***Normalization July 24, 2014***

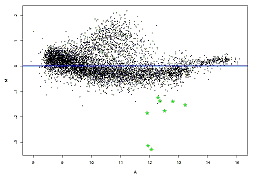
edgeR - Empirical analysis of digital gene expression data in R

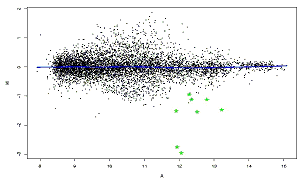
limma - Linear Models for Microarray Data

DESeq2 - Differential gene expression analysis based on the negative binomial distribution

VSN - variance stabilization normalization

RSN - Robust spline normalization





<http://www.people.vcu.edu/~mreimers/OGMDA/normalize.expression.html>

[locally weighted linear regression (loess/lowess) curve]

From Karpievitch “Normalization and missing value imputation for label-free LC-MS analysis”

*Lowess (aka robust scatter plot smoothing) performs local linear fits dependent on the user-defined fraction of points to be used for smoothing at each point. The fraction value is mostly arbitrary and suboptimal value may reduce the efficiency of the normalization and result in poorly normalized data. Empirically selected value of 0.4 has been used in several studies [7,10]. Berger et al. proposed an optimization-based procedure for estimating the fraction value [14]. Scatterplot-smoothing techniques are able to capture non-linear intensity-dependent biases and are therefore more flexible than global adjustments.*

From limma manual:

*Loess normalization assumes that the bulk of the probes on the array are not differentially*

*expressed. It doesn’t assume that that there are equal numbers of up and down regulated*

*genes or that differential expression is symmetric about zero, provided that the loess fit is*

*implemented in a robust fashion, but it is necessary that there be a substantial body of probes*

*which do not change expression levels. Oshlack et al [23] show that loess normalization can*

*tolerate up to about 30% asymmetric differential expression while still giving good results.*

*This assumption can be suspect for boutique arrays where the total number of unique genes*

*on the array is small, say less than 150, particularly if these genes have been selected for being*

*specifically expressed in one of the RNA sources.*

trimmed mean of M-values (TMM)

“normalizes for RNA composition by finding a set of scaling factors for the library sizes that minimize the log-fold changes between the samples for most genes.” This finds a correction factor which is used to compute effective library sizes, which is then entered into the statistical model. (Very important to use raw data at start.) Default: trim M-values by 30% and A-values by 5%, therefore need numbers to be correct.

From edgeR manual – since is concerned with differential expression, most technical factors will cancel out across experiment (e.g., gene length and GC content), and therefore isn’t as much of a concern. The TMM paper shows that seemingly lower false positives if don’t attempt to control for gene length.

Quantile

We assume that the underlying distributions across all of the samples should be the same, because the changes in expression across samples should be minor, i.e. there is likely a small number of genes that show moderate or large changes in expression from the rest, but the distribution of expression shouldn't change. Quantile normalization forces the distributions to be the same based on making the quantiles equal across the samples.

