Single-Cell RNA sequencing (scRNA-Seq)

Introduction

single cell RNA sequencing (scRNA-Seq) is becoming an increasingly popular technique due to advances in experimental techniques in terms of tagging cells before PCR-amplification so that its mRNA can be uniquely identified after sequenced. Importantly, microfluidic-based tagging has greatly improved the throughput of the tagging so that larger cellular population is now amenable to analysis.

scRNA-Seq is analogous to doing multiple runs of bulk RNA-Seq - that is, in both experiments, the captured RNA are reverse-transcribed (RT), amplified, pooled and seqeunced. The difference is that the tagging in bulk RNA-Seq is less vigorous and samples are usually tagged according to the tissue and the condition from which it is acquired. Whereas in scRNA-Seq, the tagging is down to the cellular level and each cell can be uniquely identified. Recently, the introduction of unique molecular identifier (UMI) seeks an even more rigorous tagging where each transcript is assigned a barcode during RT for later identifictaion.

Various protocols exists for scRNA-Seq and varies in their throughput and availablity of external standards for normalisation, including spike-ins and UMI. The microfluidic-based protocol utilises droplet to isolate cell from each other so that they can be uniquely labeled. The captured mRNA in each droplet then undergoes RT and forms "single-cell transcriptome attached to micro-particles" (STAMP), which are then pooled and undergo bulk PCR amplification. The microwell based protocol is different in it sepearates cell by pipetting or laser capturing. It is important that both protocol carried out RT at single-cell level, which potentially caused the dropout problem. After obtaining DNA reads using next-generation-sequencing (NGS), the reads are mapped back to a set of features using some reference (e.g: genome, transcriptome) and identified by their barcodes. It was noted some reads may map to pesudogenes due to sequencing error that could be incorrectly attributed as (Andrews and Hemberg (2016))

Comparison between single-cell RNA-SEQ against bulk RNA-seq

1. Obviously the scRNA-SEQ will focus on quantifying as in RNA-seq. But given the heterogeneity, more information should be kept for scRNA-Seq. "In addition to resolving cellular heterogeneity, scRNA-seq can also provide important information about fundamental characteristics of gene expression" Haque et al. (2017)
2. The mathematical relation between scRNA and bulk-RNA is very simple : bulk-RNA is simply the ensemble average of the scRNA-Seq. This is similar to the relation between gene and isoform: the gene expression is the average of different isoforms. In other words, in additional to genic expression, we now associate each read with a new variable: the cell it comes from.

Advantage of bulk RNA-Seq:

1. Less computationally intensive
2. Signal is more robust and subject to less dispersion
3. Easy experimental setup

Disadvantage of bulk RNA-Seq:

1. The data is only an ensemble average and limited the resolution.
2. Samples can be heterogeneous which introduces confounders that interfere with the interested biological signal.

One of the biggest application of bulk RNA-Seq is detecting differential expression between physiological conditions, which can be induced by altering transcription factor, pathogen presenting, application of drug, etc. More exotic questions like gene imprinting is also within the capability, in which case the transcription was altered according to the direction of breeding. In summary, as long as the transcription or the change of

transcription is homogeneous within the bulk, then bulk RNA-Seq will give enough information to validate any proposed hypothesis.

Advantage of scRNA-Seq:

1. More detailed data that permits exploration down to the cellular level. This includes
2. Larger sample size permits easier estimation for dispersion parameter

Disadvantages:

1. Computationally intensive
2. Subject to alleic dropout, a PCR artifact that greatly increased technical noise
3. Library size is more diversed and harder to model, as a result of both increased technical variability and inherent diverity of cell sizes. This could couple to cell lineage and cell cycle.
4. The signal is richer and noiser, which means more sophiscated models are required to remove confounders, as well as recovering a robust biological signal.

scRNA-Seq is unique in capable of answering questions that requires resolution at cellular level that could not be achieved easily by physical isolation. These include cell cycle, cell differentiation and tissue development, stochastic nature of gene expression. Theoretically, scRNA-Seq is capable of answering any question answerable by RNA-Seq, albeit at a higher cost and complexity.

