Lossy compression of quality scores in differential gene expression: A first assessment and impact analysis

Ana A. Hernandez-Lopez*, Jan Voges†,
Claudio Alberti*, Marco Mattavelli* and Jörn Ostermann†

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

†Leibniz Universität Hannover
Institut für Informationsverarbeitung (TNT)
Appelstr. 9A
30167 Hannover, Germany
voges@tnt.uni-hannover.de

Abstract

High-throughput sequencing of RNA molecules has enabled the quantitative analysis of gene expression at the expense of storage space and processing power. To alleviate these problems, lossy compression methods of the quality scores associated to RNA sequencing data have recently been proposed, and the evaluation of their impact on downstream analyses is gaining attention. In this context, this work presents a first assessment of the impact of lossily compressed quality scores in RNA sequencing data on the performance of some of the most recent tools used for differential gene expression.

1 Introduction

High-throughput RNA sequencing (RNA-seq) is undergoing rapid evolution since its introduction back in 2008 when several research groups, encouraged by the accessibility of novel high-throughput sequencing technologies, set out to study the transcriptome of different organisms [1–4]. It is through nucleotide sequences of RNA that information encoded in an organism's DNA is made available to the cell, and that it can be interpreted by the cell to guide the synthesis of proteins. The RNA sequences are gene readouts, i.e. copies of gene regions of DNA. These gene readouts are called transcripts and the set of all the transcripts present in a cell, or a population of cells, at a given time constitutes the transcriptome.

Researchers can gain a better understanding of the workings of cells and their connection to diseases by investigating the levels of gene activity in the transcriptome. The activity of a gene is the result of a process known as gene expression through which the DNA nucleotide sequence of a gene is converted into nucleotide sequences of RNA, and then into the amino acid sequence of a protein. The amount of gene activity can be measured by estimating the number of transcripts in a tissue sample. RNA-seq data is widely used to get quantitative information on the differences in the expression of genes between a test and control conditions. However, gene expression levels are very fragile and reflect uncertainties associated with sampling as well as technical and biological variance [5]. The certainty about the observation of a gene expression level can be improved by increasing the number of sequenced reads in a condition, which can be achieved by adding biological replicates and by deeper sequencing of existing replicates [6].

2375-0359/18/$31.00 ©2018 IEEE
DOI 10.1109/DCC.2018.00025

The test for differential gene expression (DGE) relies on the estimation of transcripts across conditions, which requires the assembly and quantification of millions of sequenced reads. The high computational cost associated to the storage and processing of millions of reads is shared by all functional genomic assays driven by high-throughput sequencing. The wealth of raw sequenced data, and the complexity of measurements to be inferred make the setup of a working bioinformatic pipeline a challenge, and an assessment of its accuracy is difficult [7–9]. Moreover, the situation aggravates in applications like DGE where multiple, deeply sequenced samples need to be analyzed.

In recent years several research groups have investigated methods to improve the effectiveness of compression technologies for the storage of high-throughput sequencing data. In particular, approaches to lossy or quasi-lossless compression of quality scores have received special attention [10–13], along with an interest to measure their impact in the calling of genomic variants [14, 15], so far the sole downstream application tested for evaluation.

In the context of gene expression (section 2) this work sets out to explore the effect of lossy compression of quality scores. For this purpose we start by observing its effect on transcript reconstruction over a simulated sample with different depths of coverage. Then, we take two real datasets of RNA-seq data and run them on a state-of-the-art DGE pipeline (section 3), and provide a first assessment of the impact (sections 4 and 5). In particular, the goal is to understand if differences arise in the calling of expressed genes, between a two-condition DGE pipeline that features full quality score scale of RNA-seq data, and a pipeline featuring reduced resolution. The focus is only on significant genes with the strongest activity and state-of-the-art tools are used to build the pipeline.

In summary, this work shows:

– that lossy quality scores marginally affect the reconstruction of transcripts in simulated data, a result that is corroborated in the calling of genes in the test for differential gene expression,

– the application of lossy compression in a pipeline for testing differential gene expression (section 3),

– how high rates of lossy compression of quality scores in RNA-seq data do not compromise, in principle, the calling of significant genes when testing for differential gene expression in a two-condition setting (section 4).

