Single Cell Genomics Data Analyses Reveal Novel Biological Insights



Xiaoming Hu

University College

University of Oxford

Thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2016

Abstract

Single cell RNA-Seq is one type of single cell –omics experiment. The technology to perform single cell RNA-Seq has recently become available, promising to enable a range of biological investigations, such as identifying cell-to-cell heterogeneity within a cell culture and novel cell types in primary tissues. Despite of these advances, data analytical aspects of this technology was proved to be challenging. Many technical issues and confounders deserve special attention. Only when these are well-understood and accounted for can bioinformatics analyses provide reliable and robust results. The aims of my research were first to comprehensively investigate the analytical aspects of single cell RNA-Seq datasets, and then to apply appropriate analytical methods best suited for each of my two research projects to identify cell-to-cell heterogeneity, in order to answer specific biomedical questions.

I investigated several technical issues that are prevalent among current single cell RNA-Seq datasets, including short read mapping biases, gene detection dropout, empty-capture sites, doublet-capture sites, cell-capture bias, coverage bias and amplification noise. In the first research project, we designed a protocol to induce human embryonic stem cells to form organoids (i.e. conceptoid bodies). We applied single cell RNA-Seq to the conceptoid bodies to identify heterogeneity among its constituent cells. My analysis revealed two main results: first, from heterogeneous cells, I identified distinct cell types; second, I showed that two cell types present in conceptoid bodies are similar to the two major cell types in Day 6 human blastocysts (i.e. inner cell mass and trophoblast, respectively). These results demonstrated that human embryonic

stem cells can give rise to the earliest cell types in a human blastocyst. This implies that human embryonic stem cells are totipotent rather than having restricted potency. Furthermore, this project introduced an organoid protocol that may have tremendous utility in understanding early human development and associated diseases. In this project, it was a major challenge to compare datasets from two different laboratories. This analytic challenge was overcome using a method that minimized such a batch effect. In the second research project, I sought to identify novel classes of postnatal Day-4 mouse midbrain dopaminergic neurons using single cell RNA-Seq. I identified at least four novel classes of dopaminergic neurons, and their respective marker genes. These predicted cell type specific marker genes have since been validated experimentally. With these markers, cell staining, sorting, conditional knockouts and opto-genetic manipulation experiments can now be performed. These studies provide a first step towards a better understanding of the repertoire of midbrain dopaminergic neurons and the mechanisms by which they are involved in neurological disease. To identify and define novel cell types presents a major challenge, because clustering of cells does not directly identify cell types. I have taken an analytical approach to iteratively split cell groups into smaller and smaller groups, to identify cell types within them. I anticipate the analytical approaches designed for these projects are also widely applicable to many other biological investigations.

List of abbreviations frequently used in this thesis

RNA-Seq: sequencing the RNA of a sample. In this thesis, it will refer to a more restricted meaning, that of sequencing the cDNA generated from poly-adenylated mRNA.

DNA-Seq: sequencing the genomic DNA of a sample.

qRT-PCR: quantitative Reverse Transcription Polymerase Chain Reaction.

FACS: fluorescence activated cell sorting

CyTOF: cytometry by time-of-flight

-omics experiment: experiment that aims to measure the entire set of protein molecules, RNA transcripts and metabolites.

non -omics experiment: experiment that aims to measure a handful pre-selected molecules

single cell -omics experiment: -omics experiment performed on individual cells

Transcriptome: the entire set of RNA transcripts.

Proteome: the entire set of proteins.

Metabolome: the entire set of metabolites.

SVM: support vector machine

QC: quality check

ERCC: external RNA controls consortium

ESC: embryonic stem cell

EpiSC: epiblast stem cell

ICM: inner cell mass

TE: trophoblast

GFP: green fluorescent protein

SNc: substantia nigra pars compacta

VTA: ventral tegmental area

mDA: midbrain dopaminergic neurons

Acknowledgement

This thesis would not be possible without the patience, clement, time, inspiration, insight, persuasiveness, knowledge and care. Especially that when I made mistakes or when I made progress, that comforting smile provided by my supervisors: Professor Christopher Paul Ponting and Doctor Wilfried Haerty.

