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

* Central dogma
  + Flow of information in cell
  + DNA - long term storage
  + Transcription to RNA - working copy, amplification
    - Messenger RNA
  + Translation to protein - functional
  + Some RNA also functional

The central dogma of biology describes the flow of information within a cell, from DNA to RNA to protein. Deoxyribonucleic acid (DNA) is the long term data storage of the cell. This molecule has a well known double strand structure. Each strand of the helix consists of a series of nucleic acid molecules linked by phosphate groups. These nucleic acids come in four species, adenosine, cytosine, tyrosine and guanine and the two strands are bound together through hydrogen bonds between matching nucleic acids known as basepairs. Guanine forms three hydrogen bonds with cytosine and adenosine forms two with tyrosine. In computing terms DNA is similar to a hard drive that provides stable, consistent storage of important information. When the cell wants to use some of this information it produces a copy of it in the form of a ribonucleic acid (RNA) molecule through a process known as transcription, similar to a computer loading information it wants to use into it’s random access memory. RNA is similar to a single strand of DNA except that the tyrosine base is replaced with another base called uracil. Because it is single-stranded RNA does not have a double helix structure but it can form complex shapes by binding to itself. There are several different types of RNA that serve different purposes. RNA molecules that are translated from genes are known as messenger RNA (mRNA). Other types of RNA include ribosomal RNA (rRNA) (that forms part of the ribosome), micro RNA which have a role in regulating gene expression and long-noncoding RNA. Genes are the sections of DNA that encode proteins and are made up of regions that code information (none as exons) with much larger non-coding regions between them (introns). When an mRNA molecule is transcribed it initially contains the intronic sequences but these are removed through a process known as splicing and a sequence of adenosine bases (a poly-A tail) is added to the end where transcription ends (the 3’ end) to mark a mature mRNA molecule. The process for producing a protein from an mRNA transcript occurs in a structure called the ribosome and is known as translation because the information encoded by nucleic acids in RNA is converted to information stored as amino acids in the protein. Proteins complete most of the work required to keep a cell functioning and can be compared to the programs running on a computer. These functions include tasks such as sensing things in the external environment, transport nutrients into the cell, regulating the expression of genes, constructing new proteins, recycling molecules and metabolism. Understanding the molecules involved in the central dogma is central to our understanding of how a cell functions.

## RNA sequencing

* Why RNA-seq?
  + What is happening?
  + High throughput
  + Easy
  + Unbiased
  + Complete
  + Different cell types

By looking at DNA we can see what genes are present in a cell but we cannot tell which of them are active and what processes they might be involved in. To do that we need to inspect the parts of the system that change dynamically. Ideally we might want to interrogate which proteins are present as they provide most of the functionality. However, while it is possible to do this using technologies such as mass spectrometry the readout they produce is difficult to interpret and the encoding is much more complex as there are 20 types of amino acids compared to only four nucleotides. In contrast RNA molecules are much easier to measure. High-throughput RNA sequencing (RNA-seq) provides a reliable method for high-quality measurement of RNA expression levels. RNA is isolated from a biological sample, converted to complementary DNA (cDNA) and provided as input to a sequencing machine. The output of an RNA-seq experiment is millions of short nucleotide sequences originating from the RNA transcripts present in the sample. In contrast to older techniques for measuring RNA, such as probe-based microarrays, RNA-seq requires no prior knowledge of existing sequences in order to measure a sample and is effective over a much greater range of expression levels.

### Library preparation

* PolyA capture
* Ribosomal depletion

The first step in preparing a sample for RNA-seq is to chemically lyse the cells, disrupting the structure of the cell wall and releasing the molecules inside. RNA molecules can then be isolated, typically using a chemical process called phenol/chloroform extraction although this can also be done by physically separating different types of molecules by passing the sample through a silica column. The majority of the RNA in a cell is ribosomal RNA, usually more than 80 percent {raz2011}. Most of the time this type of RNA is not of interest and sequencing it would reduce the ability to detect less abundant species. To select mRNA oligonucleotide probes that bind to the poly-A tail can be used but a downside of this approach is that it won’t capture immature mRNA or other types of RNA molecules. An alternative method is to use a different kind of probe specific to each species that binds to the rRNA allowing it to be removed. The choice of selection method has been shown to introduce different biases into the resulting data {sultan2014}.

