Transcriptome reconstruction with quality score distortion in reference-based alignment

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

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

Motivation

How to cope with the deluge of genomic data?

Sequencing technology

Protocols
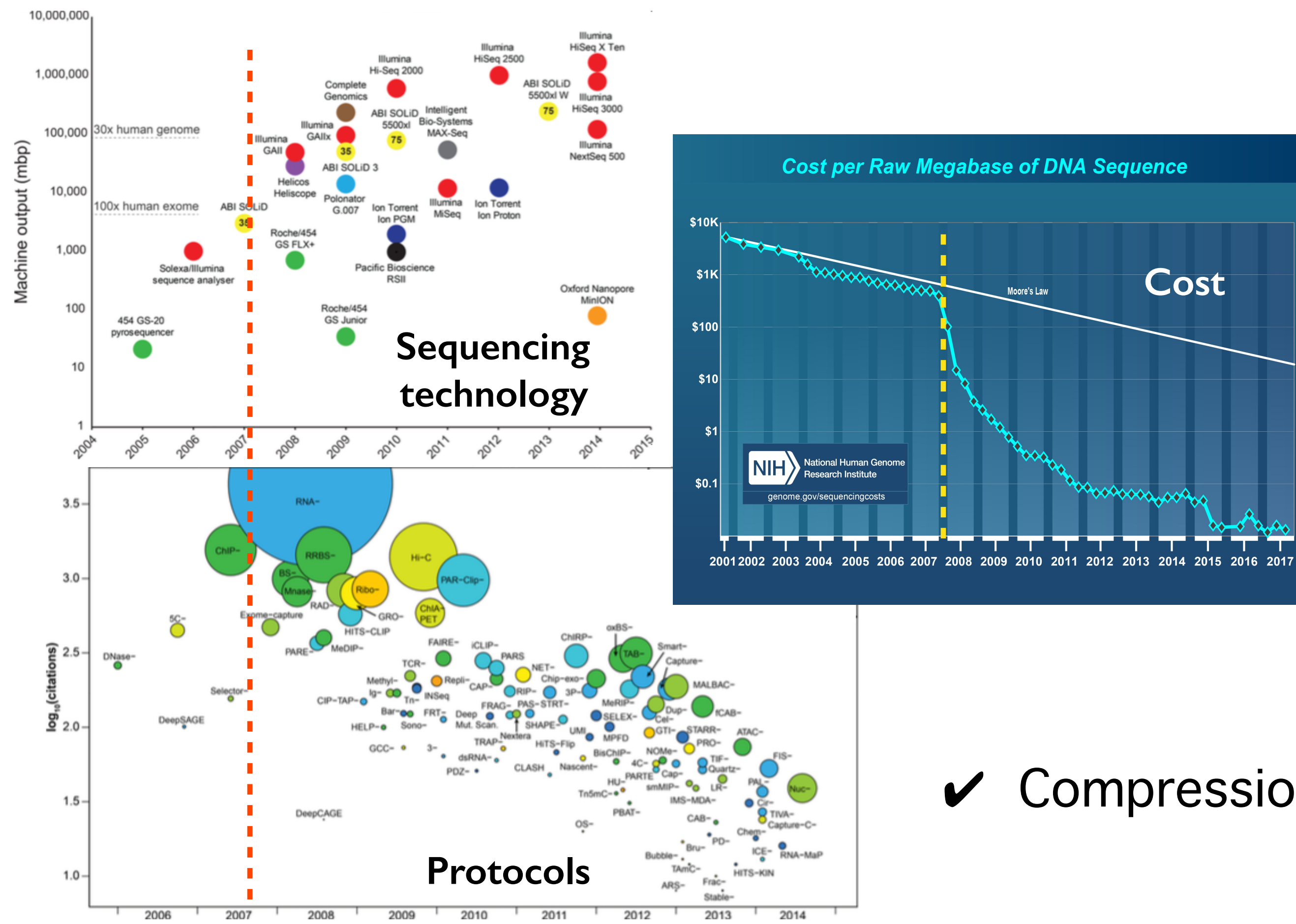
✔ Compression

Reuter A., et al. "High-throughput sequencing technologies". Molecular Cell, 2015

