Bioinformatics Advance Access published August 30, 2016

*Bioinformatics*, 2016, 1–3
doi: 10.1093/bioinformatics/btw521
Advance Access Publication Date: 11 August 2016
Applications Note

Gene expression

# NVT: a fast and simple tool for the assessment of RNA-seq normalization strategies

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

[1]Ludwig Boltzmann Institute for Cancer Research, Vienna, 1090, Austria and [2]CUBE Division of Computational Systems Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, 1090, Austria

*To whom correspondence should be addressed.

Associate Editor: Cenk Sahinalp

Received on February 18, 2016; revised on May 9, 2016; accepted on August 3, 2016

## Abstract

**Motivation:** Measuring differential gene expression is a common task in the analysis of RNA-Seq data. To identify differentially expressed genes between two samples, it is crucial to normalize the datasets. While multiple normalization methods are available, all of them are based on certain assumptions that may or may not be suitable for the type of data they are applied on. Researchers therefore need to select an adequate normalization strategy for each RNA-Seq experiment. This selection includes exploration of different normalization methods as well as their comparison. Methods that agree with each other most likely represent realistic assumptions under the particular experimental conditions.

**Results:** We developed the NVT package, which provides a fast and simple way to analyze and evaluate multiple normalization methods via visualization and representation of correlation values, based on a user-defined set of uniformly expressed genes.

**Availability and Implementation:** The R package is freely available under https://github.com/Edert/NVT

**Contact:** thomas.rattei@univie.ac.at

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

High throughput sequencing of RNA or cDNA (RNA-Seq) had an enormous impact on basic and clinical research since its introduction in the 2000s (Wang *et al.*, 2009; Wilhelm and Landry, 2009). Independently of the used sequencing technology, the vast majority of research projects attempting to measure global expression levels of features (genes, exons, small RNAs or non-coding RNAs) compare expression values of multiple samples that represent different biological states. Importantly, any such differential expression (DE) analysis requires normalized data. This means that all non-biological influence, such as potential effects of sample preparation or sequencing efficiency, has to be removed to make the data comparable in between different experiments. To balance sequencing depths several methods use scaling factors (Dillies *et al.*, 2013) (Supplementary Table S1): Total count (TC), Median (ME), Upper quartile (UQ), Trimmed mean of M-values (TMM) and the relative log expression method implemented in DESeq. Both TMM and DESeq operate under the assumption that most of the genes are not differentially expressed. Normalization methods without scaling factors are (Dillies *et al.*, 2013) (Supplementary Table S1): Quantile (Q), Reads per kilobase per million mapped reads (RPKM) and normalization by a defined gene set (G). Thus, two main concepts for data normalization in RNA-seq applications exist. While TMM and DESeq mainly consider differential library size, other normalization methods account for the distribution adjustment of read counts (TC, UQ, ME, Q, RPKM). Normalization based on RNA spike-ins (Lovén *et al.*, 2012) makes other methods obsolete but this requires the RNA spike-in which had to be planned and applied previous to the sequencing. All the previously described methods are based on specific assumptions. Thus, identifying the method(s) for which these assumptions agree with the specific experimental setting represents a significant challenge, an exception is quantro (Hicks and

Irizarry, 2015) which gives recommendations on when to use Q normalization or not. For example: if an experiment compares gene expression levels of healthy versus rapidly growing tumor cells, the assumptions of non-differentially expressed genes or equal amounts of mRNA might not apply. The decision to utilize a certain normalization method can therefore have an enormous impact on the entire downstream analysis. Also, conclusions drawn from the enrichment of differentially expressed genes with respect to functional categories might be severely affected. As RNA-Seq experiments have become popular and powerful research tools in many areas of biology and medicine, also non-specialists need to be able to explore and compare different normalization methods to select the most appropriate one. To assist researchers in these tasks we present the normalization visualization tool (NVT). NVT is a fast and simple way to visually and quantitatively assess the normalization strategy including a set of user-defined genes. This set should consist of genes that do not change their relative expression levels in the particular study, this requires preliminary knowledge or experimental data (e.g. from quantitative PCR measurements).

