Beyond comparisons of means: understanding changes in gene expression at the single-cell level

Catalina A. Vallejos[1 2], Sylvia Richardson[1] and John C. Marioni [2 3]

Summary

Single-cell RNA sequencing (scRNA-seq) can be used to characterise differences in gene expression patterns between pre-specified populations of cells. Traditionally, differential expression tools are restricted to the study of changes in overall expression between cell populations. However, such analyses do not take full advantage of the rich information provided by scRNA-seq. In this article, we present a Bayesian hierarchical model which can be used to study changes in expression that lie beyond comparisons of means. In particular, our method can highlight genes that undergo changes in cell-to-cell heterogeneity between the populations but whose overall expression is preserved. Evidence supporting these changes is quantified using a probabilistic approach based on tail posterior probabilities, where a probability cut-off is calibrated through the expected false discovery rate. Our method incorporates a built-in normalisation strategy and quantifies technical artefacts by borrowing information from technical spike-in genes. Control experiments validate the performance of our approach. Finally, we compare expression patterns of mouse embryonic stem cells between different stages of the cell cycle, revealing substantial differences in cellular heterogeneity.

Keywords: single-cell RNA-seq; differential expression; cellular heterogeneity

Background

The transcriptomics' revolution — moving from bulk samples to single-cell resolution — provides novel insights into a tissue's function and regulation. In particular, single-cell RNA sequencing (scRNA-seq) has led to the identification of novel sub-populations of cells in multiple contexts [26, 11, 20]. However, compared to bulk RNA-seq, a critical aspect of scRNA-seq datasets is an increased cell-to-cell variability among the expression counts. Part of this variance inflation is related to biological differences in the expression profiles of the cells (e.g. changes in mRNA content and the existence of cell sub-populations or transient states), which disappears when measuring bulk gene expression as an average across thousands of cells. Nonetheless, this

[1] MRC Biostatistics Unit
[2] EMBL-European Bioinformatics Institute
[3] Cancer Research UK Cambridge Institute

increase in variability is also due in part to *technical noise* arising from the manipulation of small amounts of starting material, which is reflected in weak correlations between technical replicates [4]. Such technical artefacts are confounded with genuine transcriptional heterogeneity and can mask biological signal.

Among others, one objective of RNA-seq experiments is to characterise transcriptional differences between pre-specified populations of cells (given by experimental conditions or cell-types). This is a key step for understanding a cell's fate and functionality. In the context of bulk RNA-seq, two popular methods for this purpose are edgeR [23] and DESeq [1]. Nonetheless, these are not designed to capture features that are specific to scRNA-seq datasets. In contrast, SAMstrt [13] and SCDE [14] have been specifically developed to deal with scRNA-seq datasets. All of these methods target the detection of *differentially expressed genes* based on log fold changes (LFC) of overall expression between the populations. However, restricting the analysis to changes in overall expression does not take full advantage of the rich information provided by scRNA-seq. In particular — and unlike bulk RNA-seq — scRNA-seq can also reveal information regarding cell-to-cell expression heterogeneity. Critically, traditional approaches will fail to highlight genes whose expression is less stable in any given population but whose overall expression remains unchanged.

More flexible approaches, capable of studying changes that lie beyond comparisons of means are required in order to better characterise differences between distinct populations of cells. In this article, we develop a quantitative method to fill this gap, allowing the identification of genes whose cell-to-cell heterogeneity pattern changes between pre-specified populations of cells. This analysis allows the identification of genes with less variation in expression levels within a specific population of cells, which might indicate that they are under more stringent regulatory control. Additionally, genes with increased biological variability in a given population of cells could suggest the existence of additional sub-groups within the analysed populations. To the best of our knowledge, this is the first probabilistic tool developed for this purpose in the context of scRNA-seq analyses. We demonstrate the performance of our method using control experiments and by comparing expression patterns of mouse embryonic stem cells (mESCs) between different stages of the cell cycle.

Results and discussion

A statistical model to detect changes in expression patterns for scRNA-seq datasets

We propose a statistical approach for the comparison of expression patterns between P pre-specified populations of cells. It builds upon BASiCS [25], a Bayesian model for the analysis of scRNA-seq data. As in traditional differential expression analyses, for any given gene i, changes in overall expression are identified by comparing population-specific expression rates μ_{ip} ($p = 1, \ldots, P$). However, the main focus of our approach is to assess differences in biological cell-to-cell heterogeneity between the populations. These are quantified through changes in population-specific biological *over-dispersion* parameters δ_{ip} ($p = 1, \ldots, P$), designed to capture residual variance inflation (after normalisation, technical noise removal and adjustment for overall expression), avoiding the well-known confounding relationship between mean and variance in count-based datasets [16]. Importantly, such changes cannot be uncovered by standard differential expression methods, which are restricted to changes in overall expression. Hence, our approach provides novel biological insights by highlighting genes that undergo changes in cell-to-cell heterogeneity between the populations despite the overall expression level being preserved.

To disentangle technical from biological effects, we exploit *spike-in* genes that are added to the lysis buffer and thence theoretically present at the same amount in every cell (e.g. the 92 ERCC molecules developed by the External RNA Control Consortium [12]). These provide an internal control or “gold standard”, to estimate the strength of technical variability and to aid normalisation. In particular, these control genes allow inference on cell-to-cell differences in mRNA content, providing additional information about the analysed populations of cells. These are quantified through changes between cell-specific normalising constants ϕ_{j_p} (the sub-index j_p indicates the j_p-th cell within the p-th population). A graphical representation of our model is displayed in Figure 1 (based on a two-groups comparison). It illustrates how our method borrows information across all cells and genes (biological transcripts and spike-in genes) in order to perform inference.

Posterior inference is implemented via a Markov Chain Monte Carlo (MCMC) algorithm, generating draws from the posterior distribution of all model parameters (see supplementary material). Post-processing of these draws allows quantification of supporting evidence regarding changes in expression patterns (mean and over-dispersion). These are measured using a probabilistic approach based on tail posterior probabilities, where a probability cut-off is calibrated through the expected false discovery rate (EFDR) [19]. Our strategy is flexible and can be combined with a variety of decision rules. As an illustration, the examples described in this article focus on the detection of genes whose absolute LFC in biological cell-to-cell heterogeneity between populations p and p' exceeds a minimum tolerance threshold ω_0, i.e. when $|\log(\delta_{ip}/\delta_{ip'})| > \omega_0$, for a given $\omega_0 \geq 0$. Similarly, our examples assess changes in overall expression using the decision rule $|\log(\mu_{ip}/\mu_{ip'})| > \tau_0$, where $\tau_0 \geq 0$ is an *a priori* chosen threshold for LFC in overall expression. These LFC thresholds can be used, e.g. to avoid highlighting genes with small changes in expression which are likely to be less biologically relevant.

More details regarding the model setup and the implementation of posterior inference can be found in the Methods.

