Supplementary Information to: Understanding sequencing data as compositions: an outlook and review

Thomas P. Quinn, Ionas Erb, Mark F. Richardson, and Tamsyn M. Crowley

Relative and absolute library sizes. Let y_{gj} be the raw read counts of gene g in sample j. If we denote the relative library size of sample j after sequencing by $n_j = \sum_{g=1}^{D} y_{gj}$, with D the number of genes, the parts of the read count composition are

$$x_{gj} = \frac{y_{gj}}{n_j}. \tag{1}$$

Note that the n_j have only a technical, not a biological meaning. There are also absolute counts Ka_{gj} that may be thought of as the (average) number of mRNAs from a given gene in the cells before sequencing. (Here we put the constant K to indicate that all we usually need are counts a_{gj} that are proportional to these counts, not the exact number of mRNAs.) This interpretation would only be true if each mRNA molecule produced the same number of reads. We have, however, the additional complication that longer mRNAs result in more sequencing reads, so the absolute counts Ka_{gj} are in fact a product of the number of mRNAs and the number of reads produced per gene-specific mRNA molecule. We can now express the parts x_{gj} also in terms of the unknown absolute counts:

$$x_{gj} = \frac{a_{gj}}{\sum_{g=1}^{D} a_{gj}}. \tag{2}$$

The (unknown) denominator we call the effective library size. Equating (1) and (2) relates the observed counts to the unknown absolute counts.

Size factors, normalization factors, and offsets. Comparisons between two or more samples require a transformation to a common scale. There are two strategies to achieve this. First, the CoDA strategy bases between-sample comparisons on ratios obtained from within the samples. Second, the normalization strategy attempts to transform to the common-scale counts a_{gj}. For this, we can define size factors s_j that are proportional to the ratio of relative and absolute library sizes:

$$s_j = \frac{n_j}{\sum_g a_{gj}} \tag{3}$$

Note that dividing the raw counts y_{gj} by the size factors s_j, we obtain the

common-scale counts a_{gj}:

$$\frac{y_{gj}}{s_j} = \frac{n_j x_{gj} \sum_g a_{gj}}{n_j} = a_{gj}. \tag{4}$$

Taking the log of this, we see that the logged raw counts are offset by the log of the size factor (i.e., $\log(y_{gj}) - \log(s_j)$). In the next sections, we show how size factors can be estimated. First, we consider another approach that estimates normalization factors $f_j^{(r)}$ based on the ratio between absolute library sizes

$$f_j^{(r)} = \frac{\sum_g a_{gj}}{\sum_g a_{gr}}. \tag{5}$$

Here, the sample r serves as a reference sample. Multiplying the ratio of parts between samples with these normalization factors, we get absolute count ratios:

$$\frac{x_{gj}}{x_{gr}} f_j^{(r)} = \frac{a_{gj}}{a_{gr}}. \tag{6}$$

Again, taking the log of this equation shows that the logratio of parts is offset by the log of the normalization factor (i.e., $\log(x_{gj}/x_{gr}) + \log(f_j^{(r)})$).

Unchanged genes normalize. Before we show how normalization and size factors can be estimated, let us first say a few words about reference genes u that have constant absolute counts. If known, they can be used to determine size factors

$$y_{uj} = n_j x_{uj} = \frac{n_j \text{const.}}{\sum_g a_{gj}} = \text{const.} s_j \tag{7}$$

and normalization factors

$$\frac{x_{uj}}{x_{ur}} = \frac{\text{const.}}{\sum_g a_{gj}} \frac{\sum_g a_{gr}}{\text{const.}} = \frac{\sum_g a_{gr}}{\sum_g a_{gj}} = \frac{1}{f_j^{(r)}}. \tag{8}$$

We can also see that the alr transformation using an unchanged gene in the denominator results in a normalization of the data. The per-gene components of an alr-transformed sample using an unchanged reference u evaluate to

$$\text{alr}_g^{(u)}(\mathbf{x}_j) = \log \frac{x_{gj}}{x_{uj}} = \log \frac{x_{gj} \sum_{g'} a_{g'j}}{\text{const.}} = \log\, a_{gj} - \log\, \text{const.} \tag{9}$$