The Illumina sequencing typically used for RNA-seq experiments can only read short sequences of nucleotide of approximately 40-400 basepairs. Most mRNA molecules are longer than this so to read the full length they must be fragmented into smaller parts. Most sequencing machines also only work with DNA not directly with RNA so the sample must first be reverse-transcribed using a retroviral enzyme to produce a single strand of cDNA. Many protocols have been designed for this step with each requiring a specific primer sequence to be joined to the RNA molecules {roberts2011?}. The complementary strand of cDNA is produced using a second enzyme that is usually involved in copying DNA for cell division. For some protocols fragmentation is performed after conversion to cDNA rather than at the RNA stage.

Once the cDNA has been produced it is usually necessary to attach adaptor sequences that are used to bind the molecules and initiate sequencing. They may also contain barcodes for measuring multiple samples at once. It has become standard practice to perform paired-end sequencing where a section of sequence is read from one end of a molecule before it is flipped and the other end read and this process requires an additional set of adaptors. At each of the stages of library preparation there are quality control steps to be performed to make sure a high-quality cDNA sample is loaded on to the sequencing machine.

### High-throughput sequencing

* Illumina sequencing
  + Sequence by synthesis
  + Paired end

Most RNA-seq experiments are sequenced on an Illumina maching using their Sequence by Synthesis technology. In this process the strands of cDNA fragments are separated and the adaptors bind to oligonucleotides coating a flow cell. The other end of the fragment can bind to a second oligonucleotide forming a structure where an enzyme synthesises a complementary DNA strand. This process of separation of strands and synthesis of new complementary strands is repeated until clusters of DNA fragments with the same sequence are formed. Once the clusters are significantly large the adaptor at one end of each fragment is cleaved leaving single stranded DNA attached to the flow cell and one end.

The sequencing process now begins. Nucleotides tagged with fluorescent markers are added and can bind to the next available position on a fragment if they are complementary. By adding all four nucleotides at once they compete for each position, reducing the chance of an incorrect match. Any unbound nucleotides are washed away before a laser excites the florescent tags and an image is taken. Each nucleotide is tagged with a different and the order of colours produced by a cluster shows the sequence of nucleotides in a fragment. For paired-end sequencing the fragments can be flipped and the sequencing process repeated at the other end. The images from the sequencing machine are processed to produce millions of short nucleotide reads that are the starting point for computational analysis.

### Analysis of RNA-seq data

* Experimental design
* Alignment
* Quantification
* Negative binomial
* Normalisation
* Differential expression testing
* Proportions

Many types of analyses can be performed using RNA-seq data such as identification of variants in the genetic sequence or detection of previously unannotated transcripts but the most common kind of analysis is to look for differences in the expression level of genes between groups. To do this reads are first aligned to a reference and the number of reads overlapping each genes is counted. In contrast to aligners designed for DNA sequencing RNA-seq aligners such as STAR, HiSAT2 and subread must taken into account the splicing of mRNA transcripts which causes parts of some reads to align in different locations. The alignment step is computational intensive and can take a significant amount of time. More recently tools such as kallisto and Salmon have been developed which attempt to directly quantify expression by estimating the probability that a read comes from a particular annotated transcript. These approaches are orders of magnitude faster than true alignment and potentially produce more accurate quantification at the cost of having an exact genomic position for each read.