Is all data informative? Heed the quality scores (QS)

bases 40%

quality scores 60%
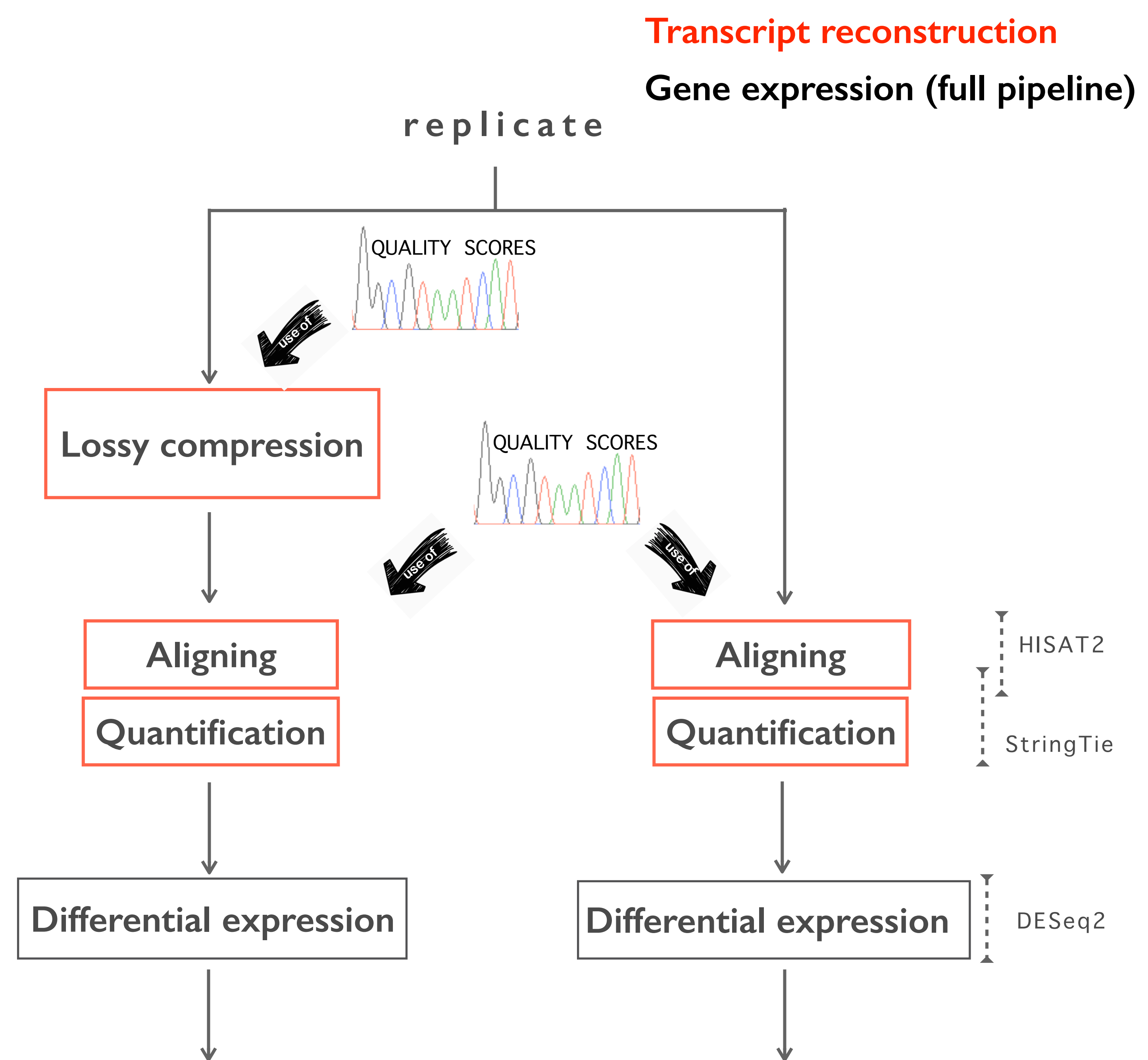
