Correction for Martinez-Jimenez, Eling et al (2017)
Additional Information

When compiling a vignette describing the statistical analysis underlying Martinez-Jimenez et al., Science, 2017 for inclusion in a Bioconductor workflow associated to the BASiCS software, we noted that the number of spike-in molecules present in the cell lysis volume had been incorrectly calculated, such that all molecule numbers were arbitrarily scaled by the same constant factor (~10^2). These values are required as an input for the computational analysis implemented in BASiCS. This changes the scale of the arbitrary units in which gene expression is measured, but does not change relative differences between conditions. Here, we provide more details about the comparison between the old and new analyses.

Changes in gene-specific parameter estimates

The arbitrary expression levels have been scaled down by a factor of ~10^2. This derives from our incorrect computation of the true spike-in concentrations – in the original publication these were *multiplied* by a factor of ~100, which explains why the endogenous gene levels were scaled by a factor of ~100. Multiplying by any fixed value has no substantial effect on the power to detect downstream changes, because all changes we interrogate (in common with most studies) are relative; in other words, we are comparing expression levels between conditions (i.e., naïve or stimulated cells) so any fixed multiplications identical on both axes cancel out. As a practical example, if the mean expression of a gene in condition one is 1000 and 100 in condition two, the Fold Change (FC) is 1000/100 = 10. If we scale all the counts by a constant factor (i.e., expression is 100,000 in condition 1 and 10,000 in condition 2), the FC remains 10. In other words our power to identify statistically significant Differentially Expressed (DE) genes is unaffected. This is illustrated in Figures R1-3.

Figure R1 shows that the posterior estimates of the mean parameter for each gene are almost perfectly correlated when computed using the old (x-axes) and corrected (y-axes) values of the ERCC spike ins. It can be observed that the estimates are offset by a constant factor, consistent with the the scaled spike in numbers used in the original publication. For genes with a higher fraction of zeros (>75%) the estimates are less stable. However, the majority of these genes are excluded from differential testing since they do not pass our minimal expression threshold.

Figure R2 compares posterior estimates for cell-specific technical scaling factors, whose role is to capture cell-to-cell biases that affect expression counts for intrinsic and spike-in genes (for instance, sequencing depth). The change in scale observed for mean expression estimates (Figure R1) is, as expected, absorbed by the normalisation parameters. In other words, if the input ERCC spike in numbers are multiplied by a common constant, technical scaling factors are divided by the same value.

