Expression of the Retinal Determination Network in multiple cell types

Matthew Gordon (mcg33@student.le.ac.uk)

The University of Leicester

Friday, 27th November 2020

Wordcount: 9,279 (excl. References)

Abstract

The Retinal Determination Network (RDN) is a small group of transcription factors originally identified for its function in the eye specification of Drosophila melanogaster. Recent studies have shown it may show expression in more tissues than previously assumed. It may also be essential for regulating proliferation, differentiation and autocrine signaling as well as interacting with other important pathways. Upregulation of RDN members DACH, EYA and SIX have been shown to contribute to tumour initiation and progression in breast cancers. Single-cell RNA-sequencing examines levels of gene expression at a cellular level looking at differences from cell to cell. It provides a higher resolution of cellular differences than traditional bulk RNA-sequencing, this gives a clearer insight into cellular behaviour. The aim of this project is to better understand the extent of the expression of RDN members in humans and mice, which may lead us to suggest a new, more ubiquitous, function for the network. During this project we will re-analyse mouse and human datasets looking for expression of RDN members in a variety of different cell types from both humans and mice. If found, expression would confirm the theory that RDN members are more common than suggested in the current published literature. Our findings show RDN expression in human & mouse retina, human embryonic glutamatergic neurons and mouse pancreatic tissue suggesting a wider function to the network. Some errors in cell filtering and potentially inconsistent syntax in FASTQ files meant human disease datasets including melanoma and two IBD datasets were unable to be analysed. Despite this we believe our findings are only the tip of the iceberg in terms of RDN interactions, with a greater number of pathway interactions present than previously thought. The full extent of the RDN is yet to be determined.

Introduction

The Retinal Determination Network

The Retinal Determination Network (RDN) is a small regulatory circuit composed of PAX6, EYA, SIX and DACH genes traditionally associated with the eye (Kong *et al.,* 2016). The RDN was originally identified for it's function in *Drosophila melanogaster* (fruit fly) eye specification (Kumar *et al.,* 2010). In *Drosophila melanogaster* it was found to be essential for the regulation of differentiation, proliferation and autocrine signalling as well as displaying multiple interactions with other key signalling pathways (Kumar *et al.,* 2010).

Whether the RDN operates in different developmental contexts remains a controversial topic. The expression of the RDN outside the eyes specification would imply that this network is required for a variety of more general developmental processes (Silver *et al.,* 2005). For example studies of *Drosophila melanogaster* gonads highlight how multiple members of the RDN play a crucial role in tissue development. It is suspected that there may be levels of RDN expression in many other tissues both mature and juvenile, however there has been little published work to confirm this hypothesis. Therefore the true scale of the impact of RDN members is largely undocumented and unknown. Furthermore, a recent paper from Kong and collaborators (*Kong et al.,* 2016) indicates that the RDN could potentially be the target of new therapeutic methods in breast cancer treatment. The authors of the paper suggested in mammalian cells that RDN members have an essential role in governing key stages of cancer development including Epithelial to Mesenchymal transition (EMT) and tumour metastasis. They suggested how in mammalian breast cancer tissues dysregulation of RDN members DACH & SIX/EYA determine cancer initiation and progression. The functional balance of DACH to SIX/EYA has been shown to maintain homeostasis in luminal cells maintaining the luminal structure. Misregulation in the form of functional loss of DACH or up-regulation of SIX/EYA drives hyper-proliferation and progression to ductal carcinoma *in situ,* with some of the most malignant of these cells undergoing EMT to acquire stem cell properties and enter blood vessels by invading through the basal

membrane. This ultimately ended in distant metastatic cancers in other tissues. This is one of very few papers linking the RDN to cancer initiation and progression so it opens the topic up to expansion by subsequent follow-up papers.

In this work I will investigate the expression of the RDN in different tissues using single-cell RNA-seq. In the following section I will introduce the single-cell RNA-sequencing's potential and pitfalls.

Single-cell RNA-sequencing

Single-Cell RNA-Sequencing (scRNA-seq) technologies allow the dissection of gene expression through measuring cellular RNA levels at single-cell resolution, this has revolutionised the field of transcriptomics (Chen *et al.,* 2019). scRNA-seq offers a better understanding of how the transcription states of neighbouring cells differ when compared to traditional bulk RNA sequencing technologies. Bulk RNA sequencing measures the gene expression levels over a large population of cells. Therefore scRNA-seq allows for more precise analysis of how each gene alters activity in each cell as well as highlighting the differences in expression across cells from the same tissue. This is especially useful in complex tissues where gene expression varies widely from cell to cell.

The initial step in scRNA-Seq is cell dissociation, this is essential to release intact individual cells from their host tissue without damaging them. The next stage is single cell isolation, the dissociated cells need to be suspended individually in solution so that they can be sequenced individually. Before sequencing the cells in solution one needs to determine which mRNA transcript came from which cell, this was a sticking point recently overcome for this technology. There are two main approaches to this, droplet based approaches and non-droplet methods. The most common two droplet-based approaches are Drop-Seq and Chromium 10X. Drop-Seq and Chromium 10X both use the principles of microfluidics, individual cells in suspension flow into a droplet containing the microparticle bead suspended in a lysis reagent at a ratio of one cell per microparticle (Macosko *et al,.* 2015). These are then pushed through oil encapsulating them in their droplet containing the cell and the microparticle suspended in lysis

reagent. Each Microparticle consists of a unique oligonucleotide bound to a primer bead, this allows identification of each cell like a barcode. Each oligonucleotide is composed of four components; PCR primer (Required for DNA amplification via PCR), Cell Barcode (Uniquely identifies cell), UMI (Unique to each oligo arm extending from the bead. For labelling each molecule) & a poly-T tail (to capture the poly-A tail of mRNA transcripts). Once individual cells are isolated with the microparticle, the lysis reagent perforates the cells releasing their contents into solution. The polyT tails on the oligomers bind to the polyA tails of the mRNA transcripts capturing them from solution. Captured transcripts form STAMPs (Single-cell transcriptomes attached to microparticles) in the droplet. The droplets are broken releasing the STAMPs into solution with each other. Reverse transcriptase enzymes in solution bind using the STAMPs as templates and convert the newly bound mRNA to cDNA strands, Second-strand synthesis completes the cDNA ready for amplification by PCR. Primer regions present as part of the oligonucleotide from the microbead bind to PCR primers in solution to initiate PCR amplifying the double-stranded cDNA. RNA is very unstable and therefore easily degraded, this makes it unsuitable to be stored for any length of time as part of a library, so it's converted into the more stable DNA for use in sequencing libraries. This stage either happens before or after PCR depending on the technology used, but the outcome is the same. Once reverse transcription and PCR are completely amplified cDNA is used to construct the sequencing library with thousands of single-cell transcriptomes, this in turn is used to create the single-cell expression profiles for each cell. Sequencing analysis maps each mRNA to it's cell and gene of origin so that each cells pool of mRNA is ready for downstream analysis (Macosko *et al.,* 2015 | Zhang *et al,.* 2019).

When comparing the two technologies, Drop-Seq offers the ability to run a far greater number of cells per run compared to Chromium 10X (150,000 cells/run vs 1,700 cells/run). This makes the cost per cell to prepare the sequencing libraries much less than that of Chromium 10X. It's therefore seen as the more customizable and cost-effective method for single-cell sequencing by many. However advantages for using Chromium 10X over Drop-seq are it has a greater capture efficiency (Percentage of cells encapsulated in a lysis buffer bubble with Microparticle) than Drop-Seq, in the

region of 65% vs. Drop-Seq's 5%. It also has a greater transcript capture rate (the percentage of each cell's transcripts caught by the polyT tails of the microbead) than Drop-seq at ~14% vs ~10.7%. Non-droplet based approaches (like SCI-Seq (Vitak *et al.,* 2017)) can achieve much higher cell starting numbers than both Drop-Seq and Chromium 10X. In SCI-Seq cells are fixed using alcohol and dispensed into wells containing a specific number of cells. As in droplet methods, the mRNA of the cells needs to be barcoded for identification and this is done via reverse transcription. Each well has a unique oligonucleotide-barcode (two rounds of FACS (fluorescence activated cell sorting) covers this). As two rounds of cell sorting are required for SCI-seq, compared to one round in droplet based methods, greater stress is imposed on the cells. This could potentially impact gene expression and the downstream analysis.

Raw results from scRNA-Seq are directly received as FASTQ files. These are mapped against a reference genome using a variety of software such as STAR Aligner (Dobin *et al,.* 2012) or Kallisto Bustools (Melsted *et al.,* 2019). After mapping against the reference genome quality control confirms the quality of the mapping. Programs like FASTQC act as tools for this process. The main functions of FASTQC are to provide a quick overview to highlight which areas of the mapping may have issues, plot graphs and tables to quickly access the data and to export results to an HTML based report. Factors such as the percentage of the genome mapped, the read depth at each position & the number of reads which are mapped to tRNA/rRNA (tRNA & rRNA not sequenced when using Chromium 10X) are all taken into account during this quality control phase.

After normalisation and quality control steps are complete the process moves onto the downstream analysis, this is essential in order to extract the relevant data from the files in order to answer the hypothesis posed. Cells are first grouped together into clusters based on their gene expression profiles. The top differentially expressed genes are the basis that determines which cell is in what cluster. Relative expressions of highly variable genes are measured to draw similarities between certain cells and differences between others. There are multiple algorithms to do this including near component analysis (NCA) and principal component analysis (PCA), both take different approaches to achieve the same aim of differentiating cells. NCA prioritises the accuracy of each

cluster to avoid cluster overlapping at the cost of a loss of cellular variance, however PCA prioritises minimising the loss of cell variance at the cost of cluster accuracy. Both methods are commonly used in papers. Combining either NCA or PCA with a t-SNE (t-distributed stochastic neighbour embedding) algorithm allows for 2D or 3D visualisation of the cell clusters. t-SNE as a method for simply viewing highly dimensional data in a 2D and 3D way to express differences in gene expression of cells was developed by Maatens et al. in 2008 (*van der Maaten et al.*, 2008). These cell clusters are defined by and often annotated with specific gene markers whose differential expression is unique to each cluster. Definition of a marker gene for a cluster is the proportion of cells highly expressing that specific gene when compared to the average expression of that gene across the whole dataset. Annotating these clusters provides information on the potential cell identity of all members of the cluster. Compositional analysis is used to determine what proportion of the cells analysed are within which cluster. The proportions of cells within each cluster can change in response to many factors, therefore it's important that this is measured. The final stage of the downstream analysis is trajectory analysis which uses the data gathered to model the future gene expression states of cells. Accurately predicting future gene expression states is essential to determine the rate of many factors such as differentiation, maturation or in the case of cancer; development and/or progression. Further details about the exact methods used during our analysis can be found within the cell clustering section of our methods.