TTGGCGTAATCATGGTCATAGCTGTTTCCTGT
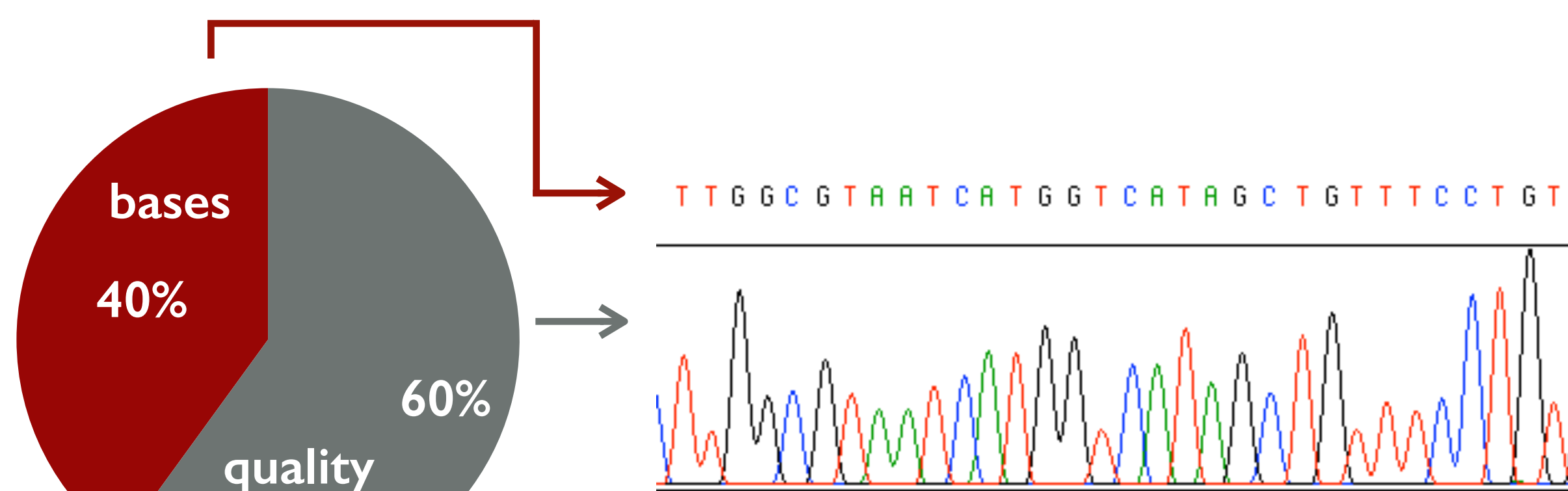
✔ Lossy compression of QS