Because of the relative recentness of scRNA-Seq, its potential is still being actively investigated. One of its fruitful application is visualising cell differentiation as trajectories in the transcription profile, which proves to be difficult to model given the highly noise nature of the data. It also allowed delineation of tumour heterogeneity, where neoplastic, non-neoplastic and immune cells can be distingushed from each other to allow better biomarker to be devised (Müller and Diaz (2017)). It remains unclear how to devise meaningful biological hypotheses at a suitable granularity that can be effectively tested with scRNA-Seq, since most of the hypotheses will be micro-based and does not easily relate to higher hierarchy like tissue/organ level.

Allelic dropout

Allelic dropout refers to the phenomena that a feature shows zero count whereas the underlying true count is non-zero in the original mRNA extraction. In other words, the assumption that read-count of this feature follows a Poisson/negative-binomial distribution is violated, and instances with zero counts are greatly enriched. One of the explanation is the probability of RT failure, a Michalie-Menten process, is greatly increased for mRNA count below a certain threshold (Andrews and Hemberg (2016)). Nevertheless, how much does this model supersedes the negative-binomial model awaits examination. The increased RT failure rate is also reminiscent of the fact that scRNA-Seq typically starts with a smaller amount of mRNA extraction (appx. 10pg) as compared to bulk RNA-Seq (appx. 100ng).

Methods and Results

Data simulation with splatter

Here μ denotes the rate parameter of the Poisson distribution from which the read count is sampled from. $m \sim Pois(\mu)$

Batch effect/group effect is simulated by mutiplying a factor sampled from 2 superposed log-normal distribution

$$\begin{aligned} \mu_{igb} &= \mu_{ig} b_b c_{ig} \\ b_b &= L_b(2s-1), where \\ L_b &\sim Lognormal(loccation, scale^2) \\ s &\sim Bernoulli(0.5) \end{aligned}$$

Differential expression is similarly simulated using a gene-specific multiplicative factor

$$\begin{aligned} \mu_{igb} &= \mu_i b_b c_{ig} \\ c_{ig} &= L_{ig}(2s_{ig}-1), where \\ L_{ig} &\sim Lognormal(loccation, scale^2) \\ s_{ig} &\sim Bernoulli(p) \end{aligned}$$

Splatter use a logistic distribution to model the probability of dropout events. That is to say, there is a threshold for mean expression, below which dropout becomes prevalent. One advantage of using logistic probability is that it allows easier fitting for the binary observation (no RNA fragment/some RNA fragment)

After examining the literature (Zappia, Phipson, and Oshlack (2017)), the most complicated step is modelling BCV in observed RNA-seq data with inverse-chi-squared distribution , which we do not intend to elaborate here

Simulation scheme:

For all simulations, we set the batch parameter to follow a dirac distribution $b_b \sim \delta(1)$ so that there is no difference between batches. Expression counts for 2000 genes are simulated for 2 conditions/groups of equal sample size using default parameters. Four batches are simulated, each containing 100 cells, which are later pooled to mimic bulk RNA-Seq results. The pooling essentially sums up the readcounts for each feature over all cells in this batch. For differential expression, the non-dropout counts (TrueCounts) is used. The dropout is set to follow a logistic distribution centered at 3, corresponding to an expression count of 20.1.

```
## Joining, by = c("bch_vct", "gp_vct")
```

General QC

Normalisation by library size is crucial

We normalised the readcounts using RLE (Relative Log Expression) See figure 1. It is evident that CPM-normalisation removes most of the within-group variance, and left the inter-group variation as the dominating variance. In the scenario where batches are subject to sampling/technical noise, potentially more normalisation is required to remove unwanted variance.