2 RNA-seq and differential gene expression

RNA-seq is a functional genomic assay based on high throughput sequencing with the primary goal of quantifying abundances of mature molecules of messenger RNA (mRNA) in a cell. The RNA in a cell is produced by DNA transcription, a process where portions of DNA are copied or transcribed into RNA nucleotide sequences. Specifically, these RNA chains transcribe segments of gene regions of DNA, also

referred to as transcripts. Many transcripts can be made from the same gene and each transcript can direct the synthesis of several protein molecules. Moreover, the cell commonly controls the production of RNA to regulate its genes whose expression can be measured by counting the abundance of transcripts present at a particular moment in time within a cell.
Different types of RNA molecules are produced during transcription but only mature mRNAs will be translated into proteins. In eukaryotic cells the mature mRNA transcripts result after RNA splicing: a process where all intron sequences are removed from mRNA transcripts and the remaining exons are joined to form a continuous sequence. Splicing can occur in different ways leaving in or out exons from the final transcript. The possibility of different splicing patterns from the same mRNA transcript is called alternative splicing and it allows the production of different proteins from the same gene during translation.
Broadly speaking, RNA-seq applications can be grouped in two categories. When the expressed transcripts are used to conduct transcriptome annotations, the application is qualitative. Other applications require some form of measuring and thus they are considered as quantitative. Examples of these applications are: the quantification of novel transcripts, alternative splicing and gene expression.
The goal of most RNA-seq experiments is to identify genes whose expression change across two experimental conditions. These differential gene expression experiments require at least six biological replicates per condition with sufficient sequencing depth [6, 16, 17].
The RNA-seq protocol is approximately the same across platforms: samples of RNA are isolated, fragmented at random positions, copied into complementary DNA, amplified and sequenced to obtain reads. The workflow described below reconstructs the transcriptome from the resulting reads and measures the expression of genes by quantifying the abundance of assembled reads.
Figure 1 shows how a pipeline for DGE can be structured in three steps:

1. Assembly. Approximately 60% of the reads are exonic and will map entirely within an exon in the reference genome. The rest of the reads come from spliced transcripts and will span two or more exons. Spliced mappers like TopHat2 [18] and HISAT2 [19, 20] can map exonic reads and identify splice junctions from reads spanning different exons. However, the assembly of exon-spanning reads requires an additional tool. According to [21] the best performing tools for this task are Cufflinks [22] and StringTie [23]. The reconstruction of the transcriptome is complete when both exonic reads and exon-spanning reads are mapped.

2. Quantification. The mapped reads are counted to measure the expressed genes in the reconstructed transcriptome. Cufflinks and StringTie simultaneously assemble and count the reads mapped to each transcript. Lightweight approaches such as Sailfish and its successor Salmon [24] and kallisto [25] bypass the assembly step and directly estimate the read count by pseudo-aligning to the reference transcriptome.

3. Estimation of magnitude and significance of differential expression. The count of reads is a relative value of the sample. Its value depends heavily on the amount of

fragments sequenced and the effective length of the genomic region in an RNA-seq experiment. Therefore read counts should be normalized to compare features (e.g. genes, isoforms and exons) within a sample; common units for normalized read counts are transcripts per million (TPM) and fragments per kilobase of exon per million reads mapped (FPKM). The magnitude of differential expression between two or more conditions is estimated by computing the fold change of normalized read counts from replicated samples. DGE tools make assumptions about the distribution of the read counts to determine the genes whose expression varies between conditions. These tools estimate the significance of expression differences by testing the null hypothesis that a gene's expression between conditions (e.g. treatment vs. control) is unaffected. In a recent publication [16] the best performing tools for estimating DGE were shown to be edgeR [26] and DESeq2 [27].