As with all cutting-edge technologies, single-cell RNA-sequencing has limitations. When compared to bulk RNA-sequencing single-cell has high levels of background noise, this can make the validity of readings unclear. The cell capture efficiency (the percentage that single-cells are suspended in droplets with uniquely barcoded beads) rate varies depending on the technology used as mentioned earlier. With a cell capture rate for Chromium 10X of ~65% and ~5% for Drop seq (AlJanahi *et al.,* 2018) many have rightly drawn attention to the fast that with so many cells not being sequenced per sample it is hard to draw definitive conclusions about specific datasets off one run. In order to cover the lost cells multiple runs have to take place, this is an issue for many cell types as the longer they're out of their optimal conditions while in a single cell suspension the more

stress the cells are under and the more this will affect their gene expression profile. There is evidence to suggest that cell specific bias can also affect the type of cell caught in the isolation phase of the scRNA-seq process, this is discussed in more detail later on. For the cells that have been successfully isolated, the next stage is to read the mRNA transcripts present in each cell. The rate at which transcripts are caught by the polyT tails of the bead-bound oligonucleotides is called the 'transcript capture' rate. With both Chromium 10X and Dropseq having very low rates at ~14% and ~10.7% (AlJanahi *et al.,* 2018) respectively, many doubt the validity of the results due to most (85%+) transcripts avoiding capture and not being sequenced. There has also been evidence that certain transcripts have a greater chance of being caught than others leading to a transcript bias that may suggest greater relative expression of certain genes compared with others. This is discussed in more detail later on. Although Dropseq has lower cell and transcript capture rates it makes up for this as it can process ~150,000 cells per run, compared to Chromium 10X's ~1,700 cells per run, this makes it still a viable option when analysing certain tissues. The high rates of both cells and transcripts lost for both methods raises an interesting dilemma regarding the results and predictions drawn from any results discovered using these technologies over traditional bulk methods.

Microfluidics are an essential part of the scRNA-seq process to isolate single cells, however this can cause shear stress on cells. This impacts mRNA readings of cells, especially adherent cells as they're more delicate when in suspension. The length of time cells are out of their optimal condition also increases stress on cells. Results of this stress can present themselves in the data in the forms of unusually high proportions of mitochondrial DNA or unexpectedly low levels of all expression as cellular contents could have leaked from the cytoplasm to the extracellular environment so were missed during the PCR phase.

Cell specific bias is a normalisation problem found in both bulk and single-cell RNA sequencing. For bulk RNA sequencing multiple runs of similar biological material are compared to the sample, however this isn't possible in single-cell RNA sequencing as each run contains multiple different cell types. Therefore alteration of the normalisation process is required to incorporate cell to cell variation. Cell specific bias also refers to

the rate of mRNA capture being inconsistent with some cells having a greater proportion caught than others, this is the main cause of sparsity in data and is known as a dropout event (AlJanahi *et al.,* 2018).

Kallisto-Bustools

Kallisto is a k-mer based method to estimate isoform abundance from RNA sequencing data (Bray *et al;* 2016). Previous methods quantify gene expression by mapping onto a reference genome with reads assigned to a position in the genome, then gene expression values are derived by counting the number of overlapping reads. Kallisto relies on pseudoalignment, this doesn't identify the positions of the reads in the transcripts but only their potential transcripts of origin. This avoids having to do an alignment of each read to a reference genome, it uses the transcriptome sequences rather than the whole genome. This reduces the time required to compute accurate estimates when compared with previous methods. Kallisto can analyse up to 30 million unaligned pair-end RNA-seq reads in less than 5 minutes on a standard laptop (Melstead *et al.,* 2019).

Bustools is a program that allows the user to manipulate BUS files from single-cell RNA sequencing datasets. It's functions include collapsing UMIs, error correcting barcodes, producing gene counts or transcript compatibility matrices among many tasks. BUS files are generated by Kallisto from raw single-cell sequencing data in FASTQ format (Melstead *et al.,* 2019).

Before analysing samples a 'Kallisto index' needs to be constructed. The program first builds a coloured de Bruijn graph from all k-mers found in the transcriptome with each node of the graph corresponding to a k-mer. Each colour represents a different transcript of origin.

Visualisation features of the Kallisto-Bustools combined module allow for visualisations of single-cell RNA-sequencing data in multiple forms. As our experiment had multiple quality-control stages, plots including Library Saturation and Knee plots allow for useful visualisations to determine the quality and spread of one's data. Knee plots were first

introduced in a Dropseq paper in 2015 (Macosko *et al.,* 2015), they are a method of ordering cells by the number of UMI counts associated to them (x-axis) and the fraction of droplets with the number of cells (y-axis). They are an important part of the quality control step as they help to visualise the proportions of the dataset that will be analysed further. Library saturation plots are another plot in the quality control stage of analysis, they help to visualise the sequencing depth of the samples by plotting UMI counts (x-axis) to the number of genes detected (y-axis).

Visualising highly dimensional data is often difficult but important to understand much of the data. Kallisto-Bustools has built in features allowing the user to make t-distributed stochastic neighbor embedding (t-SNE) plots in combination with Principal Component analysis (PCA) or Neighborhood Component analysis (NCA) clustering to give informative graphs of gene expression at a single cell level.

Aims of the project

The aim of this project is to use single-cell RNA-sequencing data to evaluate RDN expression in multiple mouse and human tissues and organs (specifically retina, embryonic glutamatergic neurons and pancreas). With the hope that finding expression will confirm our hypothesis that the RDN is an essential part of a wider variety of general developmental processes than previously thought.

METHODS

To investigate the expression of the RDN in different tissue we analysed human and mouse retina, human embryonic glutamatergic and mouse pancreatic datasets in order to get a broad view of RDN expression from very functionally different tissues with early common progenitor cells in embryonic development. Single-cell RNA-sequencing analysis will be carried out using Kallisto-Bustools to cluster cells in the datasets from their raw FASTQ files to see how gene expression determines cell types and to measure RDN expression in different datasets. Initially we try to prove RDN expression in human and mouse retina datasets, this is known to be the case so is a test of the technique. Secondly we shall test human embryonic glutamatergic neurons using a

dataset from a 2018 paper (*La Manno et al.,* 2018). Once complete we shall look at a non-neuronic mouse pancreatic dataset from a 2020 paper (*Bergen et al.,* 2020). All the analysis were performed using the University of Leicester high-performance computers (HPCs) ALICE/SPECTRE, Google Collab and local machines available in the Feuda lab. The data was visualized using Jupyter Notebooks (*Kluyver et al.,* 2016), this is an open-source web application that can be run through the terminal command line and provides a web-based interface for viewing and editing live code as well as visualising plots.

Downloading FASTQ files

FASTQ files for analysis were obtained through the 'Data Availability' section at the end of the respective research papers. The raw data files are available in the National Center for Biotechnology Information's Gene Expression Omnibus repository under unique GEO accession numbers (e.g. GSE95459) which when entered into the database (we used: EMBL-EBI (https://www.ebi.ac.uk) database) this gave links to the relevant files stored in compressed FASTQ files (.fastq.gz). The raw dataset of mouse pancreatic endocrinogenesis are deposited under the accession number **GSE132188**. The raw dataset of human embryonic glutamatergic neurogenesis are deposited under the accession code **GSE115813**. A total list of the accession numbers for the dataset used are in the 'Data Availability' section at the end of this document. Compressed FASTQ files were uploaded to our High Performance Computing (HPC) cluster directly via the Linux Mint terminal using the 'wget' command followed by the link location of the FASTQ files directly from the EML-EBI database. It's essential the compressed FASTQ files are stored and run on the HPC as they are too large and require too much memory to be run locally.

Installing Python packages

Kallisto-Bustools is a workflow for the pre-processing of scRNA-sequencing data using the Python3 programming language. The required Python 3.6.4 module is not the default of the HPC so has to be loaded before any analysis can take place, this is done through the terminal command line using the 'module load python/gcc/3.6.4' command.

To install the Kallisto-Bustools module as well as the other required python modules needed to run Kallisto-Bustools the ‘pip install’ command was used. That followed by ‘kb-python’ installed Kallisto-Bustools. The others essential for the clustering analysis are: matplotlib, scikit-learn, numpy, scipy, leidenalg & scanpy. All of which are installed the same way.

Kallisto index, Pseudoalignment and counting

A reference index relating to the species of the dataset is downloaded via Kallisto Bustools from the Caltech database. The cells we used for all our are from humans and mice so both indexes are downloaded for each dataset. The raw datasets were produced using a Chromium 10X single cell gene expression system in all but one dataset (Dropseq is the technology used for the mouse retinal dataset); this needed to be taken into account when downloading and combining the indices using Kallisto-Bustools.

Basic quality control

This is the first stage that uses python code in Jupyter Notebooks, we import the essential python3 modules using a list of import statements. Next gene names, counts and other information about the dataset are used to populate the ‘adata’ object. This ‘adata’ object principally contains gene counts. Parameters are set for the pipeline to remove inconsistent and incomplete data from our analysis which could potentially cause errors and distort the data, these consist of gene thresholds, mitochondrial percentage and cell thresholds. A maximum threshold for mitochondrial gene percentage per cell is set, if the level is over this threshold it’s filtered out of the analysis as it would suggest the cells may have burst or the sample quality is poor. Gene thresholds remove any gene with less than the set number of UMI counts per gene, with cell thresholds set to remove any cell with less than 100 genes expressed (Full list of parameters for each dataset found in commented code on our Github, link in the ‘data availability section). For cells to pass the quality control filtering they must pass the count, mitochondrial percentage and gene filtering stages to be included into the analysis.

Processing the matrix for clustering and visualisations

Set commands in Kallisto-Bustools are used to convert the counts into CPM (Counts per million) units. In each cell these counts are normalised using the total number of reads in each cell using log1p, the function log1p(x) = log(x+1). Next the Principal component analysis (PCA) algorithm is run to reduce the dimensionality of the data in order to obtain visualisations of the cells as part of a cluster of similar cells while maintaining the key variation within the dataset. This is combined with the machine learning t-distributed stochastic neighbour embedding (t-SNE) algorithm to visualise cells based off their differential expression of the same genes in order to visualise the difference in gene expression on a plot in 2D space. An alternative method for this is using a supervised learning method called neighbourhood component analysis (NCA) combined with a t-SNE algorithm to obtain more accurate cluster classifications at the cost of variance (when compared to PCA).

Finding marker genes

A key aspect of PCA/NCA with t-SNE clustering is the ability to identify differentially expressed marker genes. Expression of these genes is specific to individual clusters so is used to identify/mark them. Leiden t-tests show the gene expression score of the top 10 genes expressed in each cluster, this is visualised as a graph for each cluster (12). We also show this information as part of a table. Matplotlib allows the expression of specific genes in each cluster to be highlighted, this helps determine the expression basis in which specific clusters are unique from each other. We shall search for expression of RDN members in the NCA combined with t-SNE plot to see if there is any expression for each dataset, and if so, highlight whether it's localised to clusters or ubiquitously expressed.

Results

Cell clustering of human & mouse retinal datasets using Kallisto-Bustools

The first step of our analysis was the clustering of human and mouse retinal cells as a proof of technique to show that we can see expression of RDN members using Kallisto-Bustools in a tissue where there is substantial proof that expression occurs.

