Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data

Supplementary Materials

by

Aaron T. L. Lun[1] and John C. Marioni[1,2]

[1]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom

[2]EMBL European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

January 8, 2016

1 Details on the implementation of the DE analysis methods

For edgeR v3.12.0, DE analyses were performed with and without EB shrinkage. In the first analysis, an abundance-dependent trend was fitted to the NB dispersions, and the trended NB dispersion was used to fit a generalized linear model (GLM) for each gene (McCarthy *and others*, 2012). The GLM deviance was used to estimate the QL dispersion for each gene, which was stabilized by robust EB shrinkage towards a mean-QL dispersion trend (Lund *and others*, 2012). DE testing between groups was performed using the QL F-test. In the second analysis, the NB dispersion was estimated for each gene individually. This was used directly to fit the GLM for each gene, and DE testing for each gene was performed using the LRT.

For voom, all counts were converted to log-CPM values and precision weights were computed from a fitted mean-variance trend. A linear model was fitted to the log-CPMs and weights for each gene using methods in the limma package v3.26.3. Sample variances were stabilized by EB shrinkage, and genes were tested for DE between groups using a moderated t-test. This analysis was also repeated after using the duplicateCorrelation function (Smyth *and others*, 2005) to estimate the correlations introduced by the plate effects. For each gene, this correlation refers to the strength of the association between the count for each cell and the plate of origin for that cell. The consensus correlation across all genes was estimated and incorporated into the linear model by blocking on the plate of origin for all cells. DE testing was then performed on this refitted model.

For DESeq2 v1.10.1, the NB dispersion was estimated for each gene and a mean-dispersion trend was fitted across genes. A normal prior was applied to the log-dispersions for all genes, and the maximum *a posteriori* estimate of the NB disperison was obtained for each gene. The estimated dispersion was used to fit a GLM to the counts for each gene. DE testing was performed between groups using the Wald test.

For edgeR and voom, normalization was performed prior to modelling using the trimmed mean of M-values (TMM) method (Robinson and Oshlack, 2010). For DESeq2, normalization was performed using the size factor method (Anders and Huber, 2010). This step is not strictly necessary for analyses of simulated data, given that library-specific biases have not been introduced and all cells have the same library size. Nonetheless, normalization is included as part of the standard analysis and will be used for real data.

For Monocle v1.4.0, counts were transformed into CPM values that were considered to be roughly log-normal. These were used for DE testing between groups with the LRT in the differentialGeneTest function.

For MAST v0.933, counts were converted into log-CPMs after adding a prior count of 1. For each gene, a hurdle model was fitted to the log-CPMs across all cells. The group was used as the factor of interest and the proportion of genes with non-zero counts in each library was used as an additional covariate (Finak *and others*, 2015). The fitting method was set to “bayesglm” and EB shrinkage was turned on with the “MLE” method and the “H1” model. Putative DE genes between groups were identified using the LRT.

2 Implementing the quantile adjustment method

The quantile adjustment method models the conditional distribution of each count as a NB distribution. All cells within a single plate are defined as a separate group in a one-way layout. For each gene, means are estimated for all cells by fitting a GLM to the counts for that gene. The mean for each cell accounts for its library size as well as any gene- and plate-specific effects. A gene-specific NB dispersion is also estimated using the Cox-Reid adjusted profile likelihood in edgeR (McCarthy *and others*, 2012). The count for each cell is modelled by a NB distribution with the estimated mean and dispersion. The percentile corresponding to each count is computed, and the matching quantile for a Poisson distribution with the same mean is identified (Robinson and Smyth, 2008). This quantile is used as the Poisson-distributed pseudo-count for each cell.

This method depends on stable estimation of the mean for each cell and the dispersion for each gene. The latter should not be an issue as there should be sufficient residual d.f. across all cells to obtain a precise dispersion estimate. Note that the previously mentioned problems with residual d.f. overestimation are not applicable here. Dependencies caused by plate effects are avoided as plate-specific expression is explicitly modelled in the one-way layout. For the means, stable estimation can be ensured by having sufficient cells per plate. This is usually the case given that low-quality plates with very few cells would be removed.

