**A scaling normalization method for differential expression analysis of RNA-seq data (2010)**

Scaling to library size as a form of normalization makes intuitive sense, given it is expected that sequencing a sample to half the depth will give, on average, half the number of reads mapping to each gene. This makes sense… right?)

We believe this is appropriate for normalizing between replicate samples of an RNA population.

The number of tags expected to map to a gene is not only dependent on the expression level and length of the gene, but also the composition of the RNA population that is being sampled.

Thus, if a large number of genes are unique to, or highly expressed in, one experimental condition, the sequencing 'real estate' available for the remaining genes in that sample is decreased. If not adjusted for, this sampling artifact can force the DE analysis to be skewed towards one experimental condition. (??)

Imagine we have a sequencing experiment comparing two RNA populations, A and B. In this hypothetical scenario, suppose every gene that is expressed in B is expressed in A with the same number of transcripts. However, assume that sample A also contains a set of genes equal in number and expression that are not expressed in B.

Therefore A= 2x transcripts, B = x

Suppose that each sample is then sequenced to the same depth. Without any additional adjustment, a gene expressed in both samples will have, on average, half the number of reads from sample A, since the reads are spread over twice as many genes. Therefore, the correct normalization would adjust sample A by a factor of 2. The key here seems to be “Expression Profile” and Proportion.

The hypothetical example above highlights the notion that the proportion of reads attributed to a given gene in a library depends on the expression properties of the whole sample rather than just the expression level of that gene.

Holy shit.

**Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments (2010)**

We investigate the impact of the read count normalization method on DE results and show that the standard approach of scaling by total lane counts (e.g., RPKM) can bias estimates of DE.

As different lanes have different total read counts, i.e., sequencing depths, the usual approach is to scale gene counts within each lane by the total lane count: e.g., the now standard reads per kilobase of exon model per million mapped reads (RPKM)  (Shit this is the Mortazavi one).

We show that this form of global normalization is heavily affected by a relatively small proportion of highly-expressed genes and, as such, can give biased estimates of DE if these few genes are differentially expressed across the conditions under comparison

**Methods:**

We evaluate three types of global normalizations: (1) total lane counts, as in RPKM of [[7](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-94#CR7)], (2) per-lane counts for a "housekeeping" gene expected to be constantly expressed across biological conditions, e.g., POLR2A, (3) per-lane upper-quartile of gene counts for genes with reads in at least one lane

In essence, these global scaling factors define the null hypothesis of no differential expression: if a gene has the same proportions of counts across lanes as the proportions determined by the vector of di's, then it is deemed non-differentially expressed

The simplest form of normalization is achieved by scaling gene counts, in lane i, by a single lane-specific factor di..

**A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis(2012)**

To this end, we propose a systematic comparison of seven representative normalization methods for the differential analysis of RNA-seq data: Total Count (TC), Upper Quartile (UQ) [16], Median (Med), the DESeq normalization implemented in the DESeq Bioconductor package [14], Trimmed Mean of M values (TMM) implemented in the edgeR Bioconductor package [17], Quantile (Q) [22, 23] and the Reads Per Kilobase per Million mapped reads (RPKM) normalization [19].

Total count (TC): Gene counts are divided by the total number of mapped reads (or library size) associated with their lane and multiplied by the mean total count across all the samples of the dataset. (Note this is what Liling is doing)

Trimmed Mean of M-values (TMM): This normalization method [17] is implemented in the edgeR Bioconductor package (version 2.4.0). It is also based on the hypothesis that most genes are not DE. The TMM factor is computed for each lane, with one lane being considered as a reference sample and the others as test samples. For each test sample, TMM is computed as the weighted mean of log ratios between this test and the reference, after exclusion of the most expressed genes and the genes with the largest log ratios. According to the hypothesis of low DE, this TMM should be close to 1. If it is not, its value provides an estimate of the correction factor that must be applied to the library sizes (and not the raw counts) in order to fulfill the hypothesis. The calcNormFactors() function in the edgeR Bioconductor package provides these scaling factors. To obtain normalized read counts, these normalization factors are re-scaled by the mean of the normalized library sizes. Normalized read counts are obtained by dividing raw read counts by these re-scaled normalization factors.

