

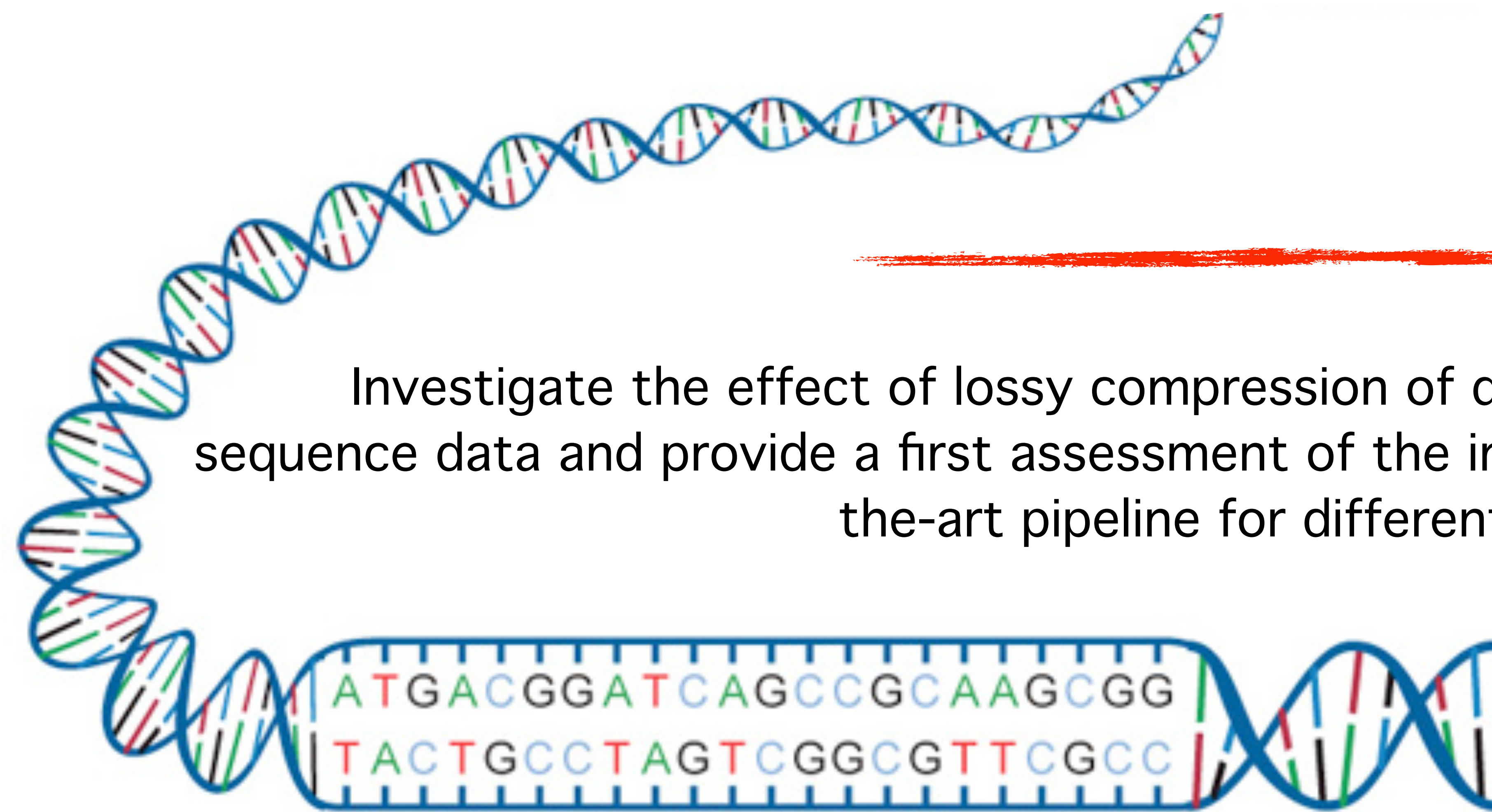
Ana A. Hernandez-Lopez¹, Jan Voges², Claudio Alberti¹, Marco Mattavelli¹ and Jörn Ostermann²

¹ École Polytechnique Fédérale de Lausanne, SCI-STI-MM, Lausanne, Switzerland.

² Leibniz Universität Hannover, Institut für Informationsverarbeitung (TNT), Hannover, Germany.

