Generalized Adaptive Shrinkage Methods and Applications in Genomics Studies

Mengyin Lu

1 Gene expression distribution deconvolution of scRNA-seq data

1.1 Introduction

In recent years, single cell RNA-seq (scRNA-seq) methods have gained substantial popularity in analyzing gene expression data. While more traditional methods like microarray and bulk RNA-seq technologies mostly depend on estimating the average gene expression level from millions of cells, individual cells often vary in their gene expression levels, which captures the dynamics of cell transcription. Typical scRNA-seq data contains the expression profile of individual cells, thereby allowing for the quantification of a much richer set of features that can capture the heterogeneity of gene expression across cells. Apart from the mean expression level, measures like dispersion (e.g. coefficient of variation) and nonzero fraction are also useful in various contexts. For instance, dispersion can provide additional information about biological states that are not captured by the population mean alone [16, 12, 20, 15]. Burstiness, on the other hand, can help us better understand transcriptional regulation at the single cell level [8, 16].

Nevertheless, isolating technical noise (henceforth simply called noise) from useful signals in scRNA-seq data and accurately estimating the relevant statistics (mean, dispersion, burstiness, etc) of true gene expression distribution is a challenging task. Compared to bulk RNA-seq, scRNA-seq techniques require more complicated procedures, resulting in elevated noise levels in the subsequent data pipeline. Unique Molecular Identifiers (UMI) [11] were introduced as a barcoding technique to reduce amplification noise, but the distribution of observed UMI counts is still insufficient for inferring the true expression distribution in many cases. Moreover, scRNA-seq data often prove challenging for classical bulk RNA-seq analysis methods to tackle primarily because of the prevalence of burstiness, which can lead to zero-inflated data.

Recently, Wang et al. [19] developed *DESCEND*, a statistical method that deconvolves the true cross-cell gene expression distribution from observed scRNA-seq UMI counts. *DESCEND* adaptively fits the true gene expression distribution using the "G-modeling" empirical Bayes distribution deconvolution approach [6]. The "G-modeling" technique only assumes the distribution to be a general exponential family distribution (with natural spline basis), which is highly flexible and avoids specifying parametric assumptions.

Inspired by *DESCEND*, we further propose a general deconvolution framework for scRNA-seq data, which decouples the technical sampling errors from UMI counts and then recover the true gene expression distribution. In addition, we also introduce three methods with different distributional assumptions: *ZINB*, Poisson *ash* and nonparametric deconvolution. For the true expression distribution, *ZINB* makes the zero-inflated gamma distribution assumption, Poisson *ash* only requires the assumption of unimodality for the non-zero part, and nonparametric deconvolution is assumption-free. Along with *DESCEND*, the four methods assumes standard Poisson sampling errors for scRNA-seq data, but retain different levels of flexibility and adaptivity due to their distributional assumptions.

We compare the estimated expression distribution, corresponding statistics and goodness of fit of the four methods on two real scRNA-seq datasets, Zeisel data [20] and Tung data [18]. We show that the methods often provide similar mean and dispersion statistics despite of the discrepancies in shape of fitted expression distribution. However, our proposed method Poisson *ash* normally achieves better fit in terms of higher likelihood, and better highlights the sub-population structure preserved in Zeisel data.

1.2 Methods

Suppose we observe UMI counts from an scRNA-seq experiment. Define observed count Y_{cg} for gene g in cell c to be the true gene expression level μ_{cg} plus additional noise. Since the underlying structure of true gene expression levels is of interest, we would like to deconvolve the variability of Y_{cg} into two parts, noise and the variability of true gene expression:

$$Y_{cg} \sim F_{cg}(\mu_{cg}), \quad \mu_{cg} \sim G_g(\cdot), \tag{1}$$

where F_{cg} captures the noise and G_g represents the true expression distribution of gene g across cells. After deconvolving G_g from the noisy observed counts Y_{cg}, we can further estimate other distribution-related quantities of interest (mean, CV, etc.).

Several studies [2, 10, 9, 7] have examined the public scRNA-seq datasets and proposed Poisson distribution to capture the noise in UMI counts after accounting for cross-cell differences in library size. Hence, the Poisson distribution is a suitable choice for F_{cg}:

$$Y_{cg} \sim \text{Poisson}(\alpha_c \lambda_{cg}), \quad \lambda_{cg} \sim G_g(\cdot), \tag{2}$$

where α_c is a cell specific scaling constant. A straightforward way to set α_c is to simply use the library size (total UMI count of cell c). Other moderated scaling factors (e.g, robust normalized scaling factors, efficiency of cell if spike-ins are available) may be suitable alternatives, depending on context. Here λ_{cg} represents the relative gene expression level for gene g in cell c. In genomic contexts, developing models for relative gene expression is typically more sensible than directly working on the absolute expression μ_{cg}.

A crucial property of scRNA-seq data is the zero-inflated pattern. The zeros could be caused by factors like the variation in expression across cells, or the bursty process of gene transcription, where periods of RNA synthesis is followed by periods of inactivity [4, 14, 5]. The UMI counts of scRNA-seq typically yield a zero-inflated gene expression distribution: zero counts in “inactivate” cells and non-zero counts in “active” cells. Therefore a point mass at zero should be included in G_g to incorporate this pattern:

$$G_g = \pi_g \delta_0 + (1 - \pi_g) H_g, \tag{3}$$

where π_g is the zero fraction (inactive cell fraction), δ_0 is the point mass at zero, and H_g is the gene expression distribution corresponding to active cells.

We then fit the deconvolution distribution G_g using the observed counts. The fitted distribution $\hat{G}_g$ captures the typical expression properties in single cell data we described above: average expression level ($\mathrm{E}(\hat{G}_g)$); zero-inflation (zero fraction π_g, nonzero mean $\mathrm{E}(\hat{H}_g)$), dispersion ($\mathrm{CV}(\hat{G}_g)$), etc.

1.2.1 Zero Inflated Negative Bimomial (*ZINB*)

Since the conjugate prior for Poisson distribution is the gamma distribution, a simple choice for H_g is $\mathrm{Gamma}(a_g, b_g)$ which results in a simple analytical form for the likelihood: after integrating out the prior distribution H_g, the marginal distribution of Y_{cg} is the mixture of a point mass at zero and a negative binomial distribution,

$$\begin{align} p(Y_{cg}) &= \int p(Y_{cg}|\lambda_{cg}) p(\lambda_{cg}) \tag{4} \\ &= \pi_g \delta_0 + (1 - \pi_g) \mathrm{NB}\left(\frac{Y_{cg}}{\alpha_c}; a_g, \frac{1}{b_g + 1}\right), \tag{5} \end{align}$$

where $\mathrm{NB}(x; r, p)$ is the probability density function for negative binomial distribution at x with parameters r (number of failures until end of experiment) and p (success probability).

For each gene, we estimate the parameters π_g, a_g, b_g by maximizing the log-likelihood L_g:

$$L_g(\pi_g, a_g, b_g) = \sum_c \log(p(Y_{cg}|\pi_g, a_g, b_g)). \tag{6}$$

The *ZINB* method is computationally efficient due to the simplicity of its statistical model. However, *ZINB* places a relatively strong assumption on the expression distribution, which is that the expression of active cells follows a gamma distribution that is by definition unimodal, and also has a certain tail decay rate. Due to this lack of flexibility, *ZINB* may fail to capture the true expression distribution if there exists complicated sub-population structures.

The *ZINB* method was implemented by Abhishek Sarkar, and the codes are available on `https://github.com/aksarkar/singlecell-qtl`.

1.2.2 *DESCEND*

Recently Wang et al. [19] proposed *DESCEND*, a method for scRNA-seq expression distribution deconvolution that adoptsg the G-modeling empirical Bayes distribution deconvolution technique in Efron [6]. *DESCEND* takes an exponential family distribution (Poisson, log-normal and gamma distributions being special cases) as H_g and estimates its shape adaptively from the observed counts using natural cubic splines.

The density of H_g has the following form:

$$p_H(x) = \exp(Q(x)^T\theta - \phi(\theta)), \tag{7}$$

where θ is a vector of parameters and $\phi(\theta)$ is the normalization factor. Specific forms of $Q(x)$ corresponds to specific parametric models (e.g. gamma, log-normal). In practice, *DESCEND* sets $Q(x)^T$ as a five-degree natural cubic spline function so that the model can learn $Q(x)$ adaptively from the data.

To estimate the parameters θ_g, DESEND maximizes the penalized likelihood

$$L_g^*(\theta_g) = \sum_c \log(p(Y_{cg}|\theta_g)) - c_0||\theta_g||^2, \tag{8}$$

where c_0 is an adaptively chosen regularization constant. Since the natural cubic spline based exponential family is highly flexible, *DESCEND* uses this regularization term to avoid over-fitting.

