Supplementary material
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]

Contents

[1] MRC Biostatistics Unit

[2] EMBL-European Bioinformatics Institute

[3] Cancer Research UK Cambridge Institute

1 Methodology

1.1 The model

We assume there are P groups of cells to be compared, each containing n_p cells ($p = 1, \dots, P$). Let X_{ij_p} be a random variable representing the expression count of a gene i ($i = 1, \dots, q$) in the j_p-th cell from group p. 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, Jiang et al., 2011). These provide an internal control or "gold standard", to estimate the strength of technical variability and to aid normalisation. Without loss of generality, we assume the first q_0 genes are biological and the remaining $q - q_0$ are technical spikes. Our model builds upon BASiCS (Vallejos et al., 2015), a Bayesian model for the analysis of single-cell RNA-seq (scRNA-seq) data. For each population of cells p, our extended model is given by

$$X_{ij_p}|\mu_{ip}, \phi_{j_p}, \nu_{j_p}, \rho_{ij_p} \overset{ind}{\sim} \begin{cases} \text{Poisson}(\phi_{j_p}\nu_{j_p}\mu_{ip}\rho_{ij_p}), & i = 1, \dots, q_0, j_p = 1, \dots, n_p; \\ \text{Poisson}(\nu_{j_p}\mu_{ip}), & i = q_0 + 1, \dots, q, j_p = 1, \dots, n_p, \end{cases} \tag{1}$$

$$\text{with} \quad \nu_{j_p}|s_{j_p}, \theta \overset{ind}{\sim} \text{Gamma}(1/\theta, 1/(s_{j_p}\theta)) \quad \text{and} \quad \rho_{ij_p}|\delta_{ip} \overset{ind}{\sim} \text{Gamma}(1/\delta_{ip}, 1/\delta_{ip}). \tag{2}$$

Here, ϕ_{j_p}'s act as cell-specific normalising constants (fixed effects), to capture differences in input mRNA content between 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 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 unique molecular identifiers (Islam et al., 2014). The random effects ν_{j_p} (with $\text{E}(\nu_{j_p}|s_{j_p}, \theta_p) = s_{j_p}$ and $\text{Var}(\nu_{j_p}|s_{j_p}, \theta) = s_{j_p}^2\theta$) capture unexplained technical noise, which leads to a variance inflation (with respect to Poisson sampling) of all expression counts within each group of cells. For each population, the strength of this technical component of variability is quantified through a single hyper-parameter θ_p, borrowing information across all genes and cells. A second set of random effects ρ_{ij_p} (with $\text{E}(\rho_{ij_p}|\delta_{ip}) = 1$ and $\text{Var}(\rho_{ij_p}|\delta_{ip}) = \delta_{ip}$), capture heterogeneous expression of a gene across cells. This is quantified through the δ_{ip}'s, capturing residual over-dispersion (beyond what is due to technical artefacts) of every gene within each group. 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 commonly used coefficient of variation — this avoids confounding effects due to changes in overall expression between the groups. Finally, 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 between groups of cells.

1.2 Prior specification

We assume prior independence between all model parameters. In Vallejos et al. (2015), an improper *non-informative* prior is assigned to the overall expression rates μ_{ip}'s. However, it does not lead to a proper posterior distribution when all the expression counts of a gene are equal to zero at all cells within a population. The latter often occurs when comparing distinct populations of cells, where population-specific markers are likely to lie within this category. As an alternative, we assign a proper prior distribution to the overall expression rates μ_{ip}'s. This is given by

$$\mu_{ip}, \overset{\text{iid}}{\sim} \text{log-N}(0, a_\mu^2) \quad \text{for } i = 1, \ldots, q_0, \tag{3}$$

A similar prior is assigned to the biological over-dispersion parameters δ_{ip}'s using

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

As discussed in the manuscript, the latter is equivalent to assigning Gaussian prior distributions for log-fold changes (LFC) in overall expression or biological over-dispersion. These are symmetric with respect to the origin, meaning that we do not *a priori* expect changes in expression to be skewed towards either group of cells. Moreover, this prior specification is helpful in situations where a gene is not expressed (or very lowly expressed) in one of the groups, where the values of a_μ^2 and a_δ^2 allow shrinkage of LFC estimates towards an appropriate range (e.g. to avoid "infinite" LCF estimates when a gene has zero total counts within one population of cells). Importantly, genes for which expression was detected in all populations are not affected by the choice of these hyper-parameter values. As a default option we use $a_\mu^2 = a_\delta^2 = 0.5$.