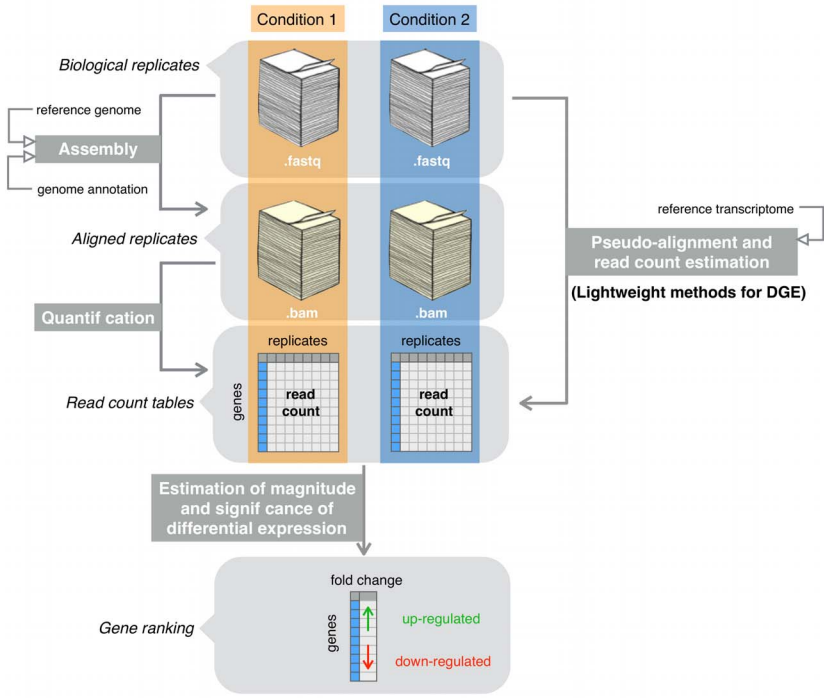

Figure 1: Organization of a pipeline for differential gene expression.

3 Experimental setting

In our first setting we investigated the effect of lossy compression of quality scores on transcript reconstruction. Using the Flux simulator [28], we generated three samples of the human chromosome 22 with one, five and ten million reads and ran them through HISAT2 and StringTie to assemble the transcripts. The samples were input in four modes: with and without quality scores, and after applying lossy compression with the tools Quartz [11] and P-/R-Block [10]. We evaluated the reconstruction of

transcripts by means of the average per-base-coverage. Because we used simulated data, the reference coverage is known and after assembly the coverage for reconstructed transcripts can be computed.
In our second setting, we focus on determining differentially expressed genes on replicates of RNA-seq data. The layout consists of three steps: assembly, quantification and the test for differential expression (see Figure 2). The sequenced reads are first mapped to the reference transcriptome guided by the genome annotation during assembly. The mapped reads are then analyzed to reconstruct the possible transcripts from which they came from; the computation of abundances of reconstructed transcripts follows. Both the assembly and quantification steps are repeated for each replicate in every condition. The test for differential expression takes place after the abundance count of all replicates of all conditions has been obtained. In this last step the magnitude and significance of expressed genes are estimated.

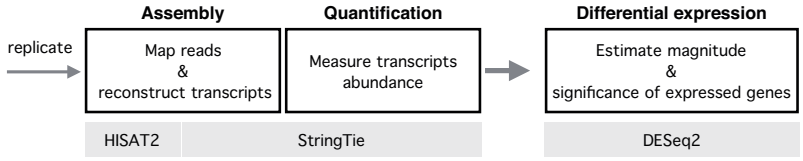

Figure 2: Steps for determining differentially expressed genes on replicates of RNA-seq data. The assembly and quantification steps are repeated for each replicate in every condition. The name of the tools used are stated below each step.

In the pipeline of Figure 2 a pre-processing step is added where lossy compression is applied to the quality scores of an input replicate (see Figure 3). In this step the sequences of nucleotides are kept intact, but their quality scores are compressed with controlled loss of information, and ultimately transformed to a coarser resolution after decompression. Three methods of lossy compression of quality scores have been applied: a uniform quantization with 2 and 8 bins (UQ2, UQ8) and the approaches proposed in the tools Quartz [11] and P-Block and R-Block, respectively [10].
A necessary pre-processing step for Quartz is the generation of a dictionary of common k-mers for each species. Then, for a given set of sequence reads, Quartz breaks these reads up into a set of overlapping k-mers. Subsequently, every position in a supporting k-mer different from a dictionary k-mer is annotated as a possible variant. Quartz assumes that these divergent bases in supporting k-mers correspond to sequencing errors or single nucleotide polymorphisms (SNPs), respectively. The corresponding quality scores are preserved by Quartz, whereas other quality scores are set to a pre-defined default value.
P- and R-Block represent quality scores by separating them into blocks of variable size, where all quality scores contained in each block comply with a chosen parameter according to some measure criterion. For each block, its length and a representative value is stored.
The pipeline under test is summarized in Figure 4. Tests on this pipeline were conducted for two organisms: the yeast S. cerevisiae[1] and the MCF-7 human breast