The *DESCEND* method is implemented in an R package *DESCEND*, which is publicly available at https://github.com/jingshuw/descend.

1.2.3 Poisson *ash*

Stephens [17] proposed *ash* (Adaptive SHrinkage) to model a list of normal observations which share a common underlying prior distribution. In the context of gene expression studies, *ash* suggests using a unimodal distribution as the prior for the true expression levels, and estimates the unimodal prior adaptively with empirical Bayes. The unimodal assumption of *ash* provides more flexibility than any specific parametric distribution assumption, but also preserves robustness against over-fitting. Here we extend the *ash* framework to tackle the gene expression distribution deconvolution problem for scRNA-seq data.

We now assume that H_g is a unimodal distribution. In practice, a mixture of uniform distributions can be used to approximate the unimodal distribution:

$$H_g = \sum_{k=1}^{K_g} p_{gk} \text{Uniform}[a_{gk}, c_g] + \sum_{l=1}^{L_g} q_{gl} \text{Uniform}[c_g, b_{gl}], \tag{9}$$

where c_g is the mode (non-negative in our setting), $\mathbf{p}_g, \mathbf{q}_g$ are mixture proportions (sum to 1), and a_{gk}, b_{gl} $(0 \leq a_{gk} < c_g,\ c_g < b_{gl})$ are pre-selected grids that cover a sufficiently wide range of values. We estimate the mode c_g and mixture proportions $\mathbf{p}_g, \mathbf{q}_g$ by maximizing the likelihood:

$$L_g(\mathbf{p}_g, \mathbf{q}_g, c_g) = \sum_c \log(p(Y_{cg} | \mathbf{p}_g, \mathbf{q}_g, c_g)). \tag{10}$$

In *ash* framework, given a fixed mode c_g, optimizing L_g over all possible $\mathbf{p}_g, \mathbf{q}_g$'s is a convex optimization problem, and we denote the optimized log-likelihood (given c_g) as $\hat{L}_g(c_g) = \arg\max_{\mathbf{p}_g, \mathbf{q}_g} L_g(\mathbf{p}_g, \mathbf{q}_g, c_g)$. Since $\hat{L}_g(c_g)$ is a function of c_g, we further use 1-d numerical optimization method to fit the mode:

$$\hat{c}_g = \arg\max_{c_g} \hat{L}_g(c_g). \tag{11}$$

With a large enough number of mixture components, any general unimodal distribution can be well approximated by the uniform mixture distribution in (9). In our applications 30-50 mixture components generally suffice.

A special case for H_g would be the non-negative unimodal distribution with mode at 0, which implies that $K_g = 0$ and all other uniform mixture components have lower limits at 0. Since the optimization procedure can be numerically unstable in such a corner case, in practice we separately fit this Poisson *ash* model with a single mode at 0, and compare its log-likelihood with the optimized log-likelihood in (10). The model with higher log-likelihood is then recommended.

Note that for scRNA-seq data, we use the "identity link" which assumes λ_{cg} unimodal distributed, instead of the "log link" which assumes $\log(\lambda_{cg})$ unimodal. Due to the zero-inflated nature of scRNA-seq, fitting the log link model is numerically unstable with extremely small grid lower bound.

The Poisson *ash* method is implemented in R package *ashr*, which is publicly available at `https://github.com/stephens999/ashr`.

1.2.4 Nonparametric deconvolution

Note that all above mentioned methods make assumptions on the expression distribution H_g for active cells: *ZINB* assumes a single gamma distribution; *DESCEND* assumes an exponential family distribution; Poisson *ash* assumes a unimodal distribution. While the distributional constraints lessens the possibility of over-fitting due to noise present in data, we still propose the fully nonparametric deconvolution method as an alternative approach.

Specifically, any distribution H_g can be approximated by a mixture of uniforms on a sufficiently dense grid [13]:

$$H_g = \sum_{k=1}^{K_g} p_{gk} \text{Uniform}[(k-1)a_g, ka_g], \tag{12}$$

where $\mathbf{p}_g$ are mixture proportions (sum to 1), and a_g is the pre-selected grid step-size. To ensure the goodness of nonparametric approximation, the number of mixture components K_g should be sufficiently large, and $[0, K_g a_g]$ covers a sufficiently wide range.

The computations necessary are essentially identical to those for general ash (Section **??**), so we reuse them here to estimate $\mathbf{p}_g$.

1.3 Applications

1.3.1 Zeisel data

Zeisel et al. [20] described a scRNA-seq dataset of mouse hippocampal region. The dataset has read counts of 12234 genes in 3005 cells from the mouse somatosensory cortex and hippocampus CA1 region. The 3005 cells have been clustered into 7 major cell types: Astrocytes-Ependymal, Endothelial-Mural, Interneurons, Microglia, Oligodendrocytes, CA1 pyramidal and S1 pyramidal, and the number cells in each cell type are 224, 235, 290, 98, 820, 939 and 399 respectively. The dataset is publicly available at `https://hemberg-lab.github.io/scRNA.seq.datasets/mouse/esc/`.

We run *ZINB*, *DESCEND*, Poisson *ash* and the nonparametric deconvolution method on a subset of this scRNA-seq dataset and compare their deconvolution results. This subset of data consists of cell types Astrocytes-Ependymal, Endothelial-Mural and Microglia (557 cells in total). We use the normalized library sizes as the scaling factors α_c for cells:

$$\alpha_c = \frac{\sum_g Y_{cg}}{\sum_{c,g} Y_{cg}/C}, \tag{13}$$

where C is the total number of cells. The numerator in (13) is the raw library size for cell c (total counts), and the denominator is the average library size across all cells.

In some applications, the properties of G_g may be further used as genetic variant specific features. For instance, the mean and spread information extracted from the expression distribution can be used in eQTL analysis. In general, we would also like to investigate if the different deconvolution methods would produce $\hat{G}_g$ with significantly distinct shapes. Hence we compare the deconvolution results in the following aspects.

Likelihood To evaluate the goodness of fit of different models, we can compare the log-likelihood of the probability models:

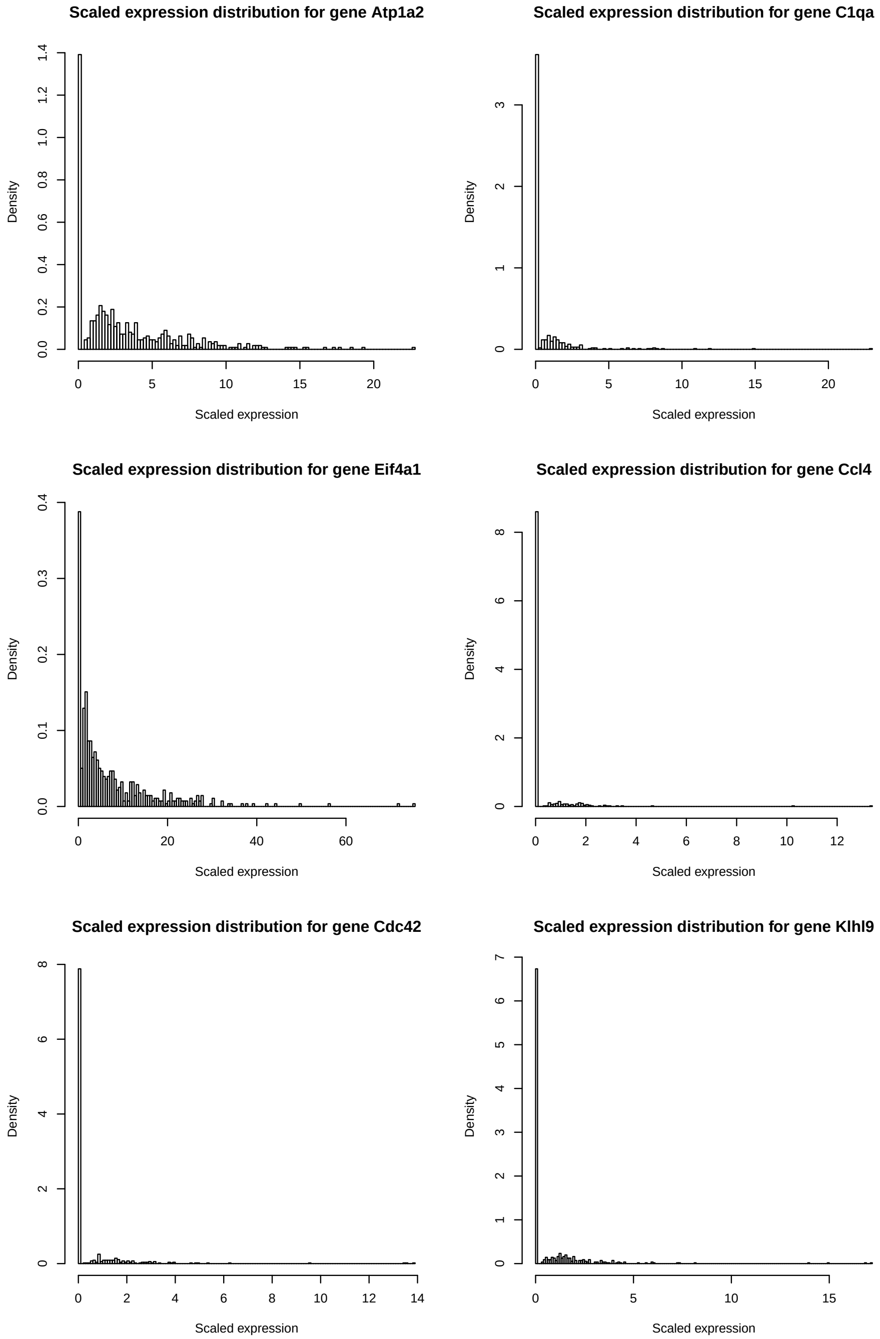
$$L_g = \sum_c \log p(Y_{cg}). \tag{14}$$