A control experiment: comparing single-cells versus pool-and-split samples

To demonstrate the efficacy of our method, we use the control experiment described in [9], where single mouse embryonic stem cells (mESCs) are compared against *pool-and-split* samples, consisting of pooled RNA from thousands of mESCs split into single-cell equivalent volumes. Such a controlled setting provides a situation where substantial changes in overall expression are not expected as, on average, the overall expression of single cells should match the levels measured on pool-and-split samples. Additionally, the design of pool-and-split samples removes biological variation, leading to a homogenous set of samples. Hence, pool-and-split samples are expected to show a genuine reduction in biological cell-to-cell heterogeneity when compared to single-cells.

Here, we display the analysis of samples cultured in a *2i* media. In these data, expression counts correspond to the number of molecules mapping to each gene within each cell. This is achieved by using unique molecular identifiers (UMIs), which remove amplification biases and reduce sources of technical variation [10]. Our analysis includes 74 single cells and 76 pool-and-split samples (same inclusion criteria as in [9]) and expression counts for 9,378 genes (9,343 biological and 35 ERCC spikes) defined as those with at least 50 detected molecules in total across all cells. As expected, our method does not reveal major changes in overall expression between single-cells and pool-and-split samples as the distribution of LFC estimates is roughly symmetric with respect to the origin (see upper panels in Figure 2) and the majority of genes are not classified as differentially expressed at 5% EFDR (see Figure 3b). However, this analysis suggests that setting the minimum LFC tolerance threshold τ_0 equal to 0 is too liberal as small log-fold changes are associated with high posterior probabilities of changes in expression (see Figure 3a) and the number of differentially expressed genes is inflated (see Figure 3b). In fact, counter-intuitively, 4,710 genes ($\sim$ 50% of all analysed genes) are highlighted to have a change in overall expression when using $\tau_0 = 0$ (a similar issue is also observed when analysing bulk RNA-seq data [17]). In contrast, this number is reduced to 559 genes ($\sim$ 6% of all analysed genes) when setting $\tau_0 = 0.4$.

Posterior inference regarding biological over-dispersion is consistent with the experimental design, where the pool-and-split samples are expected to have more homogeneous expression patterns. In fact, as shown in the right hand panel of Figure 2, the distribution of estimated LFCs in biological over-dispersion is skewed towards positive values (higher biological over-dispersion in single-cells). This is also supported by the results shown in Figure 3b, where slightly more than 2,000 genes exhibit increased biological over-dispersion in single cells and almost no genes ($\sim$ 60 genes) are highlighted to have higher biological over-dispersion in the pool-and-split samples (EFDR = 5%). In this case, the choice of ω_0 is less critical (within the range explored here). This is illustrated by the left panels in Figure 3a, where tail posterior probabilities exceeding the cut-off defined by EFDR = 5% correspond to similar ranges of LFC estimates.

mESCs across different cell cycle stages

Our second example shows the analysis of the mESC dataset presented in [5], which contains cells for which the cell cycle phase is known (G1, S and G2M). After applying the same quality control criteria as in [5], our analysis considers 182 cells (59, 58 and 65 cells in stages G1, S and G2M, respectively). To remove genes with consistently low expression across all cells, we excluded those genes with less than 20 RPM, on average, across all cells. After this filter, 5,687 genes remain (including 5,634 intrinsic transcripts, and 53 ERCC spike-in genes). As a proof of concept, to demonstrate the efficacy of our approach under a negative control, we performed permutation experiments, where cell labels were randomly permuted into 3 groups (containing 60, 60 and 62 samples, respectively). In such a case, our method correctly infers that mRNA content as well as gene expression profiles do not vary across groups of randomly permuted cells (see Figure 4).

As cells progress through the cell cycle, cellular mRNA content increases. In particular, our model infers that mRNA content is roughly doubled when comparing cells in G1 versus G2M, which is consistent with the duplication of genetic material prior to cell division (see Figure 5a). After normalisation, we rule out a global shift in expression levels between cell cycle stages (see Figure 5b and upper triangular panels in Figure 5d). Nonetheless, a small number of genes are identified as displaying changes in overall expression between cell cycle phases at 5% EFDR (see Figure 6). The actual number of differentially expressed genes depends on whether we use $\tau_0 = 0$ or $\tau_0 = 0.4$. However, the permutations described above suggest that $\tau_0 = 0.4$ is more appropriate as, just by chance, it is likely to observe LFCs below this threshold (in fact, $\tau_0 = 0.4$ roughly coincides with the 90-th percentile of the empirical distribution of posterior estimates for LFC in overall expression across all genes and permutations). Additionally, this LFC threshold is equivalent to a minimum 50% increase in overall expression (in whichever group the gene has higher expression). To validate our results, we performed Gene Ontology (GO) enrichment analysis within those genes classified as differentially expressed between cell cycle phases (see supplementary material). Here, we discuss results related to $\tau_0 = 0.4$. Not surprisingly, we found an enrichment of mitotic genes among the 545 genes classified as differentially expressed between G1 and G2M cells. In addition, the 209 differentially expressed genes between S and G2M are enriched for regulators of cytokinesis, which is the final stage of the cell cycle where a progenitor cell is divided into two daughter cells [7].

Our method uncovers a substantial decrease in biological over-dispersion when cells move from G1 to S phase, followed by a slight increase after the transition from S to G2M phase (see Figure 5c and the lower triangular panels in Figure 5d). This is consistent with the findings in [7], where the increased gene expression variability observed in G2M cells is attributed to an unequal distribution of genetic material during cytokinesis and the S phase is shown to have the most stable expression patterns within the cell cycle. Here, we discuss GO enrichment of those genes whose overall expression rate remains constant but that exhibit changes in biological over-dispersion between cell cycle stages (EFDR = 5%, $\omega_0 = 0.4$). Critically, these genes will not be highlighted by traditional differential expression tools, which are restricted to differences in overall expression rates. For example, among the genes with higher biological over-dispersion in G1 with respect to S phase, we found an enrichment of genes related to protein dephosphorylation. These are known regulators of the cell cycle [6]. Moreover, we found that genes with lower biological over-dispersion in G2M cells are enriched for genes related to DNA replication checkpoint regulation (which delays entry into mitosis until DNA synthesis is completed [3]) relative to G1 cells and mitotic cytokinesis when comparing to S cells. Both of these processes are likely to be more tightly regulated in G2M phase.

A full table with GO enrichment analysis of the results described here is provided in the supplementary material.

Conclusions

Our method provides a quantitative tool to study changes in gene expression patterns between pre-specified populations of cells. Unlike traditional differential expression analyses, our model is able to identify changes in expression that are not necessarily reflected by shifts in mean. This allows a better understanding of the differences underlying distinct populations of cells. In particular, we focus on the detection of genes whose residual biological heterogeneity (after normalisation and technical noise removal) varies between the populations. This is quantified through biological over-dispersion parameters, which capture variance inflation with respect to the level that would be expected in a homogeneous population of cells. A decision rule is defined through a probabilistic approach based on tail posterior probabilities and calibrated using the expected false discovery rate. The performance of our method was demonstrated using a controlled experiment where we recovered the expected behaviour of gene expression patterns.

