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 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

The alternative approach is to capture cells in microfluidic droplets. Cells are  
introduced into a microfluidic device while at another input beads coated in  
primers enter. At the output of the device aqueous droplets are formed in a  
mineral oil. The device is designed such that cells and beads may arrive and be  
captured simultaneously. When this occurs the cells is lysed and reverse  
transcription begins. The primary advantage of droplet-based approaches is the  
ability to capture many more cells in a single run, up to tens of thousands.  
These approaches are also less selective about cell size and much cheaper per a  
cell, although as the sequencing costs are constant studies that use  
droplet-based captures typically sequence individual cells at a lower depth.  
There are two DIY droplet-based platforms, Drop-seq[@Macosko2015-rl] and  
InDrops[@Klein2015-iw; @Zilionis2017-gu]. They differ in how the beads are made,  
when the droplets are broken and some aspects of the chemistry but they can both  
be constructed on a lab bench from syringes, plungers, a microscope and a small  
custom-made microfluidic chip (Figure <a href="#fig:indrops"><strong>??</strong></a>). More recently 10x  
Genomics have commercialised the droplet capture method as their Chromium  
device, which automates much of the process[@Zheng2016-sj].

### Unique Molecular Identifiers

* Why?
* How they work

There are various ways that protocols for these platforms can be modified, as is  
evidenced by the number of published protocols including  
Smart-seq2[@Picelli2013-or; @Picelli2014-pc], CEL-seq[@Hashimshony2012-dh],  
CEL-seq2[@Hashimshony2016-yi], Quartz-seq[@Sasagawa2013-uc] and  
SCRB-seq[@Soumillon2014-nk]. Two decisions researchers need to make when  
deciding how to process their samples is whether to use a UMI or full-length  
protocol and whether to include external spike-in controls. UMIs attempt to  
improve the quantification of gene expression by allowing the removal of PCR  
duplicates produced during amplification. Individual RNA molecules are randomly  
tagged with short nucleotide sequences as part of the capture process. 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 (Figure <a href="#fig:UMIs"><strong>??</strong></a>). These reads should be PCR duplicates rather  
than reads that truly originate from expression of that gene. While the use of  
UMIs should improve the quality of the final data used for downstream analysis  
it also introduces some difficulties. Extra processing is required to match UMIs  
with aligned reads and remove duplicates, this is further complicated by the  
potential for sequencing errors in the UMI itself[@Smith2016-bt]. Analyses that  
require full coverage of a gene such as variant detection or de-novo assembly  
are not possible and statistical analyses that were designed using full-length  
data may not reliably transfer to a UMI dataset.

### Recent advances

* New capture methods
* CITE-seq
* Cell hashing

## Features of single-cell RNA-seq data

* Why use single-cell?
* Low counts
  + Dropout
  + Bursting
  + Biology

## Analysing scRNA-seq data

Single-cell RNA-seq technologies have developed rapidly but there are still a  
number of challenges. Existing protocols are inefficient, capturing only around  
10 percent of transcripts in a cell[@Grun2014-zn]. Combined with the relatively  
low sequencing depth that is commonly used, this results in a limited  
sensitivity and an inability to reliably 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[@Liu2016-wq]. All high-throughput scRNA-seq protocols and  
platforms require tissues to be dissociated into single-cell suspensions before  
capture. This step can be non-trivial. Some tissues or cell types may not  
readily separate and the treatments used to break them down may effect the  
health of the cells and therefore their transcriptional profiles. Additionally,  
some cell types may be too big or have other physical characteristics that  
prevent them being captured using currect methods. Cells may also be damaged  
during processing, or missing or multiple cells may have been sequenced, making  
quality control an important consideration.  
  
As well as introducing technical noise, the small amounts of starting material  
and low sequencing depth mean there are many occasions where there is no  
measured expression for a particular gene in a particular cell. Some of these  
zero counts are due to the biology we wish to study, for example we expect  
different cell types to express different genes, but there are additional  
biological factors such as the cell cycle, transcriptional bursting and  
environmental interactions which cause genuine differences in expression between  
cells performing the same function. On top on this are the technical effects  
that have already been discussed including that existing protocols may not  
reliably capture all the RNA present, resulting in "dropout" events where a gene  
is expressed in a sample but not observed in the sequencing data. These zeros  
can make analysis difficult as methods must account for the missing information  
and they may violate the assumptions of existing approaches. For example the  
DEseq2 package[@Love2014-tw] has proven successful for detecting differential  
expression in bulk RNA-seq but it relies on the presence of genes without any  
zero counts, and therefore typically fails on scRNA-seq data. Examples of  
scRNA-seq analysis packages that tackle this problem are Zero Inflated Factor  
Analysis (ZIFA)[@Pierson2015-qp], which explicitly models dropout as it affects  
dimensionality reduction, and Clustering through Imputation and Dimensionality  
Reduction (CIDR)[@Lin2016-yu], which implicitly imputes zeros as it clusters  
cells.  
  
Bulk RNA-seq experiments are usually conducted on predefined groups of samples,  
for example cancer cells and normal tissue, different tissue types or treated  
and control cells. Some scRNA-seq experiments are done in a similar way where  
cells are sorted into known types based on surface markers or selected at a  
series of time points, but often they are more exploratory. Many of the current  
studies have taken samples of developing or mature tissues and attempted to  
identify what cell types are present[@Zeisel2015-rd; @Patel2014-bl;  
@Treutlein2014-wd; @Usoskin2015-fz; @Buettner2015-rq; @Klein2015-iw;  
@Trapnell2014-he]. This requires a new set of analysis techniques to be  
developed which attempt to identify cell types. Success at this task is crucial  
to the reliability of results from more developed analysis methods such as gene  
testing. Additionally is it impossible to truly replicate an individual cell and  
therefore thought must be given as to what constitutes a replicate for the  
purposes of statistical analysis.