The *ZINB*, Poisson *ash* and nonparametric deconvolution directly provide the log-likelihood in (14) when fitting the hyperparameters with empirical Bayes method. However, *DESCEND* optimizes the penalized log-likelihood L_g^* in (8) instead. Fortunately, the *DESCEND* package also gives the optimum value for unpenalized likelihood L_g (without providing the corresponding fitted deconvolution model).

Even though *ZINB*, *DESCEND*, Poisson *ash* and nonparametric deconvolution have the same Poisson likelihood, their models for G_g are not nested. Hence, using likelihood ratio tests to compare the models is statistically unsound. However, the difference in log-likelihoods of different methods still reveals the goodness of fit for scRNA-seq data. We consider the following scenarios for the difference in log-likelihood between two methods: within ±2 (insignificant difference); over ±2 but within ±5 (moderate significant difference); over ±5 (very significant difference).

We start with the restricted version of Poisson *ash* which only allows unimodal G_g: the uni-mode is either 0 or some non-zero positive value, and in the former case the density of G_g monotonically decays. We denote this model as Poisson *ash* (unimodal), which essentially assumes that one unimodal distribution is sufficient to capture the expression distribution across cells. We compare its log-likelihood with that of the the most flexible approach, fully nonparametric deconvolution. For only 0.80% genes, the nonparametric deconvolution has significantly higher log-likelihood than that of Poisson *ash* unimodal model (with difference being at least 2). For the rest over 99% genes, the log-likelihood difference is insignificant, which means the Poisson *ash* unimodal model is quite sufficient to capture the underlying gene expression distribution.

For genes where the nonparametric deconvolution achieves significantly higher log-likelihood, we further inspect those genes with largest log-likelihood difference to see if more complicated models are needed to fit the data. Figure 1 shows the scaled expression Y_{cg}/α_c distribution for six genes *Atp1a2*, *C1qa*, *Eif4a1*, *Ccl4*, *Cdc42* and *Klhl9*, where the log-likelihood difference between Poisson *ash* unimodal model and nonparametric deconvolution is 5.61, 5.25, 4.92, 4.63, 4.50 and 4.43 respectively. The six genes display a common pattern that, the scaled expression histogram split the cells into zero expression and non-zero expression parts, with noticeable gap between them. This suggests us to consider a more flexible expression distribution in form (3), with a point mass at zero δ_0 and some distribution H_g for the non-zero part.


Scaled expression distribution for gene Atp1a2

Scaled expression distribution for gene C1qa

Scaled expression distribution for gene Eif4a1

Scaled expression distribution for gene Ccl4

Scaled expression distribution for gene Cdc42

Scaled expression distribution for gene Klhl9

Figure 1: Distribution of scaled expression Y_{cg}/α_c for genes *Atp1a2*, *C1qa*, *Eif4a1*, *Ccl4*, *Cdc42* and *Klhl9*, where the nonparametric deconvolution model has significantly higher log-likelihood than that of Poisson *ash* unimodal model.

Hence, we try the models with δ_0 but different assumptions for H_g: *ZINB* (H_g is gamma distribution); *DESCEND* (H_g belongs to exponential family) and Poisson *ash* (H_g is unimodal) and check their log-likelihood improvements compared to the Poisson *ash* unimodal model (which is used as baseline model). Table 1 shows their log-likelihood improvements for the above six genes *Atp1a2*, *C1qa*, *Eif4a1*, *Ccl4*, *Cdc42* and *Klhl9*. Figure 2 shows the fitted G_g by Poisson *ash* and *DESCEND* for these six genes.

Table 1: Log-likelihood improvement upon the Poisson *ash* unimodal model for genes *Atp1a2*, *C1qa*, *Eif4a1*, *Ccl4*, *Cdc42* and *Klhl9*.

Gene	Nonparametric	Poisson *ash*	*DESCEND*	*ZINB*
$Atp1a2$	5.61	0.00	-28.12	-18.23
$C1qa$	5.25	0.00	-8.97	-7.33
$Eif4a1$	4.92	4.35	3.78	-1.98
$Ccl4$	4.63	0.00	-214.15	-1.63
$Cdc42$	4.50	2.19	3.78	1.42
$Klhl9$	4.43	4.16	1.20	-7.64

The results suggest that for genes *Atp1a2*, *C1qa* and *Ccl4*, the Poisson *ash* model with δ_0 plus unimodal H_g does not make difference from the baseline unimodal model, since the fitted H_g gets highest probability for the mixture uniform components with left limit 0, and thus makes G_g unimodal with mode at 0. On the other hand, *ZINB* and *DESCEND* yield even lower log-likelihoods than baseline. For genes *Eif4a1*, *Cdc42* and *Klhl9*, Poisson *ash* significantly increment the baseline log-likelihood of unimodal model. The *DESCEND* model also substantially improves the baseline log-likelihood for *Eif4a1* and *Cdc42*. Among these six genes, *Cdc42* is the only one where *DESCEND* achieves slightly higher log-likelihood than Poisson *ash*. In Figure 1, the

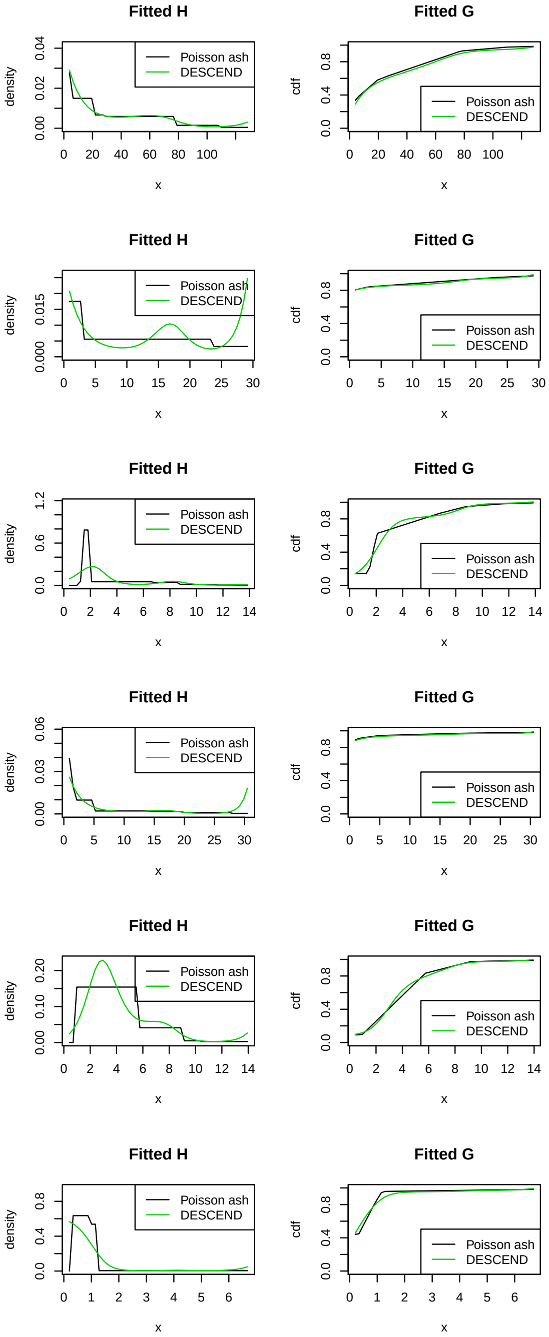


Figure 2: Fitted G_g by Poisson *ash* and *DESCEND* for genes *Atp1a2*, *C1qa*, *Eif4a1*, *Ccl4*, *Cdc42* and *Klhl9* (from top to bottom). Each row shows the pdf and cdf function of $\hat{G}_g$ for that gene.

scaled expression histogram for *Cdc42* has a more than one relatively eye-catching bumps (e.g., around 0.8 and 1.6), which might be the reason why *DESCEND* gives slightly higher log-likelihood.