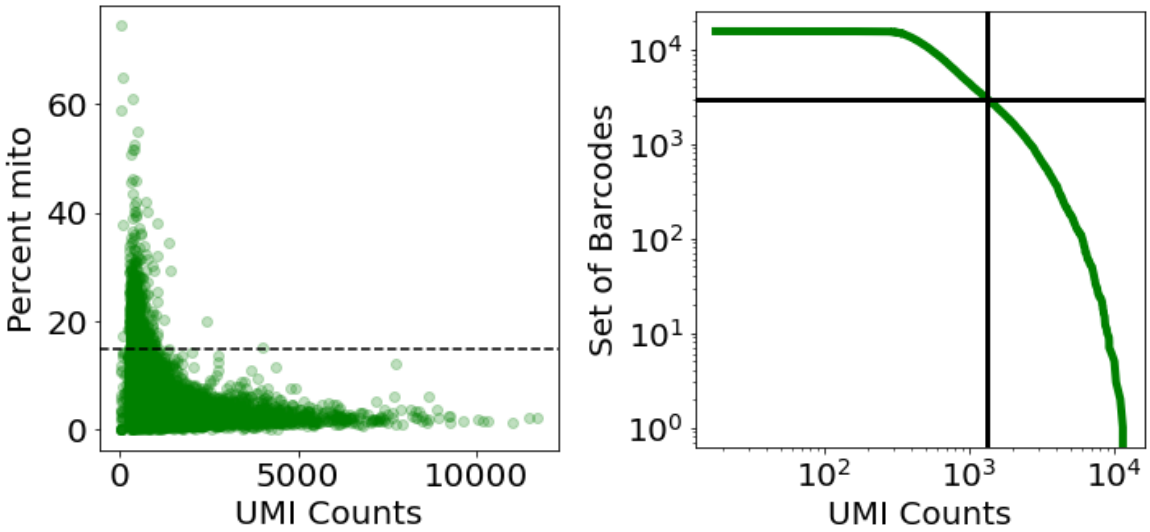

Figure 1: *Quality control steps for the mouse retinal dataset **(a)** Plot showing UMI Counts per cell with the percentage of mitochondrial genes along the y-axis. **(b)** Knee plot showing the threshold, black lines, any cells to the left of this are filtered from further analysis*

For the mouse retinal dataset there was a very high number of high quality cells, this large amount of data caused the kernel to crash and let us increase the stringency of our filtering to analyse a greater quality sample. Evidence of this increased filtering can be seen as our mitochondrial threshold was reduced to 15% among other filtering increases (Figure 1a). In figure 1b only the cells within the bottom right quadrant of the plot were analysed as they had sufficiently high UMI counts to barcode ratio.

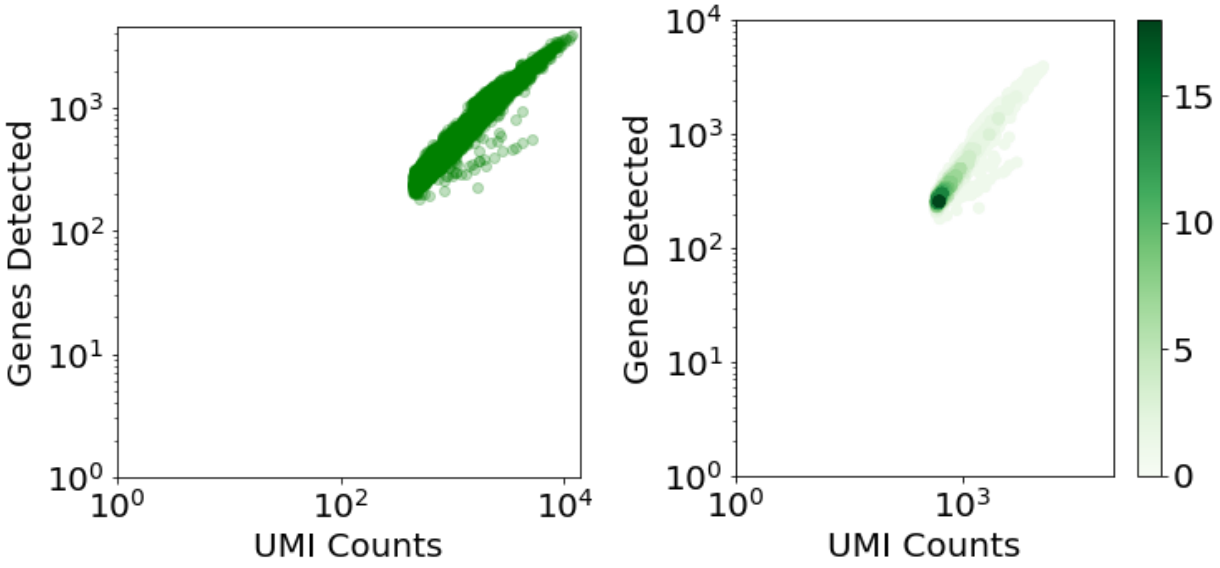

Figure 2: *Library saturation plots for the mouse retinal data.* ***(a)*** *Standard library saturation plot.* ***(b)*** *Gene density saturation plot.*

Library saturation is a good measure of the sequencing depth of a dataset. The deeper the sequencing depth the more runs needed before a new transcript is discovered, this means you can be sure you've sequenced all transcripts present after a set number of copies, to a high degree of confidence. The standard library saturation plot is a slightly misleading figure as it doesn't show how many points are overlapping each other as all are the same colour and density (Figure 2a). Therefore a density saturation plot can be used to give a better indication as to where the majority of the points are clustering (Figure 2b).

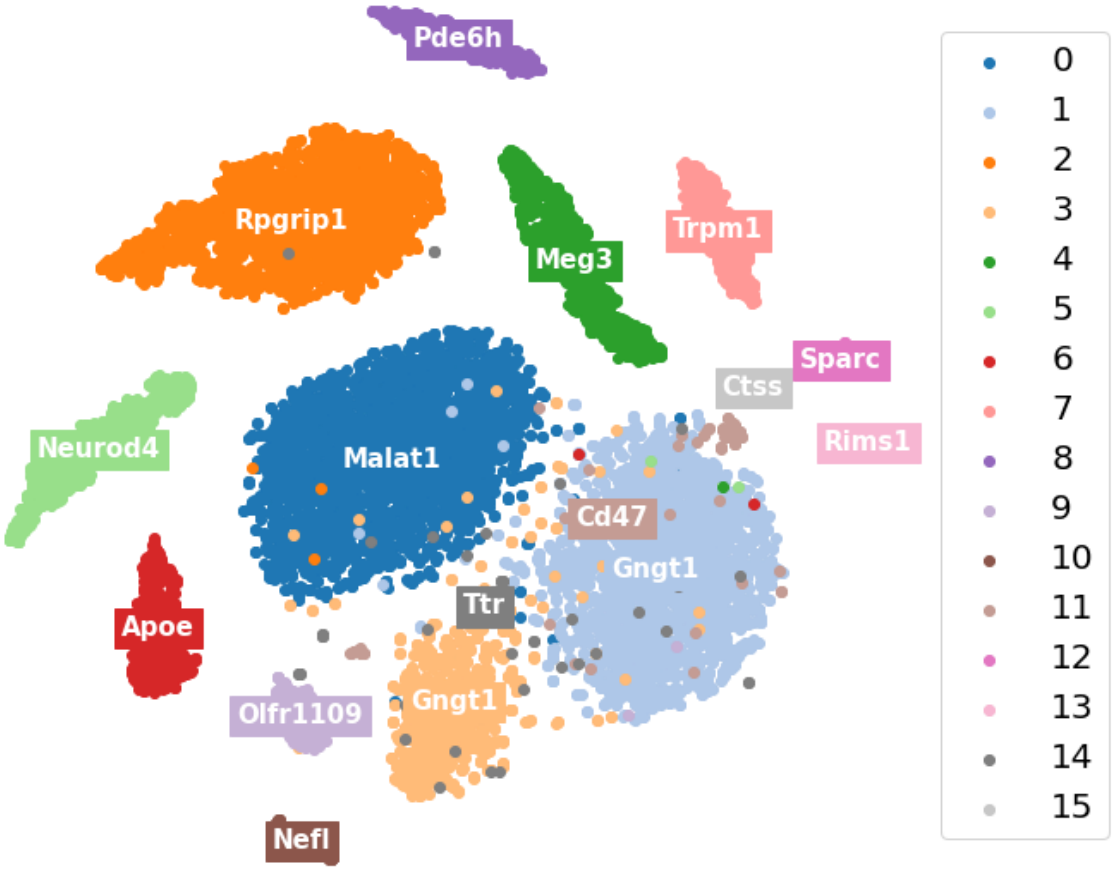

Figure 3: *NCA combined with t-SNE plot for the mouse retinal dataset, annotated with the top differentially expressed gene per cluster*

NCA analysis combined with a t-SNE plot shows that the mouse retinal dataset contains a large variety of different cell types with sufficiently different gene expression patterns to create 15 unique clusters. These clusters are composed of cells with similar gene expression patterns. As NCA clustering prioritises accuracy of classification some of the variance between cells is lost, however ir provides greater confidence that all cells in the cluster are of the same type and are therefore likely expressing similar genes at similar times. An alternative clustering method, PCA, would prioritise minimising the loss of variance in each cell providing a clearer view of the differences in expression within each cluster. However this would make specific gene expression more difficult to localise to specific clusters with a t-SNE.

The annotated marker genes are not necessarily the top expressed genes in the cluster as some genes are ubiquitously expressed in most cells, they are the top differentially expressed genes. This means they have the highest expression in each cluster relative to the average expression of that gene dataset-wide.

