main

General methodology for RNA-SEQ

1. Find out the quantity and sequences of RNA within a mixture, using RNA reads as a proxy.

Problems:

- De-novo assembly of transcripts (Trinity)
- Genome-based mapping/assembly (cufflink/tophat)
- Transcriptom-based mapping

Depending on the scientific question to be asked, methods shall be commanded as appropriate. The common granularity of mapping includes: gene, transcript/isoform, haplotype-specific isoforms. However to detect SNP one must be stringent when aligning. How to deal with un-mapped reads, is another problem.

1. preserve haplotye/isoform information is important for certain hypotheses testing
2. Multiple reads to mulitple transcript-sets: a deconvolution problem
3. Normalisation and Hypothesis testing: Obvious technical variations need to be corrected to allow for meaningful comparison, between biological samples.

- library size
- GC content/sequence composition
- transcript length

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.
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.

“In addition to resolving cellular heterogeneity, scRNA-seq can also provide important information about fundamental characteristics of gene expression” Haque et al. (2017)

“sensitivity for lowly expressed trnnscripts, full length identification, throughput/cost per cell”

Research interest:

1. Supervised testing: differential expression between given treatments
2. Unsupervised learning: detecting sub-population/patterns of expression within cell populations.
3. test-set:benchmark against bulk RNA-seq DE results. Does the scRNA allows identification of DE at a similar level? According to central-limit-theorem, using a larger sample will certainly reduce the noise.
4. Granularity: by analogy to gene-isoform problem, grouping can be done at different granularity. If we group quisecent/active T-cell together, then we will not see any difference at that level. However, this information is not required in DE analysis anyway. Think about the signal, cell differentiation/trajectory tracking is obviously one of them. DE is obviously another. But remembering the lesson from ChipSeq, the choice made in preprocessing can have an effect on the final output.
5. Be cautious about batch effect and technical noise

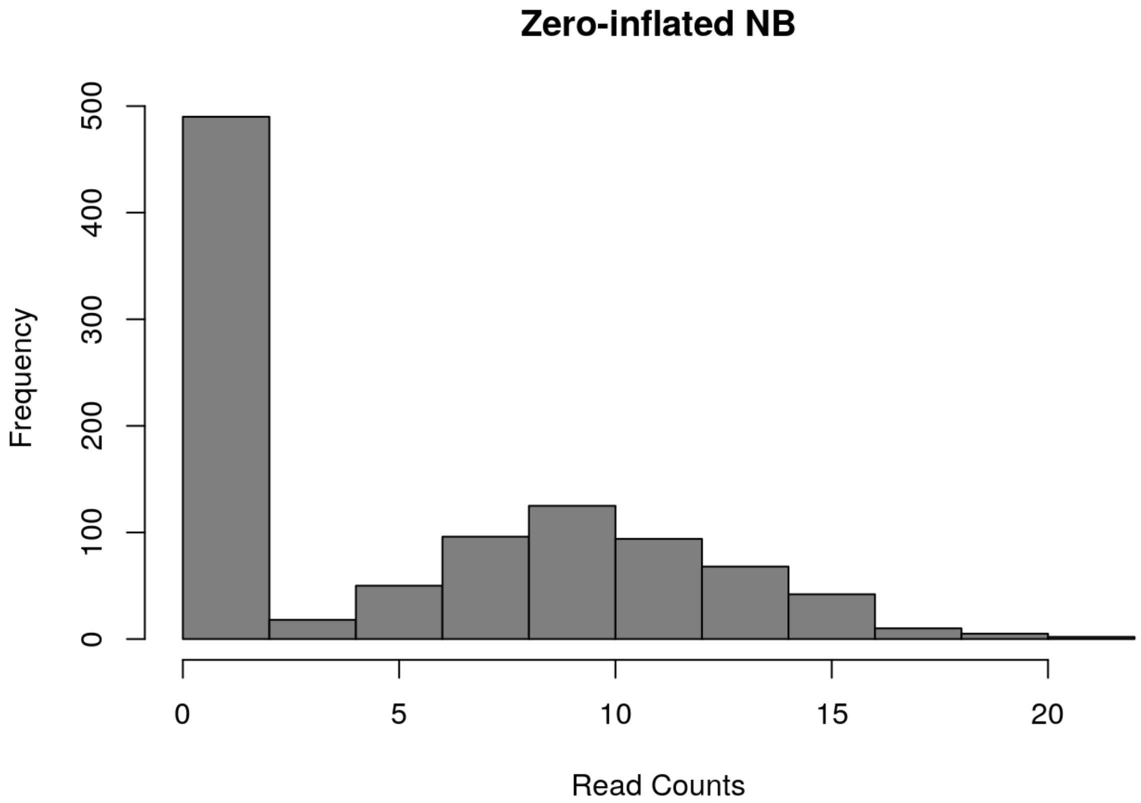

Figure 1: dropout

Batch effect is simulated by mutiplicating a factor sampled from 2 superposed log-normal distribution

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

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

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

Dropout (Zero-inflated negative binomial)

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

Since BCV represent the biologically meaningful variation, we ask how does scRNA-Seq changes our estimation for BCV.

de.facLoc

Use PRBE to becnhmark the three dispersion information.

Dispersion in differential expression

The likelihood ratio indicates how likely the

Because genes with lower average CPM tends to express more dispersion, it tends to be associated with a low confidence in LR-test. However, multiple schemes for estimating dispersion exists. If the dispersion is assumed constant across all genes (the “common” scheme), then there will be less power in separating biological signal from technical noise, causing more DE to be called during the LR-test.

However this cause problem to 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.

We are still struggling to find a good way to specify the ground truth, but the observation is that as estimation for dispersion becomes more specific, the LR-test becomes increasingly conservative.

Simple division of the multiplicative factors gives a fold-change factor, which correlates very well with the prediction, regardless of the dispersion used in LR-test.

The glmFIT would give the

Median is very stable across a group of cells, and causes dispersion to vanish.

How to normalise for dropout?

Stage Conclusion:

Single cell-based estimation is a lot more sensitive to

DE-differentially expressed

I visualised the prediction for DE by

1. Correlation plot: theoretical-logFC against fitted-logFC
2. Volcano plot of fitted-LR(likelihood ratio) against fitted-logFC

Definition of LR?

From correlation plot we identified false positives that exhibited spurious logFC with a theoretical-logFC of 0 (colored red). This group is recapitulated in the volcano plot as having low LR.

Incorrect:Side observation: single-cell result is less sensitive to FDR correction (BH usedd here), whereas simulated bulk is more prone, likely due to the fact that the signal from scRNA-seq is richer and hence more robust. It's also possible that the simulated bulk-RNA data has some statistical feature that prevented accurate estimation of P-value/likelihood ratio. However, I observed that normalising P-value by dividing it by exp(-logFC) restored the ROC curve. **Reason**: The median across cell group is taken – value is too stable for any dispersion to be estimated!!!

Further:

Normalisation by library size is crucial

We will be conducting the analysis with three packages:

1. "scater": General quality control. This package uses an assortment of statistical models to visualise the variance within the samples.
2. "edgeR" : Inference of differential expressed genes using a negative binomial model. Importantly, the dispersion parameter is fitted to indicate the inherent variance of any gene, which greatly improve the reliability of the inference.
3. "M3Drop": Investigating the dropout phenomena.

General question answering, experimental consideration of RNA-seq

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))

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 best unique in 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 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 supersede 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).

One of the consideration is the noise introduced

Addressing the DE

Start with the easiest setup with no batch effect, and evaluate

Plan: Plotting ROC/PR curve to validate the efficacy of algorithm

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.

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.