[1] https://www.ebi.ac.uk/ena/data/view/PRJEB5348

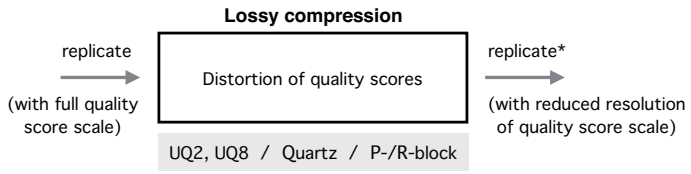

Figure 3: Lossy compression of quality scores for each replicate is prepended to the DGE pipeline. Three methods are used: uniform quantization, Quartz and P-/R-Block.

cancer cells[2]. For each sample a total of twelve biological replicates (six replicates per condition) were used. The results are presented in the following section.

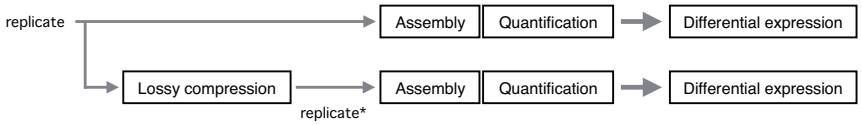

Figure 4: Layout of the pipeline for differential expression with lossy compression of quality scores. This pipeline was run for three lossy compression methods on RNA-seq data for two organisms.

4 Results

In Table 1 we report the overall alignment percentage for the four modes of the three simulated samples of the human chromosome 22. Along with the alignment rate, the bits required per quality score is shown.

Table 1: Overall alignment rate percentage with HISAT2. This value is the sum of the percentage of reads aligned exactly one time plus the percentage of reads aligned more than one time.

		1M	5M	10M	bits/QS
	full QS	77.77	78.28	79.63	3.16
	no QS	76.5	77	78.25	0
Lossy compression	Quartz	77.37	77.56	79.29	1.12
	P-Block	78.73	78.91	80.61	0.98

The distribution of transcripts ordered by coverage is shown in Figure 5(a). This data reports the coverage per reconstructed transcript in the file with 10 million reads and with full quality score scale. Figure 5(b) and (c) show in detail the coverage for the bottom and top 100 transcripts. We observe how the fluctuation of coverage is marginally different between the four modes under test.
In the analysis of gene expression the measure of change is usually reported in terms of the *fold change estimate*. This value represents how much the expression of a

