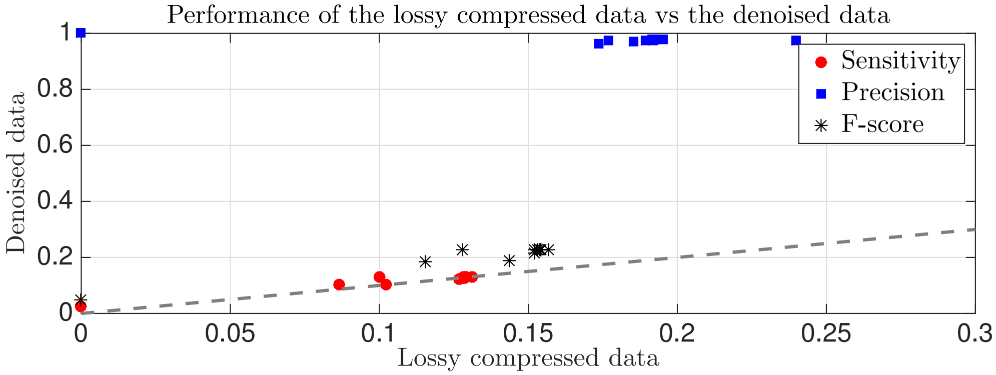

Figure 5: Improvement achieved by applying the post-processing operation. x-axis represents the performance in sensitivity, precision and f-score achieved by solely applying lossy compression, and the y-axis represents the same but when the post-processing operation is applied after the lossy compressor. Grey line corresponds to $x = y$, and thus all the points above it correspond to an improved performance.

sensitivity, precision and f-score, and an improved precision with the remaining lossy compressors. With the GATK-UG and Dindel pipelines, Rblock achieves the best performance, improving upon the original data under all three performance metrics.

4 Conclusion

We proposed a denoising scheme for quality scores composed of a lossy compressor followed by the corresponding decompressor and a post-processing operation. Experimentation on real data suggests that the proposed scheme has the potential to improve the quality of the data insofar as its effect on the downstream inferential applications, while at the same time significantly reducing the storage requirements.

Further study of denosing of quality scores is merited as it seems to hold the potential to enhance the quality of the data while at the same time easing its storage requirements. We hope the promising results presented in this paper serve as a baseline for future research in this direction. Further research should include improved modeling of the statistics of the noise, construction of denoisers tuned to such models, and performing more experimentation on real data and with additional downstream applications. A more ambitious goal for the longer run is to revisit and optimize the design of the downstream applications jointly with the processing of the quality scores.

5 Acknowledgement

This work is partially funded by a fellowship from the Basque Government, a grant from the Center for Science of Information (CSoI), the 2014-07364-01 NIH grant, the 1 U01 CA198943-01 NIH grant, the 1157849- 1-QAZCC NSF grant, the National Library of Medicine Training Grant T15 LM7033 and an NSF graduate research fellowship.

References

[1] M. L. Metzker, “Sequencing technologies - the next generation,” *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2009.

[2] J. S. Berg and et. al., “Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time,” *Genetics in Medicine*, 2011.

[3] J. K. Bonfield and M. V. Mahoney, “Compression of fastq and sam format sequencing data,” *PloS one*, vol. 8, no. 3, 2013.

[4] W.-C. Kao, K. Stevens, and et. al., “Bayescall: A model-based base-calling algorithm for high-throughput short-read sequencing,” *Genome research*, vol. 19, no. 10, 2009.

[5] S. Das and H. Vikalo, “Onlinecall: fast online parameter estimation and base calling for illumina’s next-generation sequencing,” *Bioinformatics*, vol. 28, no. 13, 2012.

[6] I. Ochoa, H. Asnani, and e. a. Bharadia, “Qualcomp: a new lossy compressor for quality scores based on rate distortion theory,” *BMC bioinformatics*, vol. 14, no. 1, 2013.

[7] R. Cánovas, A. Moffat, and A. Turpin, “Lossy compression of quality scores in genomic data,” *Bioinformatics*, 2014.

[8] Y. W. Yu, D. Yorukoglu, and B. Berger, “Traversing the k-mer landscape of ngs read datasets for quality score sparsification,” in *Research in Comp. Molecular Bio.*, 2014.

[9] G. Malysa, M. Hernaez, I. Ochoa, and et. al., “Qvz: lossy compression of quality values,” *Bioinformatics*, p. btv330, 2015.

[10] I. Ochoa, M. Hernaez, R. Goldfeder, T. Weissman, and E. Ashley, “Effect of lossy compression of quality scores on variant calling,” *Briefings in Bioinformatics*, 2016.

[11] T. Weissman and E. Ordentlich, “The empirical distribution of rate-constrained source codes,” *IEEE Trans. Inf. Theory*, vol. 51, 2005.

[12] S. Jalali and T. Weissman, “Denoising via MCMC-based lossy compression,” *IEEE Trans. Signal Process.*, vol. 60, no. 6, 2012.

[13] H. Asnani, I. Shomorony, and et. al., “Network compression: Worst-case analysis,” in *Inf. Theory Proceedings (ISIT), 2013 IEEE Intern. Symp. on*, 2013, pp. 196–200.

[14] A. McKenna and et. al., “The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data,” *Genome research*, 2010.

[15] M. A. DePristo, E. Banks *et al.*, “A framework for variation discovery and genotyping using next-generation dna sequencing data,” *Nature genetics*, vol. 43, no. 5, 2011.

[16] G. A. Auwera, M. O. Carneiro *et al.*, “From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline,” *Current Protocols in Bioinformatics*, pp. 11–10, 2013.

[17] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin *et al.*, “The sequence alignment/map format and samtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[18] A. Rimmer, H. Phan, I. Mathieson *et al.*, “Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications,” *Nature genetics*, vol. 46, no. 8, pp. 912–918, 2014.

[19] C. A. Albers, G. Lunter, and et. al., “Dindel: accurate indel calls from short-read data,” *Genome research*, vol. 21, no. 6, pp. 961–973, 2011.

[20] E. Garrison and G. Marth, “Haplotype-based variant detection from short-read sequencing,” *arXiv preprint arXiv:1207.3907*, 2012.

[21] J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit, “Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls,” *Nature biotechnology*, vol. 32, no. 3, pp. 246–251, 2014.

[22] W. Huang, L. Li, J. R. Myers, and G. T. Marth, “Art: a next-generation sequencing read simulator,” *Bioinformatics*, vol. 28, no. 4, pp. 593–594, 2012.