This thesis would not be possible without the nurture and care provided by my parents. Lena made the journey of this thesis more memorable, and with more joy. If without you, this thesis would be a lot slimmer and I would be a lot fatter; or, the thesis would be a lot fatter and I would be a lot slimmer. This thesis would not be possible without the care and understanding of Jessica. This thesis would not be possible without the financial support of ASTAR.

This thesis would not be possible without the continued help of so many of you. Just to name a few in the order of their family names: Doctor Siew-Lan Ang, Doctor Sharon Bolton, Doctor Mary Green, Doctor Peter Vasili Kharchenko, Doctor John Marioni, Doctor Shyam Prabhakar, Doctor Paul Robson, Doctor Lili Sun, Doctor Wenjie Sun. I deeply appreciate the time and effort of my thesis examiners: Professor Christopher Yau and Professor Ian Chambers. Their insightful comments helped me to further improve this thesis.

As a result, I am so indebted to all of you. This is the most enriching journey of my life.

Contents

Abstract ii

List of abbreviations frequently used in this thesis iv

Acknowledgement v

Chapter 1 Single Cell Measurements are Important for Biological Research 1

1.1 Much information is inaccessible to bulk measurements 1

1.2 Measurements performed on single cells are useful, but non -omics measurements show limitations when markers are not available 3

1.3 Molecular markers are important, are often not known and are difficult to discover 6

1.4 Single cell -omics measurements, for example single cell transcriptomics, can resolve the difficulty of identifying markers 7

1.5 For performing transcriptomic measurements, RNA-Seq is preferable to microarrays 9

1.6 Some difficulties in performing single cell RNA-Seq have been recently resolved 11

1.7 Single cell RNA-Seq has been applied to the identification of novel cell types 13

1.8 Single cell RNA-Seq is useful for quantification of biological variation 22

1.9 Single cell RNA-Seq is useful for assessing the effects of cell cycle and differentiation on the transcriptome 25

1.10 Introducing the single cell RNA-Seq protocols and platforms being used for my projects 26

1.11 Structure of the thesis 29

1.12 Summary 30

Chapter 2 Single Cell RNA-Seq Experiments and Associated Data Analyses are Challenging 32

2.1 High throughput sequencing and its associated data analysis can be complex 33

2.2 RNA-Seq and its associated data analysis can be complex 36

2.3 Single cell RNA-Seq datasets are high dimensional and require dimensionality reduction methods 44

2.4 Single cell RNA-Seq expression values are distributed with a zero-inflated pattern and require special considerations 45

2.5 A large proportion of capture units fail to capture a healthy cell 48

2.6 Single cell RNA-Seq dataset typically can also contain capture units with multiple cells 52
2.7 Single cell RNA-Seq dataset profiled cells are a biased representation of the cells present in the solution 55
2.8 Single Cell RNA-Seq exhibits strong amplification bias that can be accounted for using Unique Molecular Identifiers 56
2.9 Single cell RNA-Seq data analyses methods to identify cell types....58
2.10 Determining cell type number from a pool of cells is a subjective issue ..59
2.11 Summary 61
Chapter 3 : Human Conceptoid Bodies Derived from Human ESCs, and Insights Obtained from Associated Single Cell RNA-Seq Data Analyses 64
3.1 Abstract 64
3.2 Contributions and Publications 66
3.3 Introduction 68
3.4 Results 74
Protocol to induce trophoblast from human ESCs 74
Discovery of a transient bi-potent state co-expressing both ICM and trophoblast markers 78
Protocol to exploit the transient bi-potent state to create an organoid: a conceptoid body 84
Bioinformatics analyses on single cell RNA-Seq data reveal two distinct cell types in conceptoid body 85
Transcriptomic demonstration of that cell types in conceptoid bodies are similar to those in blastocysts 90
Examine the expression of individual genes in the cell types in conceptoid bodies and in blastocyst 96
Experimental validation of conceptoid bodies with respect to function....99
3.5 Discussion....105
The developmental position and potential of human ESCs 105
The conceptoid body is a useful model for studying human early development 109
3D environment and cell-to-cell contact are important aspects of cell differentiation....110
Studying single cell RNA-Seq data from blastocyst requires *a priori* clustering 112