[2] https://www.ebi.ac.uk/ena/data/view/PRJNA222975

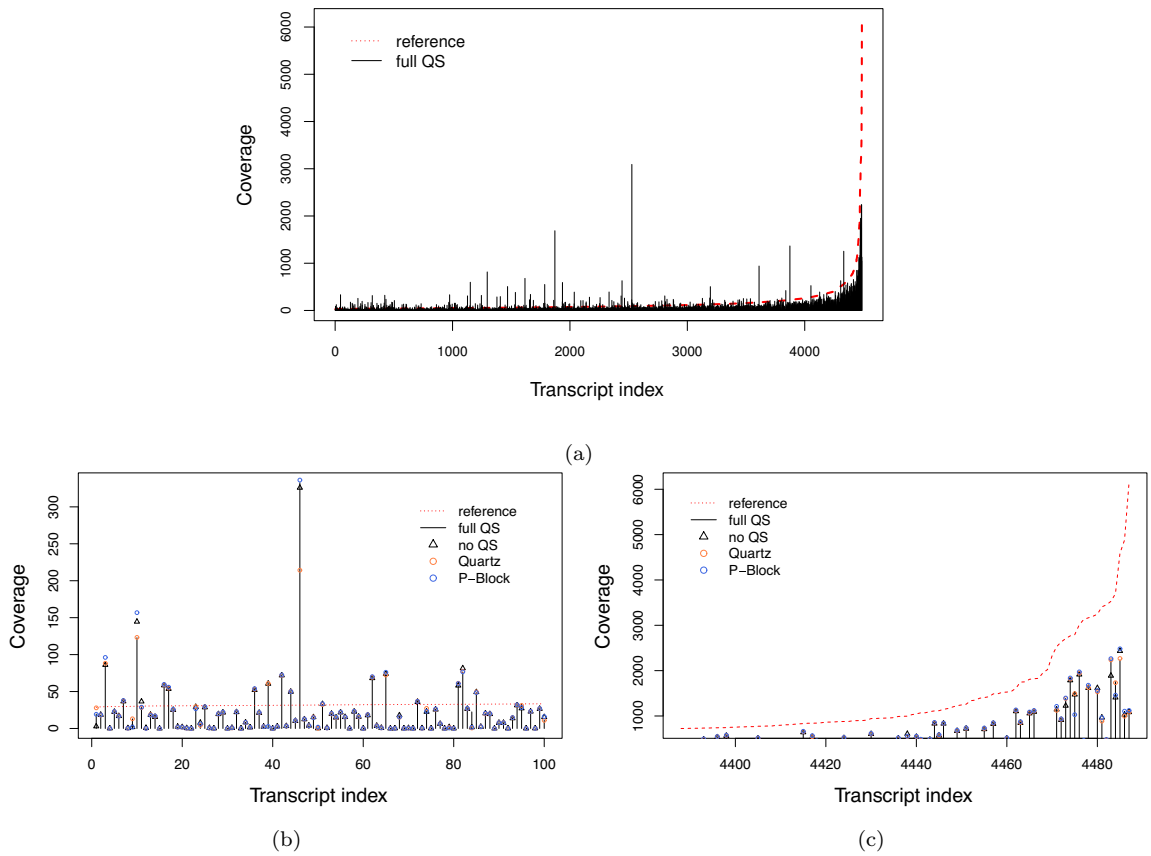

Figure 5: (a) Coverage of chromosome 22 in the file with 10 million reads. (b) Bottom and (c) top 100 transcripts in the same file.

gene seems to have changed between conditions. The fold change can be positive or negative and it is commonly transformed to log2 scale; for example, a gene with a log2 fold change of 1 means that the gene's expression increased by a factor of $2^1 = 2$. Positive values of fold change signal up-regulated genes and negative values signal down-regulated genes.

To determine the significance in the calling of expressed genes the method for differential analysis of count data proposed in DESeq2 [27] was used. For every gene a hypothesis test is conducted to decide against the null hypothesis that the variability observed of a gene's expression between conditions is the same; the result of the test is reported as a p-value. These p-values are corrected for multiple testing and adjusted to account for false positives. The false discovery rate statistic can then be used to set a threshold on the allowed percentage of false positives in the set.

For the performed tests a false discovery rate of 10% was considered and the result was sorted by the log2 fold change estimate to obtain significant genes with the strongest up- and down- regulation. This list of ranked genes is the output of the last step of the pipeline shown in Figure 4.

Table 2: Median compression rates in bits per quality score. The values are reported for both organisms and for each condition and lossy compression method.

	cond		UQ2	UQ8	Quartz	P-/R-Block
yeast	1	3.075	0.2	0.735	1.75	1.015
		[3.05, 3.10]	[0.2, 0.21]	[0.72, 0.75]	[1.66, 1.89]	[1.0, 1.04]
	2	3.08	0.205	0.735	1.025	1.015
		[3.05, 3.09]	[0.2, 0.21]	[0.72, 0.75]	[1.75, 1.85]	[1.0, 1.03]
MCF-7	1	2.21	0.16	0.70	0.57	0.975
		[1.49, 2.47]	[0.07, 0.19]	[0.35, 0.82]	[0.46, 0.61]	[0.52, 1.13]
	2	1.68	0.09	0.44	0.49	0.635
		[1.59, 1.95]	[0.08, 0.12]	[0.4, 0.57]	[0.48, 0.55]	[0.58, 0.80]