Currently, our approach requires pre-defined populations of cells (e.g. defined by cell types or experimental conditions). However, a large number of scRNA-seq experiments involve a mixed population of cells, where cell-types are not known *a priori* (e.g. [26, 11, 20]). In such cases, expression profiles can be used to *cluster* cells into distinct groups and to characterise markers for such sub-populations. Nonetheless, unknown group structures introduce additional challenges for normalisation and quantification of technical variability. A future extension of our work is to combine the estimation procedure within our model with a clustering step, propagating the uncertainty associated with each of these steps into downstream analysis.

Until recently, most scRNA-seq datasets consisted of hundreds (and sometimes thousands) of cells. However, droplet-based approaches [15, 18] have recently allowed parallel sequencing of substantially larger numbers of cells in an effective manner. This brings additional challenges to the statistical analysis of scRNA-seq datasets. In particular, current protocols do not allow the addition of technical spike-in genes. As a result, the deconvolution of biological and technical artefacts becomes less straightforward. Moreover, the increased sample sizes emphasise the need for more computationally efficient approaches that are still able to capture the complex structure embedded within scRNA-seq datasets. To this end, we foresee the use of parallel programming as a tool for reducing computing times. Additionally, we are also exploring approximated posterior inference based, for example, on an Integrated Nested Laplace Approximation (INLA) [24].

Finally, our approach lies within a generalised linear mixed model framework. Hence, it can be easily extended to include additional information such as covariates (e.g. cell cycle stage, gene length) and experimental design (e.g. batch effects) using fixed and/or random effects.

Methods

A statistical model to detect changes in expression patterns for scRNA-seq datasets

In this article, we introduce a statistical model for identifying genes whose expression patterns change between pre-defined populations (given by experimental conditions or cell types) of cells. Such changes can be reflected via the overall expression level of each gene as well as through changes in cell-to-cell biological heterogeneity. Our method is motivated by features that are specific to scRNA-seq datasets. In this context, it is essential to appropriately normalise and remove technical artefacts from the data before extracting biological signal. This is particularly critical when there are substantial differences in cellular mRNA content, amplification biases and other sources of technical variation. For this purpose, we exploit technical spike-in genes which are added at the same quantity to each cell's lysate. A typical example is the set of 92 ERCC molecules developed by the External RNA Control Consortium [12]. Our method builds upon BASiCS [25] and can perform comparisons between multiple populations of cells using a single model. Importantly, our strategy avoids stepwise procedures where datasets are normalised prior to any downstream analysis. This is an advantage over methods using pre-normalised counts, as the normalisation step can be distorted by technical artefacts.

We assume that there are P groups of cells to be compared, each containing n_p cells ($p = 1, \ldots, P$). Let X_{ij_p} be a random variable representing the expression count of a gene i ($i = 1, \ldots, q$) in the j_p-th cell from group p. Without loss of generality, we assume the first q_0 genes are biological and the remaining $q - q_0$ are technical spikes. Extending the formulation in BASiCS, we assume that

$$\mathrm{E}(X_{ij_p}) = \begin{cases} \phi_{j_p} s_{j_p} \mu_{ip}, & i = 1, \ldots, q_0; \\ s_{j_p} \mu_{ip}, & i = q_0 + 1, \ldots, q. \end{cases} \quad \text{and} \tag{1}$$

$$\mathrm{CV}^2(X_{ij_p}) = \begin{cases} (\phi_{j_p} s_{j_p} \mu_{ip})^{-1} + \theta_p + \delta_{ip}(\theta_p + 1), & i = 1, \ldots, q_0; \\ (s_{j_p} \mu_{ip})^{-1} + \theta_p, & i = q_0 + 1, \ldots, q, \end{cases} \tag{2}$$

where CV stands for *coefficient of variation* (i.e. the ratio between standard deviation and mean). These expressions are the result of a Poisson hierarchical structure (see supplementary material). Here, ϕ_{j_p}'s act as cell-specific normalising constants (fixed effects), capturing differences in input mRNA content across cells (reflected by the expression counts of intrinsic transcripts only). A second set of normalising constants, s_{j_p}'s, capture cell-specific scale differences affecting the expression counts of all genes (intrinsic and technical). Among others, these differences can relate to sequencing depth, capture efficiency and amplification biases. However, a precise interpretation of the s_{j_p}'s varies across experimental protocols, e.g. amplification biases are removed when using UMIs[10]. In addition, θ_p's are global technical noise parameters controlling the over-dispersion (with respect to Poisson sampling) of all genes within group p. The overall expression rate of a gene i in group p is denoted by μ_{ip}. These are used to quantify changes in the overall expression of a gene across groups. Similarly, the δ_{ip}'s capture residual over-dispersion (beyond what is due to technical artefacts) of every gene within each group. These so-called biological over-dispersion parameters relate to heterogeneous expression of a gene across cells. For each group, stable "*housekeeping-like*" genes lead to $\delta_{ip} \approx 0$ (low residual variance in expression across cells) and highly variable genes are linked to large values of δ_{ip}. A novelty of our approach is the use of δ_{ip}'s to quantify changes in biological over-dispersion. Importantly — and unlike the CV — this avoids confounding effects due to changes in overall expression between the groups.

A graphical representation of this model is displayed in Figure 1. To ensure identifiability of all model parameters, we assume that μ_{ip}'s are known in the case of the spike-in genes (and given by the number of spike-in molecules that are added to each well). Additionally, we impose the identifiability restriction

$$\frac{1}{n_p}\sum_{j_p=1}^{n_p}\phi_{j_p}=1, \quad \text{for } p=1,\ldots,P. \tag{3}$$

Here, we discuss the priors assigned to parameters that are gene and group-specific (see Additional file 2 for the remaining elements of the prior). These are given by

$$\mu_{ip}, \overset{\text{iid}}{\sim} \text{log-N}(0,a_\mu^2) \quad \text{and} \quad \delta_{ip} \overset{\text{iid}}{\sim} \text{log-N}(0,a_\delta^2) \quad \text{for } i=1,\ldots,q_0, \tag{4}$$

Hereafter, without loss of generality, we simplify our notation to focus on two-group comparisons. This is equivalent to assigning Gaussian prior distributions for LFCs in overall expression (τ_i) or biological over-dispersion (ω_i). In such a case, it follows that

$$\tau_i \equiv \log(\mu_{i1}/\mu_{i2}) \sim \text{N}(0,2a_\mu^2) \quad \text{and} \quad \omega_i \equiv \log(\delta_{i1}/\delta_{i2}) \sim \text{N}(0,2a_\delta^2). \tag{5}$$

Hence, our prior is *symmetric*, meaning that we do not *a priori* expect changes in expression to be skewed towards either group of cells. Values for a_μ^2 and a_δ^2 can be elicited using an *expected* range of values for LFC in expression and biological over-dispersion, respectively. The latter is particularly useful in situations where a gene is not expressed (or very lowly expressed) in one of the groups, where e.g. LFCs in overall expression are undefined (the maximum likelihood estimate of τ_i would be $\pm\infty$, the sign depending on which group expresses gene i). A popular solution to this issue is the addition of *pseudo-counts*, where an arbitrary number is added to all expression counts (in all genes and cells). While the latter guarantees that τ_i is well-defined, it leads to artificial estimates for τ_i (see Table 1). Instead, our approach exploits an informative prior (indexed by a_μ^2) to *shrink* extreme estimates of τ_i towards an expected range. This strategy leads to a meaningful shrinkage strength, which is based on prior knowledge. Importantly — and unlike the addition of pseudo-counts — our approach is also helpful when comparing biological over-dispersion between the groups. In fact, if a gene i is not expressed in one of the groups, this will lead to a non-finite estimate of ω_i (if all expression counts in a group are equal to zero, the corresponding estimate of the biological over-dispersion parameters would be equal to zero). Adding pseudo-counts cannot resolve this issue, but imposing an informative prior for ω_i (indexed by a_ω^2) will shrink estimates towards the appropriate range.

