Tools and techniques for single-cell RNA-seq data

Luke Zappia

Table of Contents

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Declaration

This is to certify that:

1. the thesis comprises only their original work towards the [name of the award] except where indicated in the preface;
2. due acknowledgement has been made in the text to all other material used; and
3. the thesis is fewer than the maximum word limit in length, exclusive of tables, maps, bibliographies and appendices or that the thesis is [number of words] as approved by the Research Higher Degrees Committee.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Acknowledgements

This template is based on thesisdown (<https://github.com/ismayc/thesisdown>) and makes use of RMarkdown (<https://rmarkdown.rstudio.com/>) and bookdown <https://bookdown.org/yihui/bookdown/>. The LaTeX template is based on John Papandriopoulos’ University of Melbourne thesis template (<https://github.com/jpap/phd-thesis-template>). Inspriation also comes from similar projects including beaverdown (<https://github.com/zkamvar/beaverdown>), aggidown (<https://github.com/ryanpeek/aggiedown>), huskydown (<https://github.com/benmarwick/huskydown>) and jayhawkdown (<https://github.com/wjakethompson/jayhawkdown>).

# Introduction

## RNA sequencing

* Central dogma
* Why RNA-seq?

### Capture and reverse transcription

* PolyA capture
* Ribosomal depletion

### High-throughput sequencing

* Illumina sequencing
  + Sequence by synthesis

### Analysis of RNA-seq data

* Experimental design
* Negative binomial
* Normalisation
* Differential expression testing

## Single-cell RNA-sequencing

Traditional bulk RNA-seq experiments average the transcriptome across the many cells in a sample but recently it has become possible to perform single-cell RNA-sequencing (scRNA-seq) and investigate the transcriptome at the resolution of individual cell. There are many situations were it is important to understand how specific cell types react and where analyses may be affected by the unknown proportions of cell types in a sample. Studies into gene expression in specific cell types previously required a way to select and isolate the cells they were interested which removed them from the other cell types they are usually associated with and made it impossible to investigate how they interact. With scRNA-seq technologies it is now possible to look at the transcriptome of all the cell types in a tissue simultaneously which has lead to a better understanding of what makes cell types distinct and the discovery of previously unknown cell types.

One area that has particularly benefitted from the rise of scRNA-seq is developmental biology. Although the genes involved in the development of many organs are now well understood arriving at this knowledge has required many painstaking experiments. During development cells are participating in a continuous dynamic process involving the maturation from one cell type to another and the creation of new cell types. Single-cell RNA-seq captures a snapshot of this process allow the transcriptome of intermediate and mature cells to be studied. This has revealed that some of the genes thought to be markers of specific cell types are more widely expressed or involved in other processes.

## Single-cell capture technologies

* First protocol
* Fluidigm

The first scRNA-seq protocol was published in 2009, just a year after the first bulk RNA-seq publication. While this approach allowed measurements of the transcriptome in individual cells it required manual manipulation and was restricted to inspecting a few precious cells. Further studies quickly showed that cell types could be identified without sorting cells and approaches were developed to allow unbiased capture of the whole transcriptome. Since then many scRNA-seq protocols have been developed including …. The first commercially available cell capture platform was the Fluidigm C1. This system uses microfluidics to passively separate cells into individual wells on a plate where they are lysed, reverse-transcribed and the collected cDNA is PCR amplified. After this stage the product is extracted from the plate and libraries prepared for Illumina sequencing. Most C1 data has been produced using a 96 well plate but more recently an 800 well plate has become available, greatly increasing the number of cells that can be captured at a time. One of the disadvantages of microfluidic cell capture technologies is that the chips have a fixed size window, meaning that only cells of a particular sizes can be captured in a single run. However, as cells are captured in individual wells they can be imaged before lysis, potentially identifying damaged or broken cells, empty wells or wells containing more than one cell. Capturing multiple cells is a known issue, with Macosko et al. finding that when preparing a mixture of mouse and human cells 30 percent of the resulting libraries contained transcripts from both species but only about a third of these doublets were visible in microscopy images[Macosko2015-rl]. The newer Polaris system from Fluidigm also uses microfluidics to capture cells but can select particular cells based on staining or fluorescent reporter expression and then hold them for up to 24 hours while introducing various stimuli. The cells can be imaged during this time before being lysed and prepared for RNA sequencing. This platform provides opportunities for a range of experiments that aren’t possible using other capture technologies.