At this stage the result is a matrix of counts known as an expression matrix where the rows are features (usually genes), the columns are samples and the values show the expression level of a particular feature in a sample. As these counts result from a sampling process that can be modeled using common statistical distributions. One option is the Poisson distribution, however this assumes that the mean and variance of each feature is equal. A better fit is the negative binomial (or Gamma-Poisson) distribution which includes an over-dispersion parameter allowing the variance to be larger than the mean. While each feature is quantified for each sample these values are not absolute measures of expression and are better understood as proportions of the total number of reads. Another complication of RNA-seq data is that the number of features (tens of thousands) is much larger than the number of samples (usually only a few per group).

The most successful methods for testing differential expression between groups of RNA-seq samples overcome this challenge by sharing information between genes. Both the edgeR and DESeq (and later the DESeq2) packages model RNA-seq data using the negative binomial distribution while Before expression levels can be tested the differences between samples must be removed through normalisation. The edgeR packaged uses the Trimmed-Mean of M values (TMM) method where… DESeq has a similar method that… When an experiment has been conducted multiple batches and there are significant differences between them alternative normalisations such as Removel of Unwanted Variation (RUV) that … may be required. The limma package uses an alternative approach where a method called voom transforms the data so that it is suitable for linear modelling methods originally designed for RNA microarray technology. Over the time the methods in these packages have been refined and new tests developed allowing for the routine analysis of RNA-seq experiments.

## 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
* Ribosomal RNA

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[Grun2014-zn]. 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[Liu2016-wq]. 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…

An alternative continuous approach is the cell velocity method in the velocyto package. RNA-seq studies typically focus on the expression of complete mature mRNA molecules but a sample will also contain immature mRNA that are yet to be spliced. Examining these reads assigned to introns can indicate newly transcribed mRNA molecules and therefore which genes are currently active. Instead of assigning cells to discrete groups or along a continuous path velocyto uses reads from unspliced regions to place them in a space and create a vector indicating the direction in which the transcriptional profile is heading. This vector can show the a cell is differentiation in a particular way or that a specific transcriptional program has been activated.

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
* Logistic regression

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

* Variant detection
* Cancer
* Immune cells

Some uses of scRNA-seq data fall outside the most common workflow and methods have been developed for a range of other purposes. For example methods have been designed for assigning haplotypes to cells, detecting allele-specific expression, identifying alternative splicing or calling single nucleotide or complex genomic variants. Other methods have been designed for specific cell types or tissues such as XXX which can assign immune cell receptors and XXX which interrogate the development of cancer samples. Most future studies can be expected to continue to follow common practice but it also expected that researchers will continue to push the boundaries of what it is possible to study using scRNA-seq technologies.

## Kidney development

### Structure and function

* Kidney structure
* Nephron structure
* Important cell types

In mammals the kidney is an organ responsible for filtering the blood in order to remove waste products. Kidneys grow as a pair with each being around the size of an adult fist and weighing about 150 g. with each being functional. Blood flows into the kidney via the renal artery and the blood vessels form a tree-like branching with ever smaller capillaries. At the end of these branches are nephrons, the functional filtration unit of the kidney. Humans can have around 1 million nephrons that are formed during development and just after birth, however they cannot be regenerated after around … of age. A capillary loop is formed inside a structure at the end of the nephron called a glomerulous and surrounded by Bowman’s capsule. Here specialised cells called podocytes create a structure called the slit diaphragm that allows water, metal ions and small molecules to be filtered while keeping blood cells and larger species such as proteins trapped within the bloodstream. The rest of the nephron is divided into segments that are responsible for balancing the concentration of these species in the filtrate. The lumen of the nephron is surrounded by capillaries which allows content to be transferred between the filtrate and blood as required. The first segment of the nephron is the proximal tubule. Here common biomolecules such as glucose, amino acids and bicarbonate are reabsorbed into the bloodstream, as is most of the water. Other molecules including urea and ammonium ions are secreted from the blood into the filtrate at his stage. This proximal tubule is followed by the Loop of Henle and the distal tubule where ions are reabsorbed including potassium, chlorine, magnesium and calcium. The final segment is the collecting duct that balances salt concentrations by exchanging sodium in the filtrate for potassium in the bloodstream using a process controlled by the hormone aldosterone. The remaining filtrate is then passed to the ureter where it is carried to the bladder and collected as urine while the blood leaves via the renal vein. In order to preform this complex series of reabsorption and and secretion each segment of the nephron is made up specialised cell types with there own set of signaling and transporter proteins. The filtration process is repeated about 12 times every hour with around 200 litres of blood being filtered every day. Aside from removing waste and maintaining the balance of species in the bloodstream the kidneys also play a role in the activation of vitamin D and synthesises the hormones erythropoietin, which stimulates red blood cell production, and renin which is part of the pathway that controls fluid volume and the constriction of arteries to regulate blood pressure.