DeSeq and edgeR normalizations. The normalization strategies that we discuss now try to estimate unchanged references by pooling information from many genes, where the underlying assumption is that the majority of genes do not change across samples. Let us denote the number of samples by N. In DeSeq [1], size factors are estimated by a median

$$s_j = \text{med}_g \frac{y_{gj}}{\left(\prod_{j'=1}^{N} y_{gj'}\right)^{\frac{1}{N}}}, \tag{10}$$

while in edgeR [2], normalization factors are determined via the trimmed mean of M-values (TMM)

$$\log_2 \frac{1}{f_j^{(r)}} = \sum_{g \in G^*} \omega_{gj}^{(r)} \log_2 \frac{x_{gj}}{x_{gr}}. \tag{11}$$

Here, G^* is the bulk set of genes that remains after ranking them both according to their log ratios and according to their expression levels, and then discarding pre-specified percentages of the highest and lowest ranked genes in both rankings. The $\omega_{gj}^{(r)}$ are precision weights that insure higher contributions of more reliable genes. The log-ratios that are summed over are known as M-values.

Relation to clr transformation. To see the connection to the clr transformation of both these normalization procedures, we specify the per-gene components of a clr-transformed sample by

$$\mathrm{clr}_g(\mathbf{x}_j) = \log \frac{x_{gj}}{\left(\prod_{g'=1}^{D} x_{g'j}\right)^{\frac{1}{D}}}. \tag{12}$$

Starting with the DeSeq size factors, we replace the median in (10) by the (mathematically more tractable) geometric mean:

$$s_j = \left(\prod_{g=1}^{D} \frac{y_{gj}}{\left(\prod_{j'=1}^{N} y_{gj'}\right)^{\frac{1}{N}}} \right)^{\frac{1}{D}} = C^{-1} \left(\prod_{g=1}^{D} y_{gj} \right)^{\frac{1}{D}}, \tag{13}$$

where the constant C evaluates to

$$C = \left(\prod_{j=1}^{N} \prod_{g=1}^{D} y_{gj} \right)^{\frac{1}{ND}}. \tag{14}$$

We can now calculate the log of the common-scale counts using (4):

$$\begin{aligned} \log\, a_{gj} = \log \frac{y_{gj}}{s_j} = \log \left(\frac{y_{gj}}{\left(\prod_{g'=1}^{D} y_{g'j}\right)^{\frac{1}{D}}} C \right) = \log \left(\frac{n_j x_{gj}}{n_j \left(\prod_{g'=1}^{D} x_{g'j}\right)^{\frac{1}{D}}} C \right) \\ = \mathrm{clr}_g(\mathbf{x}_j) + \log\, C. \end{aligned} \tag{15}$$

Thus, if we replace the median with the geometric mean in the size-factor estimate, for the log of the common-scale counts we recover the clr-transformed parts (12) plus a constant. Let us now come to the TMM normalization. If in (11) we do without trimming and weighting we obtain

$$-\log_2 f_j^{(r)} = \frac{1}{D} \sum_g \log_2 \frac{x_{gj}}{x_{gr}} = \log_2 \left(\prod_g \frac{x_{gj}}{x_{gr}} \right)^{\frac{1}{D}}. \tag{16}$$

Thus, using (6), the untrimmed and unweighted mean results in an absolute log-ratio estimate

$$\log\frac{a_{gj}}{a_{gr}} = \log\left(\frac{x_{gj}}{x_{gr}}f_j^{(r)}\right) = \log\frac{x_{gj}}{x_{gr}} - \log\left(\prod_g \frac{x_{gj}}{x_{gr}}\right)^{\frac{1}{D}} = \mathrm{clr}_g(\mathbf{x}_j) - \mathrm{clr}_g(\mathbf{x}_r). \tag{17}$$

Comparison. Comparing the last equation with (15), we see that we would get exactly the same expression using the modified DeSeq normalization. The median, trimmed mean and geometric mean lead thus to quite similar procedures, making it clear that the clr transformation can be used as a normalization. The iqlr transformation [3] as used in ALDEx2 would be even more in spirit of the trimmed mean used in edgeR. For a comparison of effective library size normalization methods, see [4]. It should be emphasized, however, that the focus in CoDA is not on scaling parts to become common-scale quantities that are comparable on absolute terms. Rather, when the within-sample ratios are compared between samples, the denominators in them have to be taken for what they are when interpreting results.