Generally, posterior estimates of τ_i and ω_i are robust to the choice of a^2_μ and a^2_δ, as the data is informative and dominates posterior inference. In fact, these values are only influential when shrinkage is needed, e.g. when there are zero total counts in one of the groups. In such cases, posterior estimates of τ_i and ω_i are dominated by the prior, yet the method described below still provides a tool to quantify evidence of changes in expression. As a default option, we use $a^2_\mu = a^2_\delta = 0.5$.

Posterior samples for all model parameters are generated via an Adaptive Metropolis within Gibbs Sampling algorithm [22]. A detailed description of our implementation can be found in the supplementary material.

Post-hoc correction of global shifts in input mRNA content between the groups

The identifiability restriction in (3) does only apply to cells within each group. As a consequence, if they exist, global shifts in cellular mRNA content between groups (e.g. if all mRNAs where present at twice the level in one population related to another) are absorbed by the μ_{ip}'s. To correct this bias, we adopt a 2-step strategy where: (i) model parameters are estimated using the identifiability restriction in (3) and (ii) global shifts in input mRNA content are treated as a fixed *offset* and corrected post-hoc. For this purpose, we use the sum of overall expression rates (intrinsic genes only) as a proxy for the overall mRNA content within each group. Without loss of generality, we use the first group of cells as a reference population. For each population p ($p = 1, \ldots, P$), we define a population specific offset effect

$$\Lambda_p = \left(\sum_{i=1}^{q_0} \mu_{ip} \right) \Big/ \left(\sum_{i=1}^{q_0} \mu_{i1} \right) \tag{6}$$

and perform the following offset correction

$$\mu^*_{ip} = \mu_{ip} / \Lambda_p, \qquad \phi^*_{j_p} = \phi_{j_p} \times \Lambda_p, \qquad i = 1, \ldots, q_0; j_p = 1, \ldots n_p. \tag{7}$$

This is equivalent to replacing the identifiability restriction in (3) by

$$\frac{1}{n_p} \sum_{j_p=1}^{n_p} \phi_{j_p} = \Lambda_p, \qquad \text{for } p = 1, \ldots, P. \tag{8}$$

Technical details regarding the implantation of this post-hoc offset correction are explained in the supplementary material. The effect of this correction is illustrated in Figure 7 using the cell cycle dataset described in the main text. As an alternative, we also explored the use of the ratio between the total intrinsic counts over total spike-in counts to define a similar offset correction based on

$$\Lambda'_p = \left(\underset{j_p=1,\ldots,n_p}{\text{median}} \left\{ \frac{\sum_{i=1}^{q_0} X_{ij_p}}{\sum_{i=q_0+1}^{q} X_{ij_p}} \right\} \right) \Big/ \left(\underset{j_1=1,\ldots,n_1}{\text{median}} \left\{ \frac{\sum_{i=1}^{q_0} X_{ij_1}}{\sum_{i=q_0+1}^{q} X_{ij_1}} \right\} \right) \tag{9}$$

In the case of the cell cycle dataset, both alternatives are equivalent. Nonetheless, the first option is more robust in cases where a large number of differentially expressed genes are present. Hereafter, we use μ_{ip} and ϕ_{j_p} to denote μ^*_{ip} and $\phi^*_{j_p}$, respectively.

A probabilistic approach to quantify evidence of changes in expression patterns

A probabilistic approach is adopted, assessing changes in expression patterns (mean and over-dispersion) through a simple and intuitive scale of evidence. Our strategy is flexible and can be combined with a variety of decision rules. In particular, here we focus on highlighting genes whose absolute LFC in overall expression and biological over-dispersion between the populations exceeds minimum tolerance thresholds τ_0 and ω_0, respectively ($\tau_0, \omega_0 \geq 0$), set *a priori*.

For a given probability threshold $\alpha_{_M}$ ($0.5 < \alpha_{_M} < 1$), a gene i is identified as exhibiting a change in overall expression between populations p and p' if

$$\pi^M_{ipp'}(\tau_0) \equiv \text{P}(|\log(\mu_{ip}/\mu_{ip'})| > \tau_0 | \{\text{data}\}) > \alpha_{_M}, \quad i = 1, \ldots q_0. \tag{10}$$

If $\tau_0 \to 0$, $\pi^M_i(\tau_0) \to 1$ becoming uninformative to detect changes in expression. As in [2], in the limiting case where $\tau_0 = 0$, we define

$$\pi^M_{ipp'}(0) = 2\max\left\{\tilde{\pi}^M_{ipp'}, 1 - \tilde{\pi}^M_{ipp'}\right\} - 1 \text{ with} \tag{11}$$

$$\tilde{\pi}^M_{ipp'} = \text{P}(\log(\mu_{ip}/\mu_{ip'}) > 0 | \{\text{data}\}). \tag{12}$$

A similar approach is adopted to study changes in biological over-dispersion between populations p and p', using

$$\pi^D_{ipp'}(\omega_0) \equiv \text{P}(|\log(\delta_{ip}/\delta_{ip'})| > \omega_0 | \{\text{data}\}) > \alpha_{_D}, \tag{13}$$

for a fixed probability threshold $\alpha_{_D}$ ($0.5 < \alpha_{_D} < 1$). In line with (11) and (12), we also define

$$\pi^D_{ipp'}(0) = 2\max\left\{\tilde{\pi}^D_{ipp'}, 1 - \tilde{\pi}^D_{ipp'}\right\} - 1 \text{ with} \tag{14}$$

$$\tilde{\pi}^D_{ipp'} = \text{P}(\log(\delta_{ip}/\delta_{ip'}) > 0 | \{\text{data}\}). \tag{15}$$

Evidence thresholds $\alpha_{_M}$ and $\alpha_{_D}$ can be fixed *a priori*. Otherwise, these can be defined by controlling the expected false discovery rate (EFDR) [19]. In our context, these are given by

$$\text{EFDR}_{\alpha_M}(\tau_0) = \frac{\sum_{i=1}^{q_0}(1 - \pi^M_i(\tau_0))\text{I}(\pi^M_i(\tau_0) > \alpha_{_M})}{\sum_{i=1}^{q_0} I(\pi^M_i(\tau_0) > \alpha_{_M})}, \quad \text{and} \tag{16}$$

$$\mathrm{EFDR}_{\alpha_D}(\omega_0) = \frac{\sum_{i=1}^{q_0}(1-\pi_i^D(\omega_0))\mathrm{I}(\pi_i^D(\omega_0) > \alpha_D)}{\sum_{i=1}^{q_0} I(\pi_i^D(\omega_0) > \alpha_D)}, \tag{17}$$

where $\mathrm{I}(A) = 1$ if event A is true, 0 otherwise.