Table 3: Ranked list of log2 fold changes for the yeast and genes associated.

	regulation	log2 fold change				gene
			UQ2	UQ8	Quartz	P-/R-Block	
yeast	up	6.0629	6.0574	5.9761	6.0631	5.9764	YOR192C-A
		5.7313	5.8074	5.8105	5.8147	5.8108	YDR034C-C
		3.6137	3.5778	5.0871	5.2070	3.5193	YHR214C-C
		2.8025	2.7971	2.7996	2.8031	2.7980	YPL025C
		2.5757	2.5702	**2.6641**	2.5764	2.5716	YOR376W
		2.4249	2.3629	**2.5722**	2.3671	2.4629	YPR158C-C
	down	-8.0886	-8.0846	-8.0834	-8.0899	-8.0844	YOR192C-B
		-8.0082	-8.0026	-8.0032	-8.0103	-8.0080	YDR034C-D
		-6.2723	-6.3004	-6.1566	**-6.6860**	**-6.1452**	YER160C
		-3.4012	-2.8554	**-6.0406**	**-6.4943**	**-6.1324**	YHR214C-B
		-2.4985	-2.5184	-4.5319	**-4.8144**	**-3.0414**	YDR210W-A
		-1.8940	-1.8929	-2.4752	**-2.5104**	**-2.5042**	YKL078W

The goal of this work consists in measuring if the calling of significant genes with the strongest up- and down- regulation in a DGE pipeline is affected by lossily compressing the quality scores associated to RNA-seq data. To get a first assessment of the impact the ranked lists computed by the pipeline for every lossy compression method were compared. In Table 2 the median compression rate in bits per quality score is shown for every compression method along with its confidence interval. Tables 3 and 4 show the ranked lists of log2 fold changes and the associated genes; values in bold are log2 fold changes for whose gene calling was different from the genes indicated in the rightmost column.

Table 4: Ranked list of log2 fold changes for the MCF-7 and genes associated.

	regulation	log2 fold change				gene
			UQ2	UQ8	Quartz	P-/R-Block	
MCF-7	up	5.2348	5.2421	5.2368	5.2324	5.2430	NM_144967
		4.2312	4.2329	4.2319	4.2312	4.2329	NM_014668
		3.8070	3.8309	3.8114	3.8058	3.8430	NM_001555
		3.7533	3.7575	3.7543	3.7516	3.7580	NM_002614
		3.6763	**3.6962**	3.6822	3.6759	3.6863	NM_001170961
		3.5690	**3.6856**	3.6276	3.5676	3.6715	NM_001202474
	down	-7.4730	-7.4970	-7.4778	-7.4722	-7.5012	NM_138780
		-4.9594	-4.9775	-4.9588	-4.9590	-4.9777	NM_001102594
		-4.2973	-4.3204	-4.3020	-4.2963	-4.3232	NM_001207059
		-3.5473	-3.5865	-3.5552	-3.5459	-3.5901	NM_014309
		-3.4331	-3.4554	-3.4369	-3.4323	-3.4581	NM_017851
		-2.5689	-2.5736	-2.5697	-2.5630	-2.5743	NR_131192

5 Discussion and conclusions

Bioinformatics pipelines driven by high-throughput sequencing data are intrinsically complex due to the need to perform measurements on the large variety of heterogeneous data they process. A systematic approach to evaluate their performance will be critically important for their implementation in genomic medicine. Moreover, the challenges associated with the manipulation of large amounts of genomic sequence data are quickly arising as a fundamental obstacle to downstream applications.
In this context, this work sets out to examine the change in the performance of a state-of-the-art bioinformatics pipeline for differential gene expression when applying lossy compression of quality scores associated to RNA-seq data. The proposed workflow enables the objective measurement of the effect of lossy compression of quality scores in the calling of genes with the strongest up- and down- regulation on two real-world sets of RNA-seq data. The obtained results show that the impact of a controlled loss of information when compressing quality scores can be minimized and reduced to zero when calling up- and down- regulated genes in RNA-seq analysis. The next step of this study could consist in investigating how quality scores are actually used by the considered pipelines, so that compression algorithms can be further tuned to generate a minimal perturbation of performance and provide smaller compressed data footprints. A principled approach to efficient compression of quality scores will improve the performance of high-throughput bioinformatics pipelines enabling applications that are not possible today due to the costs in terms of both storage space and bandwidth.