Chronic kidney disease is a major health problem in Australia with XXX percent of the population to experience it during their lifetime. Early stages of the disease can be managed but once it becomes severe the only treatment options are dialysis, which is expensive, time consuming and unpleasant, or a kidney transplant. There are also a range of developmental kidney disorders that have limited treatment options and can profoundly affect quality of life. Understanding how the kidney grows and develops is key to developing new treatments that may improve kidney function or repair damage.

### Stages of development

* Lineage
* Important genes

The kidney develops from a region of the early embryo called the intermediate mesoderm and occurs in three phases with a specific spatial and temporal order. The first phase results in the pronephros which consists of 6-10 pairs of tubules that forms the mature kidney in most primitive vertebrates such as hagfish. By about the fourth week of human embryonic development this structure dies off and is replaced by the mesonephros which is the form of kidney present in most fish and amphibians. The mesonephros is functional during weeks 4-8 of human embryonic development before degenerating although parts of it’s duct system go on for form part of the male reproductive system. The final phase of human kidney development results in the metanephros which begins developing at around five weeks to become the permanent and functional kidney. Individual nephrons grow in a similar series of stages. Cells from the duct that will become the ureter begin to invade the surrounding metanephric mesenchyme forming a ureteric bud. Interactions between these cell types, including WNT signaling, cause mesenchymal cells to condense around the ureteric bud forming a stem cell population known as the cap mesenchyme that expresses genes such as Six2 and Cited1. Cells from the cap mesenchyme first form a renal vesicle, a primitive structure with a lumen, which extends to form an s-shaped body. By this stage the lumen has joined with the ureteric bud to form a continuous tubule. The s-shaped body continues to with podocytes beginning to develop and form a glomerulus at one end and other specialised cells arising along the length of the tubule to form the various nephron segments. Several signaling pathways and cell-cell interactions are involved in this process including Notch signaling.

Most of our understanding of kidney development comes from studies using mouse models and other model species. While these have greatly added to our knowledge they do not completely replicate human kidney development and there are known to be significant differences in the developmental timeline, signaling pathways and gene expression between species. To better understand human kidney development we need models that reproduce the human version of this process.

### Growing kidney organoids

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

One alternative model of human kidney development is to grow miniature organs if a lab. Known as organoids these tissues are grown from stem cells provided with the right sequence of conditions and growth factors. Naturally occurring embryonic stem cells can be used but a more feasible approach is to reprogram mature cell types (typically fibroblasts from skin samples) using a method discovered by …. Under this protocol cells are supplied with … followed by …. The resulting cells have the ability to differentiate into any cell type and are known as induced pluripotent stem cells (iPSCs). By culturing iPSCs under the right conditions the course of differentiation can be directed and protocols for growing eye, brain and …tissues have been developed. The first protocol first protocol for growing kidney organoids was published in 2015 by Takasato et al.