Testing for differentially-expressed(DE) genes

Differential expression is tested under a model-based likelihood ratio framework. Where

$$LR = \frac{P(alternative\ hypothesis)}{P(null\ hypothesis)}$$

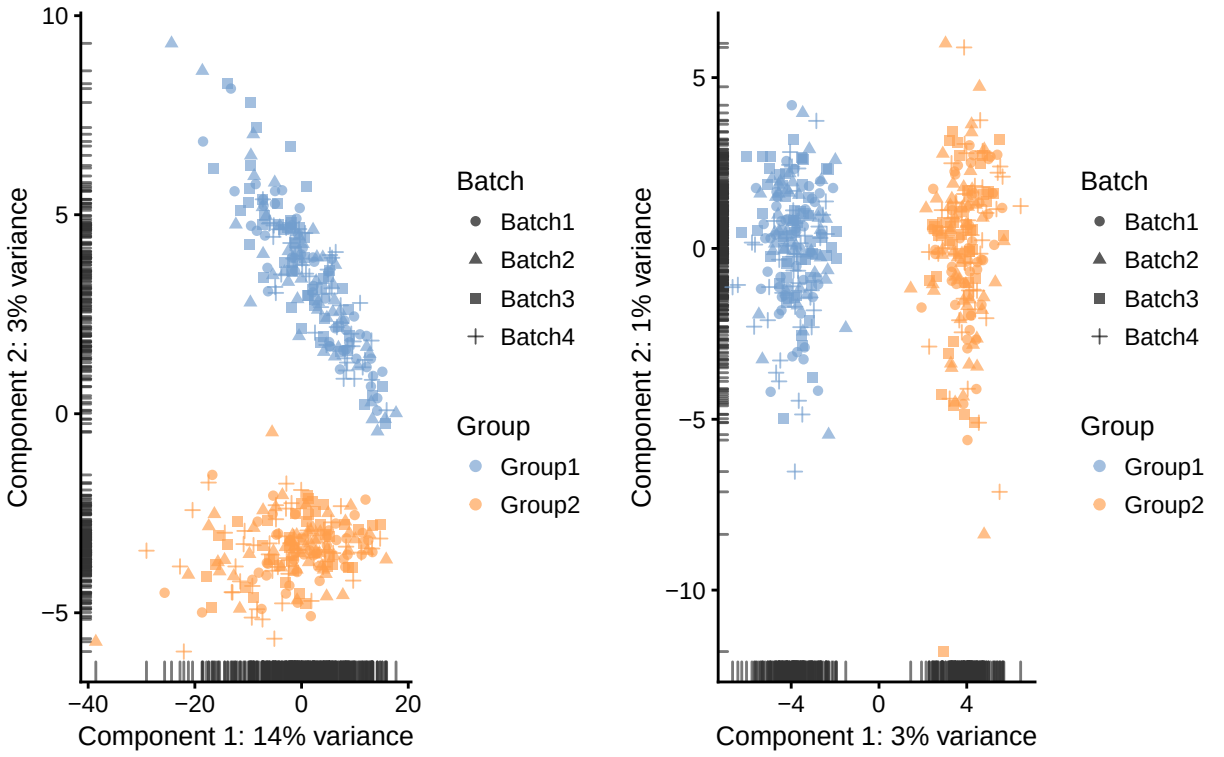


Figure 1: PCA plots of unnormalised read counts(left) and RLE-normalised CPM(right)

Where the null hypothesis the mean expressions of the sampled distributions are identical, and the alternative hypothesis assumes unequalness between these two parameters. Of note, a gamma prior is assumed for the mean expression parameter. (McCarthy, Chen, and Smyth (2012))

$$\psi \sim \Gamma(\frac{1}{\phi}, \frac{1}{\mu\phi})$$
$$m \sim Pois(\psi)$$
$$E(m) = E(\psi) = \mu$$

And the hypotheses become:

1. null: $\mu_1 = \mu_2$
2. alternative: $\mu_1 \neq \mu_2$

The parameter ϕ, is called the biological coefficient-of-variation (BCV, $CV_x = std(x)/E(x)$). It captures all the inter-library variation for any single gene. Here we focus on how BCV(dispersion) affects the result of LR-test, both in scRNA-Seq and in bulk RNA-Seq. For scRNA-Seq, the dispersion is estimated using both "common" and "common-genewise" schemes, whereas for bulk RNA-Seq, it is estimated using "common", "trended" and "trended-genewise" schemes. This is mainly because the sample size for bulk RNA-Seq is too small (4 samples to estimate 2 parameters) and additional assumptions are required to constrain the estimator.