The posterior probabilities in (10), (11), (13) and (15) can be easily estimated — as a post-processing step — once the model has been fitted (see supplementary material). In addition, our strategy is flexible and can be easily extended to investigate of more complex hypotheses, which can be defined post-hoc, e.g. to identify those genes that show significant changes in cell-to-cell biological over-dispersion but that maintain a constant level of overall expression between the groups.

Software

Our implementation is freely available as an R package [21], using a combination of R and C++ functions through the Rcpp library [8].

This can be found in `https://github.com/catavallejos/BASiCS`.

Availability of supporting data

All datasets analysed in this article are publicly available in the cited references.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

CAV, JCM and SR conceived and designed the methods and experiments. CAV implemented the methods and analysed the data. All authors were involved in writing the paper and have approved the final version.

Acknowledgements

We acknowledge all members of Richardson research group (MRC-BSU) and Marioni laboratory (EMBL-EBI) for support and discussions during the preparation of this document. In particular, we are grateful to Nils Eling (EMBL-EBI), Antonio Scialdone (EMBL-EBI) and Aaron Lun (CRUK) for numerous discussions and suggestions than enriched the final version of the manuscript. JCM and CAV acknowledge core EMBL funding. SR and CAV acknowledge core MRC funding. JCM acknowledges core support from CRUK.

References

[1] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[2] N. Bochkina and S. Richardson. Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics*, 63(4):1117–1125, 2007.

[3] M. N. Boddy and P. Russell. DNA replication checkpoint. *Current Biology*, 11(23):R953–R956, 2001.

[4] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095, 2013.

[5] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33:155–160, 2015.

[6] A. Clemens. *Protein phosphorylation in cell growth regulation*. Harwood Academic Publishers, Amsterdam, The Netherlands, 1st edition, 1996.

[7] Z. Darzynkiewicz, H. Crissman, F. Traganos, and J. Steinkamp. Cell heterogeneity during the cell cycle. *Journal of Cellular Physiology*, 113(3):465–474, 1982.

[8] D. Eddelbuettel, R. François, J. Allaire, J. Chambers, D. Bates, and K. Ushey. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[9] D. Grün, L. Kester, and A. van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, 2014.

[10] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, 2014.

[11] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

[12] L. Jiang, F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9):1543–1551, 2011.

[13] S. Katayama, V. Töhönen, S. Linnarsson, and J. Kere. SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics*, 29(22):2943–2945, 2013.

[14] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2014.

[15] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[16] A. A. Kolodziejczyk, J. K. Kim, J. C. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Bühler, P. Liu, et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485, 2015.

[17] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[18] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[19] M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

[20] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[22] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[23] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[24] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(2):319–392, 2009.

[25] C. A. Vallejos, J. C. Marioni, and S. Richardson. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Computational Biology*, 11(6):e1004333, 2015.

[26] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015.

Figures

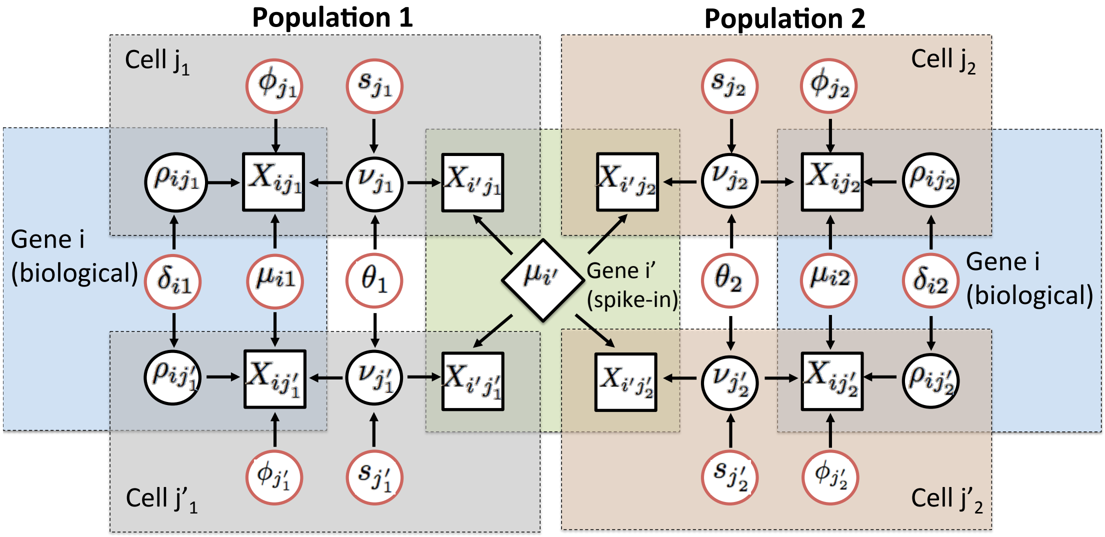

Figure 1: **Graphical representation of our model for detecting changes in expression patterns (mean and over-dispersion) based on the comparison of two pre-defined population of cells.** The diagram considers expression counts of 2 genes (i: biological and i': technical) and 2 cells (j_p and j'_p) from each population $p = 1, 2$. Observed expression counts are represented by square nodes. The central rhomboid node denotes the known input number of mRNA molecules for a technical gene i', which is assumed to be constant across all cells. The remaining circular nodes represent unknown elements, using black to denote random effects and red to denote model parameters (fixed effects) that lie on the top of the model's hierarchy. Here, ϕ_{j_p}'s and s_{j_p}'s act as normalising constants that are cell-specific and θ_p's are global over-dispersion parameters capturing technical variability, which affects the expression counts of all genes and cells within each population. Finally, μ_{ip}'s and δ_{ip}'s respectively measure overall expression of a gene i and its residual biological cell-to-cell over-dispersion (after normalisation, technical noise removal and adjustment for overall expression) within each population. Coloured areas highlight elements that are shared within a gene and/or cell. The latter emphasises how our model borrows information across all cells to estimate parameters that are gene-specific and all genes to estimate parameters that are cell-specific. More details regarding the model setup can be found in the Methods section of this article.

.