**Count based packages**

EdgeR (TMM) Trimmed Mean Method

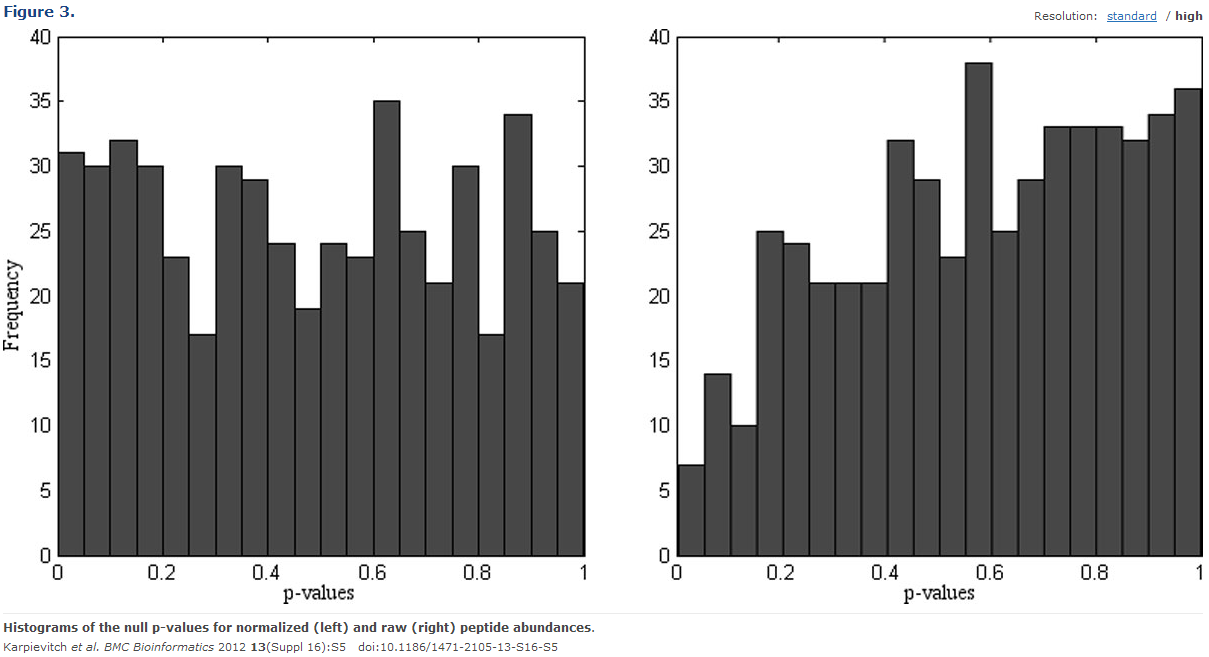
DESeq2 - median-of-ratios (term from Simon Anders)

DEGSeq (poisson)

baySeq

Limma (with voom)  
(Proteomics – if more than 1000 features most of these test are applicable)

**How to check normalization?**



PCA (plenty of arguments)

Distribution of p-values (Karpievitch)

Number of uniform genes (as metric for success, is Glusman 2013)

MSE and K-S statistics to measure the fitness of normalization methods (Garmir 2012)

From Garmire 2012 “Evaluation of normalization methods in mammalian microRNA-Seq data”

*Can be classified into two categories, according to the application of linear scaling or not. (1) The first category includes scaling, global, Lowess, and TMM. Scaling normalization assumes the ranges of data are the same and that the noise and the stochastic variations of microRNAs are proportional to the signal intensity (*[*Smyth et al. 2003*](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3358649/#B24)*). Global normalization is another linear scaling approach that scales all the data of the experimental condition against the control condition by a factor of the difference in the means of two data (*[*Smyth et al. 2003*](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3358649/#B24)*). Lowess normalization does not use a global scaling factor; instead, it calculates local scaling factors within a certain window size (*[*Smyth et al. 2003*](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3358649/#B24)*). TMM, a more recent normalization method applied to mRNA-Seq data, also assumes the majority of the mRNAs in NGS output are similar, except the data points that lie within the extreme M-value and A-value ranges. It derives a simple scaling factor after trimming the data points located in extreme M-value and A-value ranges (*[*Robinson et al. 2010*](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3358649/#B21)*). (2) The second category includes quantile, VSN, and INV. Quantile normalization is nonscaling and assumes that the overall distribution of signal intensity does not change ([Bolstad et al. 2003](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3358649/" \l "B04)). VSN assumes that most microRNAs do not change and transform the data such that the transformed variance is constant among different expression levels. Therefore, it allows better precision in low expression regions, which generally suffer from greater variance (*[*Huber et al. 2002*](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3358649/#B10)*). INV assumes that a subpopulation of expressed microRNAs does not change, and it learns a set of “invariants” through algorithms, instead of assigning “housekeeping genes” subjectively (*[*Perkins et al. 2007*](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3358649/#B17)*; [Pradervand et al. 2009](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3358649/" \l "B18)).*

Conclusion of this paper: “Lowess normalization and quantile normalization are recommended for normalizing microRNA-Seq data, whereas the TMM method should be used with caution.” Probably because of 100s (or thousands) of miRNAs. TMM needs 10s of thousands.

