Log-counts per million

We assume that an experiment has been conducted to generate a set of n RNA samples. Each RNA sample has been sequenced, and the sequence reads have been summarized by recording the number mapping to each gene. The RNA-seq data consist therefore of a matrix of read counts r_{gi}, for RNA samples $i = 1$ to n, and genes $g = 1$ to G. Write R_i for the total number of mapped reads for sample i, $R_i = \sum_{g=1}^{G} r_{gi}$. We define the log-counts per million (log-cpm) value for each count as

$$y_{ga} = \log_2 \left(\frac{r_{gi} + 0.5}{R_i + 1.0} \times 10^6 \right)$$

The counts are offset away from zero by 0.5 to avoid taking the log of zero, and to reduce the variability of log-cpm for low expression genes. The library size is offset by 1 to ensure that $(r_{gi} + 0.5)/(R_i + 1)$ is strictly less than 1 has well as strictly greater than zero.