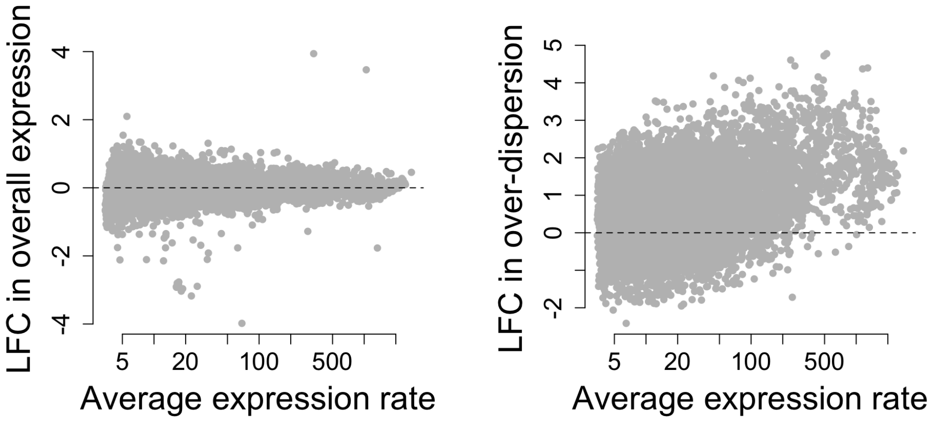

Figure 2: **Estimated LFCs in expression (mean and over-dispersion) when comparing single-cells versus pool-and-split samples (*2i* serum culture).** Posterior medians of LFC in overall expression $\log(\mu_{i\text{SC}}/\mu_{i\text{P\&S}})$ (left panel) and biological over-dispersion $\log(\delta_{i\text{SC}}/\delta_{i\text{P\&S}})$(right panel) against the average between estimates of overall expression rates for single-cells (SC) and pool-and-split (P&S) samples. Average values are defined as a weighted average between groups, with weights given by the number of samples within each group of cells. As expected, our analysis does not reveal major changes in expression levels between SC and P&S samples. In fact, the distribution of estimated LFCs in overall expression is roughly symmetric with respect to the origin. In contrast, we infer a substantial decrease in biological over-dispersion in the P&S samples. This is reflected by a skewed distribution of estimated LFCs in biological over-dispersion towards positive values.

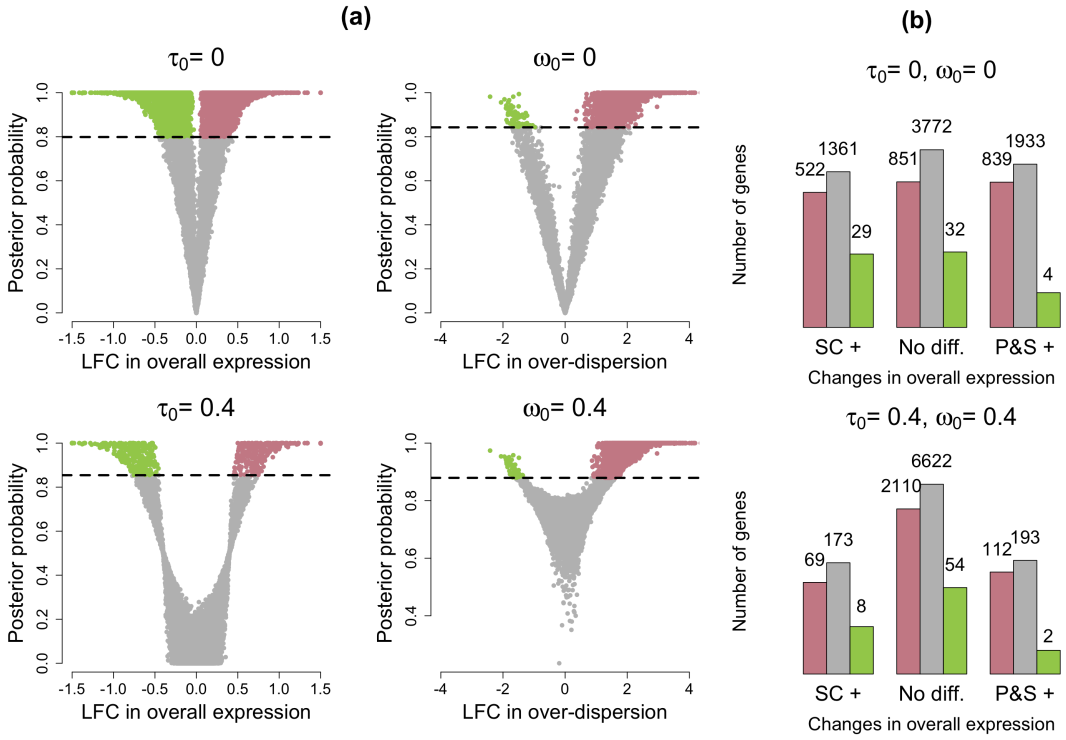

Figure 3: **Summary of changes in expression patterns (mean and over-dispersion) for single-cells versus pool-and-split samples (EFDR = 5%).** (a) Volcano plots showing posterior medians of LFCs against estimated tail posterior probabilities. Left panels relate to the test where we assess if the absolute LFC in overall expression between single-cells (SC) and pool-and-split (P&S) samples exceeds a minimum threshold τ_0. Estimates for LFCs in overall expression are truncated to the range $(-1.5, 1.5)$. Pink and green dots represent genes highlighted to have higher overall expression in the SC and P&S samples, respectively. Right panels relate to the test where we assess if the absolute LFC in biological over-dispersion between SC and P&S samples exceeds a minimum threshold ω_0. In all cases, horizontal dashed lines are located at probability cutoffs defined by EFDR = 5%. Pink and green dots represent genes highlighted to have higher biological over-dispersion in the SC and P&S samples, respectively. (b) Bins in the horizontal axis summarise changes in overall expression between the groups. We use SC+ and P&S+ to denote higher overall expression was detected in SC and P&S+ samples, respectively (the central group of bars (No diff.) corresponds to those genes where no significant differences were found). Coloured bars within each group summarise changes in biological over-dispersion between the groups. We use pink and green bars to denote higher biological over-dispersion in SC and P&S+ samples, respectively (and grey to denote no significant differences were found). Number of genes are displayed in log-scale.