3 Summation and plate-specific mean variance relationships

The robustness of direct summation to differences in the number or size of cells per plate can be explained by examining the mean-variance relationship of the simulated data. Assume that the cells on plate k can be partitioned into L_k subpopulations. Each subpopulation l contains N_l cells and differs from the other subpopulations by the library size modifier θ_l. Let $\lambda_{il} = \delta_{ik}\theta_l\mu_{ig}$ for gene i in each cell of subpopulation l. The count for each cell is independently sampled from a NB distribution with mean λ_{il} and dispersion φ_i, as described in the simulation design. The conditional variance of the count sum s_{ik} for plate k is then

$$\mathrm{var}(s_{ik}|\delta_{ik}) = \sum_l N_l\lambda_{il} + \varphi_i \sum_l N_l\lambda_{il}^2 \; .$$

Given that $E(s_{ik}) = E(\sum_l N_l\lambda_{il}) = \sum_l N_l\theta_l\mu_{ig}$, the variance of the count sum can be decomposed to

$$\begin{aligned}
\mathrm{var}(s_{ik}) &= \mathrm{var}\{E(s_{ik}|\delta_{ik})\} + E\{\mathrm{var}(s_{ik}|\delta_{ik})\} \\
&= \mathrm{var}\left(\delta_{ik}\sum_l N_l\theta_l\mu_{ig}\right) + E\left(\delta_{ik}\sum_l N_l\theta_l\mu_{ig} + \varphi_i\delta_{ik}^2\sum_l N_l\theta_l^2\mu_{ig}^2\right) \\
&= \left(\sum_l N_l\theta_l\mu_{ig}\right)^2 \mathrm{var}(\delta_{ik}) + \left(\sum_l N_l\theta_l\mu_{ig}\right) + \left(\varphi_i\sum_l N_l\theta_l^2\mu_{ig}^2\right) E(\delta_{ik}^2) \\
&= E(s_{ik})^2\mathrm{var}(\delta_{ik}) + E(s_{ik}) + \left\{\varphi_i\frac{\sum_l N_l\theta_j^2\mu_{ig}^2}{(\sum_l N_l\theta_j\mu_{ig})^2}\right\} E(s_{ik})^2 E(\delta_{ik}^2) \; .
\end{aligned}$$

Recall that $E(\delta_{ik}^2)$ and $\mathrm{var}(\delta_{ik})$ are constant for all plates. This means that only the third term is involved in defining the plate-specific aspect of the mean-variance relationship. Now, consider the behaviour of this term as the number of cells increases. If N_l increases in each of s subpopulations, the limit becomes

$$\lim_{\substack{N_{l_1}\to\infty \\ \cdots \\ N_{l_s}\to\infty}} \frac{\sum_l N_l\theta_j^2\mu_{ig}^2}{(\sum_l N_l\theta_j\mu_{ig})^2} = \sum_l \lim_{\substack{N_{l_1}\to\infty \\ \cdots \\ N_{l_s}\to\infty}} \frac{N_l\theta_j^2\mu_{ig}^2}{(\sum_l N_l\theta_j\mu_{ig})^2} = 0$$

for positive values of θ_j and μ_{ig}. This means that, for any non-zero value of $\mathrm{var}(\delta_{ik})$, the relative contribution of the plate-specific term will approach zero as the number of cells increases. Thus, the mean-variance relationship of the count sum will be similar across plates with many cells, regardless of the number or library sizes of those cells. Summation will be similarly robust to plate-specific values for the NB dispersion. Replacing φ_i with some arbitrary φ_{ik} will have little effect as the plate-specific term approaches zero.

As an aside, what happens when each plate contains several *a priori* undefined biological subpopulations with different gene expression profiles? This can be parametrized as $\lambda_{il} = \delta_{ik}\theta_l\mu_{il}$ where μ_{il} represents the subpopulation-specific mean for gene i. The above result is unaffected as one can easily replace μ_{ig} by μ_{il} in all expressions. A more subtle point concerns the independence of the count sums. If the proportion of cells in each subpopulation is the same across replicate plates, the count sums for these plates will be independent in a one-way layout. This is because the count sum for each plate in a given group will have the same expectation, i.e., $\sum_l N_l\theta_l\mu_{il}$. Each count sum can be considered to be independently sampled from a distribution with this mean, due to the independence of δ_{ik} and conditional independence of the counts. However, this will not be true if the proportion of cells in each subpopulation is different across replicate plates. In such cases, the mean for each count sum will vary for each plate in a gene-specific manner that cannot be resolved by scaling normalization. Attempting to model the count sums with a group-specific mean will lead to hidden dependencies between plates, akin to those between cells when plate effects are ignored. This situation is arguably pathological as plates with different subpopulations should not be replicates.