The gene $Cdc42$ is an example where the distinct assumptions on H_g lead to different model fits. We further investigate if the unimodal assumption of Poisson *ash* is sufficient to capture the non-zero expression distribution in contrast to *DESCEND*. Table 2 shows the distribution of the differences in log-likelihood. For each method (*DESCEND*, *ZINB*, nonparametric), we subtract the log-likelihood of Poisson *ash* from that of the target method for each gene, and then summarize the percentage of genes with log-likelihood difference falling into intervals $(-\infty, 5]$, (-5,-2], (-2,2], (2,5] and $(5, \infty)$. The five categories corresponds to the cases where: the target method gives much lower log-likelihood; significantly lower log-likelihood; similar log-likelihood; significantly higher log-likelihood; much higher log-likelihood than that of Poisson *ash*.

Table 2: Distribution (%) of log-likelihood difference between deconvolution methods and Poisson *ash*.

	$(-\infty, 5]$	(-5,-2]	(-2,2]	(2,5]	$(5, \infty)$
ZINB	2.43	12.09	85.48	0.01	0.00
DESCEND	6.26	7.71	85.93	0.09	0.01
Nonparametric	0.13	0.11	99.03	0.72	0.01

From Table 2, we see that for most genes (over 99%) , the nonparametric deconvolution and Poisson *ash* don't have significant differences in terms of the log-likelihood difference. Presumably, nonparametric deconvolution is the most flexible method and should achieve higher log-likelihood than Poisson *ash* in any scenario. However, due to numerical errors and the limitation of uniform mixtures, there are less than 1% genes where the former has noticeably higher log-likelihood. The majority of the genes result in similar log-likelihoods for Poisson *ash* and nonparametric deconvolution, so the extra flexibility provided by nonparametric deconvolution seems unnecessary here.

For *ZINB*, which puts stronger assumption on G_g (H_g is single gamma distribution) than Poisson *ash* (H_g is unimodal distribution), it is natural that Poisson *ash* gives higher log-likelihood. From Table 2, Poisson *ash* has much higher (> 5) log-likelihood compared to *ZINB* for 2.43% genes, and has significantly higher (> 2) log-likelihood for another 12.09% genes. For these genes, the zero-inflated gamma prior assumption of *ZINB* may not suffice in fitting this specific scRNA-seq dataset. The rest 85.48% genes have similar log-likelihoods for Poisson *ash* and *ZINB*, and there are only 0.01% outlier genes where the *ZINB* model gives a better fit due to numerical instability.

The comparison between *DESCEND* and Poisson *ash* is an interesting one. The prior H_g for *DESCEND* belongs to the general exponential family, so its shape is not restricted to any unimodal distribution. Nevertheless, the exponential family assumption and the limited number of basis used in g-modeling would constrain the shape of distribution. As a result, it is hard to directly compare the theoretical flexibility of the two models. Moreover, even though *DESCEND* gives the optimized unpenalized likelihood, it actually fits the model by optimizing the penalized likelihood to avoid over-fitting.

We still first compare the (unpenalized) log-likelihood of *DESCEND* with that of Poisson *ash*. Table 2 shows that Poisson *ash* has much higher (> 5) log-likelihood compared to *DESCEND* for 6.26% genes, and has significantly higher (> 2) log-likelihood for another 7.71% genes. There are 85.93% genes with similar log-likelihoods for Poisson *ash* and *DESCEND*, yet 0.1% genes have significantly higher log-likelihood for *DESCEND*. We checked the genes where *DESCEND* gives significantly higher log-likelihood: the top five genes with the highest log-likelihood differences are *Nek7*, *Agpat3*, *Fam216b*, *Tdrp* and *Gak*, with log-likelihood differences over 3.

The scaled expression histograms for these five genes are given in Figure 3. We see that these genes typically have few large outliers which make the histogram tail longer. Furthermore, the frequency of cells does not simply decrease over scaled expression level, but rather has some small bumps in the middle.

We plot the fitted prior G_g from *DESCEND* and Poisson *ash* for these 5 genes in Figure 4. The left column shows the pdf function of the fitted prior $\hat{G}_g$ (point mass at zero already absorbed) and the right column shows the cdf function of $\hat{G}_g$. The visualization indeed reveals the discrepancy between *DESCEND* fitted prior and Poisson *ash* fitted prior. Unfortunately *DESCEND* only provides density within this x-axis range so we are unable to plot the densities outside the current range in Figure 4. However, the cdf plots already indicate that for genes $Nek7$, $Agpat3$, $Tdrp$ and Gak, *DESCEND* puts a much heavier probability mass on the right tail outside the x-axis range than Poisson *ash*, which is actually unconvincing according to the original scaled expression histogram (Figure 3).

Even though *DESCEND* provides extra flexibility by allowing multimodality for H_g and hence achieves higher log-likelihood for these genes, its goodness of fit is still

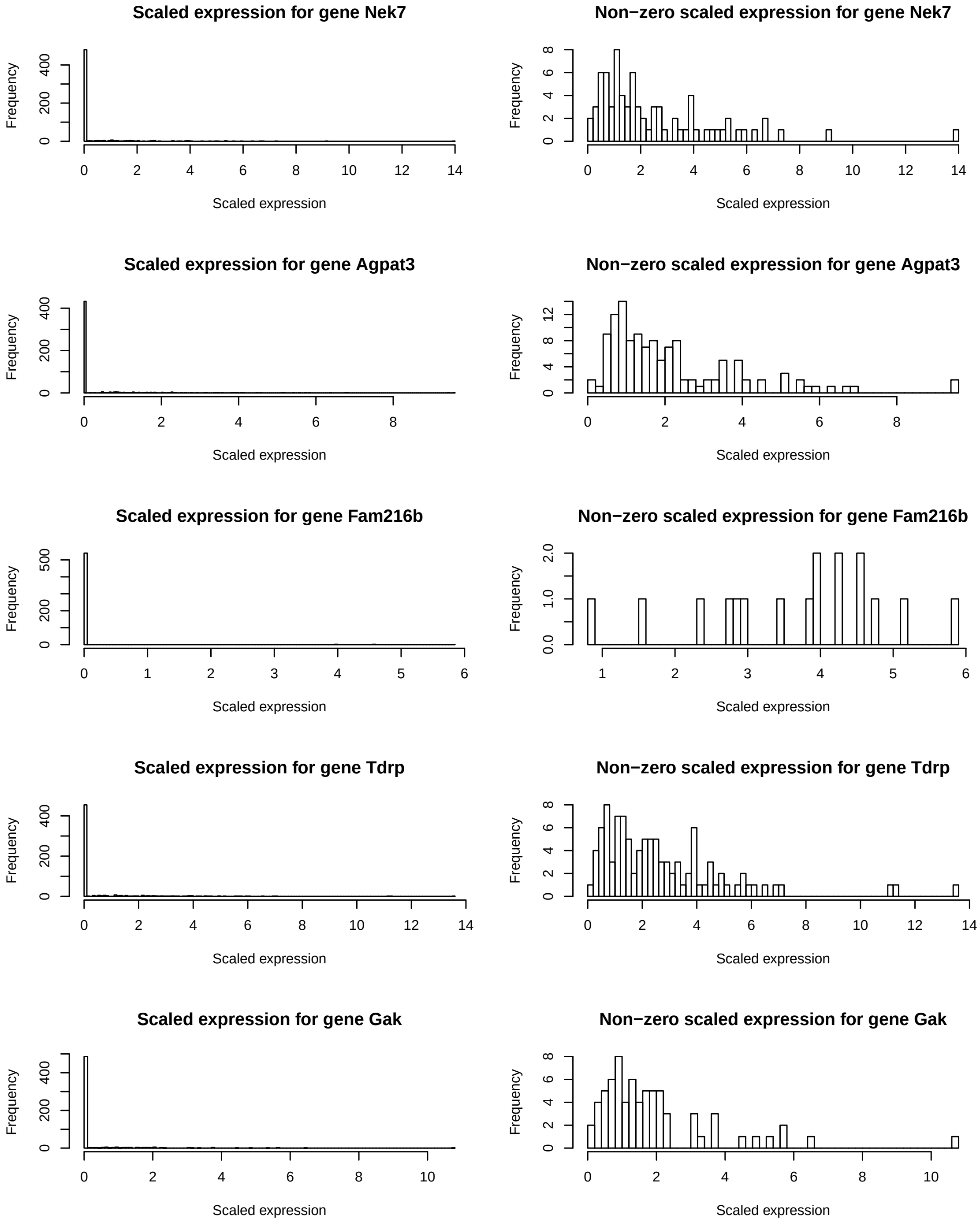


Figure 3: Distribution of scaled expression Y_{cg}/α_c for genes *Nek7*, *Agpat3*, *Fam216b*, *Tdrp* and *Gak*.

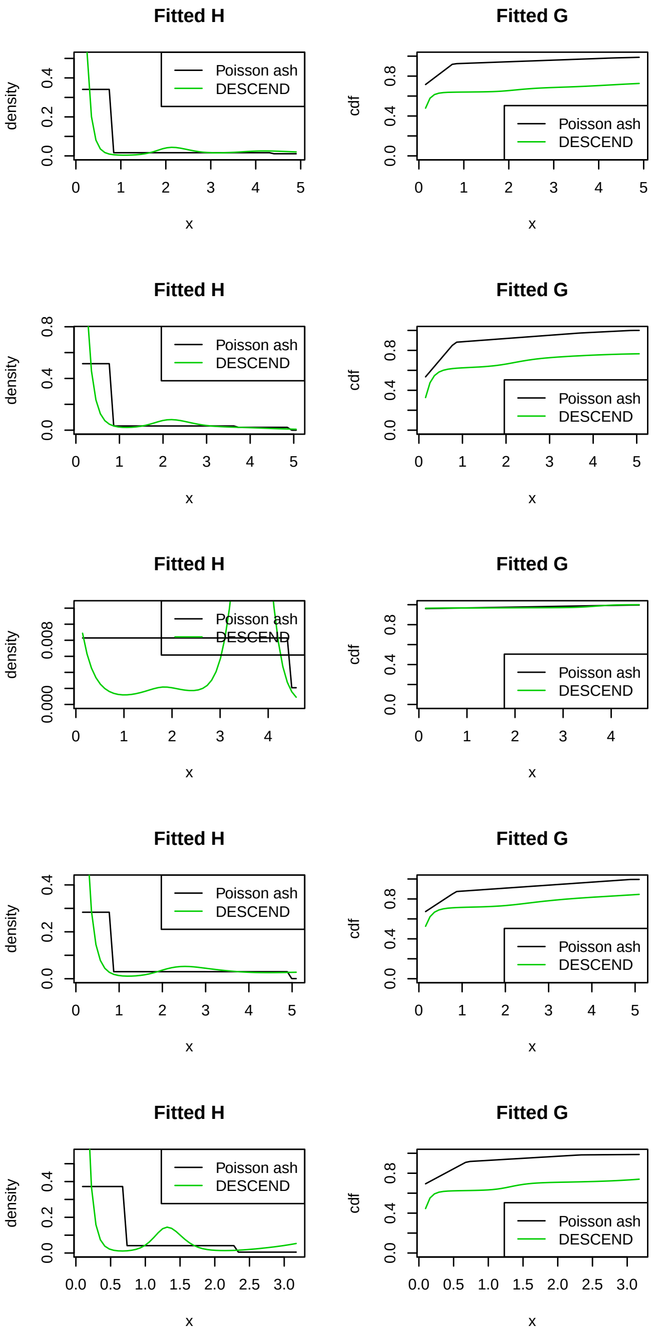


Figure 4: Fitted G_g for genes *Nek7*, *Agpat3*, *Fam216b*, *Tdrp* and *Gak*, where *DESCEND* gives much higher log-likelihood than Poisson *ash*. Each row shows the pdf and cdf function of $\hat{G}_g$ for that gene.

questionable. Looking back at Figure 3, it is insufficient to conclude that the bumps in frequency are due to underlying true expression distribution rather than Poisson noise, since the frequency bumps are not too large either. Moreover, the extremely long tailed $\hat{G}_g$ fitted by *DESCEND* seems suspicious, which suggests *DESCEND* might be over-sensitive to large outliers.

One of the original purposes of having a flexible prior G_g is to deal with the potential sub-populations in scRNA-seq data. In Zeisel data, we use the expression counts for cell types Astrocytes-Ependymal, Endothelial-Mural and Microglia. To investigate if this sub-population structure of cell types could result in multimodal $\hat{G}_g$, we further check the raw count distribution grouped by cell types (here we split Astrocytes-Ependymal into Astrocytes and Ependymal; split Endothelial-Mural into Endothelial and Mural). For gene *Agpat3* where multiple bumps exist in scaled expression distribution (Figure 3), we visualize the scaled expression distribution by cell types in Figure 5. The most notable difference in scaled expression distribution among different cell types is in the tail shape. The scaled expression distribution of Astrocytes cells yields significantly longer tail than that of Ependymal and Microglia, and also shows a rough bimodal pattern with two major parts split at 3. On the other hand, the scaled expression levels for Microglia and Ependymal cells are quite low and concentrated, and the scaled expression of Mural cells is either smaller than 3 or bigger than 9.

This example shows that the scaled expression seems to have different spreads for different cell types, rather than yielding different mode/average levels. The bimodal pattern within Astrocytes mainly contributes to the overall “bimodal” scaled expression distribution, while the differences in expression distribution among the other cell types seem to be determined mainly by spreads.

Note that in Poisson *ash*, H_g consists of many uniform mixture components with different widths (anchored at same mode). Even though the mixture components were not specifically designed for clustering sub-populations, their various spreads seem to corroborate the phenomenon described above, in that different sub-populations have tails with differing lengths.

Mean The mean of deconvolution distribution $\mathrm{E}(G_g)$ essentially represents for the expected relative expression level for gene g, after removing the technical sampling error (Poisson noise) in scRNA-seq data:

$$\mathrm{E}(G_g) = (1 - \pi_g)\mathrm{E}(H_g). \tag{15}$$

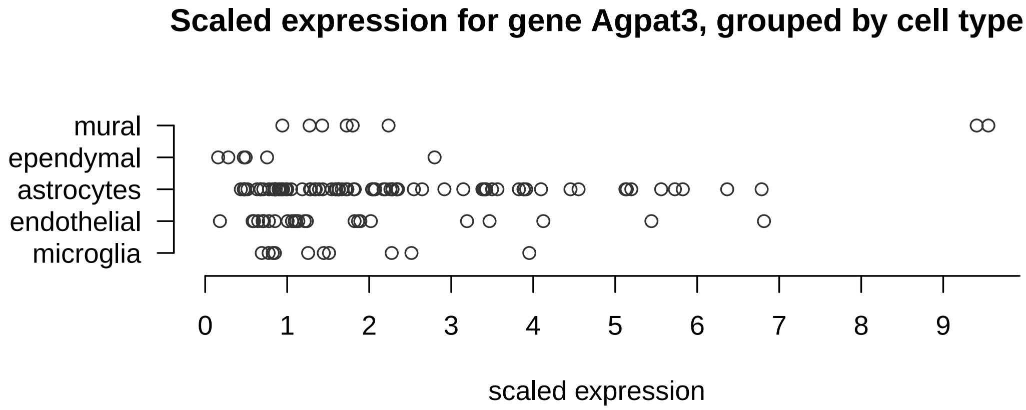


Figure 5: Non-zero scaled expression Y_{cg}/α_c for gene *Agpat3*, grouped by cell types. Each point in the figure represents for the scaled expression level of one single cell.

Figure 6 plots the gene-specific deconvolution distribution means for *ZINB*, *DESCEND* and nonparametric deconvolution against that of Poisson *ash* (on log-log scale). We see that the four methods result in very similar means for G_g. This is expected, since the average expression level for each gene after removing the Poisson noise should be relatively consistent across the deconvolution methods, even though the specific shape of G_g can be different.

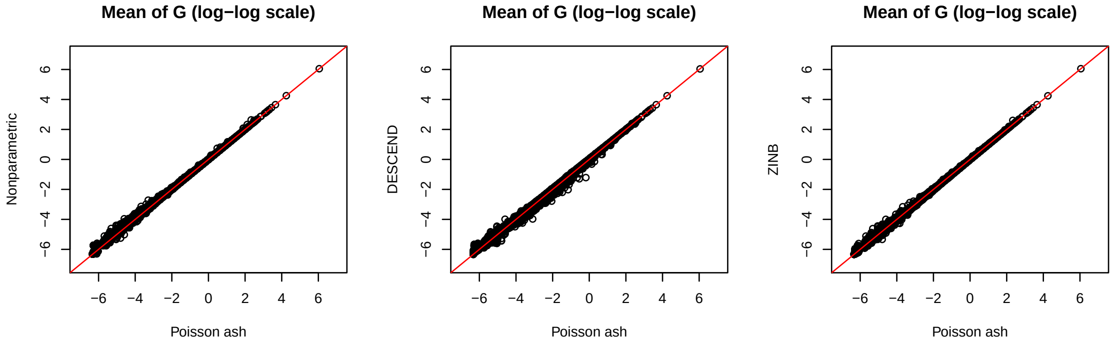


Figure 6: Mean of the deconvolution distribution G_g for methods nonparametric, *DESCEND*, *ZINB* against that of Poisson *ash*.

CV (Coefficient of Variation) Apart from the mean of G_g, the spread of the expression distribution can also useful be in some contexts (e.g. eQTL analysis). We also compute the CV (coefficient of variation) for fitted G_g:

$$\mathrm{CV}_g = \frac{\mathrm{SD}(G_g)}{\mathrm{E}(G_g)}, \tag{16}$$

where the variance of G_g is given by

$$\mathrm{Var}(G_g) = (1-\pi_g)(\mathrm{E}(H_g)^2 + \mathrm{Var}(H_g)) - \mathrm{E}(G_g)^2. \tag{17}$$

Figure 7 shows the gene-specific CV of fitted deconvolution distribution $\hat{G}_g$ for *ZINB*, *DESCEND* and nonparametric deconvolution against that of Poisson *ash*. In general, Poisson *ash* and nonparametric deconvolution have quite similar CV's. This is consistent with our previous results, where Poisson *ash* and nonparametric deconvolution have very similar performance on the Zeisel dataset .

However, the difference in CV between *DESCEND* and Poisson *ash* is much bigger. Though the CV of *DESCEND* is often comparable with that of Poisson *ash*, there are cases where *DESCEND* only results in a CV that is half that of Poisson *ash*. Note that the left bottom corner of middle plot in Figure 7 is an example where the two methods return dramatically different results. In particular, Poisson *ash* results a very small CV (close to 0) but *DESCEND* CV is almost as large as 10. The reason for this is that these outlier genes barely have any non-zero counts. For example, genes *Gm19557*, *2310002J15Rik*, *Chrdl2 Mei1* and *Speer4e* only have one active cell out of the 557 cells and the only non-zero expression count is 1. For these genes, Poisson *ash* returns $\hat{G}_g$ CV as 0.64, while *DESCEND* returns a CV over 10. We visualize the fitted distribution $\hat{G}_g$ for gene *Gm19557* to illustrate the issue. Figure 8 shows that the fitted distribution of Poisson *ash* is extremely short tailed and ends at a very small value near 0, whereas $\hat{G}_g$ of *DESCEND* has a much heavier tail and decay very slowly. For this gene with 556 zeros and 1 one, the mean of $\hat{G}_g$ for *DESCEND* and Poisson *ash* are both very small (around 0.002) but still within the same order of magnitude. However, the big difference in CVs for the two methods reveals sensitivity to model assumptions in such a corner case: a mixture of uniforms can directly cut the tail at some point, but the tail of an exponential family distribution can expand all the way to infinity. With only one non-zero count, it is extremely hard to fit the non-zero prior part H_g. The method *ZINB* shares the problem in that its CV is often much bigger than that of Poisson *ash*. The lack of flexibility of *ZINB* is more amplified here, since the fitted gamma distribution H_g has a relatively fixed decay rate and results in a heavier tail than many other exponential family distributions.

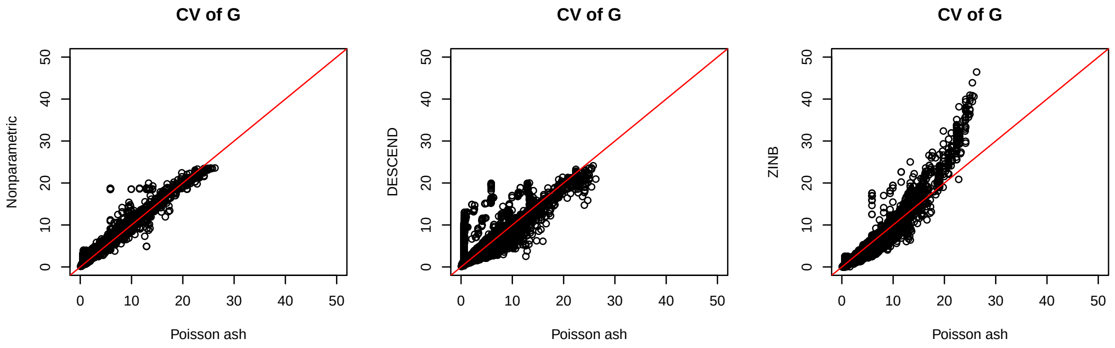


Figure 7: CV of the deconvolution distribution G_g for methods nonparametric, *DESCEND*, *ZINB* against that of Poisson *ash* on Zeisel data.

We now throw out the genes with less than 5 non-zero counts and compare the CVs using the filtered data. Figure 9 shows the gene-specific CV of $\hat{G}_g$ for *ZINB*, *DESCEND* and nonparametric deconvolution against that of Poisson *ash*. Now the CVs remain more consistent across the methods, and there are hardly any large outliers in CVs (bigger than 20).

Null proportion Another important aspect of scRNA-seq expression analysis is to capture the nonzero fraction. Specifically, the nonzero fraction $1 - \pi_g$ represents for the fraction of cells where the gene is expressed. Figure 10 shows the gene-specific

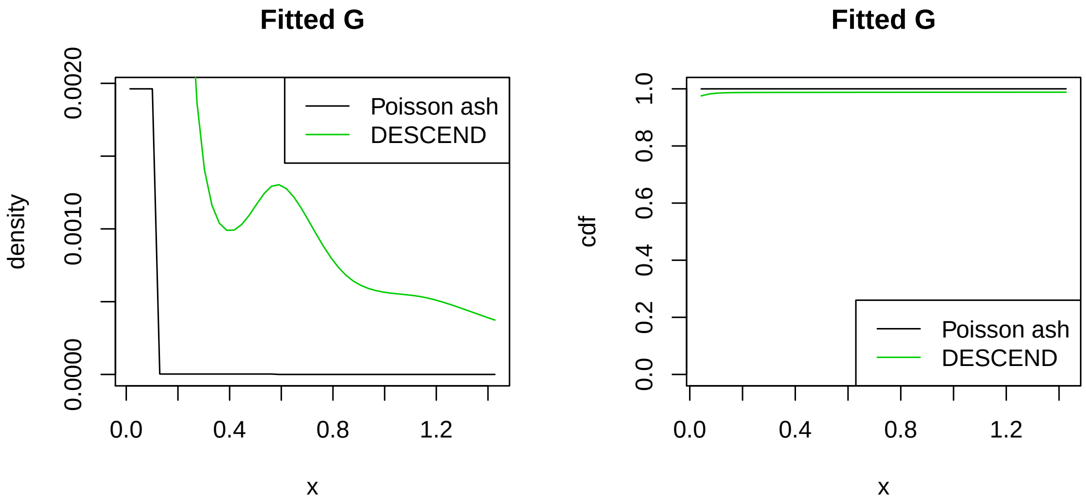


Figure 8: Fitted G_g for gene *Gm19557*, where *DESCEND* gives much higher CV than Poisson *ash*.

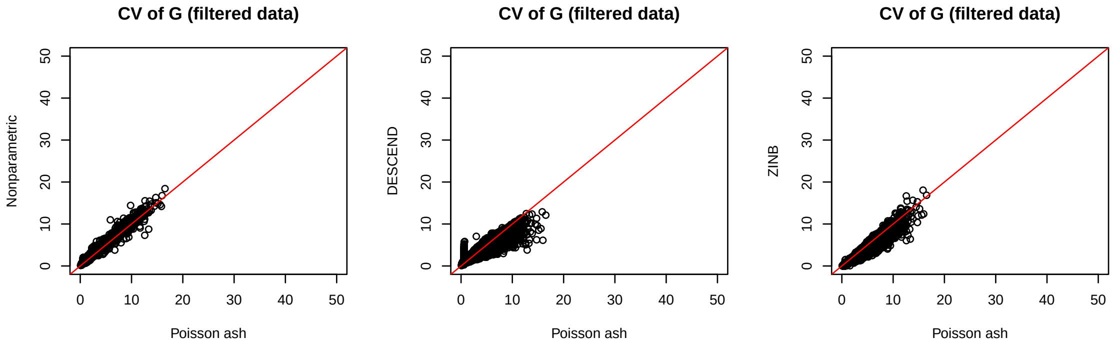


Figure 9: CV of the deconvolution distribution G_g for methods nonparametric, *DESCEND*, *ZINB* against that of Poisson *ash* on filtered Zeisel data (remove genes with less than 5 non-zero counts).

fitted null proportion $\hat{\pi}_g$ for *ZINB*, *DESCEND* and nonparametric deconvolution against that of Poisson *ash*. Unlike the mean and CV, there are no strong correlations in $\hat{\pi}_g$ among different methods. Nevertheless, this is not surprising considering the identifiability issue when estimating the null proportion: it is hard to distinguish if the low counts are due to pure Poisson noise or due to very small positive expression signal. To preserve adaptivity, none of the methods forces H_g to be far isolated from the point mass at zero, thus making it difficult to precisely capture the null proportion.