## 2 Methods and implementation

NVT is an easy to use and freely available R package that provides visualization and evaluation of 10 different normalization methods for the comparison of two RNA-Seq data samples provided by the user (use case and detailed description in vignette, see supplement). It works with raw expression values per feature, may it be genes, exons or short RNA, originating from RNA-Seq datasets. The expression data of two samples have to be provided as a table of features and their respective number of mapped reads. NVT includes the normalization methods TC, ME, TMM, UQ, the upper quartile implementation from the NOISeq (Tarazona *et al.*, 2011) package (UQ2), Q, RPKM, RPM, TPM, DESeq and G. If required for comparison reasons, also no normalization (N) can be applied. For some of these methods (RPKM and TPM), in addition to the expression values per feature, the respective gene length is also required as input. It can be provided as list or directly uploaded from a gff or gtf annotation-file via the GenomicRanges (Lawrence *et al.*, 2013) and rtracklayer (Lawrence *et al.*, 2009) packages. NVT allows to compare and evaluate normalization methods based on genes that are expected to be equally expressed in both samples. The members of this set of control genes can be visualized and used for normalization (by using 'G' as normalization parameter). The different normalization methods are evaluated via a plot function and a function which calculates correlation based on the control gene set. The correlation can serve as a main criterion for the evaluation of the performance of any normalization method implemented in NVT. Available correlation functions are the Pearson-correlation coefficient, the root-mean-square-deviation (RMSD) and the mean-absolute-error (MEA). The normalization methods can be applied individually or all methods can be applied in one step and the resulting correlation values are presented in a ranked list. If required, the normalized expression per feature can also be extracted. The basic plot function generates a scatter-plot of the normalized expression data of two RNA-Seq samples. Based on the selected control gene set, whose members are highlighted in the scatter-plot, a linear model is calculated and plotted as a red line (the linear model can also be retrieved via the respective function). If the control genes are stably expressed, the red line will overlap with the gray dashed diagonal line. The advanced plot function requires ggplot2 (Wickham, 2009) for additional density bars. This function is illustrated for human gene expression data from the airway package (Himes *et al.*, 2014) (Fig. 1).

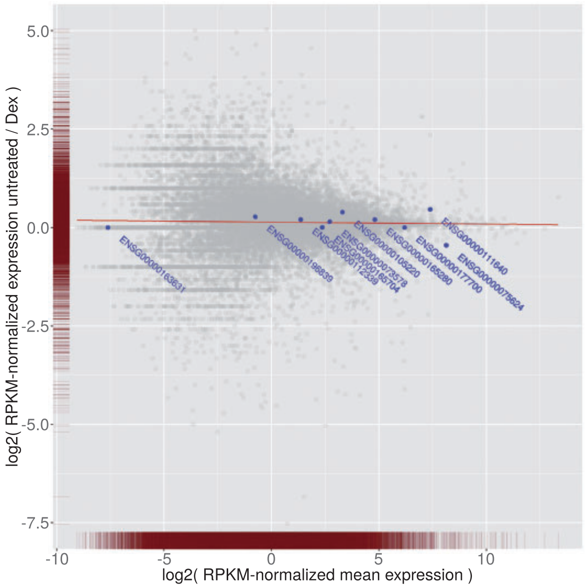

**Fig. 1.** MA-plot of RPKM normalized gene expression of N61311 cells untreated versus treated with the anti-inflammatory drug dexamethasone (Dex). The expression levels per gene are represented as gray dots, the user defined set of control genes (*GAPDH, ALB, ACTB, HPRT1, ADA, POLR2L, VCP, GPI, HBS1L* and *SDHA*) are shown in blue and the linear model calculated based on expression of the control gene set is depicted as red line

In addition, NVT also offers the possibility to compare different normalization methods in between replicates. If two biological replicates are compared, all data points (including the defined gene set) would ideally reside on the diagonal, indicated by the dashed gray line in the scatter plot. The nearer a data point is located to the diagonal line, the better its correlation of this particular feature is in the two samples.

## 3 Conclusions

The appropriate assumption(s) for correct normalization of RNA-Seq data critically depend on the nature of the particular experimental setup. NVT is a simple and fast tool to evaluate the normalization strategy for any given RNA-Seq data set. The visual comparison of normalization methods to a user-defined gene set of control genes in NVT is an efficient and intuitive way to assess the performance of different normalization methods. NVT generates publication-ready figures and also provides correlation measures. The package thereby facilitates the documentation of methodological decisions for RNA-Seq experiments. NVT is hosted on https://github.com, using its infrastructure for maintenance and bug tracking, updates and release of future versions. After assessing the demands of users, possible improvements and additional functions will be: the implementation of additional normalization methods and the possibility to include custom normalized data.

## Funding

This work was supported by European Research Council (ERC) starting grant 'ONCOMECHAML' (to T.E. and F.G.). were supported by the starting grant" ONCOMECHAML" from the European Research Council (ERC).

*Conflict of Interest*: none declared.

## References

Dillies,M.A. *et al*. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform*., **14**, 671–683.

Hicks,S. and Irizarry,R. (2015) Quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome Biol*., **16**, 117.

Himes,B.E. *et al*. (2014) RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PLoS ONE*, **9**, e99625.

Lawrence,M. *et al*. (2009) Rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, **25**, 1841–1842.

Lawrence,M. *et al*. (2013) Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol*., **9**, e1003118.

Lovén,J. *et al*. (2012) Revisiting Global Gene Expression Analysis. *Cell*, **151**, 476–482.

Tarazona,S. *et al*. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res*., **21**, 2213–2223.

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*., **10**, 57–63.

Wickham,H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*, Springer: New York.

Wilhelm,B.T. and Landry,J.R. (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, **48**, 249–257.