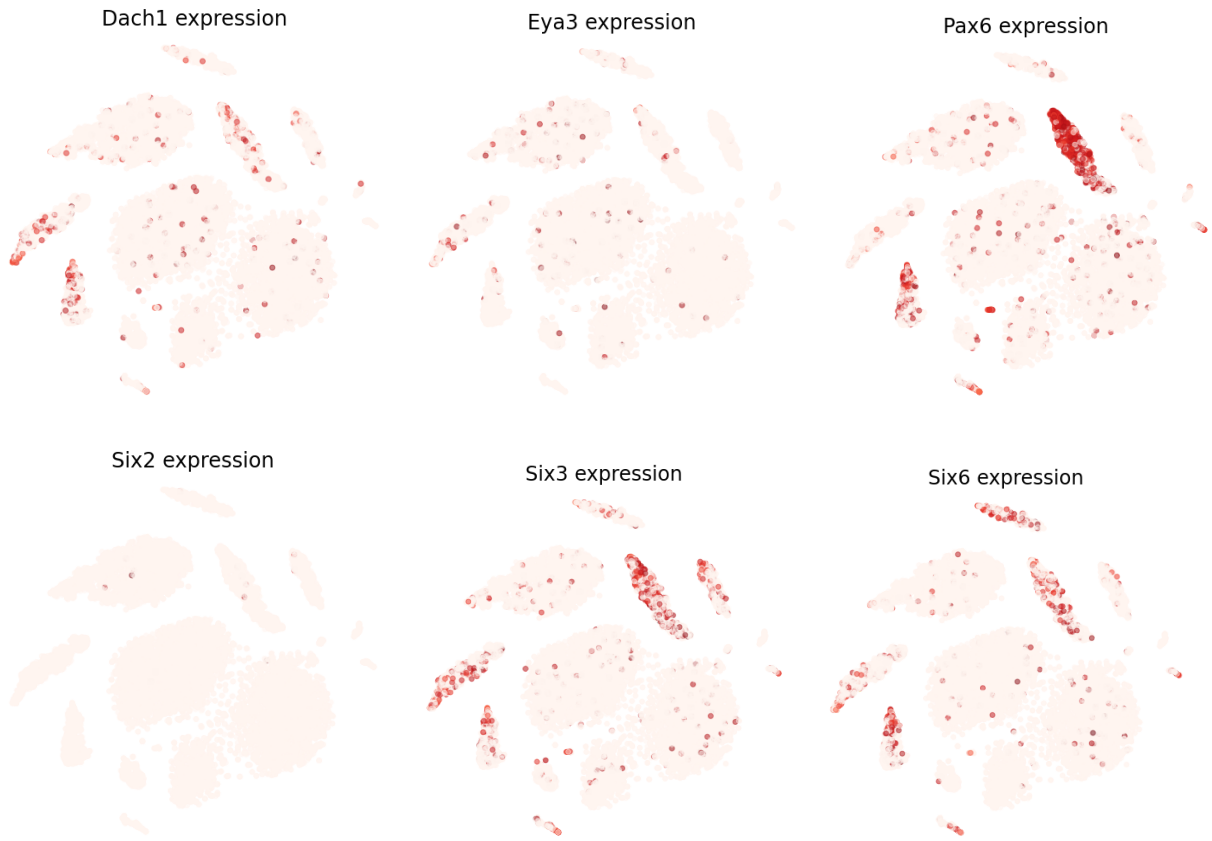

Figure 4: *NCA combined with t-SNE showing RDN member expression in the mouse retina* ***(a)*** *DACH1* ***(b)*** *EYA3* ***(c)*** *PAX6* ***(d)*** *SIX2* ***(e)*** *SIX3* ***(f)*** *SIX6*

Expression of RDN members was expected in this mouse retina dataset as it has been shown before to express high levels of member proteins in previous papers. In terms of our analysis this dataset served as a good proof of principle testing that we could use Kallisto Bustools single-cell RNA-sequencing analysis and effectively retrieve expression data from it. RDN expression is visible in the data with some clusters

expressing RDN members at greater rates than others suggesting that the RDN plays an essential role in some general functions specific to those cell types.

For the next stage of our analysis we tested a human retinal dataset also using Kallisto-Bustools to look for RDN expression. Like with the mouse retina, there is a lot of evidence to suggest that there is RDN expression in human retinas.

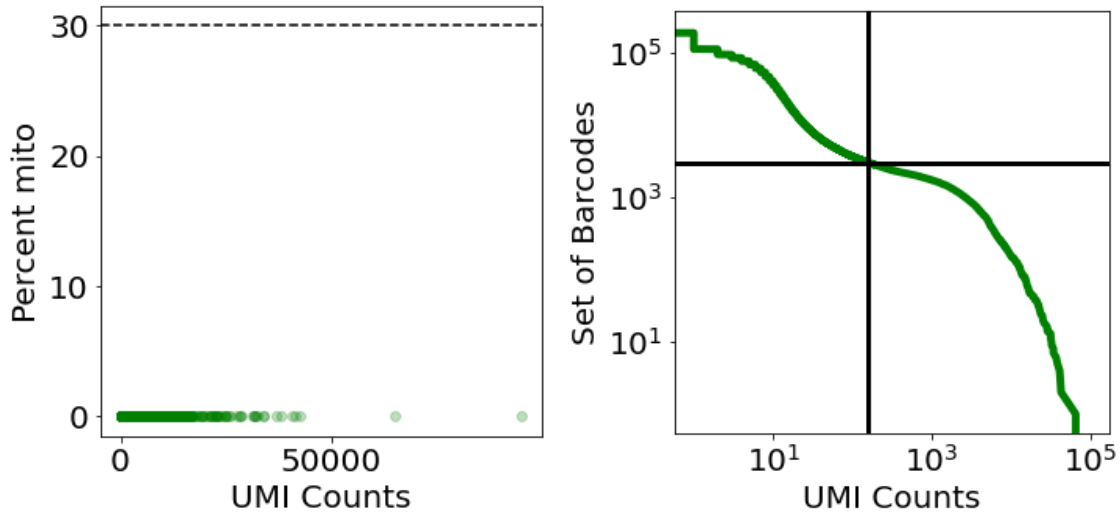

Figure 5: *Quality control steps for the human retinal dataset **(a)** Plot showing UMI Counts per cell with the percentage of mitochondrial genes along the y-axis. **(b)** Knee plot showing the threshold, black lines, any cells to the left of this are filtered from further analysis*

For the human retinal dataset there was an issue with the recognition of mitochondrial genes from our analysis, we discuss this in further detail in the 'Mitochondrial gene error for human datasets' section of our discussion. Therefore we kept the mitochondrial maximum threshold at the standard 30% as it had no effect on our filtering process (Figure 5a). In figure 1b only the cells within the bottom right quadrant of the plot were analysed as they had sufficiently high UMI counts to barcode ratio. This knee plot gave an expected shape with little interesting features to discuss about it.

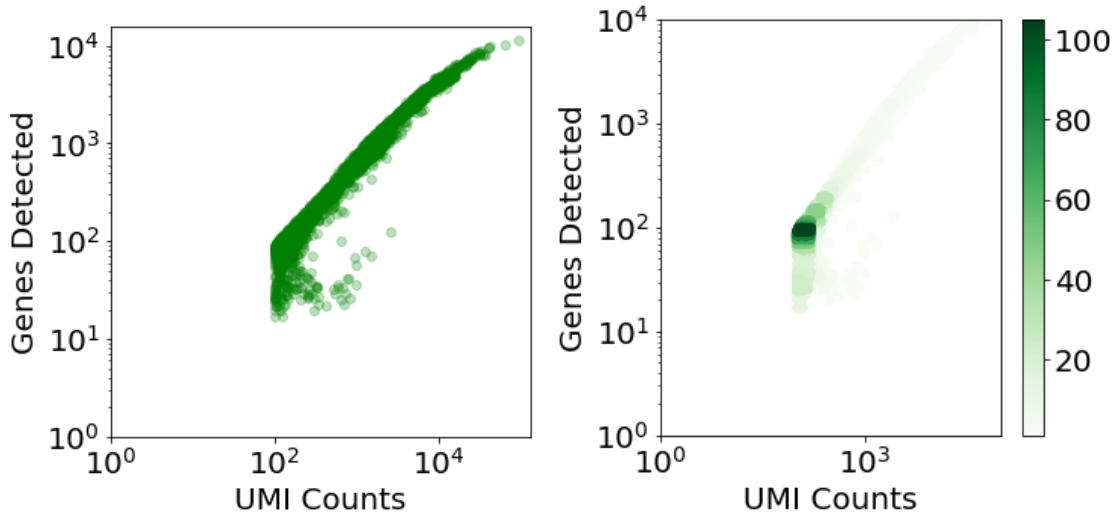

Figure 6: *Library saturation plots for the mouse retinal data.* ***(a)*** *Standard library saturation plot.* ***(b)*** *Gene density saturation plot.*

Much like with the mouse retinal dataset these library saturation plots are good measures of the sequencing depth of the datasets. With the same principles that the deeper the sequencing depth the more runs needed before a new transcript is discovered remaining the same. As before the standard library saturation plot is a slightly misleading figure as it doesn't show how many points are overlapping each other as all are the same colour and density (Figure 6a). Therefore a density saturation plot can be used to give a better indication as to where the majority of the points are clustering (Figure 6b).

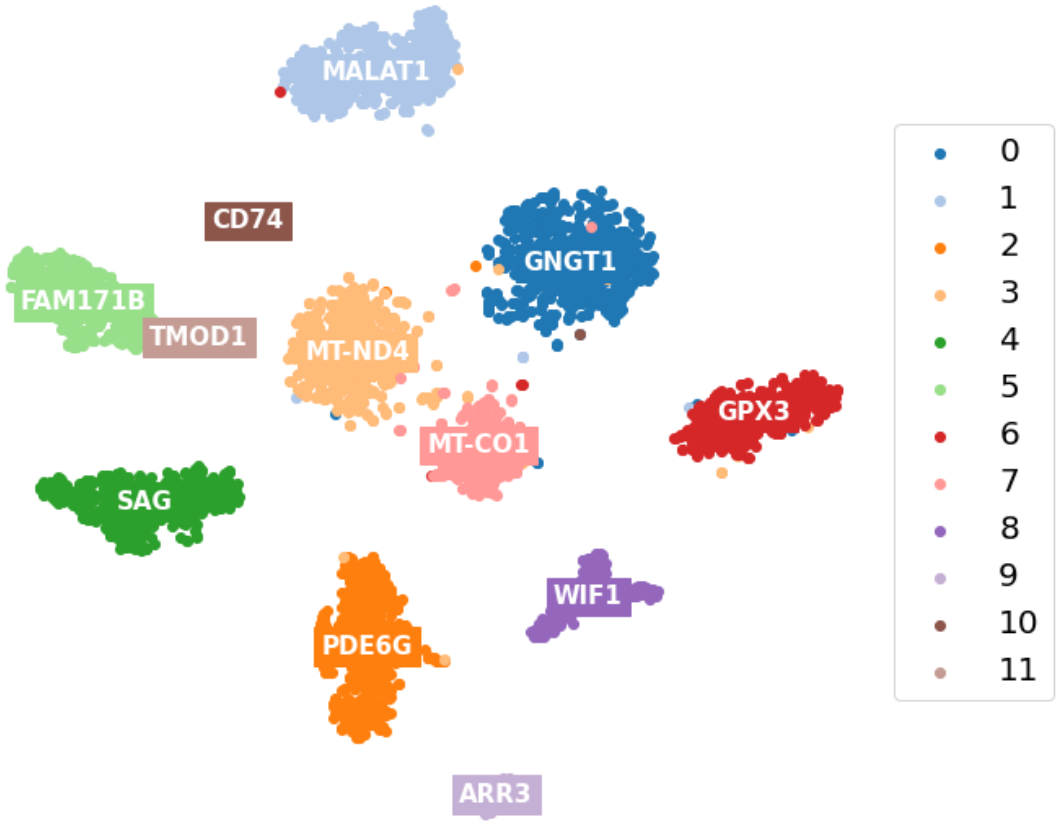

Figure 7: *NCA combined with t-SNE plot for the human retinal dataset, annotated with the top differentially expressed gene per cluster*

The human retinal dataset is smaller than that of the mouse with around 3000 cells passing the basic quality control filtering stage compared with 11000 mouse retina cells. There are less cells in each cluster, with less total clusters. This dataset displays less variety in gene expression, however the different number of cells analysed using different technologies make it hard to draw any meaningful conclusions off only the comparison of the two datasets. Especially in regards to the aims of our project.

Like with the mouse retinal data this is a good case to prove the principle that single-cell RNA-sequencing analysis is possible for human data using Kallisto-Bustools. Human data analysis did run into more problems than the mouse counterpart, this is discussed further in the discussion.