A word on RPKM and TPM. While effective library-size normalizations are specifically designed for compositional data, neither RPKM [5] nor TPM [6] are suitable for this data type under general conditions. RPKMs of a gene (or transcript) g in sample j are proportional to x_{gj}/l_g, where l_g is the length of the gene. TPMs are proportional to

$$\frac{x_{gj}/l_g}{\sum_{g'} x_{g'j}/l_{g'}}. \tag{18}$$

While TPM is now preferred to RPKM as it sums to the same number in each sample (and thus avoids some problems with inter-sample comparisons [6]), it still was designed with absolute data in mind. The definition would be compatible with compositional data if we replaced the x_{gj} by a_{gj}, but for this, we would have to use one of the normalization strategies described above. Otherwise, we assume read counts for samples are already on a common scale.

References

[1] S Anders and W Huber (2010). *Differential expression analysis for sequence count data.* Genome Biology **11** R106.

[2] MD Robinson and A Oshlack (2010). *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biology **11** R25.

[3] AD Fernandes, JN Reid, JM Macklaim, TA McMurrough, DR Edgell, GB Gloor (2014). *Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis.* Microbiome **2** 15.

[4] E Maza, P Frasse, P Senin, M Bouzayen, M Zouine (2013). *Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments* Communicative & Integrative Biology **6**(6) e25849.

[5] A Mortazavi, BA Williams, K McCue, L Schaeffer, B Wold (2008) *Mapping and quantifying mammalian transcriptomes by RNA-seq.* Nature Methods **5** 621–628.

[6] GP Wagner, KK Vincent, J Lynch (2012). *Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.* Theory in Biosciences **131**(4) 281–285.

Supplemental Figure 1

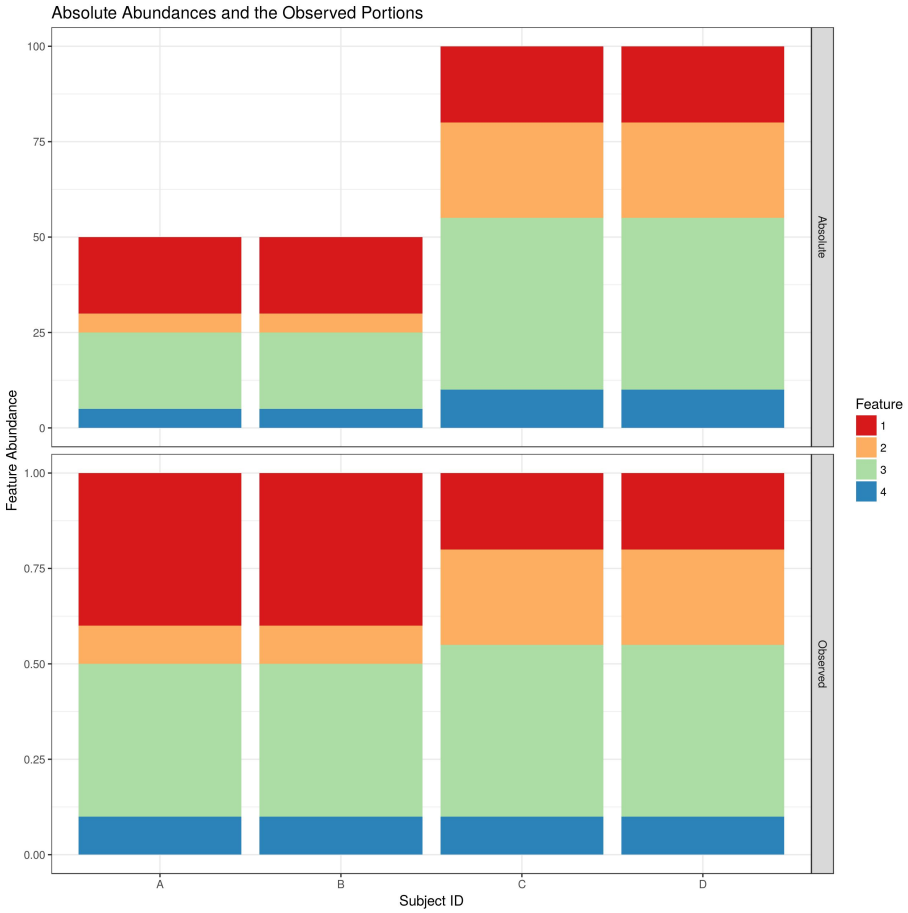

Figure 1: This figure shows a mock example of feature abundance (y-axis) for four subjects (x-axis). The top panel shows absolute feature abundance for four features (e.g., genes) as colors. The bottom panel shows relative feature abundance for the same four features. Absolute abundances and relative abundances differ. For example, although Feature 1 is equally expressed in all samples absolutely, it appears to have decreased abundance in some when measured relatively.