MOTIVATION

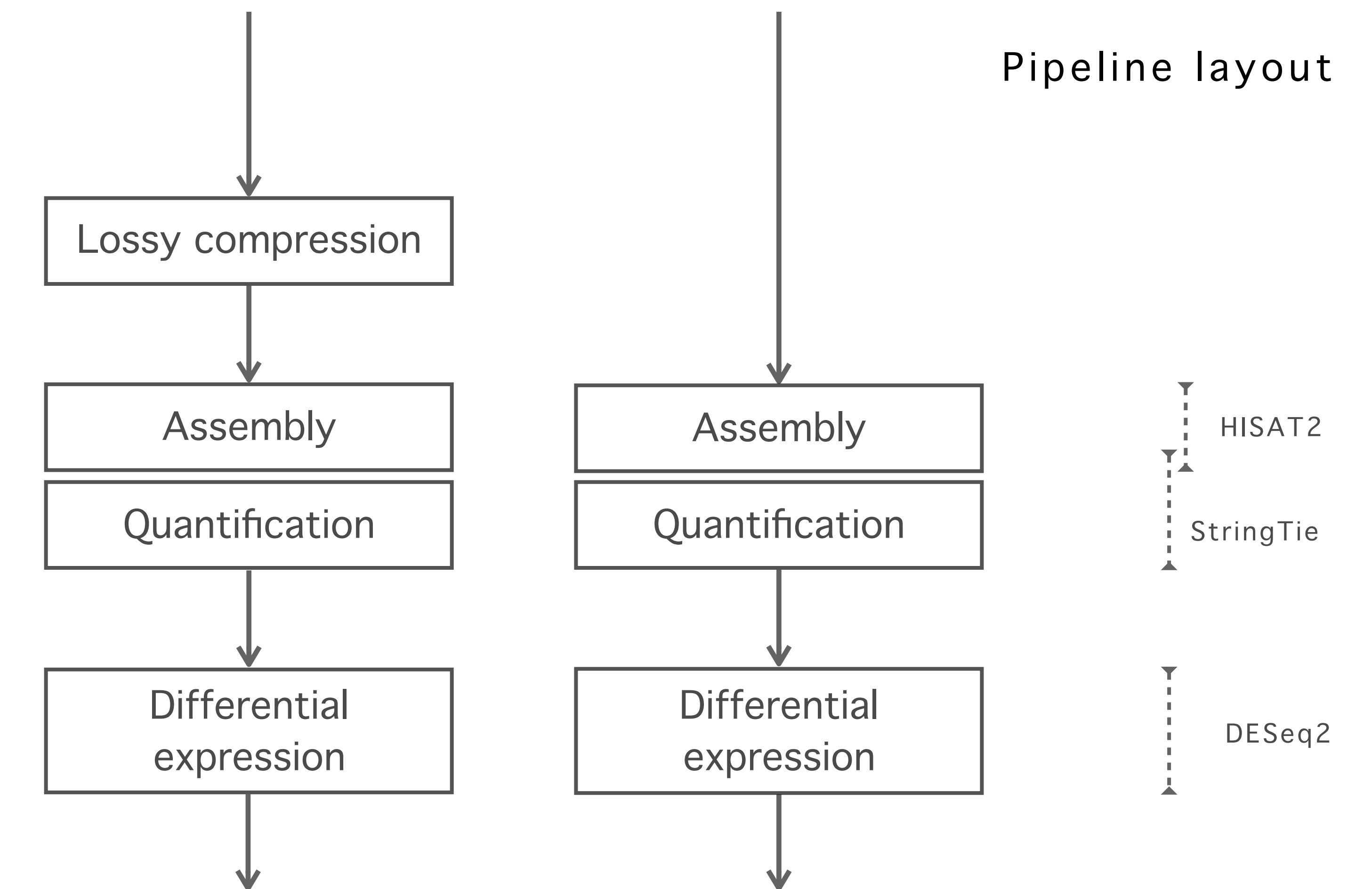
Compression strategies have recently been investigated as a mean to alleviate storage of sequence data. Lossy methods are specifically being sought after to boost compression, as is the burgeoning interest to measure their impact.



OBJECTIVE

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.

Biological replicate



regulation		log2 fold change				gene
		UQ2	UQ8	Quartz	P-/R-Block	
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
down	3.5690	3.6856	3.6276	3.5676	3.6715	NM_001202474
	-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

METHODOLOGY

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.

CONCLUSION

Our results suggest that high rates of controlled loss of information of the quality scores do not compromise, in principle, the calling of significant genes and show how the impact can be reduced to zero.