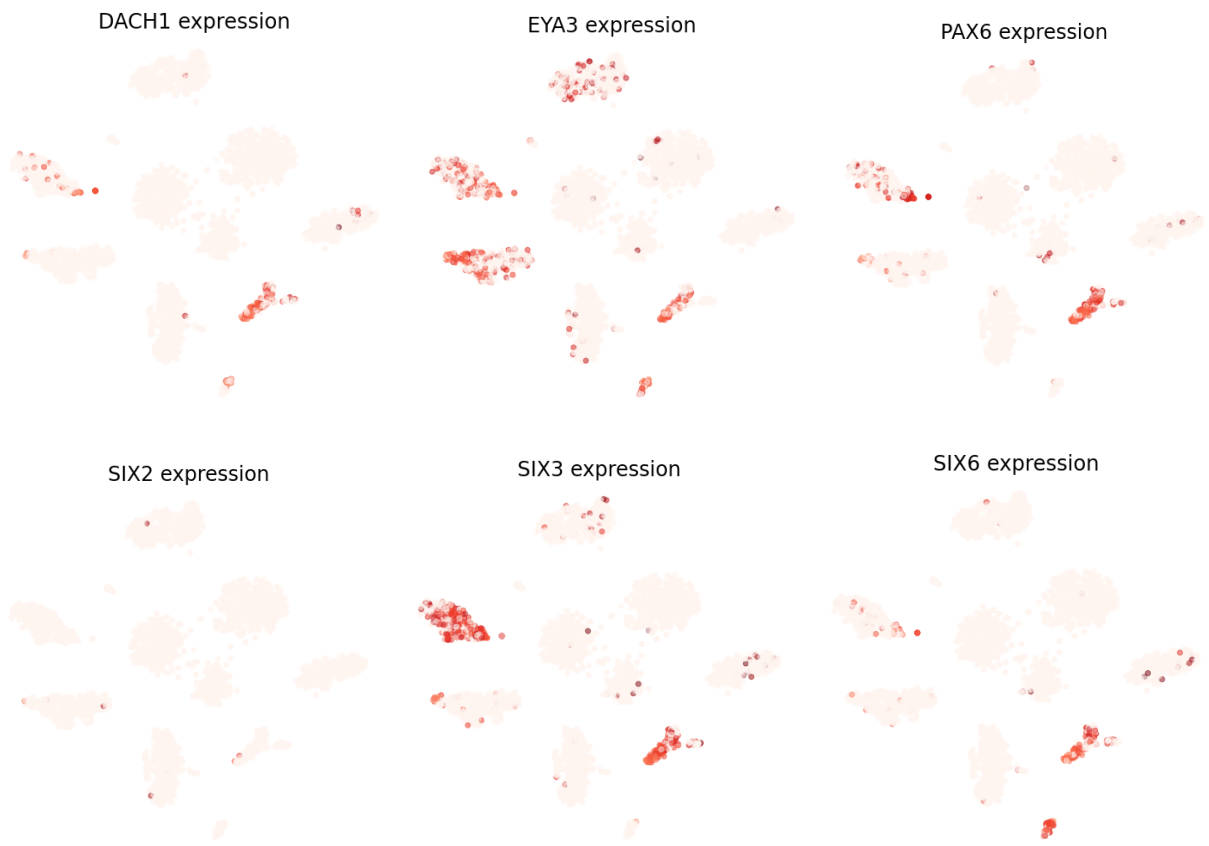

Figure 8: *NCA combined with t-SNE showing RDN member expression in the human retina.* ***(a)*** *DACH1* ***(b)*** *EYA3* ***(c)*** *PAX6* ***(d)*** *SIX2* ***(e)*** *SIX3* ***(f)*** *SIX6*

Figure 8 shows us that there is expression of RDN members in the human retina dataset as was expected. This confirms evidence from previous published work that human retinal tissue expresses high levels of RDN member proteins. For our analysis this dataset is a proof of principle to show that single-cell RNA-sequencing analysis for human data is effective at highlighting differentially expressed genes. It's visible in the data that some clusters expressed RDN members at a greater rate than others, suggesting that the RDN plays an essential role in some functions specific to those cell types.

When analysing figure 8 it's clear that some clusters expressed all RDN members while others showed minimal expression, these differences in gene expression signatures

suggest a functional role the RDN plays in certain types of cell found in the human retina. The significance of this is explored further in the discussion.

Cell clustering of human embryonic glutamatergic neurons using Kallisto-Bustools

Next we ran the same analysis on a human embryonic glutamatergic neuron dataset from the paper from La Manno and collaborators (*La Manno et al.,* 2018). Filtering parameters including Cell threshold, Gene threshold, Top genes were at their default settings, the same as before, this reduced the number of cells down from 382,000 cells to 9,500 cells. These 9,500 consisted of the highest quality reads while anomalous and low quality samples were removed from the analysis. The purpose of this analysis was to see if the RDN was expressed in non-retinal human tissues as this would suggest greater functionality of the circuit.

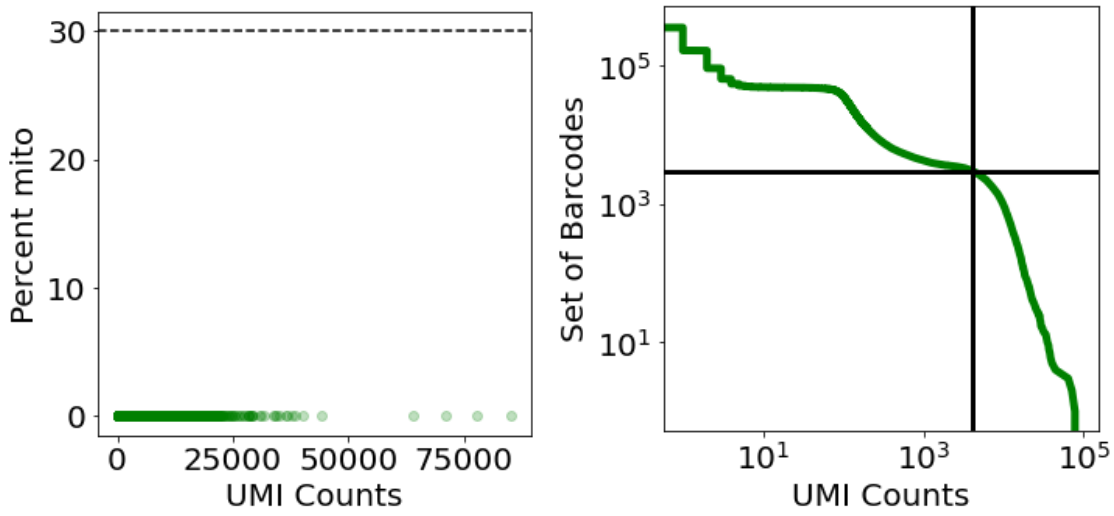

Figure 9: ***(a)*** *Plot showing UMI Counts per cell with the percentage of mitochondrial genes along the y-axis.* ***(b)*** *Knee plot showing the threshold, black lines, any cells to the left of this are filtered from further analysis*

The knee plot was first introduced as a method to plot cells and order them by the number of UMI counts associated with them and the fraction of droplets with the threshold number of cells on the y-axis (Macosko *et al.,* 2015). The number of UMI

counts per cell was plotted against the percentage of mitochondrial genes (Figure 9a). The threshold used for this dataset has been set at a standard maximum of 30% mitochondrial genes, this excludes any data with values greater as the cause is likely down to either poor sample quality or cell stress causing cell lysis. This would result in a loss of cytoplasmic mRNA leading to what appears to be a greater proportion of mtRNA caused by a loss of cytoplasmic RNA rather than an abundance of mtRNA. There was sufficient data to use the standard recommended filtering settings for this dataset. Like with the previous human dataset figure 9a shows 0% mitochondrial genes within the whole dataset, we believe this is due to either an error in the Kallisto human index or within the labelling of mitochondrial genes within the FASTQ file. This is discussed in further detail as part of the 'Mitochondrial gene error for human datasets' section of the discussion.

Cells to the right of the threshold (displayed by black lines running horizontally and vertically across the plot) in the knee-plot are excluded from the data leaving only the cells with a high number of UMI counts per unique barcode, this are suggestive of good quality data (Figure 9b).

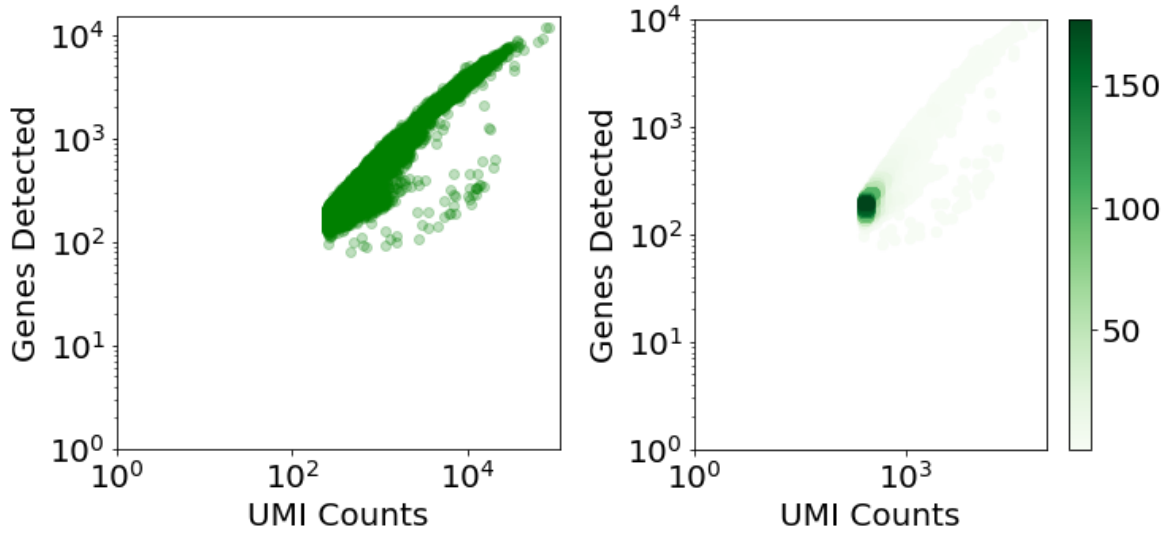

Figure 10: *Library saturation plots for the human embryonic glutamatergic neuron dataset.* ***(a)*** *Standard library saturation plot.* ***(b)*** *Gene density saturation plot.*

Like the plots for previous datasets both graphs in figure 10 are to demonstrate the sequencing depth of the dataset. As before figure 10a is slightly misleading however as it doesn't accurately show how many points are stacked on-top of each other as there is no density gradient. This gradient is shown by the density map in figure 10b showing how tightly packed most of the points are, this density of points at a similar level is lost in figure 10a (Figure 10).