Supplemental Figure 2

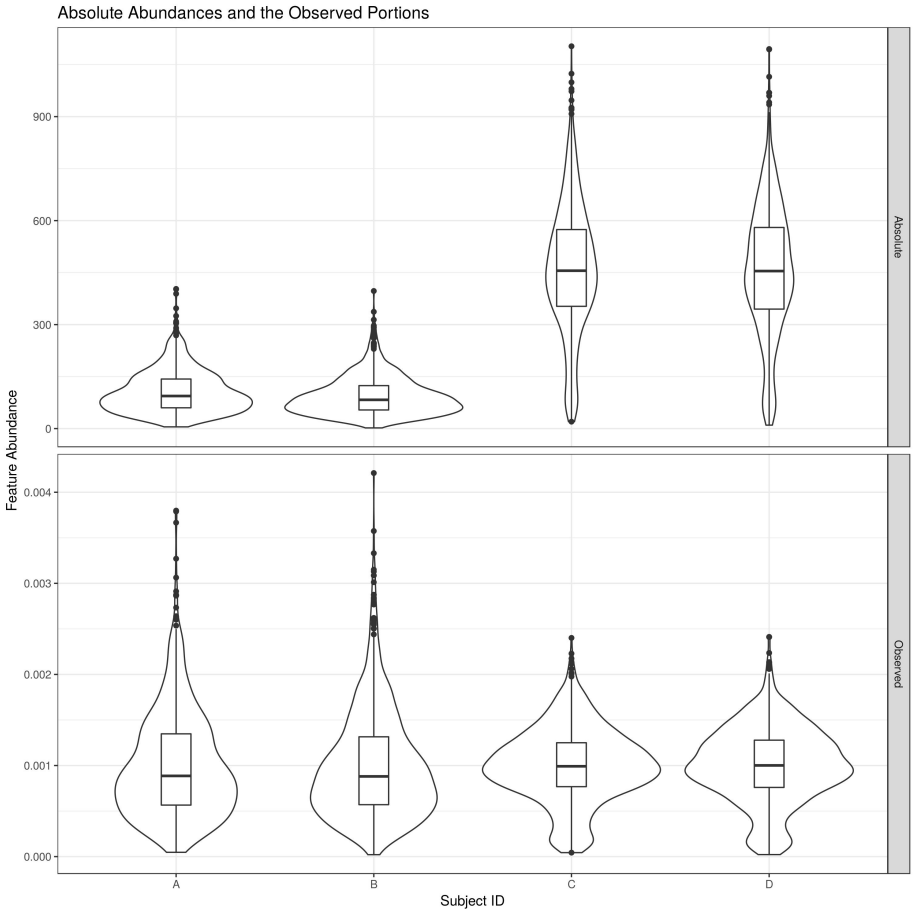

Figure 2: This figure shows another mock example of feature abundance (y-axis) for four subjects (x-axis). The top panel shows absolute feature abundance for 1000 features (simulated based on a negative binomial distribution). The bottom panel shows relative feature abundance for the same 1000 features. Absolute abundances and relative abundances differ. In absolute terms, 900 features have increased abundance in Subjects C and D and 100 features have equal abundance across all Subjects. Yet, some features in Subjects C and D appear to have decreased abundance when measured relatively.

Supplemental Figure 3

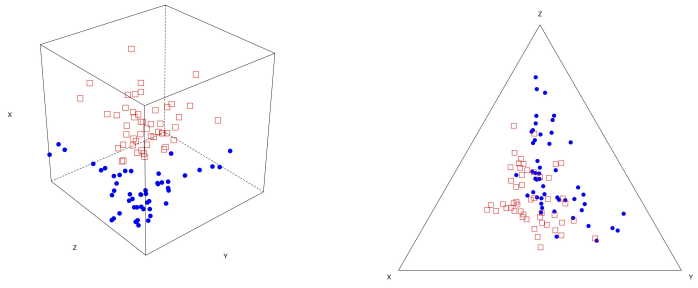

Figure 3: This figure shows another mock example of three variables measured across 100 subjects (as points) belonging to one of two groups (as colors). The left panel shows absolute abundance visualized with a 3D scatter plot. The right panel shows relative abundance visualized with a ternary diagram. Although the two groups are linearly separable in absolute space (left), the boundary is blurred in relative space (right).