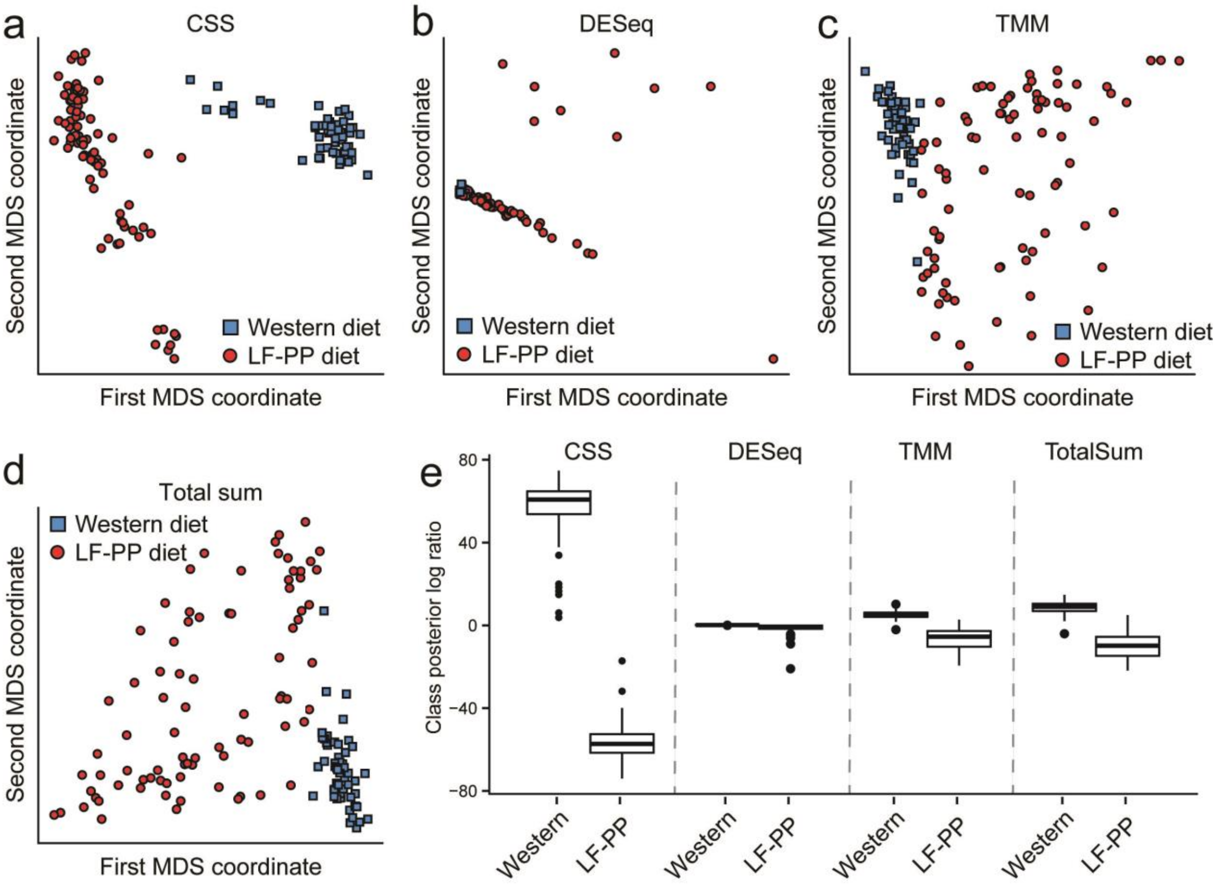
Supplementary Note for: A fair comparison

Authors: Paul I. Costea, Georg Zeller, Shinichi Sunagawa, Peer Bork

European Molecular Biology Laboratory, Heidelberg, Germany.