DESeq: This normalization method [14] is included in the DESeq Bioconductor package (version 1.6.0) [14] and is based on the hypothesis that most genes are not DE. A DESeq scaling factor for a given lane is computed as the median of the ratio, for each gene, of its read count over its geometric mean across all lanes. The underlying idea is that non-DE genes should have similar read counts across samples, leading to a ratio of 1. Assuming most genes are not DE, the median of this ratio for the lane provides an estimate of the correction factor that should be applied to all read counts of this lane to fulfill the hypothesis. By calling the estimateSizeFactors() and sizeFactors() functions in the DESeq Bioconductor package, this factor is computed for each lane, and raw read counts are divided by the factor associated with their sequencing lane.

Results/ Discussion:

Based on three real mRNA and one miRNA-seq datasets, we confirm previous observations that RPKM and TC, both of which are still widely in use [40, 41], are ineffective and should be definitively abandoned in the context of differential analysis

The TC approach, on the other hand, ignores the fact that different biological samples may express different RNA repertoires. In addition, it may too often be biased by the behavior of a relatively small number of high-count genes that are not guaranteed to have similar levels of expression across different biological conditions (Evaluation of statsitcal methods, 2010).

Bottom line, both DE and TC are good to go.

**Normalization of RNA-seq data using factor analysis of control genes or samples (2014) (This is the spike-in one)**

Although their approach does not make any assumptions concerning differences in gene expression between samples, it relies on another equally important assumption: technical effects should affect the spike-ins in the same way as they do the genes.

If, for instance, some library preparation step affects spike-in and gene counts differently, then normalization based on the spike-ins may incorrectly adjust the expression measures for the bulk of the genes.

We propose a normalization strategy for RNA-seq, remove unwanted variation (RUV), that uses factor analysis to adjust for nuisance technical effects, based on counts (or residuals counts) for either negative control genes or negative control samples, that is, genes or samples that are not expected to be influenced by the biological covariates of interest.

We further demonstrate that RUV, whether based on controls or not, generally outperforms state-of-the-art normalization approaches in the context of differential expression inference. It improves upon other control-based methods.

RUV makes use of a subset of the data to estimate the factors of unwanted variation and adjusts for these in the model for differential expression analysis.

We propose three alternative approaches for estimating the factors of unwanted variation: (i) RUVg uses negative control genes, assumed not to be differentially expressed with respect to the covariates of interest (e.g., ERCC spike-ins); (ii) RUVs uses negative control samples for which the covariates of interest are constant (e.g., centered counts for technical replicates of sample A and of sample B in the SEQC data set); (iii) RUVr uses residuals from a first-pass GLM regression of the unnormalized counts on the covariates of interest.

Results:

Strong strong library preparation effects were observed for spike ins?

The proportion of reads mapping to the ERCC spike-ins was highly variable between samples and deviated markedly from the nominal value.

The proportion of reads mapping to the spike-ins was stable between sequencing runs of the same library, but was very variable between libraries and exhibited a strong treatment effect (being consistently higher in treated than in control samples). These distributional properties of the spike-ins have important implications for inferring differential gene expression.

Properly behaved spike-ins could be a valuable resource for normalization: by design, their read counts are expected to be constant (or to have known fold-changes) between samples and hence any deviations from nominal fold-changes should reflect nuisance technical effects. One could therefore use functions of the spike-in counts to scale gene-level read counts, using existing procedures such as upper-quartile or trimmed mean of M values (TMM)[4](https://www.nature.com/articles/nbt.2931#ref4) normalization.

OOOO interesting.

RUV is implemented in the open-source R package *RUVSeq*, with source code freely available through the Bioconductor Project .