Potential doublet issue does not confound the results we obtained so far 113
Comparing data from two different labs is challenging and requires special attention 114
3.6 Materials and Methods 117
M1 Single cell RNA-Seq assays 117
M2 RNA-Seq reads mapping and quality filtering on each cell 117
M3 Genes expression quantification 119
M4 Cell type identification in hESC-derived conceptoid bodies 119
M5 Differential expression analysis 120
M6 Quality Check and Cell type identification in IVF-derived blastocyst 121
M7 Data-driven analysis to examine the ICM and the trophoblast within conceptoid bodies as well as the ICM and the trophoblast within human blastocyst 125
M8 Expression similarity to a panel of human tissues 125
M9 Similarity of *in vivo* and *in vitro* blastocystic trophoblast 126
M10 Protocol differences 127
M11 Differential expression analysis between blastocystic trophoblast and ICM cells derived from *in vivo* blastocyst 128
M12 Differential expression analysis between blastocystic trophoblast and ICM cells derived from *in vitro* conceptoid bodies 129
M13 Differentially expressed genes between human ICM cells from in vivo blastocyst and protocol-matched human ESCs 129
M14 Differentially expressed genes between human ICM cells from in vitro conceptoid bodies and protocol-matched human ESCs 129
M15 Principal Component Analysis 129
M16 Analysis of single cell expression detected by qRT-PCR 130
M17 Statistical analysis of attachment assay and cell migration assay 131
Chapter 4 Identification of Different Classes of Dopaminergic Neurons in the Mouse Midbrain 133
4.1 Abstract 133
4.2 Publication and Contributions 134
4.3 Introduction 135
Motivation 135
Experimental Design 143

4.4 Results 148
Overview of all subpopulations of cells that will be discussed in this chapter 148
The first iteration of analyses split all 172 pass QC cells into two major groups: 110 dopaminergic neurons and 62 other GFP+ cells 150
Analysis of the group of 110 mature midbrain dopaminergic neurons recapitulated known classes: VTA and SNc dopaminergic neurons 163
Differential expression analysis identified novel markers for SNc dopaminergic neurons at postnatal Day 4 173
Differential expression analysis identified novel markers for VTA dopaminergic neurons at postnatal Day 4 176
Differential expression analysis on the group of 110 dopaminergic neurons identified novel classes of midbrain dopaminergic neurons 178
Among the 28 SNc dopaminergic neurons, clustering revealed two sub-classes of SNc dopaminergic neurons that can be validated 183
Explore further heterogeneities among the 28 SNc dopaminergic neurons 191
Among the 21 VTA dopaminergic neurons, clustering revealed two sub-classes of VTA dopaminergic neurons that can be validated.....196
Explore further heterogeneities among the 21 VTA dopaminergic neurons 201
Among the group of 62 other GFP+ cells, two different classes of neurons were identifiable.....203
The group of *Crhbp*+ cells are located at two anatomical regions: the ventral midline region and scattered at the VTA of mouse midbrain.....208
Heterogeneous expression of ephrins and ephrin receptors among 110 dopaminergic neurons.....210
4.5 Discussion.....214
Rationale and importance of this research 214
Four validated subpopulations of dopaminergic neurons configured by two orthogonal parameters 215
Subpopulations of dopaminergic neurons that can potentially co-transmit glutamate and GABA 218
The *Crhbp+* class of neurons are potentially capable of co-transmitting dopamine and GABA 221

The benefits of entirely data-driven approaches for analysing single cell RNA-Seq dataset 222
The superior resolution when identifying subpopulations of midbrain dopaminergic neurons and the small number of cells required to identify their marker genes 223
Limitation of the clustering based classification method 223
Limitation of the pseudo-random capture process 226
Limitation of molecular information based classification 227
4.6 Materials and Methods 228
D1 Single cell RNA-Seq assays 228
D2 RNA-Seq reads mapping and quality filtering on each cell 228
D3 Expression estimation for genes 229
D4 Assess whether any batch failed completely 229
D5 Clustering of cells *a priori* 231
D6 Differential expression analysis 232
D7 Iterative splitting strategy and highly variable gene strategy to identify novel cell groups and their respective markers 232
Chapter 5 Conclusion 234
5.1 Summary of the results presented in my thesis 234
5.2 Summary of the importance of my thesis 238
5.3 Suggestions for future research directions 239
References 241