BCV/Dispersion

It is evident that tagwise/trend scheme better captures the BCV for both granularities (see figure 2), lowly expressed genes tend to express higher variability. However, we notice cellwise samples show a higher average CPM (count per million), reminiscent of its reduced library size. It also shows a higher BCV (~5 as compared

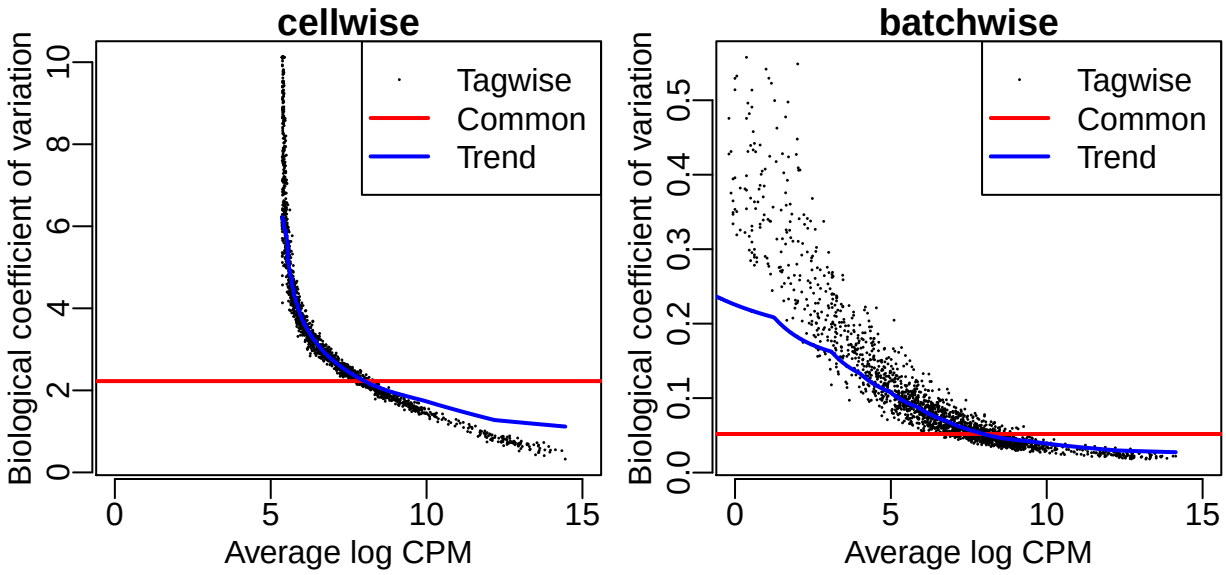
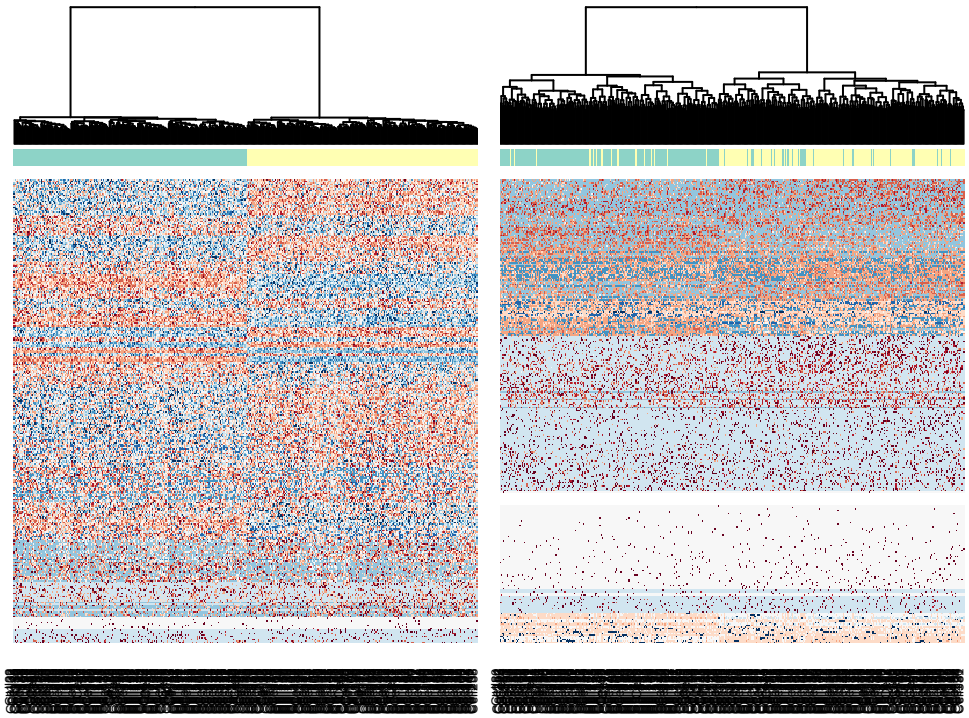