What downstream applications can deal with QS distortion?

Objectives

- Application of lossy compression in a candidate application for differential gene expression (DGE)
- Evaluate impact of lossy QS on transcript reconstruction
- Measure the effect of lossy QS on the calling of significant genes

Methods

Transcript reconstruction

Gene expression (full pipeline)

replicate

QUALITY SCORES

Lossy compression

QUALITY SCORES

Aligning — Quantification → Differential expression

Aligning — Quantification → Differential expression

HISAT2

StringTie

DESeq2

Results

Transcript reconstruction — simulated sample —

	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
Lossy compression — P-Block	78.73	78.91	80.61	0.98

Alignment percentage
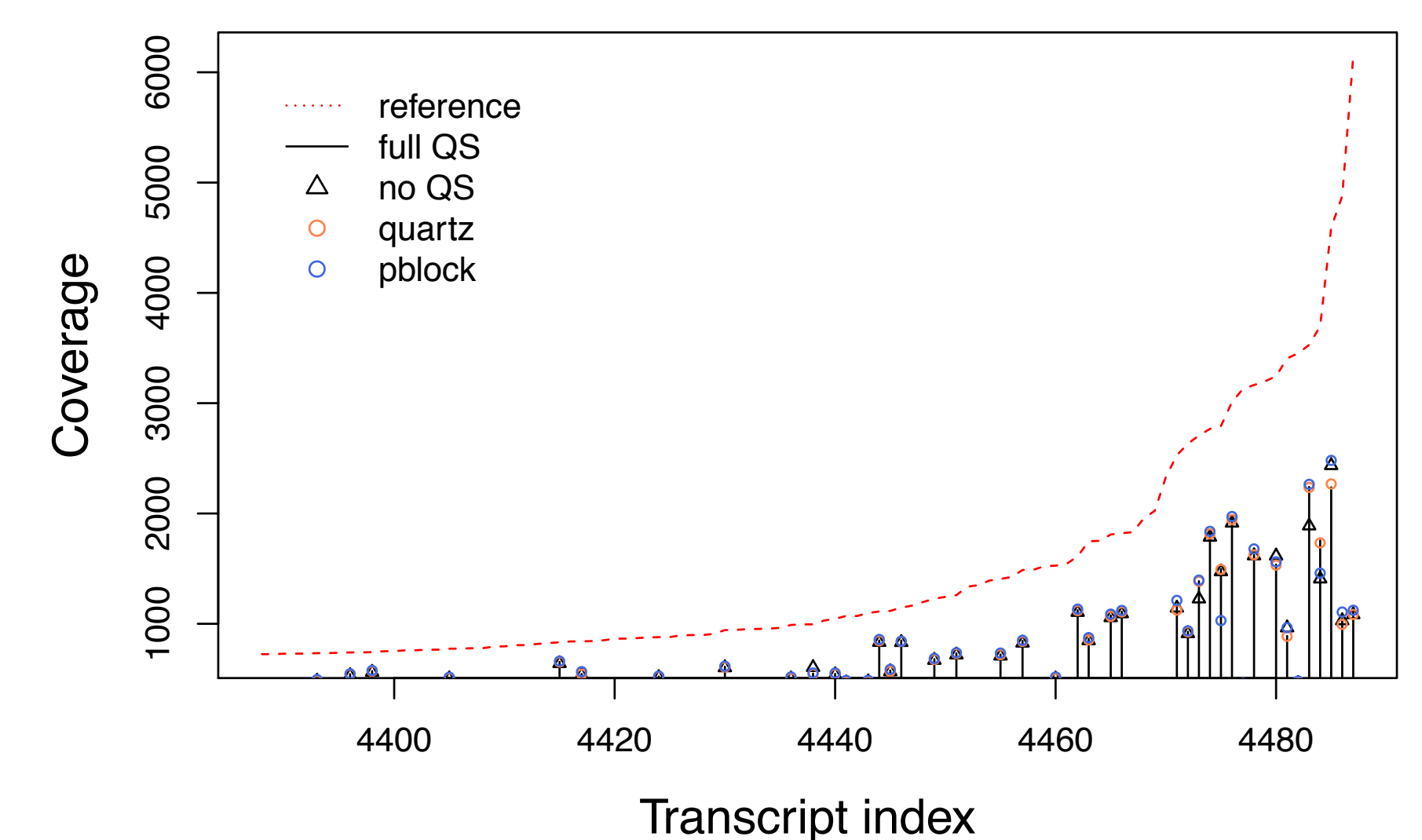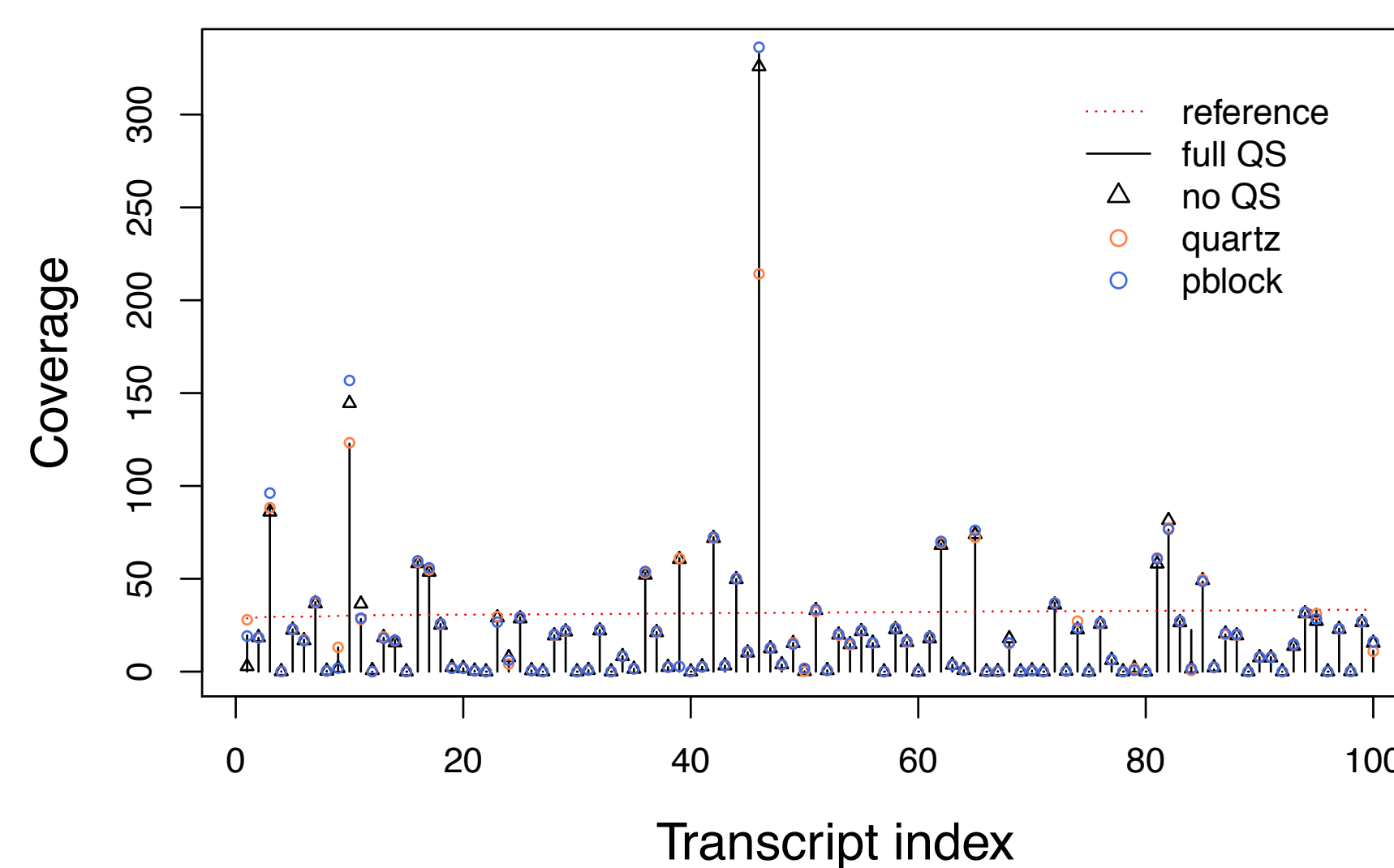
Gene expression — real samples, six replicates —

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

Compression rate

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

Ranked list of expressed genes

Conclusions

We find that differential gene expression (with a quality-aware assembler) is a promising application over which to use lossy quality score compression. This is supported by the observation that the calling of the most salient and discernible genes are affected by their compression.