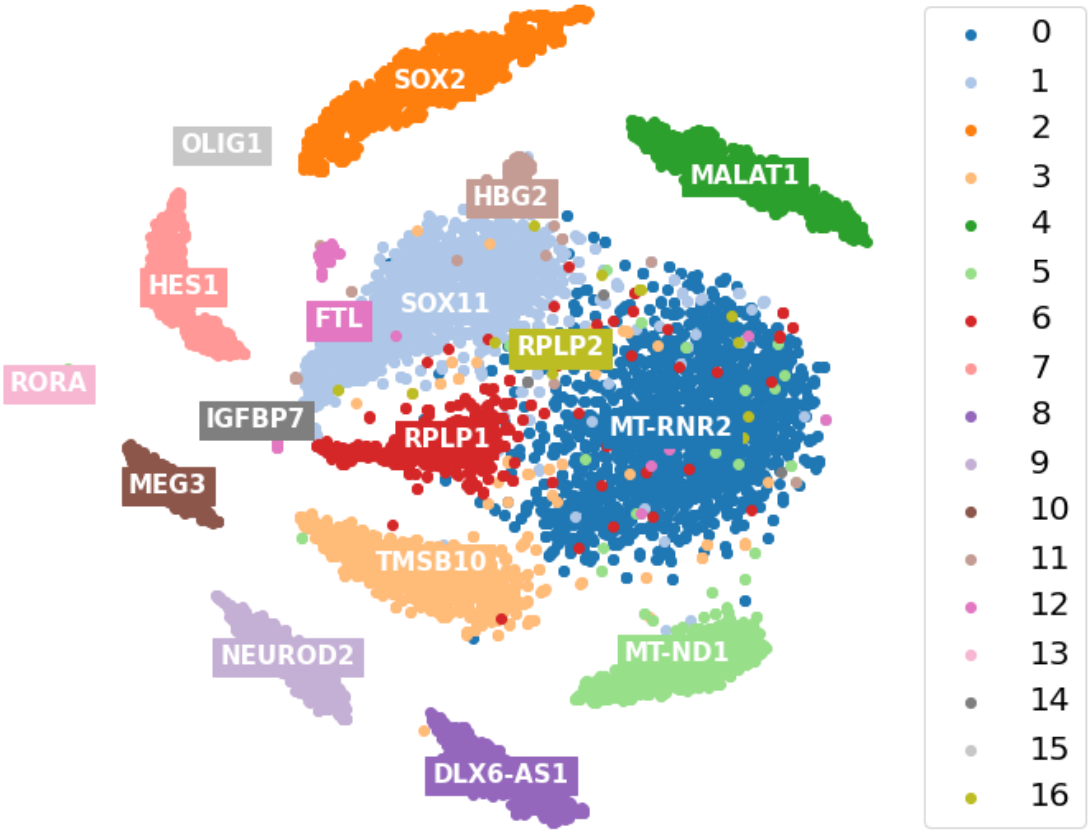

Figure 11: *NCA combined with t-SNE plot for the human embryonic glutamatergic neuron dataset, annotated with the top differentially expressed gene per cluster*

Some genes such as MALAT1 were highly expressed in all clusters, while also being the annotated marker gene of differential expression in cluster 4 (Figure 11). Figure 11 shows that despite being the marker gene for only one cluster, MALAT1 is highly

expressed in all clusters and almost all cells of the human embryonic glutamatergic neuron dataset.

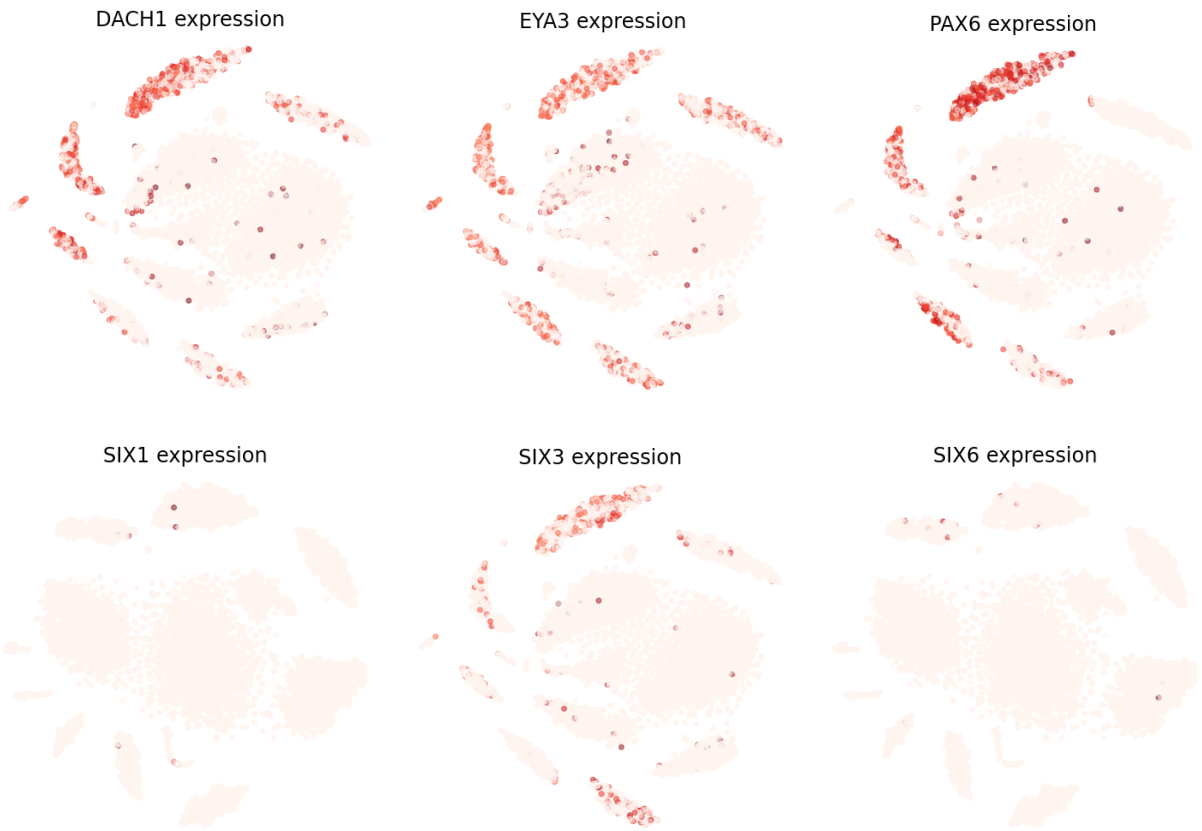

Figure 12: *NCA combined with t-SNE showing RDN member expression in the mouse retina* ***(a)*** *DACH1* ***(b)*** *EYA3* ***(c)*** *PAX6* ***(d)*** *SIX1* ***(e)*** *SIX3* ***(f)*** *SIX6*

With this dataset we were initially unsure whether there would be any RDN expression in any of the clusters, which is shown to be noticeably concentrated in a few specific clusters where present. As this dataset didn't contain any cancer or retina cells we don't yet have an explanation for the clear expression of RDN in some of the clusters. Testing multiple different genes in the network allowed us to confirm the reliability of this data, it also helped us when speculating the cause of this expression by ruling out anomaly as the solution (Figure 12). Expression of all RDN members was localised to specific clusters with some clusters expressing little to no RDN members. As the cells are embryonic there's a high likelihood that there are a large number of stem or progenitor

cells present in the sample, pluripotency of these cells could be the cause behind the expression in this non-retinal dataset. Further discussion as to the cause of expression is speculated further in the discussion under the ‘Human embryonic glutamatergic neuron expression of the RDN’ section.

Cell clustering of Mouse pancreatic dataset using Kallisto Bustools

As with the previous datasets analysed we ran cell clustering using Kalliston-Bustools, this time on a large mouse pancreatic dataset from the paper “Generalizing RNA velocity to transient cell states through dynamical modeling” by V.Bergen and others published in Nature journal in 2020. We wanted to test whether there was RDN expression in non-neuronic mature cells to show, if expression was present, how the RDN is very likely to have a function in a variety of different tissues through interactions with multiple pathways. Due to the large number of cells, filtering parameters were more stringent than those used for the neuron and both retina datasets in order to cut the cell number down from 688k cells to 12.9k cells. This was also the case for the mouse retina data. Therefore only samples of a higher quality would pass the filtering, the potential implications of this to the results are discussed in the ‘Mouse pancreas expression of RDN’ section of the discussion.

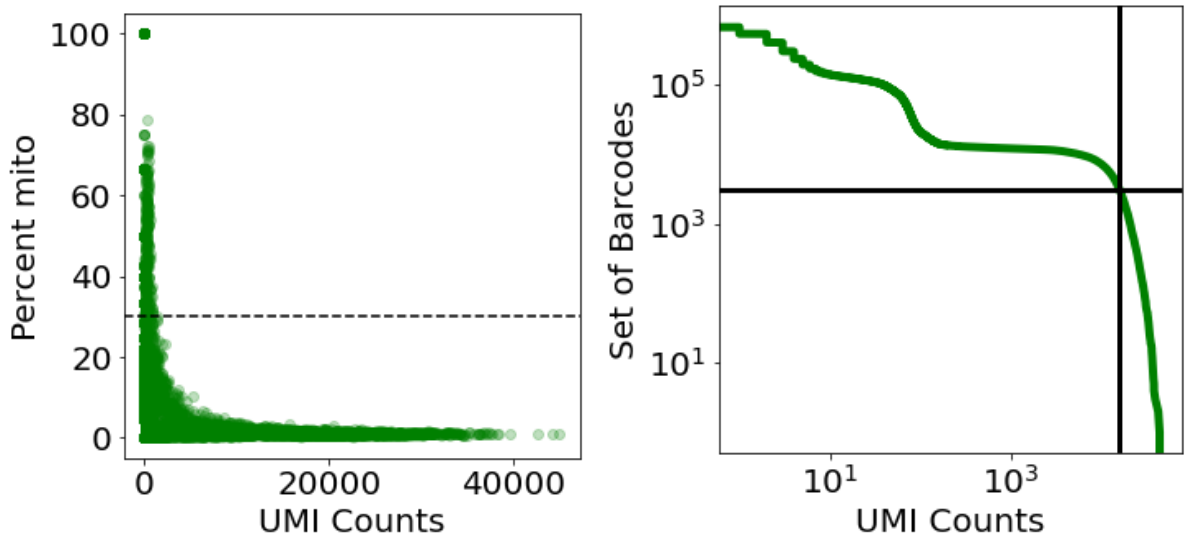

***Figure 13**: **(a)** Plot showing UMI Counts per cell against the percentage of mitochondrial genes **(b)** Knee plot showing the threshold (black lines) any cells to the left or above this line are filtered from further analysis*

Like with previous analysis the mitochondrial gene percentage threshold is set at the standard 30%, any cells with higher percentage are filtered out. This is displayed on the plot as a dotted line where all cells above are removed and has been kept the same for consistency across the project (Figure 13a). The majority of cells have less than 50% mitochondrial genes, too much higher than is suggestive of an error in the dataset. Causes of high mitochondrial gene percentage are the same as with the human embryonic glutamatergic neuron dataset. The knee plot shows the number of barcodes compared against the number of UMI counts. Cells to the left and above the two threshold lines are filtered out from further analysis due to an insufficient ratio of UMI counts to barcodes(Figure 13b). Due to the number of high quality reads in this sample the threshold, as shown by the knee plot, is much stricter. High numbers of UMI codes per barcode is indicative of a good quality sample that is very likely to be an actual cell with minimal signs of damage or stress.