References

[1] U. Nagalakshmi et al., "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science*, vol. 320, no. 5881, pp. 1344–1349, May 2008.

[2] R. Lister et al., "Highly integrated single-base resolution maps of the epigenome in Arabidopsis," *Cell*, vol. 133, no. 3, pp. 523–536, May 2008.

[3] B. Wilhelm et al., "Dynamic repertoire of a eukaryotic transcriptome surveyed at single nucleotide resolution," *Nature*, vol. 453, no. 7199, pp. 1239–1243, June 2008.

[4] A. Mortazavi et al., "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, July 2008.

[5] Li. Sheng et al., "Detecting and correcting systematic variation in large-scale RNA sequencing data," *Nature Biotechnology*, vol. 32, pp. 888–895, 2014.

[6] Y. Liu et al., "RNA-seq differential expression studies: more sequence or more replication?," *Bioinformatics*, vol. 30, no. 3, pp. 301–304, 2014.

[7] Q. Liu et al., "Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data," *International Conference on Intelligent Biology and Medicine*, December 2012.

[8] J. O'Rawe et al., "Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing," *Genome Medicine*, vol. 5, no. 3, March 2013.

[9] S. Hwang et al., "Systematic comparison of variant calling pipelines using gold standard personal exome variants," *Scientific reports*, December 2015.

[10] R. Cánovas et al., "Lossy compression of quality values in genomic data," *Bioinformatics*, vol. 30, no. 15, pp. 2130–2136, August 2014.

[11] Y. Yu et al., "Quality score compression improves genotyping accuracy," *Nature Biotechnology*, vol. 33, no. 3, pp. 240–243, March 2015.

[12] G. Malysa et al., "QVZ: lossy compression of quality values," *Bioinformatics*, vol. 31, no. 19, pp. 3122–3129, May 2015.

[13] D. Greenfield et al., "GeneCodeq: quality score compression and improved genotyping using a Bayesian framework," *Bioinformatics*, June 2016.

[14] C. Alberti et al., "An Evaluation Framework for Lossy Compression of Genome Sequencing Quality Values," *Proceedings 2016 Data Compression Conference (DCC)*, pp. 221–230, March 2016.

[15] I. Ochoa et al., "Effect of lossy compression of quality scores on variant calling," *Briefings in Bioinformatics*, March 2016.

[16] N. Schurch et al., "How Many Biological Replicates Are Needed in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?," *RNA*, vol. 22, pp. 1–13, 2016.

[17] SEQC/MAQC-III Consortium, "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium," *Nature Biotechnology*, vol. 32, no. 9, pp. 903–914, 2014.

[18] D. Kim et al., "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, no. 4, pp. R36, 2013.

[19] D. K̇im et al., "HISAT: a fast spliced aligner with low memory requirements," *Nature Methods*, vol. 12, no. 4, pp. 357–360, 2015.

[20] M. Pertea et al., "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown," *Nature Protocols*, vol. 11, no. 9, pp. 1650–1667, 2016.

[21] K. Hayer et al., "Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data," *Bioinformatics*, vol. 31, no. 24, pp. 3938–3945, 2015.

[22] C. Trapnell et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[23] M. Pertea et al., "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads," *Nature Biotechnology*, vol. 33, no. 3, pp. 290–295, 2015.

[24] R. Patro et al., "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms," *Nature Biotechnology*, vol. 32, no. 5, pp. 462–464, 2014.

[25] N. Bray et al., "Near-optimal probabilistic RNA-seq quantification," *Nature Biotechnology*, vol. 34, no. 5, pp. 525–527, 2016.

[26] M. D. Robinson et al., "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[27] M. I. Love et al., "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, pp. 550, 2014.

[28] T. Griebel et al., "Modelling and simulating generic RNA-Seq experiments with the flux simulator.," *Nucleic Acids Res.*, vol. 20, no. 40, pp. 10073–10083, 2012.