We also assign proper prior distributions to the remaining model parameters, using

$$\begin{aligned} s_{j_p} &\sim \text{Gamma}(a_s, b_s), \quad j_p = 1, \ldots, n_p; p = 1, \ldots, P \quad (5) \\ \theta_p &\sim \text{Gamma}(a_\theta, b_\theta), \quad p = 1, \ldots, P \quad (6) \\ \Phi_p &\sim n_p\text{Dirichlet}(a_{\Phi_p}), \quad \Phi_p = (\phi_{j1}, \ldots, \phi_{j_{n_p}})'; p = 1, \ldots, P \quad (7) \end{aligned}$$

By default, we set $a_s = b_s = a_\theta = b_\theta = 1$ and $a_{\Phi_p} = \mathbf{1}_{n_p}$, where $\mathbf{1}_{n_p}$ denotes an n_p-dimensional vector of ones.

1.3 Markov Chain Monte Carlo implementation

Posterior inference is implemented via a Markov Chain Monte Carlo (MCMC) algorithm, generating draws from the posterior distribution of all model parameters. In particular, we use an

Adaptive Metropolis within Gibbs Sampling algorithm (Roberts and Rosenthal, 2009), where the variance of the proposal distributions are internally tuned to achieve an optimal acceptance rate. However, sampling the random effects ρ_{ij_p}'s throughout the algorithm results in a slow convergence (despite allowing conjugate updates of other model parameters). This is particularly critical when the sample size increases. To overcome this problem, we implemented Bayesian inference based on the marginal model obtained after integrating out the ρ_{ij_p}'s, i.e.

$$X_{ij_p}|\mu_{ip},\delta_{ip},\phi_{j_p},\nu_{j_p},\theta \sim \begin{cases} \text{Neg-Binomial}\left(\delta_{ip}^{-1}, \frac{\phi_{j_p}\nu_{j_p}\mu_{ip}}{\phi_{j_p}\nu_{j_p}\mu_{ip}+\delta_{ip}^{-1}}\right), & i=1,\ldots,q_0, j_p=1,\ldots,n_p; \\ \text{Poisson}(\nu_{j_p}\mu_{ip}), & i=q_0+1,\ldots,q, j_p=1,\ldots,n_p, \end{cases} \tag{8}$$

for which the associated likelihood function is given by

$$\left[\prod_{i=1}^{q_0}\prod_{p=1}^{P}\prod_{j_p=1}^{n_p}\frac{\Gamma(x_{ij_p}+1/\delta_{ip})}{\Gamma(1/\delta_{ip})x_{ij_p}!}\left(\frac{1/\delta_{ip}}{\phi_{j_p}\nu_{j_p}\mu_{ip}+1/\delta_{ip}}\right)^{1/\delta_{ip}}\left(\frac{\phi_{j_p}\nu_{j_p}\mu_{ip}}{\phi_{j_p}\nu_{j_p}\mu_{ip}+1/\delta_{ip}}\right)^{x_{ij_p}}\right] \times \tag{9}$$

$$\left[\prod_{i=q_0+1}^{q}\prod_{p=1}^{P}\prod_{j_p=1}^{n_p}\frac{(\nu_{j_p}\mu_{ip})^{x_{ij_p}}}{x_{ij_p}!}e^{-\nu_{j_p}\mu_{ip}}\right].$$

