Differential gene expression with lossy compression of quality scores in RNA-seq data

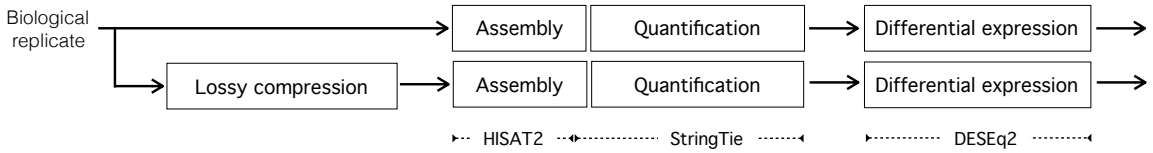
Ana A. Hernandez-Lopez*, Jan Voges†,
Claudio Alberti*, Marco Mattavelli* and Jörn Ostermann†

*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

In the last decade genome sequencing has been the target of prolific and fruitful advances. The development of new methods and the advent of their enabling technology has made possible to sequence trillions of nucleotides from samples of organisms; an increase in machine output by a factor of 10^5 nucleotides. Simultaneously, the cost per genome is on the wane, which is driving the creation of new protocols and applications that attempt to exploit the wealth of genomic data in prospect of biological discovery. The measurements to be inferred from the deluge of sequence data pose important bioinformatic challenges on two fronts in particular. The first pertains to the assessment of effectiveness of the computational methods proposed by bioinformatic tools. The second bears on the manipulation, processing and storage of genomic data in the workflow of bioinformatic pipelines.

Compression strategies have recently been investigated as a mean to alleviate storage of sequence data. Lossy methods are specifically being sought after to boost compression, as is the burgeoning interest to measure their impact [1]. In this work, we present a pilot study to investigate the effect of lossy compression of quality scores in RNA sequence data, and provide a first assessment of the impact on a state-of-the-art pipeline for differential gene expression. The pipeline's layout is shown below.



We ran tests on two real datasets with 12 biological replicates each and measured the calling of significant genes with the strongest up- and down- regulation using the log2 fold change estimate. Our results suggest that high rates of controlled loss of information do not compromise, in principle, the calling of significant genes and show how the impact can be reduced to zero. The next step of this study will investigate the precise use of quality scores in the methods developed by bioinformatic tools for differential gene expression.

[1] 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.