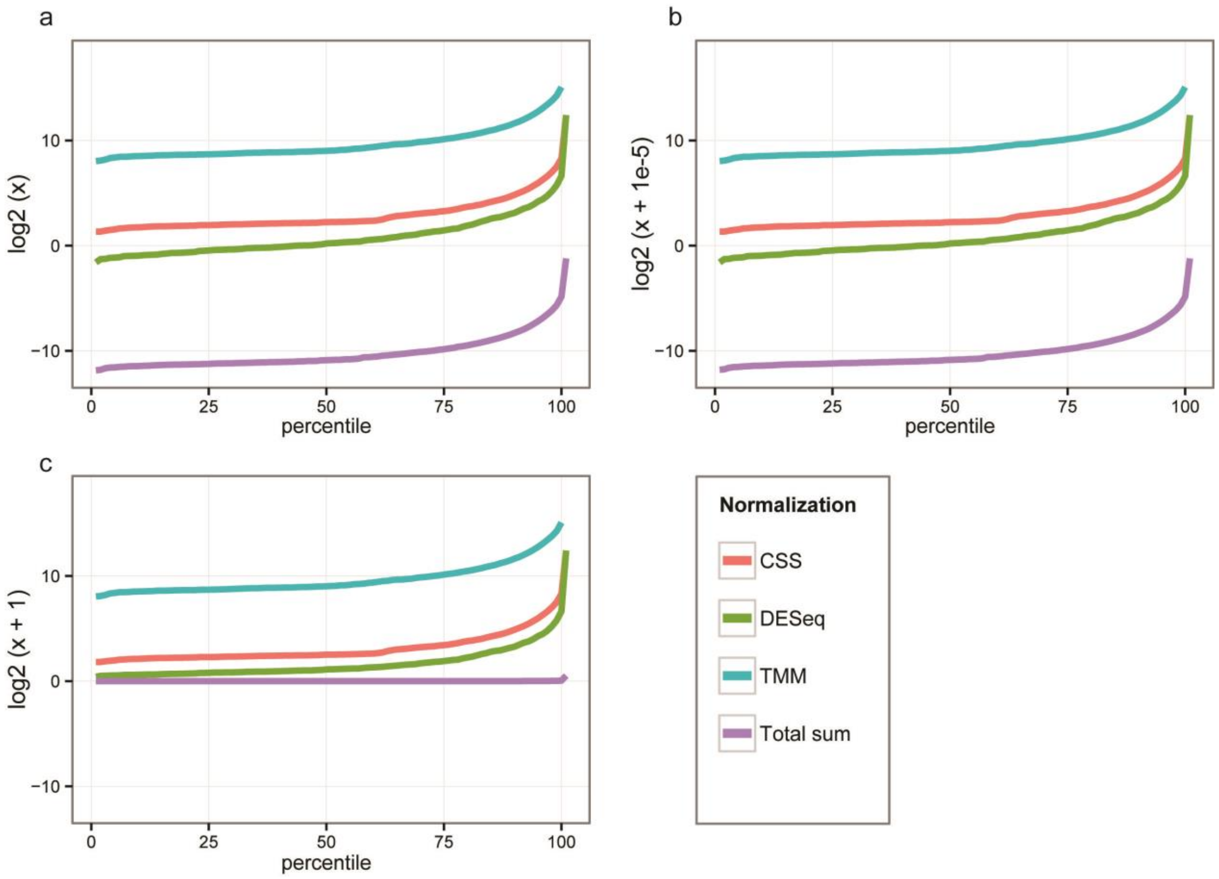
Supplementary Figure 1	Reproduction of Fig. 1 by Paulson et al.
Supplementary Figure 2	Illustration of the effect of the generalized log transform for different pseudo-counts.
Supplementary Note	Explanation of additional analyses

Supplementary Figure 1



Supplementary Figure 1 | Reproduction of Figure 1 by Paulson et al.[1] (see Supplementary Note 1 for differences), originally described as follows: (a–d) We plot the first two principal coordinates in a multidimensional scaling (MDS) analysis of mouse stool data normalized by CSS (a), DESeq size factors (b), trimmed mean of M-values (TMM) (c) and total-sum scaling (d). Colors indicate clinical phenotype (diet). LF-PP, low-fat, plant polysaccharide–rich diet. CSS normalization of data successfully separates samples by diet while controlling within-group variability. (e) Class posterior probability log-ratio for Western diet obtained from linear discriminant analysis. Each box corresponds to the distribution of leave-one-out posterior probability of assignment to the “Western” cluster across normalization methods (whiskers indicate 1.5× interquartile range). Samples were best distinguished by phenotypic similarity using CSS normalization.

Supplementary Figure 2



Supplementary Figure 2 |Illustration of the effect of the generalized log transform for different pseudo-counts. (a-c) Effect of the log transform, log(x + z) on non-zero values resulting from the four normalization methods depends on pseudo-count choice; no pseudo-count, z = 0 (a), z = 1e-5 (b) and z = 1 (c). The x-axis represents the percentiles of the distribution of non-zero values after each normalization. A precentile representation was chosen because the normalized counts differ considerably in magnitude.

Supplementary Note

Reproducing Figure 1

Supplementary Fig. 1 shows a reproduction of Fig. 1 in Paulson et al.[1] with the following minor correction: diet labels in (e) are corrected (they were swapped in the original figure).

Data dependent adjustment of the generalized log transform

To be able to apply the log transform, despite it not being defined at 0, a pseudo-count is added to allow transformation of the entire count matrix, using the form log(x + z), where z is the pseudo-count. When comparing the effect of this log-transform on several data sets whose range of values may differ by orders of magnitudes (as is the case in the comparison by Paulson et al.[1]), it is important that the transformation effect of applying the logarithm is nonetheless kept similar across data sets.

Applying the log transformation to the non-zero values, without a pseudo-count (**Supplementary Fig. 2a**), shows the baseline effect of the transformation. Choosing the pseudo-count z in a data-dependent manner, here by setting it to a value that is smaller than the minimum input value (**Supplementary Fig. 2b**) will yield a comparable transformation across all normalizations. As shown in **Supplementary Fig. 2c**, adding a count of z = 1 to all data sets (each resulting from application of a different normalization) is approximately turning the logarithm of the total sum scaled data into a constant transformation function with a dramatic difference to the non-linear transform that is effectively applied to the other data sets.

Thus, in **Figure 1**, a pseudo-count of 1 was used for CSS and TMM normalization as the minimum non-zero values of CSS and TMM normalizations are 2.5 and 262.98, respectively. For the DESeq normalization, the minimum non-zero value is 0.33 we thus chose a smaller value of 0.01. For total-sum a pseudo-count of 0.00001 was used as the minimum normalized value is 0.00027.

References

1. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. *Nat. Methods* **10,** 1200–2 (2013).