Under this specification, the full conditionals required for the implementation correspond to

$$\pi(\mu_{ip}|\cdots) \propto \frac{\mu_{ip}^{\sum_{j_p=1}^{n_p}x_{ij_p}}}{\prod_{j_p=1}^{n_p}(\phi_{j_p}\nu_{j_p}\mu_{ip}+1/\delta_{ip})^{x_{ij_p}+1/\delta_{ip}}} \times \exp\left\{-\frac{1}{2a_\mu^2}(\log(\mu_{ip}))^2\right\}, \tag{10}$$

$$\pi(\delta_{ip}|\cdots) \propto \left[\prod_{j_p=1}^{n_p}\frac{\Gamma(x_{ij_p}+1/\delta_{ip})}{\Gamma(1/\delta_{ip})}\frac{(1/\delta_{ip})^{1/\delta_{ip}}}{(\phi_{j_p}\nu_{j_p}\mu_{ip}+1/\delta_{ip})^{x_{ij_p}+1/\delta_{ip}}}\right] \times \exp\left\{-\frac{1}{2a_\delta^2}(\log(\delta_{ip}))^2\right\}, \tag{11}$$

$$\pi(s_{j_p}|\cdots) \propto s_{j_p}^{a_s-(1/\theta_p)-1}\exp\left\{-\frac{\nu_{j_p}}{s_{j_p}\theta_p}-s_{j_p}b_s\right\}, \tag{12}$$

$$\pi(\nu_{j_p}|\cdots) \propto \left[\prod_{i=1}^{q_0}\frac{\nu_{j_p}^{x_{ij_p}}}{(\phi_{j_p}\nu_{j_p}\mu_{ip}+1/\delta_i)^{x_{ij_p}+1/\delta_{ip}}}\right]\left[\prod_{i=q_0+1}^{q}\nu_{j_p}^{x_{ij_p}}e^{-\nu_{j_p}\mu_i}\right]\nu_{j_p}^{(1/\theta_p)-1}e^{-\nu_{j_p}/(\theta_p s_{j_p})}, \tag{13}$$

$$\pi(\theta_p|\cdots) \propto \frac{\left(\prod_{j_p=1}^{n_p}(\nu_{j_p}/s_{j_p})\right)^{1/\theta_p}}{\Gamma^{n_p}(1/\theta_p)}\theta_p^{a_\theta-(n_p/\theta_p)-1}e^{-(1/\theta_p)\sum_{j_p=1}^{n_p}(\nu_{j_p}/s_{j_p})-b_\theta\theta_p} \tag{14}$$

$$\pi(\Phi_p|\cdots) \propto \frac{\prod_{i=1}^{q_0}\phi_{j_p}^{\sum_{j_p=1}^{n_p}x_{ij_p}}}{\prod_{i=1}^{q_0}\prod_{j_p=1}^{n_p}(\phi_{j_p}\nu_{j_p}\mu_{ip}+1/\delta_{ip})^{x_{ij_p}+1/\delta_{ip}}} \times \pi(\Phi_p), \text{ with } \pi(\Phi_p) \text{ as in (7)}, \tag{15}$$

for $i=1,\ldots,q_0$, $j_p=1,\ldots,n_p$ and $p=1,\ldots,P$. To sample from the posterior distribution of all model parameters, we use Gaussian Random walks for these full conditionals (10)-(14) and Dirichlet proposals for the full conditional in (15).

Our implementation is freely available as an R package R Core Team (2014), using a combination of R and C++ functions through the Rcpp library (Eddelbuettel et al., 2011). This can be found in `https://github.com/catavallejos/BASiCS`.

1.4 Post-hoc offset correction of global shifts in mRNA content between groups

To ensure identifiability of all model parameters, we introduce the identifiability restriction

$$\frac{1}{n_p}\sum_{j_p=1}^{n_p}\phi_{j_p}=1, \quad \text{for } p=1,\ldots,P. \tag{16}$$

This restriction 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 the 2-step strategy described below.

(i) **Estimation step.** Model parameters are estimated under the identifiability restriction in (16), using the MCMC algorithm described in Section 1.3. For each parameter, this algorithm generates a sample of N random draws from the associated posterior distribution. In particular, for each μ_{ip} and ϕ_{j_p}, we denote these samples by $\{\mu_{ip}^{(1)},\ldots,\mu_{ip}^{(N)}\}$ and $\{\phi_{j_p}^{(1)},\ldots,\phi_{j_p}^{(N)}\}$, respectively.