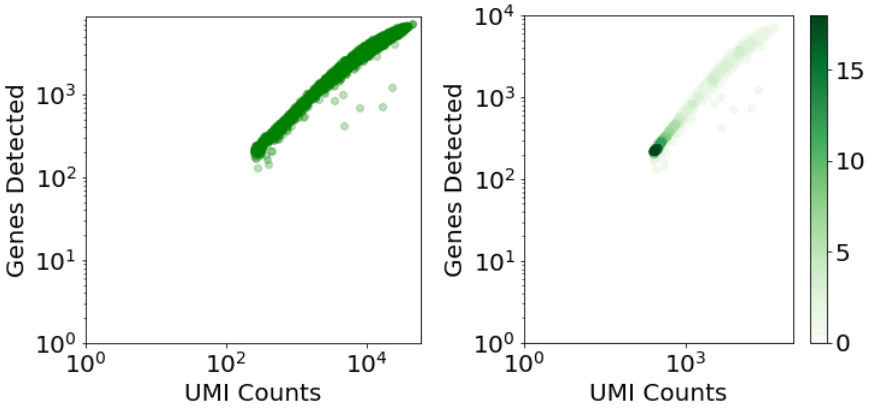

Figure 14: (a) *Library saturation plot, a sign of sequencing depth* ***(b)*** *Gene density plot, shows more detail as to where cells are clustering vs. Library saturation plot*

Library saturation plots show the depth of sequencing, the deeper the sequencing the more times the same transcripts are run, this reduces the chance that transcripts are missed by random chance. The required depth of sequencing is dependent on the library complexity, the more complex the library the deeper the sequence saturation needed to detect a new transcript and the deeper one would need to sequence to be sure of missing minimal gene expression. Basic saturation plots show each cell the same (Figure 14a), the same as in figure 6 this is a misleading figure as it hides the density of hits with similar UMI counts and genes detected. A density plot helps to overcome this (Figure 14b).

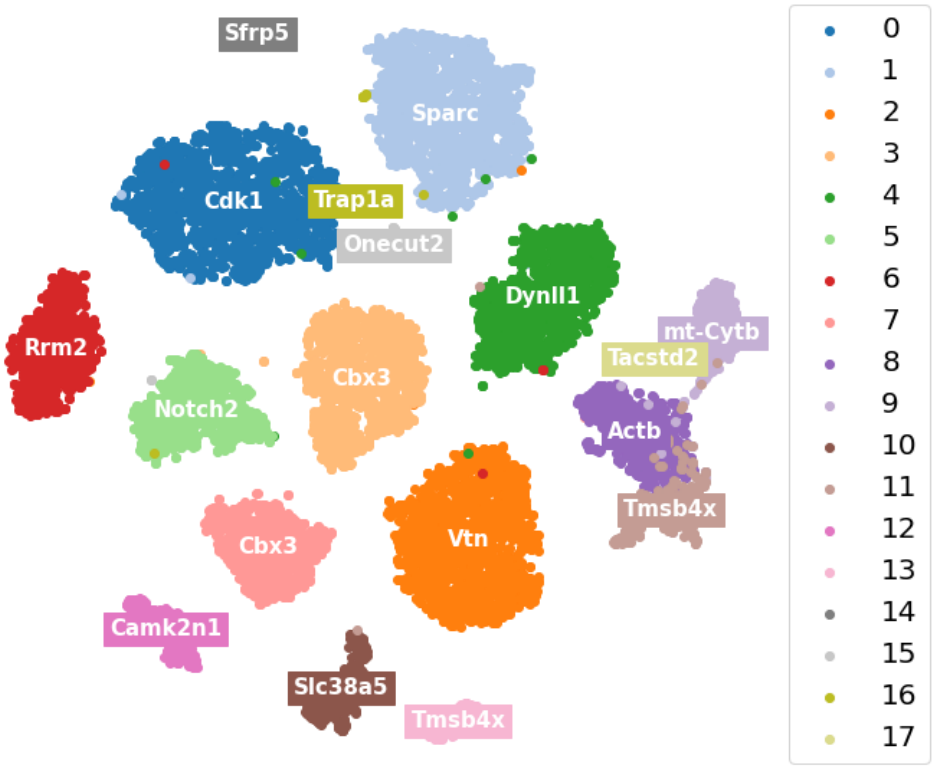

Figure 15: *NCA combined with t-SNE plot for the mouse pancreatic dataset, annotated with the top differentially expressed gene per cluster*

NCA combined with t-SNE for the mouse dataset, each cluster is annotated with their ‘marker gene’, this is the top differentially expressed gene in that cluster and is what each member of that cluster has in common (Figure 15). For example Cbx3 is the top differentially expressed marker gene for two of the clusters,

Figure 16: *NCA combined with t-SNE showing RDN member expression in the mouse pancreas.* ***(a)*** *DACH1* ***(b)*** *EYA3* ***(c)*** *PAX6* ***(d)*** *SIX2* ***(e)*** *SIX3* ***(f)*** *SIX6*

Expression of the RDN in mature mouse pancreas cells was an unexpected result. This is because there is very little published evidence that it had been shown to be expressed in non-retinal tissues. Linked into the fact that it hadn't been hypothesised to be this commonly expressed in mature healthy pancreatic tissues makes this result so interesting. While expression of some RDN members is greater than others and expression is sometime localised to individual clustering (in the case of PAX6), proteins like DACH1 & EYA3 appear to be evenly expressed to a noticeable level in thousands of cells (Figure 16a & 16b). The high levels of PAX6 expression observed in one cluster suggests it as a potential lead marker gene due to it's high differential expression when compared to other clusters. This evidence helps suggest the RDN has a more general function through interactions with multiple pathways.

Human melanoma cancer dataset analysis using Kallisto-Bustools

For the final stage of our analysis we decided to look for RDN expression in human cancer tissues. This dataset is from the paper published from a 2017 paper (Aibar *et al,.* 2017), the data consists of a NFATC2 knockdown A375 human skin melanoma cell line. The aim of these results was to build on the work by Kong and collaborators (Kong *et al.,* 2016) that showed high levels of RDN members due to dysregulation in breast cancer cells. Our hope was to show RDN expression in human melanoma cells too, this would suggest that the RDN plays a functional role in the initiation and/or progression of skin carcinoma *in situ.*

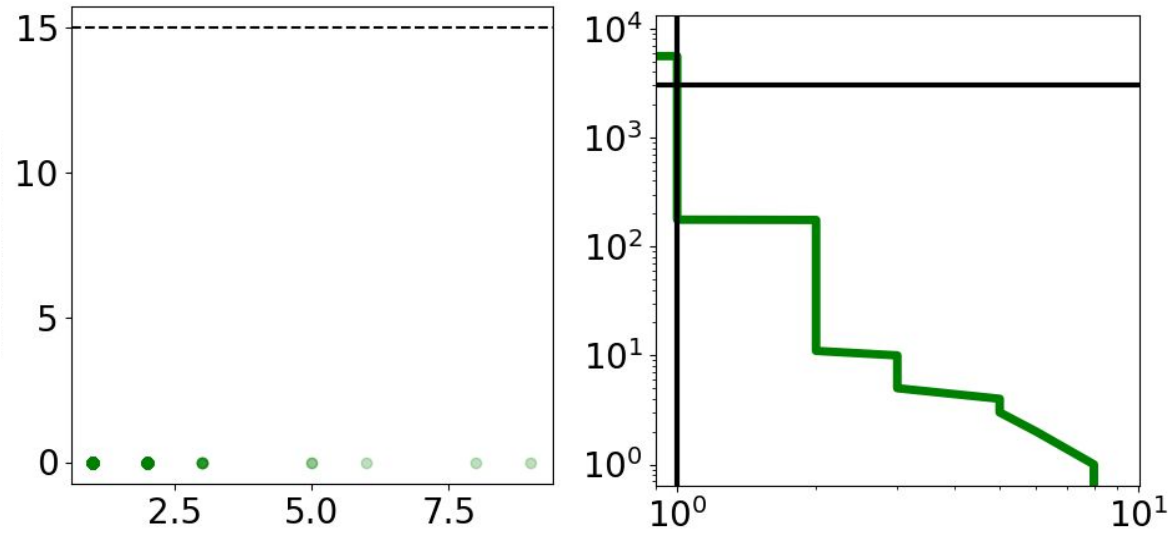

Figure 17:*(a)* *Plot showing UMI Counts per cell (x-axis) the percentage of mitochondrial genes (y-axis).* ***(b)*** *Knee plot showing the threshold, black lines, any cells to the left of this are filtered from further analysis. UMI counts (x-axis) compared with Set of Barcodes (y-axis)*

Issues with the filtering of the melanoma dataset removed all the cells and almost all the genes from the analysis. Before filtering there were 60,623 genes across 8,395 cells, when using our standard cell filtering parameters used on all human datasets before to remove anomalous data from the analysis. In the case of this dataset all but 276 of the genes were removed and none of the cells survived the filtering process. This left the above knee and library saturation plots with very little data to show, just that of the few

remaining genes. Therefore further analysis including that of PCA and NCA clustering with both combined with t-SNE plots to visualise highly dimensional data was impossible due to the lack of cells in the new shape (Figure 17). This meant that we were unable to look for RDN expressions from this dataset.

In order to make sure that these results weren't caused by an error in those specific FASTQ files uploaded to the EMBL-EBI we decided to run more from diseased tissues. We attempted to analyse a dataset from a paper published in 2018 by James Kinchen and collaborators (Kinchen *et al.,*2018), this data contained Human Colonic Mesenchymal cells present in a patient with Inflammatory Bowel Disease (IBD). Unfortunately we had the same issue with this data as we did with the NFATC2 knockdown melanoma cell line. Our third and final attempt to analyse diseased human cells looking for RDN expression was with the data released from a 2019 paper published in nature by Kaushal Parikh and collaborators (Parikh *et al.,* 2019). This dataset contained human colonic epithelial cells from biopsies of both healthy patients and those suffering from IBD, like the two previous papers this also became unusable due to the issues with the cell filtering.

In the 'Issues with cancer dataset filtering' section of the discussion we go into detail discussing potential causes of the filtering error and suggest future ways to potentially avoid the issue for any upcoming work on the topic.

Discussion

Human and mouse retinal expression of RDN

For the majority of the RDN expression in the mouse retinal clusters it's localised to a couple clusters, PAX6, SIX3 and SIX6 especially (Figure 4). This suggests that RDN expression serves a specific function for those cell types. While expression of RDN members was expected in the mouse retinal data, such localisation wasn't. Relevance of RDN expression in certain clusters of the datasets relative to clusters that have little to no expression can be better understood with further research. RDN expression in the human retinal dataset was similar to that of the mouse dataset. Expression was

expected but much more ubiquitously rather than in very localised clusters of cells highly expressing, this is suggestive of an essential function played by the RDN in certain mature retinal cells.

Human embryonic glutamatergic neuron expression of the RDN

When analysing the human embryonic glutamatergic neuron dataset to look for RDN expression, we were unsure whether there would be any expression in a healthy non-retinal tissue. Such high levels of localised expression in specific tissues are suggestive that RDN members have a role in the differentiation of maturing neuron cells. A second potential explanation for these levels of localised expression could be down to the pluripotency displayed by many progenitor cells before they fully differentiate into mature neurons.