4 Benefits of hiding the variability between cells

If more cells contribute to the sum, variability between cells will be hidden as it will no longer affect the total variance. However, this is not undesirable for single-cell analyses. Genes with reproducible DE between plates should not be penalized for having high variability within plates, e.g., due to cellular heterogeneity or the presence of subpopulations. This philosophy is almost the exact opposite of that in microarray analyses involving technical replicates, where variability between replicates is explicitly modelled rather than being hidden by averaging the replicate signals (Smyth *and others*, 2005). The difference in strategies is due to the fact that the technical replicates in microarray analyses are expected to be similar. Genes with large differences between replicates are considered to be unreliable and should have reduced significance for DE. In contrast, cells are not expected to be similar due to biological heterogeneity within populations. Large cell-to-cell variability has no bearing on the reliability of DE between populations when many cells are present. For example, if the subpopulation structure is the same in each replicate plate, one would have a situation with large variability between cells in different subpopulations but reproducible count sums across plates.

5 Avoiding zero counts during normalization

One particular benefit of using summed counts with existing analysis methods is in the normalization step. Methods like size factor and TMM normalization are based on ratios between counts in different cells, but will be compromised for noisy data that contain many zero and near-zero counts. For example, size factor normalization constructs an average reference library by taking the geometric mean of all counts for each gene, and normalizes each library against this reference by computing the median ratio of the count to the geometric mean across all genes. In the presence of many zeroes, the geometric mean itself becomes zero such that many ratios will be undefined. Even if a more stable average is used (e.g., the arithmetic mean), the median ratio for many libraries will be equal to zero if more than 50% of the counts of that library are zero. This is not a sensible scaling factor for normalizing expression values. Summation reduces the incidence of zero counts so that these normalization methods can be properly applied. In particular, normalization of the summed counts will remove any plate-specific differences that affect each gene in the same manner, e.g., differences in capture efficiency or sequencing depth between plates. Removal of cell-specific biases within each plate is not required as these will average out when many cells are summed together. Indeed, recall that type I error control is maintained in the simulations where the library size differs across cells.

