Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Gene expression

SUPPLEMENTARY INFORMATION TO NVT: a fast and simple tool for the assessment of RNA-Seq normalization strategies

Thomas Eder [1,2], Florian Grebien [2] and Thomas Rattei [2,*]

[1]Ludwig Boltzmann Institute for Cancer Research, Währinger Straße 13A, 1090 Vienna, Austria.
[2]CUBE Division of Computational Systems Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Althanstraße 14, 1090 Vienna, Austria.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

SUPPLEMENTARY TABLES

Table S1: List of normalization methods.

Table 1. List of normalization methods.

Method	Abbreviation	Description
Total count	TC	This is one of the simplest normalization methods,which divides the number of mapped reads per gene (i.e. the number of sequence reads that could be aligned to the reference sequence of a gene) by the total number of mapped reads and multiplies the result by the mean of total mapped reads of the compared samples.
Median	ME	Very similar to TC, ME uses the median of mapped reads without zero counts, instead of the total number of mapped reads.
Upper quartile (Bullard *et al*., 2010)	UQ	Similar to TC, UQ uses the 75th percentile excluding zero read genes, instead of the total number of mapped reads.
Trimmed mean of M-values (Robinson and Oshlack, 2010)	TMM	After removal of genes with the highest log expression ratios between samples and the genes with highest expression, the weighted mean of log ratios between the compared samples is used as scaling factor.
Relative log expression method implemented in DESeq (Anders and Huber, 2010)	DESeq	This method uses a scaling factor for a sample (or lane). Per gene, the ratio read count to its geometric mean across all samples (lanes) is calculated and the median of this ratio is used.
Quantile (Bolstad *et al*., 2003; Smyth *et al*., 2005)	Q	Uses the distributions of the mapped reads per gene counts and equalizes them across lanes or sequencing runs. The assumption of this method is that the distributions can be matched.
Reads per kilobase per million mapped reads (Mortazavi *et al*., 2008)	RPKM	In addition to correcting for different library sizes, RPKM also corrects for gene length. The number of mapped reads per gene is divided by the total number of million mapped reads (which represents RPM) and then divided by the gene length in kilobases to yield RPKM. This method assumes that the total mRNA amount was identical in the compared samples. The fragments per kilobase million (FPKM) value is similar, but mainly used for applications of paired-end sequencing, where two paired-reads can map.
Reads per million mapped reads	RPM	The number of mapped reads per gene is divided by the total number of million mapped reads.
Transcripts per million mapped reads (Li *et al*., 2010)	TPM	For TPM normalization, the read counts are divided by the gene length (in kilobases) and then divided by a scaling factor, which is the sum of the read per kilobase values.
Normalization by a defined gene set (Bullard *et al*., 2010)	G	The use of housekeeping genes is standard for normalizing qRT-PCR expression data. The underlying assumption is that the expression of the chosen gene set is not influenced by the conditions that change in the experiment.

References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Bolstad, B.M., Irizarry, R., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185-193.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493-500.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621-628.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.

Smyth, G.K. (2005). limma: Linear Models for Microarray Data. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit (ed.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, New York, pp. 397-420.