Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level – Supplementary Materials

Yang Liao [1,2] and Wei Shi [1,2,3,4]

[1]Olivia Newton-John Cancer Research Institute, Heidelberg, Victoria 3084, [2]School of Cancer Medicine, La Trobe University, Heidelberg, Victoria 3084, [3]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052 and [4]School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria 3010, Australia

This document includes Supplementary Tables S1 - S6.

Supplementary Table S1. Comparison of read bases trimmed off by read trimmers and read bases soft-clipped by Subread. Subread was run on the untrimmed reads. Columns 'Trimmed' give the percentages of read bases that were trimmed off by each trimming method, columns 'Soft-clipped in trimmed' the percentages of trimmed bases that were also soft-clipped by Subread and columns 'Adapter in soft-clipped & trimmed' the percentages of adapter bases found in all the read bases that were commonly removed by a read trimmer and Subread. Adapter percentage of a trimmer was calculated using the adapter bases reported by the same trimmer.

Method	UHRR			HBRR		
	Trimmed (%)	Soft-clipped in trimmed (%)	Adapter in soft-clipped & trimmed (%)	Trimmed (%)	Soft-clipped in trimmed (%)	Adapter in soft-clipped & trimmed (%)
Trimmomatic–adapters and SW	4.6	19.0	10.3	4.4	17.8	11.3
Trimmomatic–adapters and MI	4.0	20.7	11.0	3.9	19.2	11.7
TrimGalore	2.3	29.2	26.9	2.3	27.7	26.6

Supplementary Table S2. Adapters reported by each read trimmer and percentages of adapter bases soft-clipped by Subread. Columns 'Adapter bases(%)' give the percentage of read bases in a library that were called as adapter bases by a trimmer, and columns 'Soft-clipped adapters(%)' give the percentage of adapter bases that were soft-clipped by Subread.

Method	UHRR		HBRR	
	Adapter bases(%)	Soft-clipped adapters(%)	Adapter bases(%)	Soft-clipped adapters(%)
Trimmomatic–adapters and SW	0.10	93.8	0.09	93.5
Trimmomatic–adapters and MI	0.10	93.8	0.09	93.5
TrimGalore	0.57	32.0	0.55	29.9

Supplementary Table S3. Mapping concordance of reads before and after trimming (or trimmed by different methods). The Subread aligner was used in the read mapping. Reads were mapped to the human reference genome GRCh38/hg38. Concordantly mapped reads are those reads that have a <100bp change in their mapping start position after trimming (or trimmed with a different method), or have a mapping start change of 100bp or more but still map to the same gene (ie. map to a different exon of the same gene).

Method 1	Method 2	% Concordantly mapped reads	
		UHRR	HBRR
No trimming	TrimGalore	98.4	98.3
No trimming	Trimmomatic-adapters and SW	96.1	96.2
No trimming	Trimmomatic-adapters and MI	97.4	97.5
TrimGalore	Trimmomatic-adapters and SW	96.2	96.3
TrimGalore	Trimmomatic-adapters and MI	97.8	97.8
Trimmomatic-adapters and SW	Trimmomatic-adapters and MI	97.5	97.7

Supplementary Table S4. Correlation of trimmed and untrimmed RNA-seq data with the truth in the simulation data. Shown are the coefficients of Pearson correlation between $\log_2$-RPKM expression values of 28,288 genes calculated from the RNA-seq data processed by each method and the true $\log_2$-RPKM expression values of these genes generated in the simulation. Three simulation datasets with different levels of adapter contamination were included in this analysis. Percentage of adapter bases (adapter-containing reads) in each dataset is indicated in the headers of last three columns of the table.

	% adapter bases (% adapter reads)		
Method	0.1(1.1)	0.5(5.5)	1(11.0)
No trimming + Subread	0.957	0.957	0.957
Trimmomatic–adapters and SW + Subread	0.957	0.957	0.957
Trimmomatic–adapters and MI + Subread	0.957	0.957	0.957
TrimGalore + Subread	0.957	0.957	0.957

Supplementary Table S5. Total amount of disk space used by each method. A UHRR library containing 15 million 100bp read pairs was used in this evaluation. The consumed disk space includes the storage of raw reads, trimmed reads and mapping results (BAM file). Raw reads and trimmed reads are both in gzipped FASTQ format.

Method	Disk usage (GB)
No trimming + Subread	5.5
Trimmomatic–adapters and SW + Subread	7.6
Trimmomatic–adapters and MI + Subread	7.7
TrimGalore + Subread	7.8

Supplementary Table S6. Correlation of SEQC RNA-seq data mapped by STAR with the TaqMan RT-PCR data. Untrimmed SEQC RNA-seq reads were provided for mapping. STAR was run with default setting or with the option –alignEndsType EndToEnd. Shown are the coefficients of Pearson correlation between $\log_2$ expression values of 949 genes measured by the TaqMan RT-PCR technique and their RNA-seq expression levels generated from using each method ($\log_2$-RPKM).

Method	UHRR	HBRR
STAR (default)	0.848	0.867
STAR End-to-End	0.847	0.866
Subread	0.851	0.870