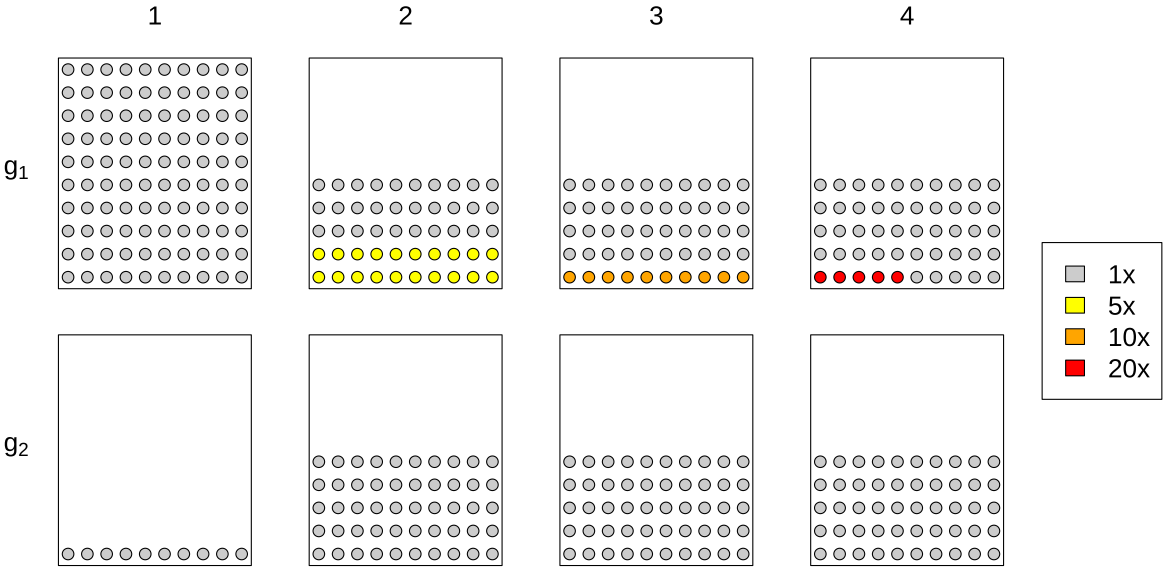

Figure S1: Simulation scenarios with different number of cells per plate and different library sizes per cell. Each scenario is represented by two plates where the set-up for each plate is used for all plates in group g_1 or g_2. Scenario 1 contains different numbers of cells, while scenarios 2 to 4 contain plates with different library sizes between cells. The relative library size θ_j of each cell is indicated by the colour of the well.

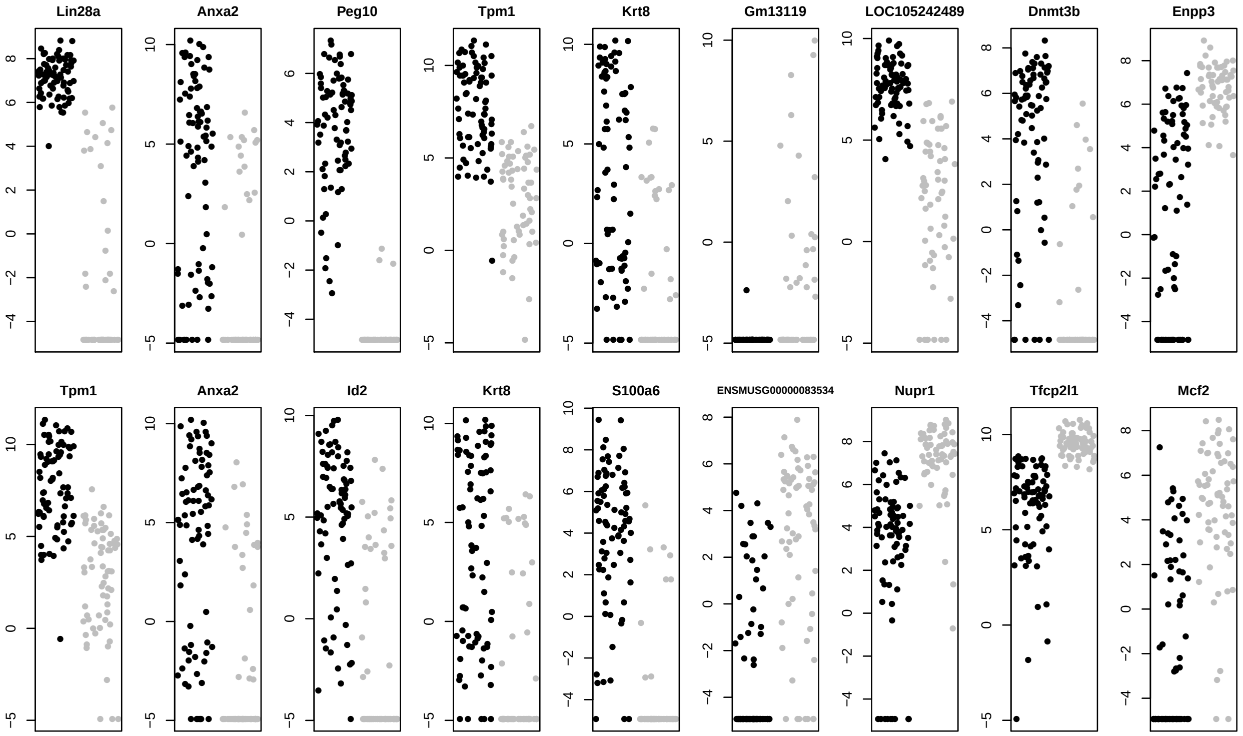

Figure S2: Expression profiles of DE genes between 2i and serum (top) or between a2i and serum conditions (bottom) in the mESC data set, for the top genes detected by QL edgeR on summed counts. Each plot contains the expression profile for a gene, where the y-axis represents the log-CPMs and each point represents a cell in the serum (black) or a2i/2i conditions (grey). A prior of 0.25 was added to each count to avoid undefined log-CPMs from zeroes. For simplicity, only cells in one batch are shown for both contrasts.