### Droplet based cell capture

* Drop-seq
* Indrop
* 10x Chromium

An alternative to using microfludics to capture cells in wells is to capture them in nano-droplets. A dissociated cell mixture is fed into a microfluidic device while at another input beads coated in primers enter. The device is designed to form aqueous droplets within mineral and the inputs are arranged so that cells and beads can be simultaneously captured within a droplet. When this happens the reagents carried along with the bead lyse the cell and any PolyA tagged RNA molecules present can bind to the capture probes on the bead. Reverse transcription and PCR amplification then begins and an individual cDNA library is produced for each cell, tagged with the unique barcode sequence present on the bead. The main advantage of droplet-based capture technologies is the ability to capture many more cells at one time, up to tens of thousands. These approaches are also less selective about cell size and produce less doublets. As a result they are much cheaper per a cell, although as sequencing costs are fixed studies using droplet-based captures typically sequence individual cells at a much lower depth.

Droplet-based capture was popularised by the publication of the Drop-seq and InDrop platforms in 2015. This are both DIY systems and although they differ in how the beads are produced, when the droplets are broken and some aspects of the chemistry they can both be constructed on a lab bench from syringes, automatic plungers, a micro scope and a small custom-made microfluidic chip. A similar commercially available platform is the 10x Genomics Chromium device which automates and streamlines much of the process. This device uses droplet-based technologies for a range of applications including capture of cells for scRNA-seq. More specialised captures, such as those aimed at profiling immune cell receptors are also possible and the company has recently announced kits for single-cell ATAC-seq capture.

### Unique Molecular Identifiers

* Why?
* How they work

In contrast to plate-based capture methods, which often provide reads along the length of transcripts, droplet-based capture methods typically employ protocols which include short random nucleotide sequences known as Unique Molecular Identifiers (UMIs). Individual cells contain very small amounts of RNA and to obtain enough cDNA a PCR amplification step is necessary. Depending on their nucleotide sequence different transcripts may be amplified at different rates which can distort their relative proportions within a library. UMIs attempt to improve the quantification of gene expression by allowing the removal of PCR duplicates produced during amplification. The nucleotide probes used in droplet-based capture protocols include a PolyT sequence which binds to mature mRNA molecules, a barcode sequence which is the same for every probe on a bead and 8-10 bases of UMI sequence which is unique to each probe. The UMI sequences are long enough that the probability of capturing two copies of a transcript on two probes with the same UMI is extremely low. After reverse-transcription, amplification, sequencing and alignment de-duplication can be performed by identifying reads with the same UMI that align to the same position and therefore should be PCR duplicates rather than truly expressed copies of a transcript. For this method to be effective each read must be associated with a UMI which means that only a small section at the 3’ end of each transcript is sequenced. This has the side effect of reducing the amount of cDNA that needs to be sequenced and therefore increasing the number of cells that can be sequenced at a time. While the improvement in quantification of gene expression levels is useful for many downstream analyses it comes at the cost of coverage across the length of a gene which is required for applications such as variant detection and de-novo assembly. **READS ALONG GENE** Statistical methods designed for full-length data may also be affected by the difference properties of a UMI dataset. Datasets with UMIs also need extra processing steps which can be complicated by the possibility of sequencing errors in the UMI itself.

### Recent advances

* New capture methods
* CITE-seq
* Cell hashing
* CRISPR
* Multiple measurements, same cell

Although droplet-based techniques are currently the most commonly used cell capture technologies other approaches have been proposed that promise to capture even more cells even more cheaply. These include approaches based around nanowells…

Extensions to the standard protocols have also been proposed that allow extra measurements from the same cell. One such protocol is CITE-seq which enables measurement of the levels of selected proteins at the same time as the whole transcriptome. Antibodies for the proteins of interest are labelled with short nucleotide sequences. These antibodies can then be applied to the dissociated cells and any that remain unbound washed away before cell capture. The antibody labels are then captured along with mRNA transcripts and a size selection step is applied to separate them before library preparation. Similar antibodies can be used to allow multiplexing of samples through a process known as cell hashing. In a typical scRNA-seq experiment each batch corresponds to a single sample. This complicated analysis as it is impossible to tell what is noise due to cells being processed in the same way and what is true biological signal. Cell hashing uses an antibody to a ubiquitously expressed protein but with a different nucleotide sequence for each sample. The samples can then be mixed, processed in batches and then the cells computationally separated based on which sequence they are associated with. An added benefit of this approach is the simple detection of doublets containing cells from different samples.