**Dillies**

1-5 are count based; 6 and 7 are based on library size

Not included is GC content (as well as gene length shouldn’t be important in context of DE)

Normalization Methods Assessed:

1. Total Count (TC)

2. Upper Quartile (UQ)

3. Median (Med)

4. Quantile (Q)

5. Reads per Kilobase per Million mapped reads (RPKM)

6. DESeq

7. Trimmed Mean of M-values (TMM)

**1. Total Count (TC)**

𝑔𝑒𝑛𝑒 𝑐𝑜𝑢𝑛𝑡 × 𝑚𝑒𝑎𝑛 𝑡𝑜𝑡𝑎𝑙 𝑐𝑜𝑢𝑛𝑡 𝑎𝑙𝑙 𝑠𝑎𝑚𝑝𝑙𝑒𝑠

𝑡𝑜𝑡𝑎𝑙 𝑐𝑜𝑢𝑛𝑡𝑠 𝑡ℎ𝑖𝑠 𝑠𝑎𝑚𝑝𝑙𝑒

Assumes the total counts should have been the same; heavily influenced by high-count genes that may not be consistent across samples.

**2. Upper Quartile (UQ)**

Same as TC, but “total counts are replaced by the upper quartile of counts different from 0”

After throwing out the top 25% of highly expressed genes, assumes the total remaining counts should have been the same

**3. Median (Med)**

“Also similar to TC, the total counts are replaced by the median counts different from 0”

Assumes the median gene count value should have been the same.

note: differences with median and UQ due to differences in numbers of high counts or zeros.

**4. Quantile (Q)**

Taken from microarray analyses, each sample is adjusted to have the same distribution of counts.

Assumes the overall distribution of counts should have been the same (can lead to increased within-condition variability)

note: tended to identify weakly expressed genes as DE

**5. Reads Per Kilobase per Million mapped reads (RPKM) [aka FPKM]**

𝑔𝑒𝑛𝑒 𝑐𝑜𝑢𝑛𝑡 × 103 x 106

𝑔𝑒𝑛𝑒 𝑙𝑒𝑛𝑔𝑡ℎ 𝑡𝑜𝑡𝑎𝑙 𝑙𝑖𝑏𝑟𝑎𝑟𝑦 𝑠𝑖𝑧𝑒

Assumes the total counts should have been the same and that gene length differences should be accounted for

note: tend to identify weakly expressed genes as DE

**6. DESeq – median-of-ratios approach**

“…the median of the ratio, for each gene, of its read count over its geometric mean across all [samples]”

Assumes that most genes are not changing, so the measured difference of most genes can be used to adjust the bias in total library size

**7. Trimmed Mean of M-values (TMM)**

Pick one sample as the reference and calculate the M-values (log ratio) for each of the other samples. Calculate the weighted mean of the M-values after trimming out the genes with the highest expression (A-value) and largest M-values. (relative to total read count)

Assumes that most genes are not changing, so the measured difference of most genes can be used to adjust the bias in total library size

**Performance on simulated data**

If library size is equivalent and there aren’t a lot of high counts, all methods (including no method), perform similarly. When library size is different, DESeq and TMM maintain low FDR and good power. TC falls apart with high-count genes, and Q can actually cause more variability when libraries have different distributions, as these can be over-corrected. This leaves UQ, Med, DESeq and TMM for dealing well with variations (again, library size differences are best with DESeq and TMM). Based on miRNA paper though, probably stick with UQ or Med for small library (100s).

**General Suggestions:**

* Try multiple methods and chose common DE (compare rankings too… binned?)
* Eliminate DE prior to normalization approach (Median or UQ method), then re-incorporate. TbT by Kadota is norm with TMM, analyze with baySeq, then remove DE and re-normalize (TMM-baySeq-TMM). Is now implemented in TCC (Sun et al. 2013)