### Pre-processing and quality control

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

As scRNA-seq data has become available there has been a rapid development of new  
bioinformatics tools attempting to unlock its potential. Currently there are at  
least 80 available software packages that have been designed specifically for  
the analysis of scRNA-seq data, the majority of which have been published in  
peer-reviewed journals or as preprints. A table of scRNA-seq software is  
available at [https://goo.gl/4wcVwn](). Prior to analysis the sequencing reads  
from an scRNA-seq experiment are processed in much the same way as a bulk  
experiment. Typically there is some quality control of the raw reads, reads are  
aligned to a reference genome and the number of reads overlapping annotated  
features (genes or transcripts) is counted. Alternatively, for full-length  
reads, probabilistic quantification methods such as kallisto[@Bray2016-tm] or  
Salmon[@Patro2015-kl] can be used. These approaches can greatly improve  
processing time which is important when there may be tens of thousands of  
samples but they are currently incompatible with UMI protocols. When using  
conventional alignment UMI samples need extra processing with tools such as  
UMI-tools[@Smith2016-bt] or umis[@Svensson2016-eg]. The resulting gene by cell  
matrix of expression values is the starting point for most analysis.  
  
Quality control of individual cells is important as most experiments will  
contain some low-quality cells that could be uninformative or misleading.  
Quality control can be performed on various levels: on the quality scores of the  
reads themselves, how or where the reads align or features of the expression  
matrix such as the total expression, expression of spike-ins or expression of  
particular genes. The Cellity package attempts to do this by inspecting a series  
of biological and technical features and using principal component analysis or  
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. Jiang, Thomson and Stewart take a different approach,  
assuming that expression outliers are associated with poor sequencing  
quality[@Jiang2016-ys]. The scater package[@McCarthy2016-cw] emphasises a more  
exploratory approach to quality control. While it cannot automatically detect  
low-quality cells, scater provides a convenient object for storing scRNA-seq  
data with functions for plotting associated features, making it easy for the  
user to define their own filtering thresholds. Plate-based platforms such as the  
Fluidigm C1 can have additional biases based on the location of individual  
wells. The OEFinder package attempts to identify and visualise these "ordering  
effects"[@Leng2016-it].

### Integrating multiple datasets

* Why?
* Seurat CCA
* New methods

Technical variation is a known problem in high-throughput studies and Kim et al.  
predict 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, but how normalisation should be be applied to single-cell datasets is  
yet to be clearly established. Some 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 library size and  
gene length. Alternatively, normalisation methods designed for detecting  
differential expression in bulk samples such as the Trimmed Mean of M-Values  
(TMM)[@Robinson2010-ll] or the DESeq method[@Anders2010-pq] can be applied, but  
it is unclear how suitable they are for the single-cell context. Most of the  
methods that have been developed specifically for estimating technical variance  
in scRNA-seq data make use of spike-ins. 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].  
However, using spike-ins for normalisation relys on the assumption that they  
properly capture the dynamics of the underlying datasets, and even if this is  
the case is it restrictive as they are not compatible with all current  
sequencing protocols. Lun, Bach and Marioni don't make use of 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  
normalisation factors that are deconvolved back to the individual cell  
level[@Lun2016-mq].

### Grouping cells

* Clustering
* Seurat
* Other approaches

Once a set of high-quality cells has been established the true analysis can  
begin. Many of the current packages focus on the task of assigning cells to  
groups before applying more traditional differential expression testing. This  
approach makes sense for a sample with a defined set of mature cell types and is  
taken by tools such as SINgle CEll RNA-seq profiling Analysis  
(SINCERA)[@Guo2015-mf], Single-Cell Consensus Clustering (SC3)[@Kiselev2016-fa],  
Seurat[@Satija2015-or], 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]. These tools attempt to cluster similar  
cells together based on their expression profiles, forming groups of cells of  
the same type. Often a dimensionality reduction step is included which can help  
to remove some of the noise present in scRNA-seq data. Once groups of cells are  
identified many of these packages can test genes for changes in expression,  
identifying genes that are differentially expressed across the groups, or marker  
genes that are expressed in a single group. These genes can be used to identify  
which cell types each group represents, or alternatively known marker genes can  
be inspected.

### Ordering cells

* Pseudotime
* Monocle
* Other approaches

In other situations, for example where stem cells are differentiating into  
mature cell types, it may be more appropriate to order cells along a continuous  
trajectory from one cell type to another. Trajectory analysis was pioneered by  
Monocle 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 others developed, including TSCAN[@Ji2016-ws],  
SLICER[@Welch2016-cw], CellTree[@DuVerle2016-ni], Sincell[@Julia2015-zc] and  
Mpath[@Chen2016-kx]. In their review of methods for trajectory inference,  
Cannoodt, Saelens and Saeys break the process into two parts, dimensionality  
reduction and then trajectory modelling (Figure  
<a href="#fig:trajectory-inference"><strong>??</strong></a>)[@Cannoodt2016-iv]. Dimensionality reduction  
consists of calculating similarities between cells, projecting onto lower  
dimensions using manifold learning techniques such as PCA or  
t-SNE[@Maaten2008-ne] then clustering cells or constructing a graph between  
then. The trajectory is then formed by finding a path between cells and ordering  
the cells along it. Once a path has been inferred important genes can be  
identified by looking for those that change expression over the course of the  
path. These genes can be important in their own right as they describe the  
biology of the path, but that can also be used to identify cell types at the end  
points of the path or where the path branches. Deciding which assignment  
approach is most appropriate depends on the source of that data and the  
questions you intend to ask. There are currently no studies comprehensively  
comparing the performance of different methods for each approach.

### Gene detection and interpretation

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

## 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