Chapter 1 Single Cell Measurements are Important for Biological Research

Measuring the abundance of biological molecules for individual cells is important for biological research. In this chapter, I will start by describing the utility of performing single cell measurement on a handful pre-selected molecules (i.e. non -omics measurement, such as FACS, qRT-PCR, CyTOF). This is followed by describing some drawbacks of these non -omics single cell measurements. These drawbacks motivate the development of methods for performing single cell -omics measurements, such as proteomics methods, transcriptomics methods and metabolomics methods. I will then describe in detail a highly successful single cell transcriptomics method: single cell RNA-Seq. This is followed by describing published examples of applying single cell RNA-Seq to uncover novel biology, such as identifying novel cell types in tissue samples and in organoid cell cultures. The research aims, approaches and technologies of these published examples are similar to those of my research projects described in Chapter 3 and 4. The structure of the thesis will be briefly introduced in the last section of the chapter.

1.1 Much information is inaccessible to bulk measurements

To measure the abundance of biological molecules, for example proteins, DNA molecules, RNA molecules, and other metabolites in a sample of interest is an important activity of molecular biology research. This is because the abundances of biological molecules may reveal the function and processes

occurring in the sample. Many studies perform such measurements on bulk samples containing many cells, because more reliable measurements can typically be obtained with larger amount of starting material obtained from bulk samples.

However, in the scenarios described below, the desirable information is inaccessible to such bulk measurements. (i) In most situations, a bulk sample will contain cells of different cell types. Hence any measurements, for example protein abundance measurement, obtained from the bulk sample only reflect the total amount of protein present in the entire bulk sample. Information regarding how these proteins are distributed among the different cell types is inaccessible. (ii) Even if the bulk sample contains cells of the same type, these cells might vary in their differentiation statuses or their levels of maturity, leading to cell-to-cell differences. For example, experiments have shown that some totipotent blastomere cells from the 4-cell stage of embryogenesis can tend towards differentiation into a specific cell fate, although they remain totipotent and are able to acquire other cell fates [1]. Information regarding the varying levels of differentiation status is inaccessible to bulk sample measurements. (iii) Even for a homogeneous pool of cells, the cells can be in different cell cycle stages, causing their molecular constituents to differ from one another. (iv) Even among synchronized cells or post-mitotic cells of the same cell type, gene expression might be heterogeneous for biological reasons. In this scenario, a research aim may be to identify such heterogeneously expressed genes. For example, the expression of the pluripotency regulator *NANOG* can be

heterogeneous among a group of apparently equivalent embryonic stem cells [2]. Additionally, the abundance of ALDH1A1 protein can be different among an apparently homogeneous population of midbrain dopaminergic neurons, suggesting underappreciated heterogeneity are present among these cells [3]. These observations, although not fully understood, suggest that these cells despite being one cell type, may function heterogeneously. If we measure only the total abundance of biological molecules in a bulk sample, these biologically relevant heterogeneity will remain inaccessible. Thus, it is highly desirable to measure the abundance of biological molecules for individual cells.

1.2 Measurements performed on single cells are useful, but non -omics measurements show limitations when markers are not available

Due to the disadvantages of bulk measurements explained previously, many methods have been developed to perform measurements on single cells. However, most of these popular single cell measurement methods are non -omics experiments, meaning that instead of measuring the entire set of biological molecules, a small subset were selected beforehand.

Below, I will introduce three examples of popular single cell non -omics measurement methods, namely: fluorescence-activated flow cytometry (FACS), cytometry by time-of-flight (CyTOF), and qRT-PCR. (i) FACS is a method to measure the fluorescence levels of every individual cell for thousands of cells when they pass the fluorescence detector one-by-one. When the fluorescence level of a cell is indicative of the abundance of a specific protein in the cell, the

quantification of such fluorescence level then provides a proxy of the protein's abundance in the cell. There are several limitations to this approach. It only detects the proteins that can be linked to a detectable fluorescence protein (e.g. using good antibodies). The total number of protein species measurable in one single experiment is limited by the number of fluorophores that have mutually non-overlapping colour spectrum that are reliably discernible. (ii) CyTOF experiment is similar to FACS, where a panel of rare earth metals serve the same function as fluorophores in FACS. The experiment measures the abundance of each of the rare earth metals in individual cells. The abundance of each rare earth metal is a proxy for the abundance of the corresponding antibody protein that is attached to it, which in turn is a proxy for the abundance of the corresponding antigen protein in the cell being measured. This enables the abundance of several pre-selected proteins in a single cell to be measured for thousands of cells. Much like the limitations of the FACS approach, the CyTOF approach is limited by the availability of antibodies. Furthermore, the total number of proteins that can be measured for each cell in the CyTOF approach is limited by the number of distinct rare earth metals that are discernible in such time-of-flight measurements. For example, in the CyTOF system commercially available from Fluidigm, this limit is about 40 proteins. (iii) Single cell qRT-PCR is a method to measure the abundance of several pre-selected RNA species in single cells. The method splits the cytoplasm of one cell into several equal aliquots. For example, the Biomark microfluidic system from Fluidigm splits it into 96 equal aliquots. In each of the 96 aliquots, the microfluidic system performs a different qRT-PCR experiment to measure the abundance of a pre-selected mRNA

transcripts. There are several limitations to this approach. For example, the PCR efficiencies for different transcripts can be different, and if so, the estimated abundances will be greatly affected. Furthermore, because each measurement requires one aliquot of a cell's cytoplasm, the number of RNA transcripts being detectable in a single cell is limited by the number of separate experimentable aliquots that we can divide its cytoplasm into. The Biomark system (i.e. a commercially available system from Fluidigm) is able to estimate the relative abundances of up to 96 RNA transcripts concurrently for a single cell. It is worth noting that the Biomark system can experiment on 96 single cells in parallel (i.e. in total performing 9216 qRT-PCR experiments at a single run), greatly enhancing throughput and convenience.

As can be seen from above, there is a common limitation among all single cell non -omics approaches: they can only measure a limited number of pre-selected molecules (e.g. approximately 10 proteins in flow cytometry; approximately 40 proteins in CyTOF; approximately 96 RNA transcripts in qRT-PCR). This lack of completeness is a major disadvantage: a set of proteins or RNAs need to be chosen beforehand, typically according to existing literature. This implies that the experiments can only investigate the cell types whose marker molecules are already known. This disadvantage limits the application of non -omics based single cell experiments, because it hinders their ability to discover novel cell types when a marker molecule is not known beforehand.

1.3 Molecular markers are important, are often not known and are difficult to discover

For centuries, morphology was the only way to distinguish cell types. A molecular marker is a molecule that is specifically present in one cell type but not in other, perhaps neighbouring, cell types. A molecule that is commonly referred to as a "marker" is so-called, because it *marks* a cell type out from its surrounding cell types. In the era of rapidly advancing molecular methods for biological investigations, knowing a molecule marker of the cell type of interest is extremely important. This is because it enables the application of a continuously expanding suite of molecular methods to investigate a biological sample. For example to identify cellular localisation using protein staining, or to perturb a specific group of cells for characterising their function (e.g. using conditional knockout and opto-genetic techniques). However, for many cell types, especially novel cell types, a marker molecule is not known *a priori*. Furthermore, depending on the molecular experiment in mind, a good marker also shall possess many other characteristics. For example, if the experiment in question uses an antibody to stain for the marker, it will be preferable if the marker is abundantly present on the surface of the cell, because it is more accessible to antibodies compared to a marker located in the nucleus. As a result, it is important to identify many candidate marker molecules for different purposes.

To discover many candidate marker molecules for a cell type of interest is an important but difficult task in molecular biology research. The only feasible way to achieve this goal is to perform -omics measurements on the cell type of interest, and separately on its surrounding cell types. A contrast between the

two -omics measurements would then reveal the molecules that are only abundantly present in the cell type of interest, and that are not present in any surrounding cell types. This comparison requires two conditions to be satisfied: (i) being able to isolate the cell type of interest from its surrounding cell types; (ii) being able to perform -omics measurements for the isolated cell type, and its surrounding cell types separately. For cell types that can be isolated from their surrounding cell types based on anatomical, morphological or other physical properties, marker discovery experiments were possible. For the remaining cell types that cannot be isolated from their surrounding cell types based on these physical properties, historically we have not been able to identify molecular markers for these cell types. These cell types will remain unknown, because (i) a molecular marker is required to purify the cell type; (ii) a purified population of this cell type is required to discover its molecular markers. When we neither know a molecular marker of the cell type, nor are able to purify the cell type, little progress can be made.

1.4 Single cell -omics measurements, for example single cell transcriptomics, can resolve the difficulty of identifying markers

When discovering novel cell types and their respective molecular markers, a good approach to perform single cells -omics measurements. Single cell -omics experiment is defined as an experiment that measures the abundance of the complete set of biological molecules, including all proteins, RNAs, DNAs, metabolites etc. In a single cell -omics experiment, the investigator no longer

uses a molecular marker to isolate the cell type of interest from its surrounding cell types. Instead, the investigator simply isolates individual cells regardless of cell type, and then performs -omics experiments to measure the abundance of all biological molecules of each individual cell. It is important to note that at this point the cell type identity of each cell remains unknown. To discover the cell type identity without relying on any prior knowledge, cells are clustered into groups based on their pair-wise similarity and their cell types inferred from their clustering. After the cell type identities are determined, candidate marker molecules are revealed by contrasting the abundance of biological molecules in the cell type of interest against that of other surrounding cell type(s).

It is desirable to measure all biological molecules in single cell -omics experiment. Unfortunately, for those biological molecules that are not nucleic acids, the current technology is only able to detect biological molecules that are of high abundances. One example of such technology is mass spectrometry. Because most biological molecules are not present in one cell at sufficient quantities, mass spectrometry cannot detect their abundances in single cell samples. Furthermore, many types of biological molecules have overlapping mass spectra, making them not discernible by the current mass spectrometry technology. For nucleic acid molecules, the current technology is dramatically different. One is able to economically and conveniently estimate the abundance of a large number of nucleic acid molecule species in a single cell sample using microarrays or using high throughput sequencing. Considering the relative convenience and lower cost of measuring nucleic acids, it would be desirable if

the need to measure abundances of all molecules can be fulfilled by only measuring the abundances of nucleic acids. For abundances of proteins, It has been shown that the abundance of mRNA can be used to approximate the abundance of the proteins that they encode, because the correlation between mRNA abundance and the abundance of protein it encodes ranges between 0.45-0.74 [4]. This wide range is due to many reasons such as the regulation of translational rate and degradation rate which are not accessible via the measurement of mRNA abundance. For abundances of other molecules, such as metabolites, there is no evidence suggesting that their abundances are correlated with the abundance of RNAs, hence their abundance has so far been ignored in single cell studies.

In summary, an ideal single cell -omics experiment that measures the set of all proteins, transcripts, metabolites and other molecules could identify novel cell types and their respective molecular markers. In the absence of such an ideal single cell -omics experiment, single cell transcriptomics that aim to measure all transcripts is a pragmatic compromise.

1.5 For performing transcriptomic measurements, RNA-Seq is preferable to microarrays

For performing transcriptomic experiments, both microarray and RNA-Seq platforms are able to achieve this economically. However, the RNA-Seq platform is preferable to the microarray platform for several reasons, including, but not limited to, the four reasons listed below. (i) The microarray platform rely on a fixed panel of probes that are pre-defined at the time of manufacture. This

prevents the investigation of un-annotated or incorrectly annotated genes and transcripts. On the other hand, the RNA-Seq platform is capable of recovering reads from the entire transcript body regardless of whether it is annotated, making it possible to determine and distinguish novel transcripts [5]. This was shown to be important because nearly all genes in the human genome are alternatively spliced [6], and it is highly likely that in any genome, many isoforms are erroneously annotated or still not annotated yet. (ii) With RNA-Seq, if the coverage over a specific transcript is sufficiently high and if the sequencing reads are tiled over the entire transcript, it is possible to assemble the full length transcript, even without gene annotation or a reference genome (see [7] and references therein). Consequently, RNA-Seq can for instance be applied to cancer samples to detect fusion genes created by genomic re-arrangements, or be used to analyse samples without a high-quality reference genome [8]. (iii) It has been shown that microarrays have a limited measurement dynamic range, which typically covers about three orders of magnitude. On the contrary, the dynamic range of RNA-Seq measurement is much wider, which typically covers about six to nine orders of magnitude [9]. (iv) The detection sensitivity of RNA-Seq is higher than that of microarrays, enabling the detection of lowly abundant gene expression. These technical benefits have made RNA-Seq a technology of choice in many biological research studies, including in single cell transcriptomics experiments.

1.6 Some difficulties in performing single cell RNA-Seq have been recently resolved

The amount of DNA required for a high throughput sequencing experiment was about 500ng. The amount of mRNA present in a typical diploid human cell is about 10pg [10] (one pg is one trillionth of a gram; 500ng = 10pg x 50000). Traditionally, transcriptomics experiments such as RNA-Seq experiments were performed using bulk samples. The reason that RNA-Seq has not been successfully performed on single cells is related to the inability of generating 500ng of cDNA from 10pg of mRNA, while maintaining an accurate representation of the underlying mRNA. Because amplifying cDNA typically uses PCR based exponential amplification methods, each round of PCR amplification is biased towards particular cDNA fragments. If the amplification step requires too many rounds of PCR amplification, the amplified cDNA fragments does not re-capitulate the starting mRNAs because the bias is compounded multiplicatively. Thus, it is not possible to attain 500ng of cDNA from a typical single cell [11] using conventional (i.e. bulk sample) RNA-Seq protocols. As a result, single cell RNA-Seq was not previously possible, despite traditional RNA-Seq being successfully performed by various labs for years.

Recently, there have been several breakthroughs in RNA-Seq experimental protocols that enabled sufficient amount of cDNA to be derived from a single cell that accurately represent the starting mRNA. The most important one being the efficiency of reverse transcription from negligible to about 5% [12-14], which is then further increase to about 40 percent[15, 16]. Other breakthroughs include library preparation protocols, such as a transposase based protocol (e.g. Nextera

from Illumina) that yields higher amount of sequencing-suitable DNA fragments relative to traditional sonication based library preparation protocols [17]. Furthermore, breakthroughs with microfluidic devices made parallel sample handling, and automated reaction in micro-volumes possible. Such microfluidic devices enabled single cell RNA-Seq experiments to be performed conveniently for hundreds to thousands of cells in parallel. For example, the Fluidigm C1 system enabled up to 96 cells to be sequenced in parallel in one experiment [18]; the Fluidigm C1-800 system increased the number of cells to 800. More recently, a micro-droplet based method described in [19] supplied by the company 10X Genomics, further enabled single cell RNA-Seq to be performed on about 1000-8000 cells in parallel in a single experiment. These breakthroughs together not only made single cell RNA-Seq possible but also made it able to be conveniently performed on hundreds to thousands of single cells in parallel. Furthermore, because commercially available single cell RNA-Seq kits and microfluidic systems made the adoption easy, many researcher are now able to perform single cell RNA-Seq experiments in their own labs.

The recently available single cell RNA-Seq technology has been exploited for making a plethora of biological discoveries. For example, the discovery of novel cell types [20], the identification of cell-to-cell heterogeneity among a pool of otherwise homogeneous cells cultured in 2i medium and alt-2i medium [21] , the discovery of key genes controlling the cell cycle in different cell types [22, 23] and the immune response to external threat where not all immune cells respond

in exactly the same manner [24]. