Cell, Volume *181*

Supplemental Information

Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients

Pierre Bost, Amir Giladi, Yang Liu, Yanis Bendjelal, Gang Xu, Eyal David, Ronnie Blecher-Gonen, Merav Cohen, Chiara Medaglia, Hanjie Li, Aleksandra Deczkowska, Shuye Zhang, Benno Schwikowski, Zheng Zhang, and Ido Amit

Methods S1. scRNA-seq differential expression analysis using cloglog regression. Related to Figure 2.

May 1, 2020

1 scRNAseq data set : Mathematical primer

1.1 A quick review on scRNAseq data set modeling

Since the first generation of scRNAseq technologies, several attempts were made to model and analyze scRNAseq data sets using rigorous probabilistic models. scRNAseq data are hard to deal with for several reasons : they are high-dimensional (thousands of genes expressed) and highly sparse (more than 90 percents of the values are zeros). Moreover, intrinsic dimensionality and complexity of the data can be high due to the presence of dozens of cell types and states.

The first developed models were based on highly parametrized Zero-Inflated Negative Binomial (ZINB) distribution and were fitted using Bayesian derived methods (Finak et al., 2015; Kharchenko et al., 2014). However, the exponential increase in the number of sequenced cells made such techniques too computational heavy. Moreover, recent papers suggested that the sparsity observed in scRNAseq data is mostly due to limited sampling of the original RNA molecule pool and can be modeled using simple Negative Binomial (NB) distribution without using any zero inflation (Svensson, 2019; Silverman et al., 2018). Attempts were also made to fit multinomial distribution with success but at the cost of an increased mathematical and computational complexity (Townes et al., 2019).

As the global trend in the scRNAseq field favors in sequencing more cells very shallowly (15.000 reads per cells, corresponding to a thousand of Unique Molecular Identifiers (UMIs) per cell), development of robust and simple methods that can deal with such sparse data without using complex and over-fitting models is required (Svensson et al., 2019).

1.2 Mathematical notations

We will consider a scRNAseq experiment with m cells and p genes. Let U a m*p matrix where U_{ij} corresponds to the number of UMIs identified for gene j in cell i. Total number of UMIs for cell i is called L_i. Directly $L_i = \sum_{j=1}^{p} U_{ij}$. We will refer to this value as the cell library size in the manuscript.

ScRNAseq data can be normalized using different strategies to mitigate the variations of cell library sizes. The most common one consist in computing the ratio of UMIs of a cell coming from a given gene. Then the normalized expression of gene j in cell i is therefore $\pi_{ij} = U_{ij}/L_i$. π_{ij} is usually multiplied by 10^6 to provide the *count per million (CPM)* of the gene. Log-transformation is then applied with a pseudo count of 1. The transformed variable is then expected to follow a normal distribution and is therefore equal to $LCPM_{ij} = log(\pi_{ij} * 10^6 + 1)$.

In the next part of the manuscript we will use discretized gene expression. If not stated otherwise, the discretized gene expression of gene j in cell i is D_{ij} and is equal to :

$$D_{ij} = \begin{cases} 1 & \text{if } U_{ij} > 0 \\ 0 & \text{if } U_{ij} = 0 \end{cases} \tag{1}$$

1.3 Shallow scRNAseq data are mostly binary

While first scRNAseq data sets consist in few cells sequenced with millions of reads, the global trend is to sequence more cells very shallowly (Svensson et al., 2019). We therefore wondered if such approach did not result in data sets that are mostly binary, i.e quantification of gene expression can be summarized as distinguishing cells expressing or not the gene of interest. To test this hypothesis we looked at the distribution of non-null values from a massively parallel single-cell RNA sequencing (MARS-seq, (Keren-Shaul et al., 2019)) experiment that consist in single-cell sequencing of different zones of the mouse spleen using a technique called NICHE-seq (Medaglia et al., 2017). MARS-seq relies on polyT primers and therefore generates 3' biased but also shallow (mean library size around 15.000 reads) and can therefore be considered as representative of other methods such as the commercial 10X Chromium technology. The data set was processed using PAGODA2 pipeline (Lake et al., 2018) and cells were clustered using Louvain's graph clustering method (Blondel et al., 2008).

For each of the 22 identified cell cluster types, we computed the mean proportion of non-null UMI counts that are bigger than one. While this value varies between clusters, we observed values between 46% and 65%, suggesting that the majority of the non-null values are equal to one. Moreover, these values were computed by using all genes, including highly expressed genes such as

ribosomal genes, and exogenous spike-in (ERCC). In practice, the genes of biological interest are not the most highly expressed but the most variable, thus reinforcing our hypothesis of binary data.

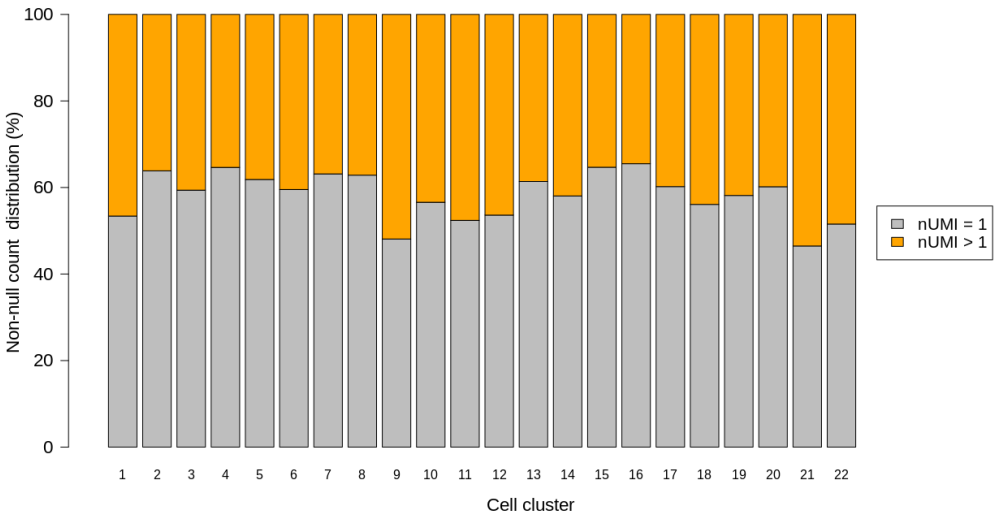

Figure 1: Proportion of non-null values equal or bigger than one across cell clusters

We then looked at the relationship between the cell library size and two features describing the amount of information available in each cell : the number of genes detected in the cell and the proportion of non-null UMI counts that are bigger than one. We observed for both measurements a linear relationship with library size when using logarithmic scales. However while a 40 fold increase in cell library size only doubles the amount of count values bigger than 1 (30% to 60%), it multiplies by 10 the amount of genes detected. Moreover, Pearson correlation was significantly higher between library size and the number of genes detected than between library size and the proportion of non-null count values bigger than one. We therefore conclude that an increase in cell library size results in a dramatic increase in the number of genes detected but only marginally affects our ability to distinguish different level of expression. Therefore the use of a discretized version of the data fully makes sense.

2 Description of the statistical models

2.1 First model : homogeneous gene expression

Let's consider a gene expressed in a homogeneous cell population of m cells. The gene of interest represents a proportion α of the original RNA molecule pool and each cell i has L_i UMIs sampled. We will consider that when one UMI is sampled from the RNA pool, it has a probability α of coming