Using this protocol iPSCs are first grown on a plate where Wnt signaling is induced by the presence of CHIR, an inhibitor of glycogen kinase synthase 3. After several days of growth the growth factor FGF9 is added which is required to form the intermediate mesoderm. Following several more days of growth the cells are removed from the plate and formed into three dimensional pellets. A short pulse of CHIR is added to again induce Wnt signaling and the pellets continue to be cultured in the presence of FGF9. Growth factors are removed after about five days of 3D culture and the organoids continue to grow for a further two weeks at which point tubular structures have formed. These kidney organoids have been extensively characterised using both immunofluorescence imaging and transcriptional profiling by RNA-seq. Imaging showed that the tubules are segmented and express markers of podocytes, proximal tubule, distal tubule and collecting duct, however individual tubules are not connected in the same way they would be in a real kidney. By comparing RNA-seq profiles with those from a range of developing tissues the organoids from this protocol were found to be most similar to trimester one and two fetal kidney. While the bulk transcriptional profiles may be similar this analysis does not confirm that individual cells types the same lab-grown kidney organoids and the true developing kidney. Further studies using this protocol have shown that it is reproducible with organoids grown at the same time being having very similar transcriptional profiles however organoids from different batches can be significantly different, potentially due to differences in the rate at which they develop.

While they are not a perfect model of a developing human kidney organoids have several advantages over other models. In particular they have great potential for uses in the modeling of developmental kidney diseases. Cells from a patient with a particular mutation can be reprogrammed and used to grow organoids that can then be used for functional studies or drug screening. Alternatively gene editing techniques can be used to insert the mutation into an existing cell line or correct the mutation in the patient line allowing comparisons on the same genetic background. Modified versions of the protocol that can produce much larger numbers of organoids, for example by growing them in swirler cultures, could potentially be used to produce cells in sufficient numbers for cellular therapies. Extensive work is been done to improve the protocol in other ways as well such as improving the maturation of the organoids or directing them more towards particular segments. Overall kidney organoids open up many possibilities for studies to better help use understand kidney development and potentially help develop new treatments for kidney disease.

# The scRNA-seq tools landscape

## Introduction

When I began my PhD in early 2016 single-cell RNA-sequencing technologies were just beginning to become widely available. Since then there has been a rapid uptake and there are now many studies using this approach. Along with the growth in the adoption of scRNA-seq technologies there has been an explosion in the number of software tools for analysing these datasets. This chapter charts the growth in the scRNA-seq analysis landscape over time.

In 2016 there were relatively few analysis methods available and to answer questions like how many tools perform a particular task or which areas were developers focusing on or was there a tool for doing this I began to record details about them. Inspired by similar projects such as Sean Davis’ Awesome Single Cell page I decided to make this collection public. This turned out to be useful to other researchers and over time a simple spreadsheet became the scRNA-tools database and website (<https://scRNA-tools.org>). A paper published in *PLoS Computational Biology* describing this resource forms the main part of this chapter.

By having access to details about existing analysis tools we were able to explore how the field has developed. We found that computational researchers had focused their efforts on analysis tasks specific to scRNA-seq data such as clustering and ordering of cells or handling the larger numbers of zeros. We also saw that many of the tools performed tasks common to several stages of analysis including dimensionality reduction of various kinds and visualisation. Developers of scRNA-seq analysis tools tend to embrace a open-source and open-science approach. Most tools are developed on GitHub were others can ask questions and submit improvements. The majority are also available under open-source licenses allowing their code to be reused for other purposes, although there is also a significant proportion that do not have any associated license. Tools are commonly made public by releasing a preprint publication, making them available to the community much more quickly and giving early adopters a chance to contribute to their development.

A section at the end of this chapter presents an updated version of some of this analysis based on the most recent version of the database.

# Simulating scRNA-seq data

## Introduction

To be accepted and used any computational method for data analysis needs to demonstrate that it is effective at the task it aims to complete. Ideally this can be done by evaluating performance on a real dataset where the results are already known. Unfortunately in many case such gold standard datasets are not available. This is particular true for genomic data where it is difficult to know what the truth is or it is limited to only small sections of the genome. It is possible to create some genomic datasets where the truth is known, for example through carefully performed mixing experiments, but these often do not capture the true biological complexity. In many cases the most effective way to evaluate an analysis method is by testing it on a simulated datasets. Simulations have the additional advantage of relatively cheap and easy to produce allowing exploration of a wide range of possible parameters. This is the approach taken by many early methods for scRNA-seq analysis but often the simulations they used where not well explained, code for reproducing them was not available and perhaps most importantly they didn’t show that the synthetic datasets were similar to real scRNA-seq data.