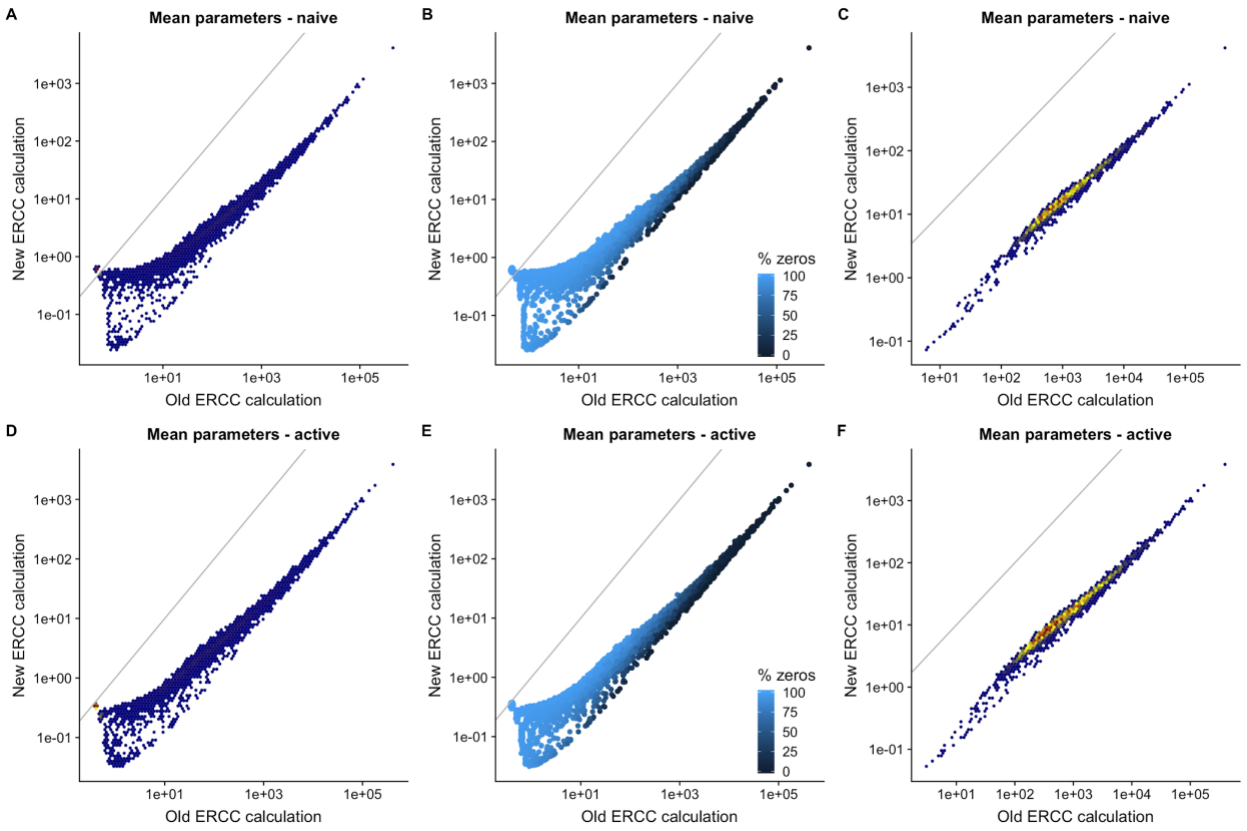

Figure R1. Comparison of posterior estimates for gene-specific mean expression parameters, before and after updating the input ERCC calculation for naive cells (A-C) and active cells (D-F). (A and D) All genes. Pearson's correlation between old and new estimates for naive cells is 0.996 (0.926 if correlation is calculated after log-transformation). For active cells, this correlation is equal to 0.995 (or 0.954 after log-transformation). (B and E) All genes are coloured according to the percentage of zero-counts observed across all cells within the group. (C and F) Only genes with less than 75% of non-zero-counts in naive and active cells, respectively. A high correlation is observed between the posterior estimates. Changing the input ERCC calculation introduces a global offset, which equally affects both groups of cells, as expected. The largest differences between the estimates are observed for those whose percentage of zero-counts is above 75%, most of which are excluded when performing differential testing.