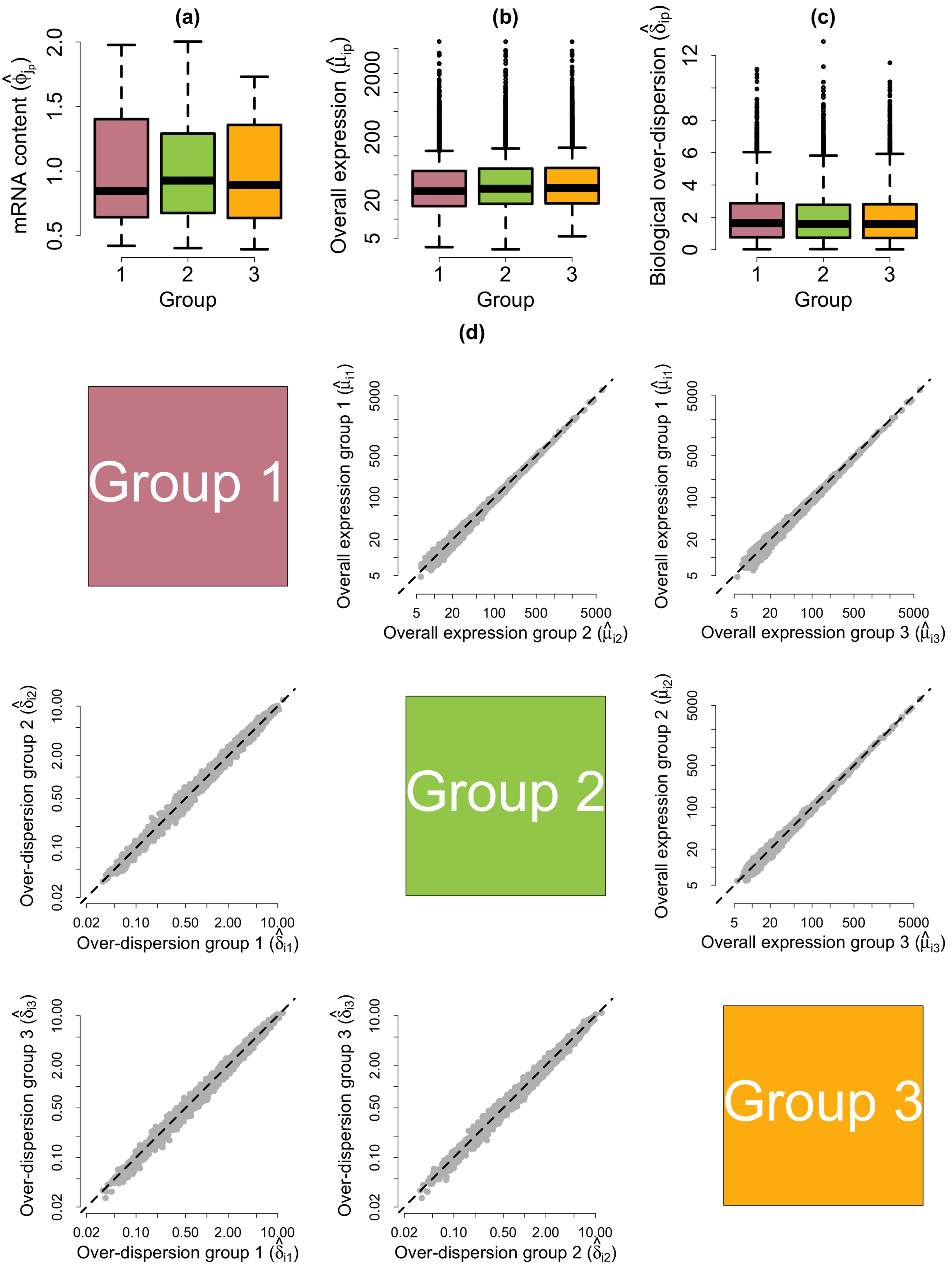

Figure 4: **Posterior estimates of model parameters based on random permutations of the mESC cell cycle dataset.** For a single permuted dataset (a) Empirical distribution of posterior medians for mRNA content normalising constants ϕ_{j_p} across all cells. (b) Empirical distribution of posterior medians for gene-specific expression rates μ_{ip} across all genes. (c) Empirical distribution of posterior medians for gene-specific biological over-dispersion parameters δ_{ip} across all genes. (d) As an average across 10 random permutations. Upper diagonal panels compare estimates for gene-specific expression rates μ_{ip} between groups of cells. Lower diagonal panels compare gene-specific biological over-dispersion parameters δ_{ip} between groups of cells.

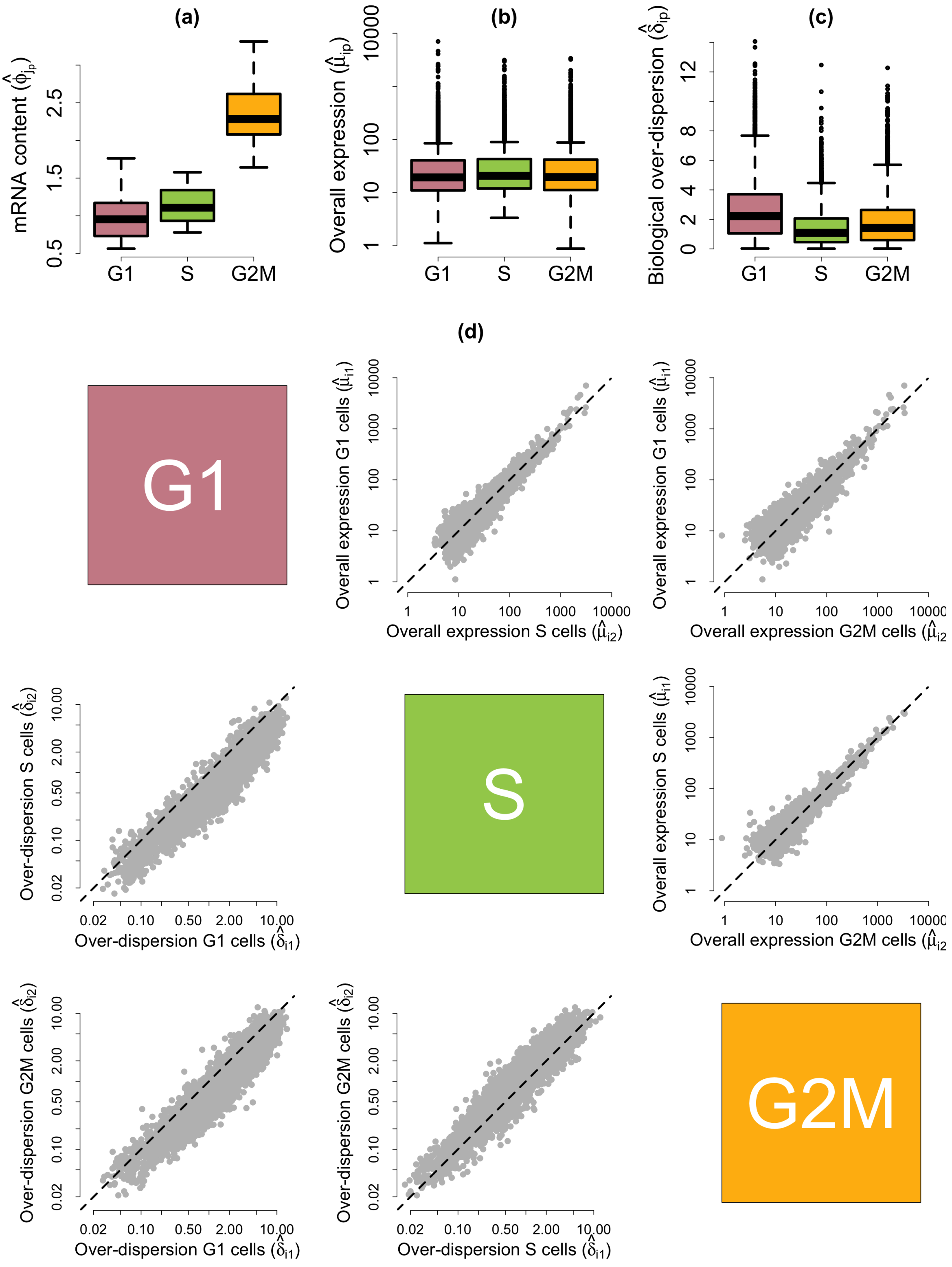

Figure 5: **Posterior estimates of model parameters for mESCs across different cell cycle phases.** (a) Empirical distribution of posterior medians for mRNA content normalising constants ϕ_{j_p} across all cells. (b) Empirical distribution of posterior medians for gene-specific expression rates μ_{ip} across all genes. (c) Empirical distribution of posterior medians for gene-specific biological over-dispersion parameters δ_{ip} across all genes. (d) Upper diagonal panels compare estimates for gene-specific expression rates μ_{ip} between groups of cells. Lower diagonal panels compare gene-specific biological over-dispersion parameters δ_{ip} between groups of cells.