Mouse pancreas expression of RDN

While we were unsure about whether we'd find RDN expression in the human embryonic glutamatergic neuron dataset, we were skeptical about finding any expression in the mature, fully differentiated mouse pancreatic tissue data. SIX2/3 & 6 expressions were very minimal whereas DACH1 and EYA3 were expressed in moderate levels across almost all clusters. The most interesting result was PAX6 which showed minimal expression in all but one cluster where expression was very high. In this cluster it was one of the top differentially expressed genes, this is suggestive to PAX6 playing an important role for that specific cell type within the pancreas (Figure 16). A possible explanation for the high levels of PAX6 within one specific cluster could suggest that the cluster contains progenitor or stem cells displaying pluripotency. A recent study published in March 2020 suggests stem cells are present in adult pancreatic tissue (*Jebaraj et al.,* 2020), this supports that as a possible cause of PAX6 expression.

In order to prevent Jupyter Notebooks on SPECTRE from crashing due to the large cell numbers we needed to increase the filtering parameters to let fewer cells through. Under the standard conditions 15,000 cells would have passed filtering, this was too

much data for the system to run. Increasing filtering narrowed the number of cells passing down to 11,900 cells which ran the code without crashing. If greater processing power were available then altering the filtering parameters may not have been necessary. However we believe that this didn't affect results enough to invalidate them, as all cells analysed would have been so anyway with standard filtering parameters.

Mitochondrial gene error for human datasets

While processing our FASTQ file data using Kallisto-Bustools with the standard human clustering index provided by Kallisto-Bustools we encountered a problem where it appeared there was no mitochondrial RNA present in any of the samples. As this issue was present with all human data from multiple papers using many FASTQ files we ruled that out as the cause of the problem and instead suggested the issue is linked to the human index. This error could be as simple as a basic syntax error where the index, the code and the FASTQ files all use slightly different methods to denote a transcript of mitochondrial origin. Or it could be something larger and harder to fix.

The scale at which this has affected the results for all human datasets is hard to measure. High proportions of mitochondrial RNA are indicative of poor quality cells due to possible loss of cytoplasmic RNA through perforations in a lysed cell (*Illicic et al.,* 2016). Those cells that would have been removed by the 30% mitochondrial gene percentage filter are now in the analysis for the human samples when they were removed from the analysis of the mouse data. Whether the number of cells that have got through filtering when they should have been removed is large enough to significantly affect the results is hard to say, but worth noting when looking at the results.

Kallisto-Bustools lets the user make their own indexes for running clustering analysis on organisms which don't yet have a pre-made index. Using this feature to remake the human index may be a way around the mitochondrial gene percentage error, this would only work if the cause of the issue is due to an error in the standard matrix. Failing this,

consulting the authors of the Kallisto-Bustools module could lead to alterations fixing the error and preventing it from affecting further analysis of human datasets.

Issues with cancer dataset filtering

Initially multiple human cancer datasets were to be analysed in an attempt to measure gene expression levels of members of the Retinal Determination Network. However filtering errors meant that all but a few cells were removed from the analysis in the filtering stage, this left insufficient numbers of cells to gather meaningful clustering and velocity information. The cause of this could be caused by several factors. A likely scenario being inconsistencies in the formatting of the FASTQ files that may have affected how Kallisto-Bustools combines the file with the index to create Anndata objects. Many other papers before have used Kallisto-Bustools and have not reported similar errors suggesting it may be bad luck that the FASTQ files we analysed were inconsistent. Solutions to this issue may be solved by either consalting the authors of authors of Kallisto-bustools to discuss the issue, alternatively, one could analyse the data set using different techniques such as the Seurat module for R which is commonly used in many papers for similar quality control filtering procedures.

Further work & improvements

Further work directly off the findings of this project would be to go over errors found and look for solutions to the issues. This would open up the ability to analyse multiple cancer datasets to look for expression of RDN members to see whether there is upregulation in cancer tissues, potentially opening up avenues of research into the use of RDN as therapeutic targets for cancer treatments if found to be important putative drivers of cancer initiation and progression.

Many of the papers we used analysed their data using the Seurat module in R, the module is designed for quality control, analysis and exploration of single-cell RNA-sequencing data. Analysis using this could be an improvement as it may overcome some of the errors we faced using Kallisto-Bustools such as filtering in

human cancer datasets and the loss of mitochondrial gene recognition in all human datasets.

Provided further work into RDN expression in cancers finds evidence that it exists, there could be scope to look at using RNA velocity analysis to analyse some cancer datasets. If we assume that RDN expression is expressed in multiple types of cancer, RNA velocity analysis will allow us to observe whether the expression of the RDN can be linked to the initiation or progression of the cancer in that tissue. RNA velocity analysis looks at the ratio of unspliced to spliced mRNA present in each cell, if one assumes that the rate of transcription, splicing and degradation are steady states across all cells within the dataset one would be able to predict the future gene expression profile of the cell many hours into the future. This would allow analysis to map the current state of a cell's gene expression and it's future one. If members of the RDN were to be highly expressed in cancer cells showing a large change in gene expression it could be suggested that the RDN member was a driving force for that initiation or progression of the cancer. If proven this could yield a potential therapeutic target for future cancer treatments.

Conclusion

The aim of our project was to re-analyse mouse and human datasets to look for expression of RDN members in an attempt to discover whether there was expression in tissues previously assumed to not express members of the RDN. This was largely successful until, due to an issue with the filtering of human cancer datasets we were unable to perform clustering analysis for any diseased datasets. Despite this we did uncover some very interesting results as RDN expression appears common in mouse pancreatic and human embryonic glutamatergic neurons where little published work has shown expression. This was a surprise to see in a non-retinal and non-neuronal tissue. Therefore we suggest that if it can be present in datasets that are, on the surface, so diverse then surely RDN expression is more common than previously thought. If so it potentially has a bigger role in the day to day functioning of many cell types than previously given credit for. Further analysis is needed to confirm expression in a wider

variety of tissues before any more assumptions can be made. After which analysis can then move onto the role of this expression especially in the case of expression in non-neuronal tissues. This may eventually lead the scientific community to view the RDN as a new, more ubiquitous, regulatory circuit.

References

Aibar, S. *et al.* (2017) 'SCENIC: single-cell regulatory network inference and clustering', *Nature methods,* 14(11), pp. 1083-1086.

Aldridge, S. and Teichmann, S.A. (2020) 'Single cell transcriptomics comes of age', *Nature communications,* 11(1), pp. 4307.

AlJanahi, A.A., Danielsen, M. and Dunbar, C.E. (2018) 'An Introduction to the Analysis of Single-Cell RNA-Sequencing Data', *Molecular therapy. Methods & clinical development,* 10, pp. 189-196.

Bergen, V. *et al.* (2020) 'Generalizing RNA velocity to transient cell states through dynamical modeling', *Nature biotechnology,* .

Bray, N. *et al.* (ed.) (2015) *Near-optimal RNA-Seq quantification.*

Charlotte Soneson *et al.* (ed.)*1 Preprocessing choices affect RNA velocity results 2 for droplet scRNA-seq data.*

Chen, G., Ning, B. and Shi, T. (2019) 'Single-Cell RNA-Seq Technologies and Related Computational Data Analysis', *Frontiers in genetics,* 10, pp. 317.

Cui, H. *et al.* (2013) 'Association of decreased mitochondrial DNA content with the progression of colorectal cancer', *BMC cancer,* 13(1), pp. 110.

Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Computer applications in the biosciences,* 29(1), pp. 15-21.

Gehring, W.J. and Ikeo, K. (1999) 'Pax 6: mastering eye morphogenesis and eye evolution', *Trends in genetics,* 15(9), pp. 371-377.

Ilicic, T. *et al.* (2016) 'Classification of low quality cells from single-cell RNA-seq data', *Genome biology,* 17(1), pp. 29.

Kinchen, J. *et al.* (2018) 'Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease', *Cell,* 175(2), pp. 372-386.e17.

Kluyver, T. *et al.* (2016) 'Jupyter Notebooks-a publishing format for reproducible computational workflows', *IOS Press*.

Kong, D. *et al.* (2016) 'The retinal determination gene network: from developmental regulator to cancer therapeutic target', *Oncotarget,* 7(31), pp. 50755-50765.

Kumar, J.P. (2010) 'Retinal determination the beginning of eye development', *Current topics in developmental biology,* 93, pp. 1-28.

La Manno, G. *et al.* (2018) 'RNA velocity of single cells', *Nature,* 560(7719), pp. 494-498.

Luecken, M.D. and Theis, F.J. (2019) 'Current best practices in single‐cell RNA‐seq analysis: a tutorial', *Molecular systems biology,* 15(6), pp. e8746-n/a.

Lukowski, S.W., Lo, C.Y., Sharov, A.A., Nguyen, Q., Fang, L., Hung, S.S., Zhu, L., Zhang, T., Grünert, U. and Nguyen, T. (2019) 'A single‐cell transcriptome atlas of the adult human retina', *The EMBO journal*, 38(18), pp. e100811.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N. and Martersteck, E.M. (2015) 'Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets', *Cell*, 161(5), pp. 1202-1214.

Melsted, P. et al. (2019) Modular and efficient pre-processing of single-cell RNA-seq. Available at: https://search.datacite.org/works/10.1101/673285.

Parikh, K. *et al.* (2019) 'Colonic epithelial cell diversity in health and inflammatory bowel disease', *Nature (London),* 567(7746), pp. 49-55.

Ren, X., Kang, B. and Zhang, Z. (2018) 'Understanding tumor ecosystems by single-cell sequencing: promises and limitations', *Genome biology,* 19(1), pp. 211.

Serena J. Silver and Ilaria Rebay (2005) 'Signaling circuitries in development: insights from the retinal determination gene network', *Development,* 132(1), pp. 3-13.

Valencia, J.E. *et al.* (ed.) (2019) '*Ciliary photoreceptors in sea urchin larvae indicate pan-deuterostome cell type conservation.'*

van der Maaten, L. J. P and Hinton, G.E. (2008) 'Visualizing High-Dimensional Data Using t-SNE', Journal of machine learning research, 9(nov), pp. 2579-2605.

Vitak, S.A. *et al.* (2017) 'Sequencing thousands of single-cell genomes with combinatorial indexing', *Nature methods,* 14(3), pp. 302-308.

Zhang, X. *et al.* (2019) 'Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems', *Molecular cell,* 73(1), pp. 130-142.e5.

Data availability

The datasets analysed in our project are all publicly available and published. All files were downloaded from the EBI server (https://www.ebi.ac.uk/) in FASTQ format. Search for files by relevant GEO accession number: Raw human retina dataset - **E-MTAB-7316**, Mouse retina dataset - **SRR1853178**, Human embryonic glutamatergic neuron dataset - **SRP129388**, Mouse pancreatic dataset - **GSE132188**, Human Melanoma dataset - **GSE99466**.

The code we used to perform both the clustering and the velocity analysis are stored in our Github repository:

- Link to our Github repository : https://github.com/mcg33/IRP_2020_Code

Acknowledgements

I would like to thank Dr. Roberto Feuda for his guidance and support throughout this project. I would also like to thank all team members of Dr. Feuda's lab for their assistance with this project.