Figure 2: The inter-library variation under different granularity

to ~0.3 in batchwise), reflecting that scRNA-Seq is capturing more within-batch heterogeneity. As a result of accurate BCV estimation, the robustness of DE-prediction is greatly improved (see MA-plot, figure 3).

Notice we have yet to quantify this robustness in DE-precdition, due to difficulty in constructing the ground truth set: how to derive the set of true DE genes from the simulation parameters? The estimated fold-change is produced from fitting the glm and correlates very well with the theoretical fold-change, but does not tell anything about the quality of the LR-test. At the moment, we use heatmap to allow visually evalute the DE-prediction. Striking, batchwise model produces much better prediction of DE than cellwise model (see figure 5,4). Whether this is an artifact of my computation or a genuine difference between the two models, requires a further investigation.



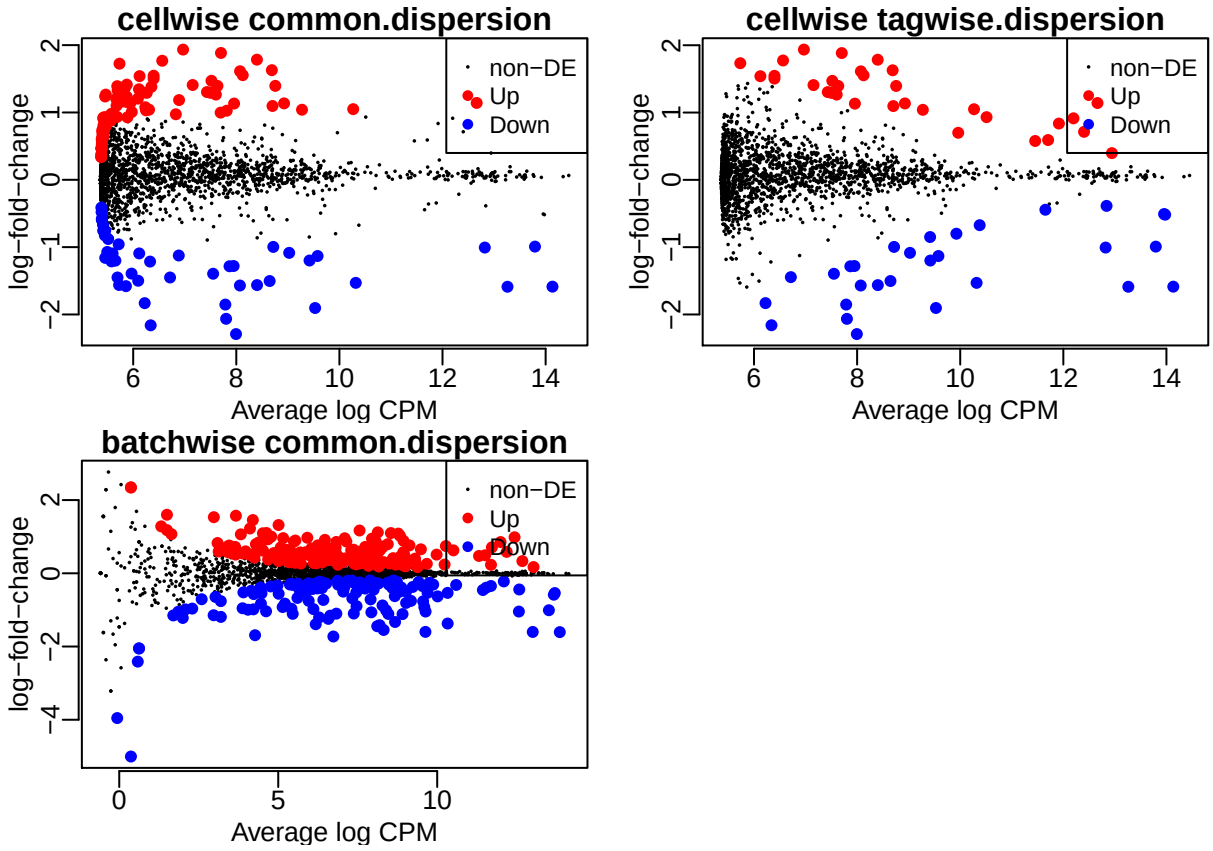


Figure 3: MA-plots colored at FDR=0.05 (under BH correction)

Figure 4: CPM matrix after dropout of DE genes predicted at batchwise level using tagwise dispersion. Left: without dropout.Right:with dropout

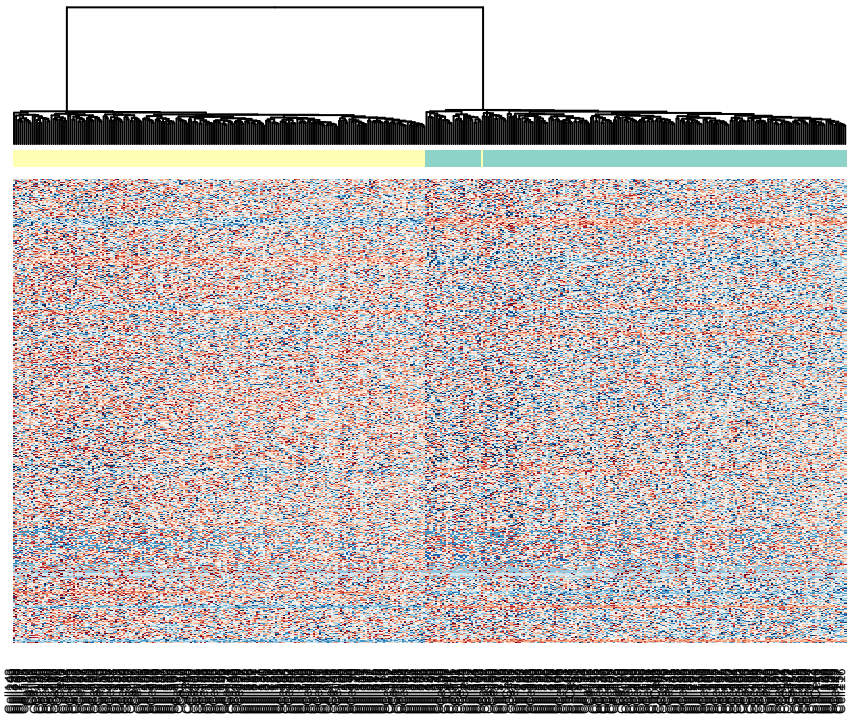


Figure 5: CPM matrix of DE genes predicted at cellwise level using tagwise dispersion