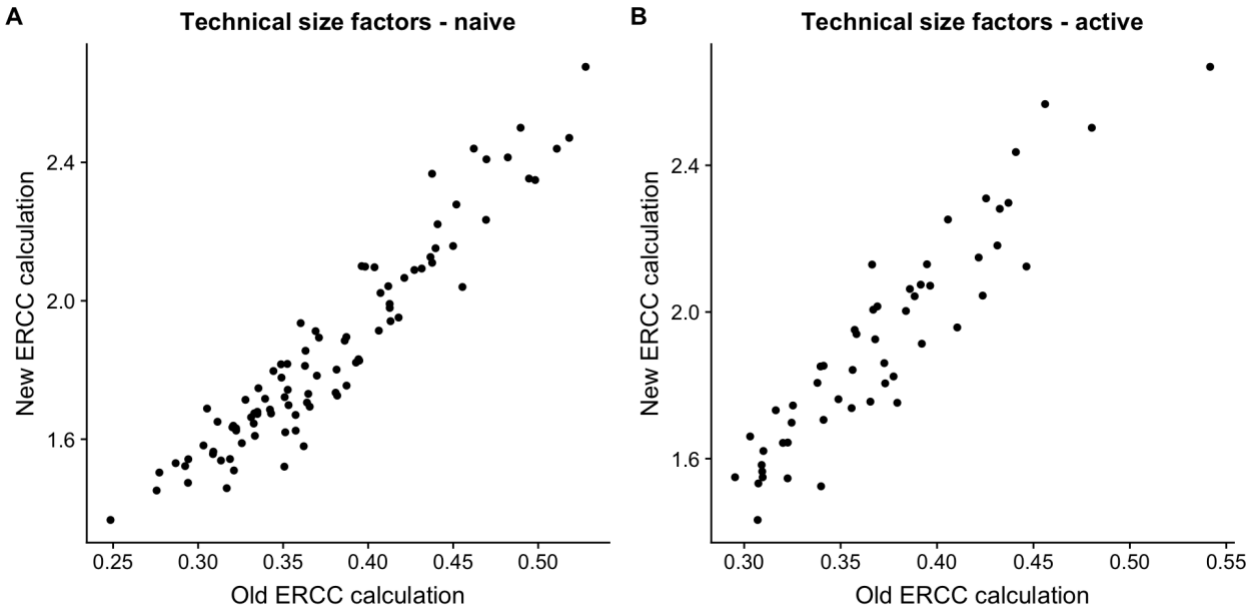

Figure R2. Comparison of estimated cell-specific technical scaling factors, before and after updating the input ERCC calculation for naive cells (A) active cells (B). The role of these technical scaling factors is to capture cell-to-cell biases that affect both intrinsic and technical spike-in genes. For naive and active cells, the correlation between old and new estimates is 0.950 and 0.929, respectively. The effect observed in Figure R1 is absorbed by a change in the scale of these normalisation parameters, as discussed above.

Figure R3 shows the estimated log2-FC in expression for each gene between the naive and stimulated cells using the old (x-axis) or corrected (y-axis) ERCC spike in molecule numbers. It can be observed that the Fold Changes are highly concordant, with most points falling on the diagonal, showing strong concordance in the set of genes that are DE in both the original and the corrected analysis. As before, the estimates are noisiest for genes that are lowly expressed. On the right hand side, we can observe extremely strong concordance for genes that are expressed above a moderate level in both naive and activated cells.

Figure R4 further demonstrates that the genes identified as DE using both the old and the corrected analyses are highly consistent. Specifically, we rank genes by significance (i.e., we order them by the posterior probability of DE) separately for the old (x-axis) and new (y-axis). If similar sets of genes are DE in both analyses we would expect the points to fall along the diagonal - this is precisely what we observe. Again this demonstrates that although the mean expression counts are scaled down in the corrected analysis, the set of DE genes is highly consistent between the original and corrected analyses, demonstrating the robustness of our results.