This chapter presents Splatter, a software package for simulating scRNA-seq datasets presented in a publication in *Genome Biology*. Splatter is designed to provide a consistent, easy-to-use interface for multiple scRNA-seq simulation models previously used to develop analysis tools. We do this by providing two functions for each model, one which estimates parameters from a real dataset and a second that generates a synthetic dataset using those parameters. Each model has different assumptions and reproduces different aspects of scRNA-seq data and we explain these differences in the paper. We also present Splat, our own simulation model based on the Gamma-Poisson distribution. This model includes several aspects of scRNA-seq data including highly expressed outliers genes, differences in library sizes between cells a relationship between the mean and the variance of each gene and the ability to add a dropout effect linked to gene expression. When designing the Splat simulation our goal was to reproduce scRNA-seq data as well as possible rather than test a specific method with the result being that the model is highly flexible and able to generate a range of scenarios including datasets with multiple groups of cells, batch effects and continuous trajectories.

In the paper we compare how well each simulations reproduces a range of scRNA-seq datasets including UMI and full-length protocol, different capture methods and homogenous and complex tissues. We found that the Splat simulation was a good match for some of these simulations across a range of methods, however it was also clear that some models more faithfully reproduced different aspects of the data, particularly for datasets from different sources. The Splatter R package is available for download from Bioconductor (<https://bioconductor.org/packages/splatter>).

## Splatter publication

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

# Visualising clustering across resolutions

## Introduction

Clustering of cells to form groups is a common task when analysing scRNA-seq data that is not required for bulk RNA-seq experiments and one that has received a lot of attention from analysis methods developers. The need to group samples is not unique to genomic data and clustering techniques are used in many other fields for a wide variety of purposes. Whatever kind of data you are interested in and whatever clustering method is being used a question that commonly comes up is how many clusters do we want to have? This can be controlled by setting an exact value, changing a parameter that indirectly controls the clustering resolution or affected by the values of other parameters and the number of clusters that are used can often have a profound affect on how the results are interpreted. Existing measures of clustering typically only consider a single clustering resolution at a time or require multiple rounds or permutations and clustering which can be infeasible for large datasets. In this chapter I propose an alternative visualisation-based aid for deciding which clustering resolution to use.

Clusterings of the same dataset at different resolutions are often related and it is common for new clusters formed at higher resolutions to be formed by splitting existing clusters. However when comparing clusterings it is not always clear what those relationships are and how significant they might be. The method I describe here was published in *GigaScience* and suggests clustering datasets at multiple resolutions then considering the overlap in samples at neighbouring resolutions. By doing this we can build a graph structure we call a “clustering tree”. Visualising this tree allows us to see where new clusters form, how they are related and the stability of particular clustering resolutions. In the paper we demonstrate this approach using simulated datasets, a simple dataset commonly used as an example for machine learning techniques and a complex scRNA-seq dataset from blood.

While the structure of clustering trees can help choose a clustering resolutions to use they are more generally a compact, information dense visualisation that can show information across clustering resolutions. This is something that is not possible with traditional visualisations used for clustering results such as t-SNE projections and is achieved by trading individual information about each sample for summarised information about clusters and adding a resolution dimension. Overlaying important domain knowledge (such as the expression of known marker genes) onto these visualisations can be particularly informative and we also demonstrate this in our paper.

Clustering trees can be produced using the clustree R package which is built on the tidygraph and ggraph packages and is available from CRAN (<https://cran.r-project.org/package=clustree>).

# Analysis of kidney organoid scRNA-seq data

# Conclusion

# References