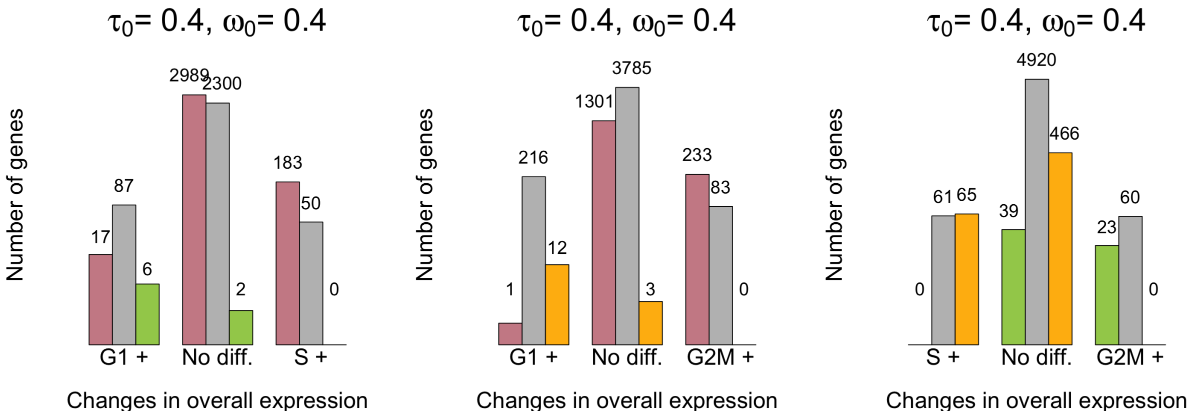

Figure 6: **Summary of changes in expression patterns (mean and over-dispersion) for the mESC cell cycle dataset (EFDR = 5%).** Bins in the horizontal axis summarise changes in overall expression between each pair of groups. We use G1+, S+ and G2M+ to denote higher overall expression was detected in cell cycle phase G1, S and G2M, respectively (the central group of bars (No diff.) corresponds to those genes where no significant differences were found). Coloured bars within each group summarise changes in biological over-dispersion between the groups. We use pink, green and yellow bars to denote higher biological over-dispersion in cell cycle phases G1, S and G2M, respectively (and grey to denote no significant differences were found). Number of genes are displayed in log-scale.

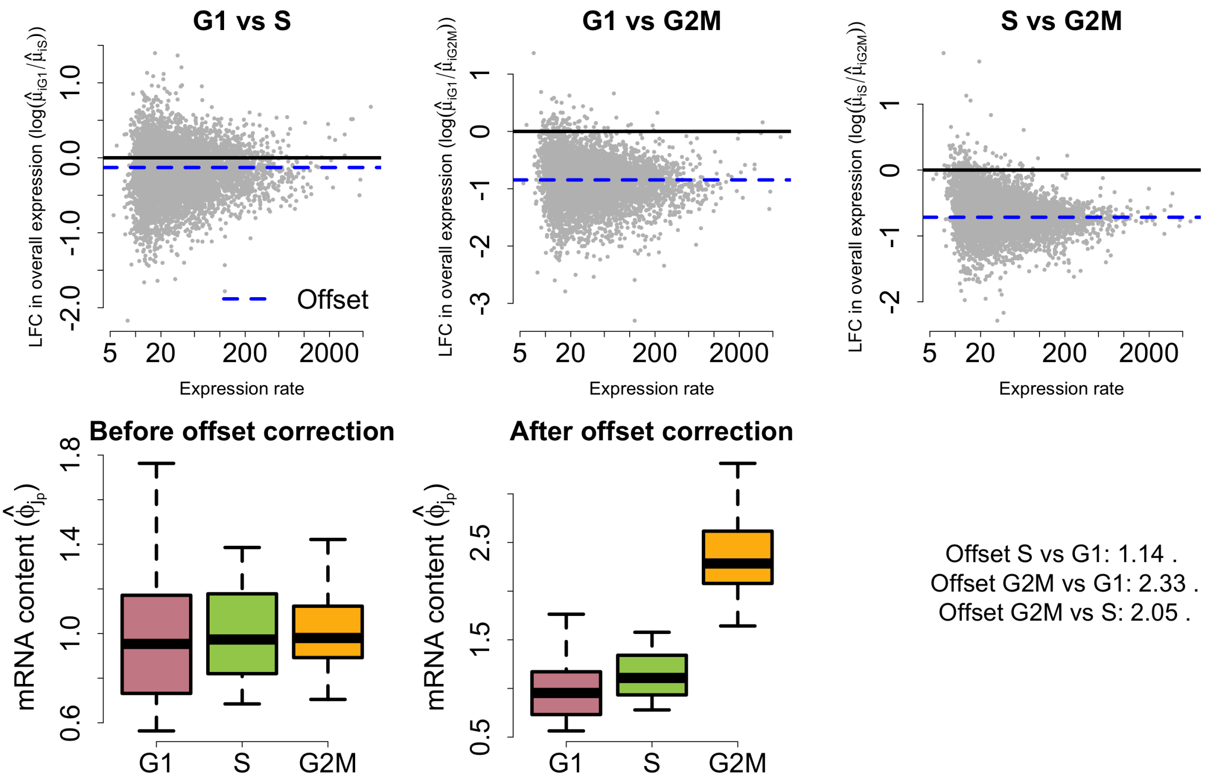

Figure 7: **Post-hoc offset correction for cell cycle dataset.** Upper panels display posterior medians for LFC in overall expression against the weighted average between estimates of overall expression rates for G1, S and G2M cells (weights defined by the number of cells in each group). Lower panels illustrate the effect of the offset correction upon the empirical distribution of posterior estimates for mRNA content normalising constants ϕ_{j_p}. These figures illustrate a shift in mRNA content throughout cell cycle phases. In particular, our model infers that cellular mRNA is roughly duplicated when comparing G1 to G2M cells.

Tables

Table 1: Synthetic example to illustrate the effect of pseudo-counts addition over the estimation of log-fold changes in overall expression. For simplicity, we assume that normalisation is not required so that pseudo-counts are linearly reflected in the overall expression rates. While pseudo-counts introduce an additive effect, log-fold change estimates measure changes on a multiplicative scale. Hence, pseudo-counts addition leads to an artificial deflation of log-fold change estimates. As a consequence, such estimates cannot be meaningfully interpreted.

	Empirical	Adding 0.5	Adding 1
	estimate	pseudo-counts	pseudo-count
Overall expression rate in population 1 (μ_{i1})	10	10.5	11
Overall expression rate in population 2 (μ_{i2})	0	0.5	1
Log-fold change in overall expression 1 vs 2	$+\infty$	3.04	2.40

Additional Files

Additional file 1 — Supplementary material

This document contain additional details regarding the statistical model presented in this article and the implementation of Bayesian inference. Additionally, it displays the R code used for the data analysis.