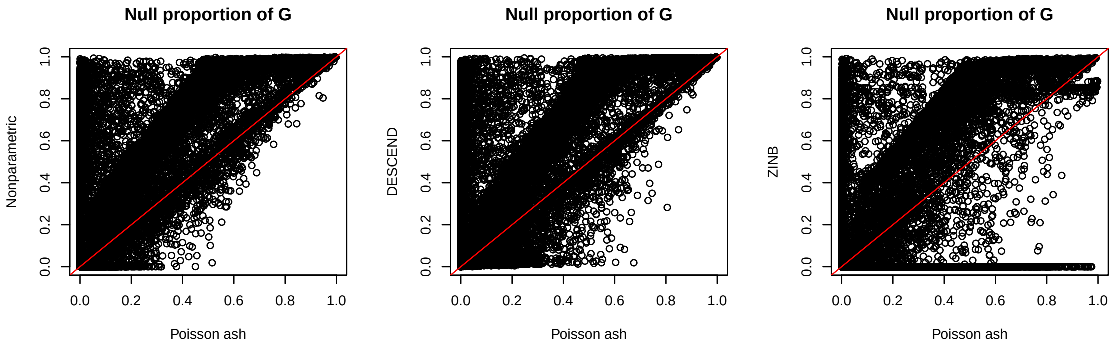


Figure 10: Null proportion π_g of the deconvolution distribution G_g for methods nonparametric, *DESCEND*, *ZINB* against that of Poisson *ash* on Zeisel data.

Shape For the shape of distribution G_g, one major difference in distribution assumptions between *DESCEND* and Poisson *ash* is the ability to handle multimodality. Poisson *ash* only allows for unimodal H_g, but *DESCEND* does not put any constraints on its number of modes, and sometimes does result in multimodal $\hat{H}_g$ in practice. However, whether or not the multimodality assumption is necessary in practice would be up for debate. For example, as we discussed in Section 1.3.1, the extra flexibility of allowing multimodality could lead to over-fitting issues instead of improving the fitted expression distribution when large outliers are present in scRNA-seq data.

To detect potential multimodal pattern for scRNA-seq expression distribution, Bacher and Kendziorski [1] proposed the following approach: genes for which at least 75% of cells showed non-zero expression are selected. For each gene, zeros were removed and the R package *Mclust* was applied to log expression to estimate the number of modes (because zeros were removed prior to *Mclust*, a mode at zero will not contribute to the total number of modes). The *Mclust* package fit a Gaussian mixture model, and the output optimal number of mixture components was used as the estimate of the number of modes. Note that this approach is particular different from our previous discussed deconvolution methods: it directly fits a Gaussian mixture model for the observed expression (on log scale), but does not involve any deconvolution step which estimates the true expression distribution from the noisy observations. Bacher and Kendziorski [1] analyzed three scRNA-seq datasets and visualized the number of modes in Figure 1c of their paper, showing that at least 60% genes have more than two modes.

We applied the same *Mclust* method on Zeisel data, among the 3215 genes with at least 75% non-zero expressed cells, 1666 (51.82%) genes just have one mode, 681 (21.18%) and 61 (1.9%) genes have 2 and 3 modes respectively, and only 1 gene, *Gpr37l1*, has 4 modes. We visualize the scaled expression distribution and the fitted deconvolution distribution of *DESCEND* and Poisson *ash* for gene *Gpr37l1* in Figure 11. The log-likelihood of nonparametric deconvolution, Poisson *ash* and *DESCEND* models are given by -1328.454, -1329.898 and -1338.806 respectively. For this gene, Poisson *ash* is sufficient to capture the underlying expression distribution since its log-likelihood is very close to that of the nonparametric deconvolution. Figure 11 also indicates that the Poisson *ash* fitted distribution seems more sensible compared to *DESCEND*, that the density substantially drops beyond 20. Therefore, even though the observed expression histogram has quite a few bumps and results in 4 modes in the *Mclust* fitted model, it is nevertheless convincing that the underlying true expression distribution is multimodal.

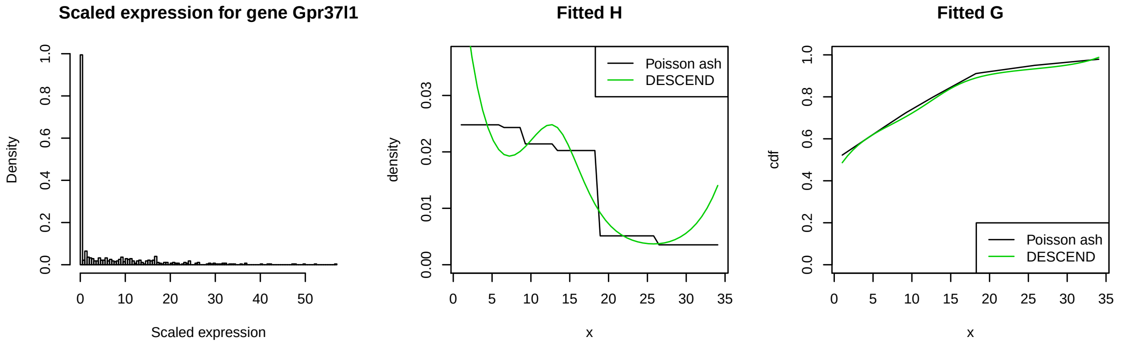


Figure 11: Scaled expression distribution, fitted expression distribution $\hat{G}_g$ by *DESCEND* and Poisson *ash* of gene *Gpr37l1*.

1.3.2 Tung data

The data in Tung et al. [18] have three C1 replicates from three human induced pluripotent stem cell lines and UMI were added to all samples. One replicate of the individual NA19098.r2 was removed from the data due to low quality and 564 cells are kept after filtering. The dataset is publicly available at `https://github.com/jdblischak/singleCellSeq`. Each replicate is a batch with less than 100 cells.

Unlike the Zeisel data, the sample size of Tung data is relatively small. We run *DESCEND* and Poisson *ash* on one replicate NA19091.r2 (96 cells in total) and compare their results on this dataset with a limited number of cells.

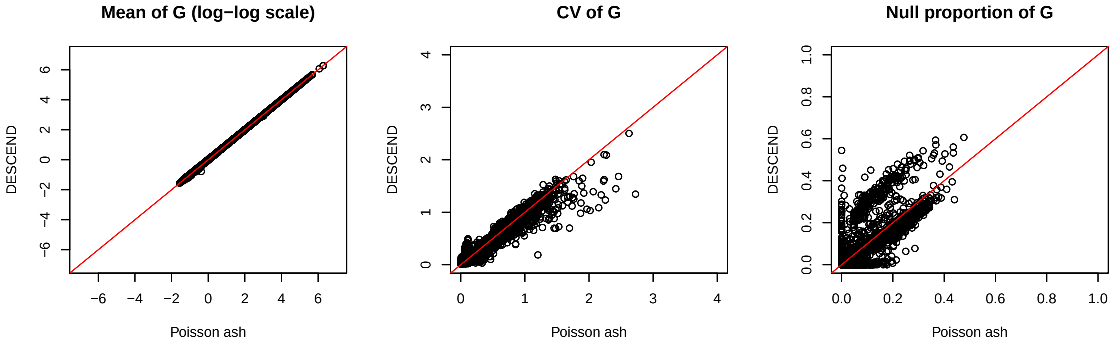


Figure 12: Mean, CV and null proportion of the fitted distribution $\hat{G}_g$ for methods *DESCEND* and Poisson *ash* on Tung data.

Figure 12 shows the mean, CV and null proportion $\hat{\pi}_g$ of the fitted distribution $\hat{G}_g$ given by *DESCEND* and Poisson *ash*. For the CV plot we also filter out the genes with less than 5 non-zero counts. The patterns of $\hat{G}_g$ mean/CV for the two methods are similar to what we observed from Zeisel data: the means are almost the same (Pearson correlation 99.99%), the CVs are very similar (Pearson correlation 96.66%), but the null proportions are less correlated due to the identifiability issue. For Tung data, most genes have very small zero fractions, which actually makes the null proportion easier to estimate compared to Zeisel data.