(ii) **Offset correction step.** Once the model has been fitted, 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 $\sum_{i=1}^{q_0}\mu_{ip}$ (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 and define population-specific offset effects as

$$\Lambda_p=\left(\sum_{i=1}^{q_0}\mu_{ip}\right)\Big/\left(\sum_{i=1}^{q_0}\mu_{i1}\right), \quad p=1,\ldots,P \tag{17}$$

To estimate these quantities, we firstly create MCMC samples for each Λ_p ($p=1,\ldots,P$), denoted by $\{\Lambda_p^{(1)},\ldots,\Lambda_p^{(N)}\}$ with $\Lambda_p^{(m)}=\sum_{i=1}^{q_0}\mu_{ip}^{(m)}$. Secondly, for each population p ($p=1,\ldots,P$), we estimate these offset effects as the posterior medians of each Λ_p, i.e.

$$\hat{\Lambda}_p=\underset{m=1,\ldots,M}{\text{median}}\left\{\Lambda_p^{(m)}\right\} \tag{18}$$

Finally, offset corrected MCMC samples for each μ_{ip} and ϕ_{j_p} are generated as

$$\mu_{ip}^{*(m)}=\mu_{ip}^{(m)}/\hat{\Lambda}_p \quad \phi_{j_p}^{*(m)}=\phi_{j_p}^{(m)}\hat{\Lambda}_p, \tag{19}$$

for $m=1,\ldots,M$, $i=1,\ldots,q_0$, $j_p=1,\ldots,n_p$ and $p=1,\ldots,P$. These *offset corrected* chains are returned as an output of the MCMC sampler implemented in our R library.

1.5 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 $\alpha_{_D}$ ($0.5 < \alpha_{_D} < 1$), a gene i is identified to exhibit a change in biological over-dispersion between populations p and p' if

$$\pi^D_{ipp'}(\omega_0) \equiv \mathrm{P}(|\log(\delta_{ip}/\delta_{ip'})| > \omega_0|\{\text{data}\}) > \alpha_{_D}, \quad i = 1, \ldots q_0. \tag{20}$$

An empirical estimate of this tail posterior probability can be easily obtained using an MCMC sample from the posterior distribution of the δ_{ip}'s. In fact, this quantity can be estimated as

$$\hat{\pi}^D_{ipp'}(\omega_0) = \frac{1}{N}\sum_{m=1}^{N} \mathbf{I}\left(|\log\left(\delta^{(m)}_{ip}/\delta^{(m)}_{ip'}\right)| > \omega_0\right), \tag{21}$$

where $\delta^{(m)}_{ip}$ denotes the m-th posterior sample from δ_{ip} and where $\mathbf{I}(A)$ is an indicator function equal to 1 if the event A is true and 0 otherwise. If $\omega_0 \to 0$, $\pi^D_i(\omega_0) \to 1$ becoming uninformative to detect changes in biological over-dispersion. As in Bochkina and Richardson (2007), in the limiting case where $\tau_0 = 0$, we define

$$\pi^D_{ipp'}(0) = 2\max\left\{\tilde{\pi}^D_{ipp'}, 1 - \tilde{\pi}^D_{ipp'}\right\} - 1 \text{ with } \tilde{\pi}^D_{ipp'} = \mathrm{P}(\log(\delta_{ip}/\delta_{ip'}) > 0|\{\text{data}\}). \tag{22}$$

In line with (21), the latter can be estimated as

$$\hat{\tilde{\pi}}^D_{ipp'} = \frac{1}{N}\sum_{m=1}^{N} \mathbf{I}\left(\log\left(\delta^{(m)}_{ip}/\delta^{(m)}_{ip'}\right) > 0\right). \tag{23}$$

Analogous expression are used for the detection of genes whose overall expression rate changes between the populations. In such a case, we replace $\log(\delta_{ip}/\delta_{ip'})$, ω_0 and α_D by $\log(\mu_{ip}/\mu_{ip'})$, τ_0 and α_M, respectively.

References

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

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

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

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

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

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