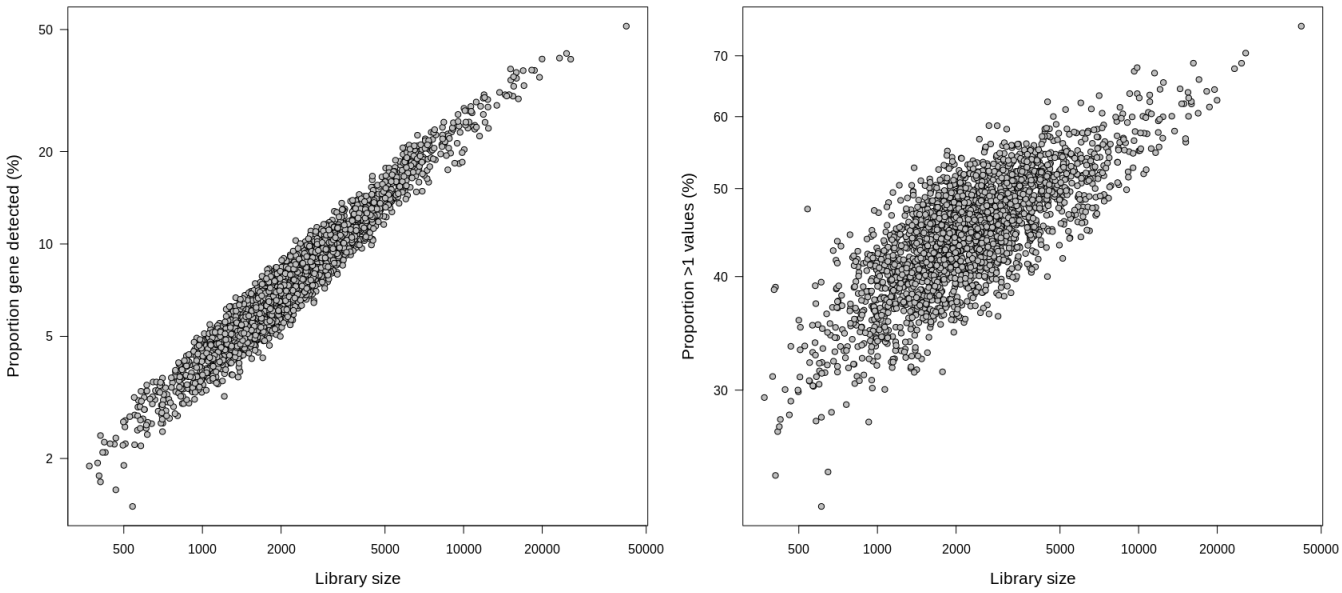

Figure 2: Effects of library size increase on the number of genes detected (left) and non-null values distribution (right).

from the gene of interest and $1-\alpha$ of not coming from that gene. As we assume here independent sampling, the probability of not detecting the gene in cell i across all L_i sampled molecules is equal to $(1-\alpha)^{L_i}$. Therefore the probability of detecting it is equal to :

$$P(\text{Gene detected in cell } i) = 1 - (1-\alpha)^{L_i} \tag{2}$$

This generative model can be used to identify the most suited statistical model. In our case we used a binomial regression with a Complementary Log-Log (cloglog) link function instead of the conventional logistic link function. The cloglog function is equal to :

$$F(x) = 1 - exp(-exp(x)) \tag{3}$$

Instead of using directly the library cell size as an explanatory variable, we used the log transformed library cell size with a coefficient set to 1 in addition to a simple intercept termed μ. When injected to 3 we obtained the following results :

$$\begin{aligned} F(\mu + \log(L_i)) &= 1 - exp(-exp(\mu + \log(L_i))) \\ &= 1 - \exp(-exp(\mu) * L_i) \end{aligned} \tag{4}$$

Equation 2 can be rewritten :

$$P(Gene\,detected\,in\,cell\ i) = 1 - exp(log(1-\alpha) * L_i) \tag{5}$$

And therefore the α parameter can be derived from the intercept μ of the fitted model as :

$$\alpha = 1 - exp(-exp(\mu)) \tag{6}$$

2.2 Second model : heterogeneous gene expression

Now that we have established a link between a simple generative model and the cloglog regression, we can develop more complex models where gene expression can change across cells due to different variables (cell type, cellular stimulation...). To do so, we can simply use the formalism of Generalized Linear Models (GLM) and add a linear additive term before applying the cloglog link function. In the case of a single discrete explanatory variable with p different values, we will use a coefficient variable β, a p row vector, and a design matrix M of size $n \times p$. Therefore :

$$P(Gene\,detected\,in\,cell\ i) = 1 - exp(-exp(\mu + log(L_i) + \beta * M)) \tag{7}$$

In the case of simple variables which takes only two different values (stimulated and control cells), the ratio of gene expression level between the two groups (i.e of the α parameter) can be derived from equation 6 :

$$Expression\ ratio = \frac{1 - exp(-exp(\mu + \beta))}{1 - exp(-exp(\mu))} \tag{8}$$

3 Implementation of the model

3.1 Model fitting with R

R is a powerful programming language and environment that is able to efficiently fit a large variety of statistical models, including GLMs, and to test statistical significance of variables used in the model. The models described in the previous sections can be easily fitted with the following command :

```
model = glm(D~1+offset(log(L))+Variable, family = biomial(link="cloglog")))
```

Here we consider that **D** corresponds to the discretized gene expression level, **L** the total UMI count and **Variable** a vector describing to which group belong each cell. This model needs to be fitted for each gene recursively in order to detect the most differentially expressed genes.

3.2 Testing for statistical significance

Now that the model is fitted, the statistical significance of the variable contribution can be tested using a Likelihood Ratio Test (LRT). This is done simply by writting :

$\mathrm{Term\_significance} = \mathbf{summary}(\mathbf{model})$

The statistical significance of the intercept and variable can then be extracted from the summary object. It is possible to remove genes whose model fitting did not succeed by throwing out models where the intercept term is not significant, i.e the detection probability does not increase with library size. Lastly, when the p-values associated with the variable are computed for each gene of interest, they are corrected using Benjamini Hochberg correction (Benjamini and Hochberg, 1995) to remove false positive.