1.3.3 Buettner data

The data in Buettner et al. [3] were used in Bacher and Kendziorski [1] to estimate the number of modes of observed log expression distribution. Individual *Mus muscu-*

lus embryonic stem cells were sorted using fluorescence-activated cell sorting (FACS) for cell-cycle stage, then single cell RNA-seq was performed using the C1 Single Cell Auto Prep System (Fluidigm). The scRNA-seq dataset is consisted of 96 *Mus musculus* embryonic stem cells in the G2M stage of the cell cycle. This dataset is publicly available at `https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2805/`.

Bacher and Kendziorski [1] applied *Mclust* on 1000 genes with at least 75% non-zero expressed cells and thus estimate the number of modes for log expression. There are 343, 614, 41 and 2 genes with 1, 2, 3 and 4 modes respectively.

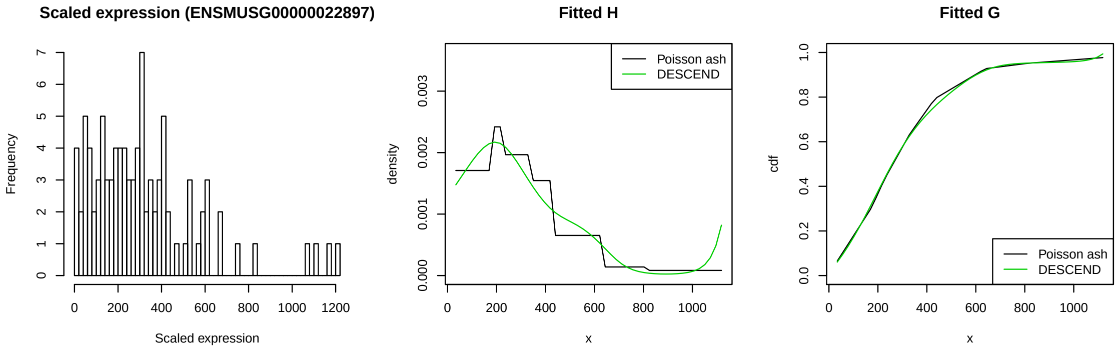


Figure 13: Scaled expression distribution, fitted expression distribution $\hat{G}_g$ by *DESCEND* and Poisson *ash* of gene *Dyrk1a*, which has 4 *Mclust* modes for log expression.

Gene *Dyrk1a* (ENSMUSG00000022897) has 4 *Mclust* modes. We apply *DESCEND*, Poisson *ash* and the nonparametric deconvolution on this gene and the log-likelihoods are -638.36, -630.71 and -621.51 respectively. The notable difference between log-likelihoods of Poisson *ash* and nonparametric deconvolution implies that a unimodal distribution might be insufficient to capture the non-zero expression distribution. Figure 13 shows the scaled expression distribution and fitted G_g by

DESCEND and Poisson *ash* for *Dyrk1a*. In this case, a multimodal H_g might be needed.

1.4 Discussion

For scRNA-seq data, Wang et al. [19] proposed *DESCEND*, a method that deconvolves the true cross-cell gene expression distribution from observed UMI counts. In this project, we generalize a deconvolution framework for scRNA-seq data that removes the noise from UMI counts and then estimate the true gene expression distribution. The method *DESCEND* is a special case which assumes the expression distribution to be an flexible exponential family distribution (with natural spline basis). We also propose the methods *ZINB*, Poisson *ash* and nonparametric deconvolution with different assumptions on the expression distribution: besides the point mass at zero to account for burstiness in scRNA-seq data, *ZINB* assumes the active cell expression distribution to be a single gamma distribution; Poisson *ash* only requires the active cell expression distribution to be unimodal; and nonparametric deconvolution does not even have specific distributional assumptions. The nonparametric deconvolution method is the most flexible one among the four methods, while *ZINB* is the method with the most strict constraints. On the other hand, an overly flexible method may have overfitting issues due to the noise present in data. The ideal method would strike a balance between adaptivity and robustness. The relative flexible assumptions of *DESCEND* or Poisson *ash* make them adaptive to a wide range of data, but the constraints (exponential family for the former, unimodality for the latter) also prevent them from overfitting to some extent.

We compared the performances of the four methods on real scRNA-seq datasets Zeisel data and Tung data, and saw that the four deconvolution methods *ZINB*, *DESCEND*, Poisson *ash* and nonparametric deconvolution typically produce nearly identical means for the fitted expression distribution $\hat{G}_g$. The coefficient of variations of $\hat{G}_g$ can differ at times, but are generally similar across the methods when there are sufficient active cells to well estimate the expression distribution. Nevertheless, it is difficult to accurately estimate the null proportion (fraction of inactive cells) no matter which method is used, because accurately separating out noise from true signals on extremely low expressed genes is nearly infeasible.

Even though some summary properties (mean, CV) are relatively consistent across the methods, the fitted expression distribution $\hat{G}_g$ itself can look noticeably different. In theory, the nonparametric deconvolution model should always achieve the highest log-likelihood. In practice however, we discover that the log-likelihood of Poisson *ash* model is often comparable to that of nonparametric deconvolution, and occasionally much higher than that of *ZINB* or *DESCEND*. Although likelihood is an intuitive criterion to evaluate the goodness of fit of the model, it neglects the potential overfitting possibility and the justification of model based on domain knowledge. In applications, we should carefully examine the fitted distributions by different methods and choose the one with proper distributional assumptions and reasonable results, depending on context. For example, there are some cases in Zeisel data where *DESCEND* achieves higher log-likelihood than Poisson *ash*, yet the expression distribution does not tend to be multimodal.

Note that for the expression distribution of active cells H_g, the four methods do not make strong assumptions that rely on the biological nature of scRNA-seq data. *ZINB* uses the conjugate prior gamma distribution for computational convenience, while *DESCEND* and Poisson *ash* use flexible distribution families (exponential family, unimodal family) to guarantee adaptivity. However, the biological backgrounds and properties of the scRNA-seq dataset can be considered during the model selection procedure. For example, on Zeisel data we find that the sub-population structure due to different cell types actually contributes to differences in distribution tail lengths, rather than multimodality corresponding to cell type groups. In this case, the Poisson *ash* model with uniform mixture components in various widths would naturally be suited, and also has better interpretability.

References

[1] Bacher, R. and C. Kendziorski (2016). Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology 17*(1), 63. 28, 30, 31

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

[3] Buettner, F., K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle (2015). Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology 33*(2), 155. 30

[4] Chubb, J. R., T. Trcek, S. M. Shenoy, and R. H. Singer (2006). Transcriptional pulsing of a developmental gene. *Current biology 16*(10), 1018–1025. 4

[5] Dar, R. D., B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum, C. D. Cox, M. L. Simpson, and L. S. Weinberger (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences 109*(43), 17454–17459. 4

[6] Efron, B. (2016). Empirical bayes deconvolution estimates. *Biometrika 103*(1), 1–20. 2, 6

[7] Grün, D., L. Kester, and A. Van Oudenaarden (2014). Validation of noise models for single-cell transcriptomics. *Nature methods 11*(6), 637. 4

[8] Kærn, M., T. C. Elston, W. J. Blake, and J. J. Collins (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics 6*(6), 451. 2

[9] Kim, J. K., A. A. Kolodziejczyk, T. Ilicic, S. A. Teichmann, and J. C. Marioni (2015). Characterizing noise structure in single-cell rna-seq distinguishes genuine from technical stochastic allelic expression. *Nature communications 6*, 8687. 4

[10] Kim, J. K. and J. C. Marioni (2013). Inferring the kinetics of stochastic gene expression from single-cell rna-sequencing data. *Genome biology 14*(1), R7. 4

[11] Kivioja, T., A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods 9*(1), 72. 2

[12] Klein, A. M., L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell 161*(5), 1187–1201. 2

[13] Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association 109*(506), 674–685. 9

[14] Raj, A., C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi (2006). Stochastic mrna synthesis in mammalian cells. *PLoS biology 4*(10), e309. 4

[15] Shaffer, S. M., M. C. Dunagin, S. R. Torborg, E. A. Torre, B. Emert, C. Krepler, M. Beqiri, K. Sproesser, P. A. Brafford, M. Xiao, et al. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature 546*(7658), 431. 2

[16] Shalek, A. K., R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, et al. (2014). Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature 510*(7505), 363. 2

[17] Stephens, M. (2016). False discovery rates: a new deal. *Biostatistics*, kxw041. 7

[18] Tung, P.-Y., J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific reports 7*, 39921. 3, 29

[19] Wang, J., M. Huang, E. Torre, H. Dueck, S. Shaffer, J. Murray, A. Raj, M. Li, and N. R. Zhang (2017). Gene expression distribution deconvolution in single cell rna sequencing. *bioRxiv*, 227033. 2, 6, 32

[20] Zeisel, A., A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science 347*(6226), 1138–1142. 2, 3, 10