CRISPR-Cas9 gene editing has also been developed as an extension to scRNA-seq protocols. One possibility is to introduce a mutation at a known location that can then be used to demultiplex samples processed together. It is possible to do this with samples from different individuals or cell lines but the advantage of a gene editing based approach is that the genetic background remains similar between samples. It is also possible to investigate the effects of introducing a mutation. Protocols like Perturb-Seq introduce a range of guide RNA molecules to a cell culture, subject the cells to some stimulus then perform single-cell RNA sequencing. The introduced mutation can then be linked to the response of the cells to the stimulus and the associated broader changes in gene expression.

Other approaches that allow multiple measurements from individual cells include…

## Analysing scRNA-seq data

* Low counts
  + Dropout
  + Bursting
  + Biology

Cell capture technologies and scRNA-seq protocols have developed rapidly but there are still a number of challenges with the data they produce. Existing approaches are inefficient, capturing around 10 percent of transcripts in a cell(**???**). When combined with the low sequencing depth per cell this results in a limited sensitivity and an inability to detect lowly expressed transcripts. The small amount of starting material also contributes to high levels of technical noise, complicating downstream analysis and making it difficult to detect biological differences(**???**). In order to capture cells they must first be dissociated into single-cell suspensions but this step can be non-trivial. Some tissues or cell types may be more difficult to separate than others and the treatments required to break them apart may effect the health of the cells and their transcriptional profiles. Other cell types may be too big or have other characteristics that prevent them being captured. In these cases related techniques that allow the sequencing of RNA from single nuclei may be more effective. Cells may be damaged during processing, multiple cells captured together or empty wells or droplets sequenced making quality control of datasets an important consideration.

As well as increasing technical noise the small amounts of starting material and low sequencing depth mean there are many occasions where zero counts are recorded, indicating no measured expression for a particular gene in a particular cell. These zero counts often represent true biological signal we are interested as we expect different cell types to express different genes. However they can also be the result of confounding biological factors such as stage in the cell cycle, transcriptional bursting and environmental interactions which cause genuine changes in expression but that might not be of interest to a particular study. On top of this there are effects that are purely technical factors in particular sampling effects which mean result in “dropout” events where a transcript is truly expressed in a sample but is not observed in the sequencing data. In bulk experiments these effects are limited by averaging across the cells in a sample but for single-cell experiments they can present a significant challenge for analysis as methods must account for the missing information and they may cause the assumptions of existing methods to be violated. One approach to tackling the problem of too many zeros is to use zero-inflated versions of common distributions but it is debatable whether scRNA-seq datasets are truly zero-inflated or the the additional zeros are better modeled with standard distributions with lower means. Another approach is to impute some of the zeros, replacing them with estimates of how expressed those genes truly are based on their expression in similar cells. However imputation comes with the risk of introducing false structure that is not really present in the data.

Bulk RNA-seq experiments usually involve predefined groups of samples, for example cancer cells and normal tissue, different tissue types or treatment and control groups. It is possible to design scRNA-seq experiments in the same way for example by sorting cells into known groups based on surface markers, sampling them at a series of time points or comparing treatment groups but often they are more exploratory. Many of the single-cell studies to date have sampled developing or mature tissues and attempted to profile the cell types that are present[Zeisel2015-rd; Patel2014-bl; Treutlein2014-wd; Usoskin2015-fz; Buettner2015-rq; Klein2015-iw; Trapnell2014-he]. This approach is best exemplified by the Human Cell Atlas project which is attempting to produce a reference of the transcriptional profiles of all the cell types in the human body. Similar projects exist for other species and specific tissues. As scRNA-seq datasets have become more widely available a standard workflow has developed which can be applied to many experiments. This workflow can be divided into four phases: 1) Data acquisition, Pre-processing of samples to produce a cell by gene expression matrix, 2) Data cleaning, quality control to refine the dataset used for analysis, 3) Cell assignment, grouping or ordering of cells based on their transcriptional profile, and 4) Gene identification to find genes that represent particular groups and can be used to interpret them. Within each phase a range processes may be used and there are now many tools available for completing each of them, with over XXX tools currently available. An introduction to the phases of scRNA-seq analysis is provided here but the analysis tools landscape is more fully explored in Chapter X.

### Pre-processing and quality control

* Alignment
* Droplet selection
* UMIs
* Doublet detection
* Bad cells
* Gene filtering
* Cell ranger
* scater
* cell free DNA

The result of a sequencing experiment is typically a set of image files from the sequencer or a FASTQ file containing nucleotide reads but for most analyses we use an expression matrix. To produce this matrix there is a series of pre-processing steps, typically beginning will some quality control of the raw reads. Reads are then aligned to a reference genome and the number of reads overlapping annotated features (genes or transcripts) is counted. In recent years probabilistic quantification methods such as kallisto[Bray2016-tm] or Salmon[Patro2015-kl] that estimate transcript expression directly without requiring complete alignment have become popular as they dramatically reduce processing time and potentially produce more accurate quantification. These can be applied to full-length scRNA-seq datasets but have required adaptations such as the Alevin method for UMI-based datasets. When using conventional alignment UMI samples need extra processing with tools like UMI-tools[Smith2016-bt] or umis[Svensson2016-eg] in order to assign cell barcodes and deduplicate UMIs. For datasets produced using the Chromium platform the Cell Ranger software is a complete preprocessing pipeline that also includes an automated downstream analysis. Other packages such as scPipe also aim to streamline this process with some such as XXX designed to work on scalable cloud based infrastructure which may be required as bigger datasets continue to be produced.

Quality control of individual cells is important as experiments will contain low-quality cells that can be uninformative or lead to misleading results. Quality control can be performed on various levels, from the quality scores of the reads themselves, how or where reads align to features of the expression matrix. Particular types of cells that are commonly removed include damaged cells, doublets where multiple cells have been captured together and empty droplets or wells that have been sequenced but do not contain a cell. The Cellity package attempts to automate this process by inspecting a series of biological and technical features and using machine learning methods to distinguish between high and low-quality cells[Ilicic2016-wy]. However the authors found that many of the features were cell type specific and more work needs to be done to make this approach more generally applicable. The scater package[McCarthy2016-cw] emphasises a more exploratory approach to quality control at the expression matrix level but providing a series of functions for visualising various features of a dataset. These plots can then be used for selecting thresholds for removing cells. Plate-based capture platforms can produce additional biases based on the location of individual wells, a problem which is addressed by the OEFinder package which attempts to identify and visualise these “ordering effects”[Leng2016-it].

Filtering and selection of features also deserves attention. Genes or transcripts that are lowly expressed are typically removed from datasets in order to reduce computational time and multiple-testing correction but it is unclear how many counts indicate that a gene is truly “expressed”. Many downstream analysis operate on a selected set of genes which can have a dramatic effect on their results. These features are often selected based on how variable they are across the dataset but this may be a result of noise rather than biological importance. Alternative selection methods have been proposed such as M3Drop which…

### Normalisation and integration

* Why?
* Seurat CCA
* New methods
* Tung?
* Different data types

Technical variation is a known problem in high-throughput genomics studies, for example it has been estimated that only 17.8 percent of allele-specific expression is due to biological variation with the rest being technical noise[Kim2015-mo]. Effective normalisation has been shown to be a crucial aspect of analysis for bulk RNA-seq datasets and similarly this is true for single-cell experiments. Some full-length studies use simple transformations like Reads (or Fragments) Per Kilobase per Million (RPKM/FPKM)[Mortazavi2008-vu] or Transcripts Per Million (TPM)[Wagner2012-qf] which correct for the total number of reads per cell and gene length. For UMI data the gene length correction is not required as reads only come from the ends of transcripts. Normalisation methods designed for detecting differential expression between bulk samples such as Trimmed Mean of M-Values (TMM)[Robinson2010-ll] or the DESeq method[Anders2010-pq] can be applied, but is is unclear how suitable they are for the single-cell context. Most of the early normalisation methods developed specifically for scRNA-seq data made use of spike-ins, synthetic RNA sequences added to cells in known quantities such as the ERCC…. Brennecke et al.[Brennecke2013-pt], Ding et al.[Ding2015-ht] and Grün, Kester and van Oudenaarden[Grun2014-zn] all propose methods for estimating technical variance using spike-ins, as does Bayesian Analysis of Single-Cell Sequencing data (BASiCS)[Vallejos2015-ef]. Using spike-ins for normalisation assumes that they properly capture the dynamics of the underlying dataset and even if this is the case it is restricted to protocols where they can be added which does not include droplet-based capture techniques. The scrna package implements a method that doesn’t rely on spike-ins, instead using a pooling approach to compensate for the large number of zero counts where expression levels are summed across similar cells before calculating size factors that are deconvolved back to the original cells[Lun2016-mq]. The BASiCS method has also been adapted to experiments without spike-ins by…, but only for designed experiments where groups are known in advance.

Early scRNA-seq studies often made use of only a single sample but as technologies have become cheaper and more widely available it is common to see studies with multiple batches or making use of publicly available data produced by other groups. While this expands the potential insights to be gained it presents a problem as to how to integrate these datasets and a range of computational approaches for doing this have been developed. The alignment approach in the Seurat package uses Canonical Correlation Analysis (CCA) to identify a multi-dimensional subspace that is consistent between datasets. Dynamic Time Warping (DTW) is then used to stretch and align these dimensions so that the datasets are similarly spread along them. Clustering can then be performed using these aligned dimensions but as the original expression matrix is unchanged the integration is not used for other tasks such as differential expression testing. The authors of scran using a Mutual Nearest Neighbours (MNN) approach that… A recent update to the Seurat method combines these approaches by identifying “anchors” that…Alternative integration methods such as…

### Grouping cells

* Clustering
* Seurat
* Other approaches
* Comparison
* Classification

Grouping similar cells is a key step in analysing scRNA-seq datasets that is not usually required for bulk experiment and as such it has been a key focus of methods development with over XXX tools released for clustering cells. Some of these methods include SINgle CEll RNA-seq profiling Analysis (SINCERA)[Guo2015-mf], Single-Cell Consensus Clustering (SC3)[Kiselev2016-fa], single-cell latent variable model (scLVM)[Buettner2015-rq] and Spanning-tree Progression Analysis of Density-normalised Events (SPADE)[Anchang2016-vo], as well as BackSPIN which was used to identify nine cell types and 47 distinct subclasses in the mouse cortex and hippocampus[Zeisel2015-rd]. All of these tools attempt to cluster similar cells together based on their expression profiles, forming groups of cells of the same type. One clustering method that has become popular is that included in the Seurat package. This method begins by selecting a set of highly variable genes then performing PCA on them.**NEW GENE SELECTION** A set of dimensions is then selected that contains most of the variation in the dataset. Alternatively if Seurat’s alignment method has been used to integrate datasets the aligned CCA dimensions are used instead. Next an MNN graph is constructed by considering the distance between cells in this multidimensional space. In order to separate cells into clusters a community detection algorithm such as Louvain optimisation is run on the graph with a resolution parameter that controls the number of clusters that are produced. Seurat’s clustering method has been shown too….

For tissue types that are well understood or where comprehensive references are available an alternative is to directly classify cells. This can be done using a gating approach based on the expression of known marker genes similar to that commonly used for flow cytometry experiments. Alternatively machine learning algorithms can be used to perform classification based on the overall expression profile. Methods such as … take this approach. For example… Classification has the advantage of making use of existing knowledge and avoids manual annotation and interpretation of clusters which can often be difficult and time consuming. However it is biased by what is present in the reference datasets used typically can not reveal previously unknown cell types or states. As projects like the Human Cell Atlas produce well-annotated references based on scRNA-seq data the viability of classification and other reference-based methods will improve.

### Ordering cells

* Pseudotime
* Monocle
* Other approaches
* Comparison

In some studies, for example in development where stem cells are differentiating into mature cell types, it may make sense to order cells along a continuous trajectory from one cell type to another instead of assigning them to distinct groups. Trajectory analysis was pioneered by the Monocle package which used dimensionality reduction and computation of a minimum spanning tree to explore a model of skeletal muscle differentiation[Trapnell2014-he]. Since then the Monocle algorithm has been updated and a range of other developed including TSCAN[Ji2016-ws], SLICER[Welch2016-cw], CellTree[DuVerle2016-ni], Sincell[Julia2015-zc] and Mpath[Chen2016-kx]. In their comprehensive review and comparison of trajectory inference methods Cannoodt, Saelens and Saeys break the process into two steps. In the first step dimensionality reduction techniques such as PCA or t-SNE[Maaten2008-ne] are used to project cells into lower**[?]** dimensions where the cells are clustered or a graph constructed between them. The trajectory is then created by finding a path through the cells and ordering the cells along it. This review compares the performance on a range of datasets… They found that…

Deciding on which assignment approach to use depends on the source of the data, the goals of the study and the questions that are being asked. Both grouping and ordering can be informative and it is often useful to attempt both on a dataset and see how they compare.

### Gene detection and interpretation

* DE
* Marker genes
  + Alternatives - Gini, classifiers
* Reviews
* Classification

Once cells are assigned by clustering or ordering the problem is to interpret what these groups represent. For clustered datasets this is usually done by identifying genes that are differentially expressed across the groups or marker genes that are expressed in a single cluster. Many methods have been suggested for testing differential expression some of which take in to account the unique features of scRNA-seq data. For example…The large number of cells in scRNA-seq datasets mean that some of the problems that made standard statistical tests unsuitable for bulk RNA-seq experiments do not apply and simple methods like the unpaired Wilcoxon rank-sum test (or Mann-Whitney U test) may give reasonable results in this setting. Methods originally developed for bulk experiments have have also been applied to scRNA-seq datasets. Some of these methods have well understood statistical frameworks and have been shown to perform well in multiple comparisons. However the assumptions they make may not be appropriate for single-cell data and methods such as ZiNB-WaVe may be required to transform the data that is appropriate for their use.

Often the goal is not to find all the genes that are differentially expressed between groups but to identify genes which uniquely mark particular clusters. This goal is open to alternative approaches such as the Gini coefficient which measures unequal distribution across a population. Another approach is to construct machine learning classifiers for each genes to distinguish between one group and all other cells. Genes that give good classification performance should be good indicators of what is specific to that cluster.

When cells have been ordered along a continuous trajectory the task is slightly different. Instead of testing for a difference in means between two groups the goal is to find genes that have a relationship between expression and pseudotime. This can be accomplished by fitting splines and testing the coefficients. For more complex trajectories it can also be useful to find genes that are differently expressed along each side of a branch points. Monocle’s BEAM method does this by… Genes that are associated with a trajectory are important in their own right as they describe the biology along a path but they can also be used to identify cell types at end points.

Interpreting the meaning of detected markers genes is a difficult task as is likely to remain so. Gene set testing to identify related categories such as Gene Ontology terms can help but often it is necessary to rely the results of previous functional studies. This can only be reliably done by working closely with experts who have significant domain knowledge in the cell types being studied. An additional concern for unsupervised scRNA-seq studies is that the same genes are used for clustering or ordering and determining what those clusters or trajectories mean. This is a problem addressed by XXX who suggest a differential expression test using a long-tailed distribution for testing genes following clustering.

### Alternative analyses

* Cell velocity
* Variant detection
* Cancer
* Immune cells

## Kidney development

### Structure and function

* Kidney structure
* Nephron structure
* Important cell types

### Stages of development

* Lineage
* Important genes

### Growing kidney organoids

* Why?
  + Disease modelling
* Protocol
* Growth factors
* Characterisation
* Reproducibility

# The scRNA-seq tools landscape

# Simulating scRNA-seq data

## Introduction

## Splatter publication

You can read the Splatter paper [here](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1305-0).

# Visualising clustering across resolutions

# Analysis of kidney organoid scRNA-seq data

# Conclusion

# References