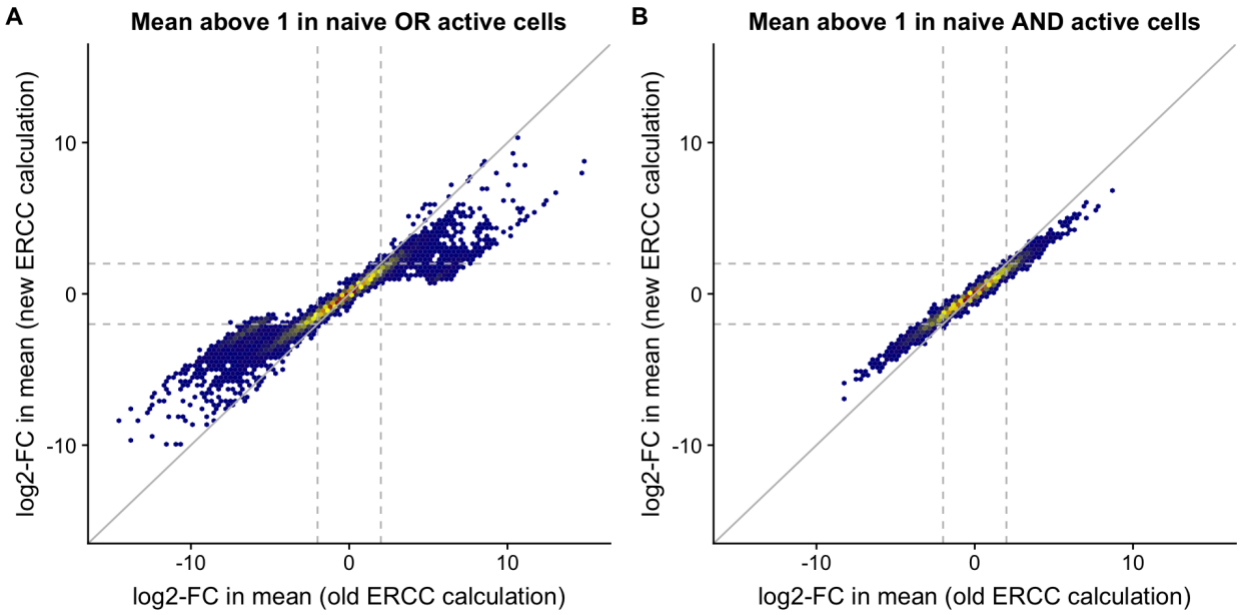

Figure R3. Comparison of log$_2$-FC in mean expression between the groups, before and after updating the input ERCC calculation. Dashed lines denote the log2-FC threshold that was used in our DE analysis. As expected, the effect observed in Figure 1 cancels out when calculating the log2-FC between the groups. The largest differences are observed for the most lowly expressed genes, whose mean expression estimate is below 1 in either group of cells (such genes are excluded in panel B).

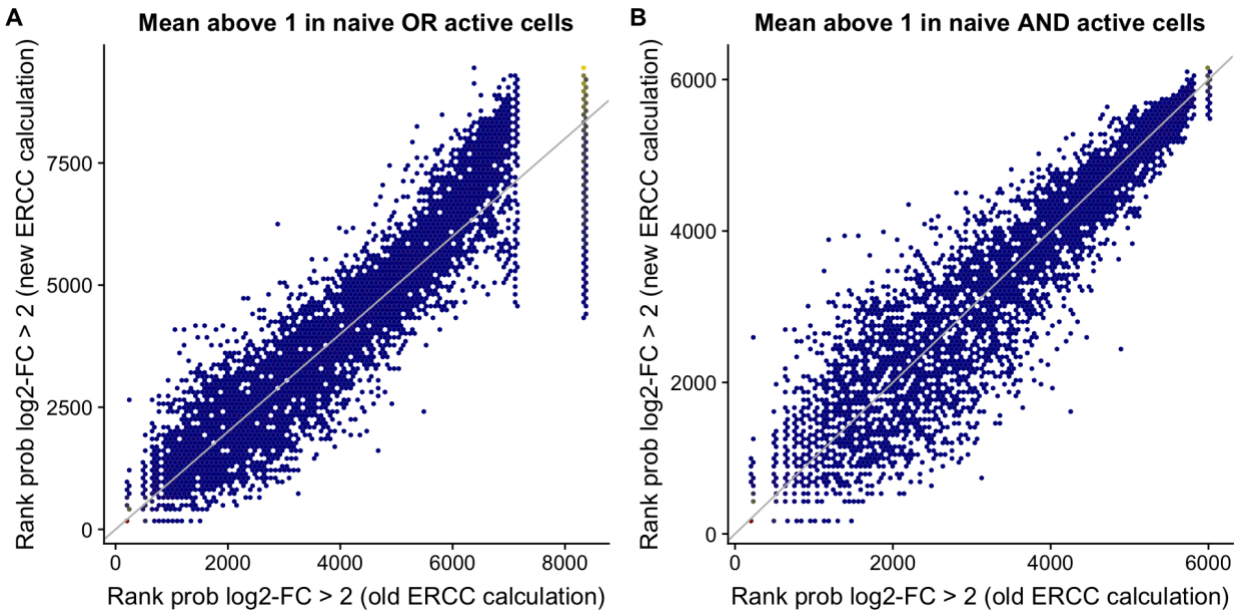

Figure R4. Comparison of tail posterior probabilities associated to the DE test (log2-FC > 2, as applied in our analysis), before and after updating the input ERCC calculation. Strong concordance in terms of the ranking of genes based on their posterior probabilities is observed, illustrating that both analysis lead to highly similar sets of DE genes.

Figures R5 and R6 provide a similar analysis for gene-specific over-dispersion estimates.

Figure R5 shows the posterior estimates of the dispersion parameter for each gene when computed using the old (x-axes) and corrected (y-axes) values of the ERCC spike ins. Whilst a global offset was observed in terms of mean expression parameters (Figure R1), this effect is captured by the normalization parameters (Figure R6) and does not affect the scale of the dispersion estimates. For genes with a higher fraction of zeros (>75%) the estimates are less stable. However, the majority of these genes are excluded from differential testing since they do not pass via our minimal expression threshold. For the remaining genes, the over-dispersion parameter estimates are almost perfectly correlated between the old and the new analyses.

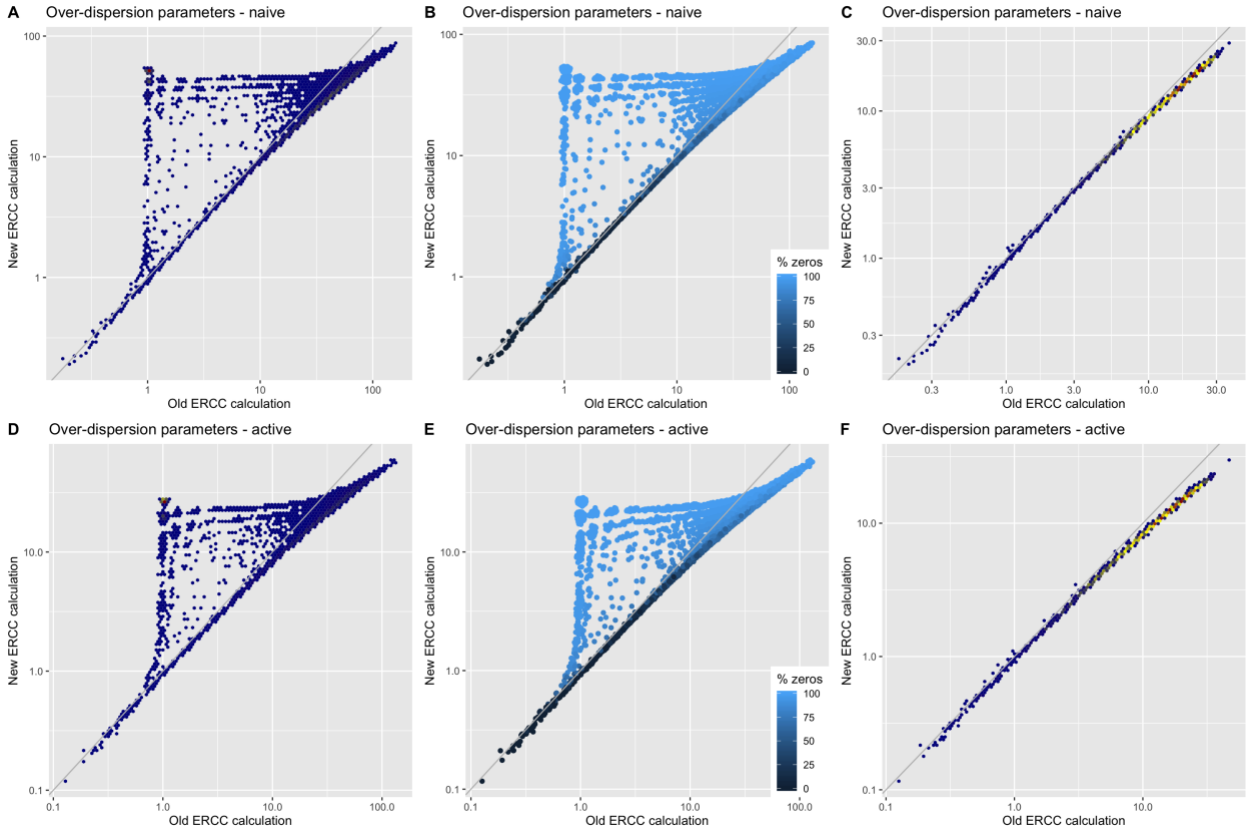

Figure R5. Comparison of posterior estimates for gene-specific dispersion parameters, before and after updating the input ERCC calculation for naive cells (A-C) active cells (D-F). (A and D) All genes. (B and E) All genes are coloured according to the percentage of zero-counts observed across all cells within the group. (C and F) Only genes with less than 75% of non-zero-counts in naive and active cells, respectively. A high correlation is observed between the posterior estimates among these genes. Pearson's correlation between old and new estimates for naive cells is 0.997 (0.999 if correlation is calculated after log-transformation). For active cells, this correlation is equal to 0.996 (or 0.999 after log-transformation). The largest differences between the estimates are observed for

those whose percentage of zero-counts is above 75%, most of which are excluded when performing differential testing.

Figure R6 shows the estimated log2-FC in over-dispersion for each gene between the naive and stimulated cells using the old (x-axis) or corrected (y-axis) ERCC spike in molecule numbers. It can be observed that the FC are highly concordant, with most points falling on the diagonal, showing strong concordance in the dispersion analysis in both the original and the corrected analysis. As before, the estimates are noisiest for genes that are lowly expressed. On the right hand side, we can observe extremely strong concordance for genes that are expressed above a moderate level in both naive and activated cells.

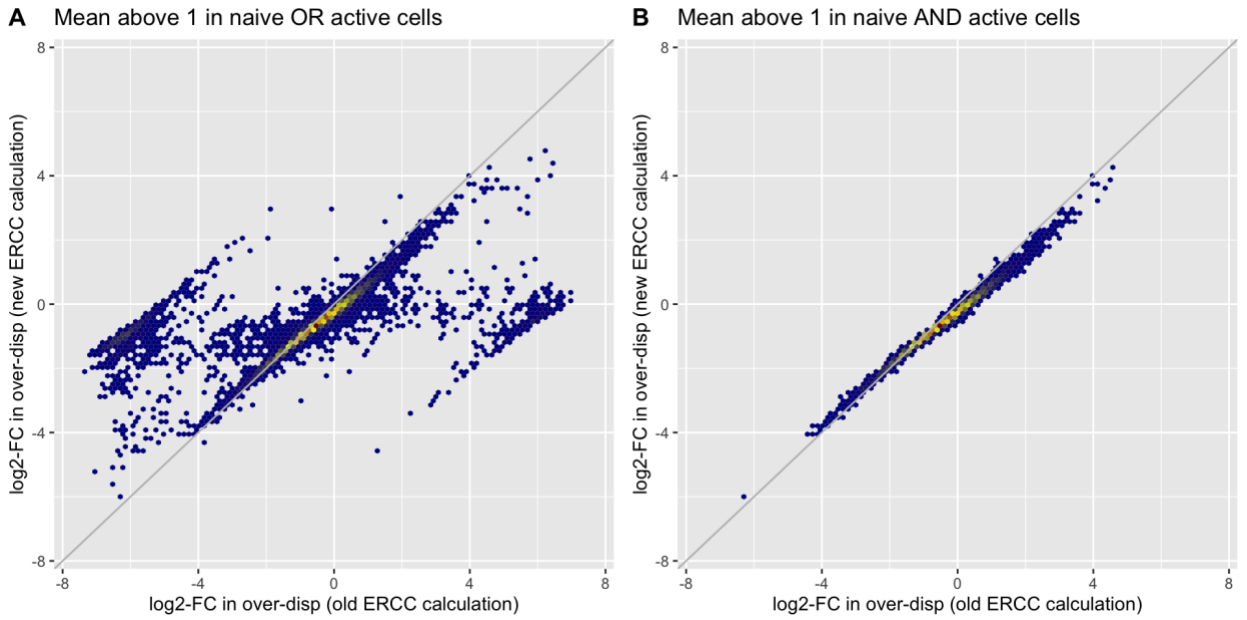

Figure R6. Comparison of log2-FC in over-dispersion between the groups, before and after updating the input ERCC calculation. The largest differences are observed for the most lowly expressed genes, whose mean expression estimate is below 1 in either group of cells (such genes are excluded in panel B).

Finally, to further illustrate the consistency between the old and the new analysis, we also calculated the overlap between the lists of genes displayed in Figure 4. When visualizing log2 fold changes (LFC) in mean expression between the old and young animals, Figure 4 focuses on the core immune activation programme (i.e. shared activation genes in Supplementary Table 3). In our revised analysis, this list includes 138 genes (the same genes are displayed for B6 and CAST). The majority of these genes (91%) were also included in the original analysis. A similar behavior is observed for those genes used to display LFC in over-dispersion (i.e. shared activation genes that are non-differentially expressed between old and young animals). In this case, these include 84 and 77 genes for B6 and CAST animals, respectively. In both cases, more than 90% of these genes were also included in the original version of Figure 4.