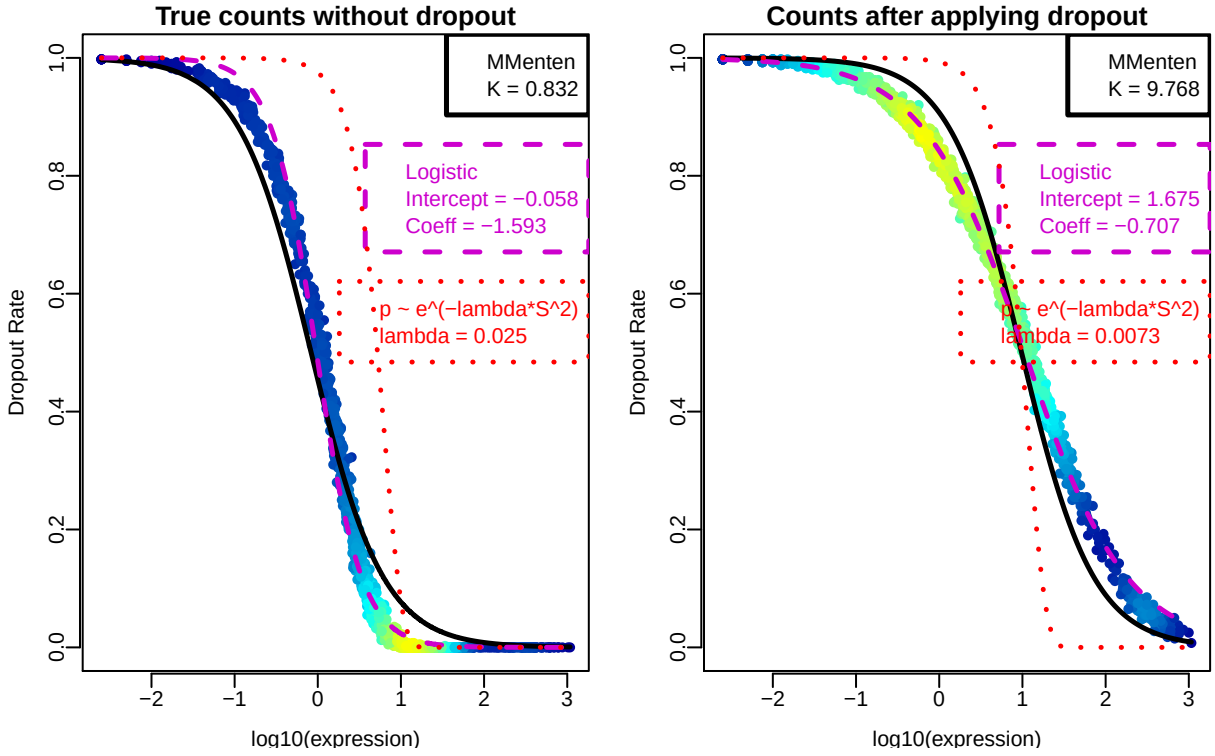


Figure 6: Distribution of dropout events

Dropout

We adjusted the dropout midpoint so that the simulated data resembles real data (@) in their distribution of dropout probabilty, often modeled as a function of average CPM of non-zero emissions:

$$P(m = 0) = f(E(m|m > 0))$$

Importantly, after application of dropout, the previously identified DE-genes are masked by noise on the CPM matrix (figure 4right), possibly due to it compromising the estimation for library size and adding noise to the biological signal. The dropout should be taken as a phenomena that breaks the Poisson/NB distribution assumption, and may be visualised by fitting a Michalis-Menten curve to the dropout probability distribution (figure 6), with a bigger K indicating more severe dropout. Whether other packages like zinbwave/BASiCS is capable of improving this situation requires further investigation.

Future

1. Explore how true batch effects can be removed.
2. Seek a quantitaive evaluation of DE-prediction
3. Explore DE-prediction with noisy dataset.

References:

Andrews, Tallulah S., and Martin Hemberg. 2016. "Modelling Dropouts Allows for Unbiased Identification of Marker Genes in scRNASeq Experiments." *bioRxiv*. Cold Spring Harbor Laboratory. doi:10.1101/065094.

Haque, Ashraful, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. 2017. "A Practical Guide to Single-Cell Rna-Sequencing for Biomedical Research and Clinical Applications." *Genome Medicine* 9 (1): 75. doi:10.1186/s13073-017-0467-4.

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. "Differential Expression Analysis of Multifactor Rna-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10): 4288–97. doi:10.1093/nar/gks042.

Müller, Sören, and Aaron Diaz. 2017. "Single-Cell mRNA Sequencing in Cancer Research: Integrating the Genomic Fingerprint." doi:10.3389/fgene.2017.00073.

Zappia, Luke, Belinda Phipson, and Alicia Oshlack. 2017. "Splatter: Simulation of Single-Cell Rna Sequencing Data." *Genome Biology*. doi:10.1186/s13059-017-1305-0.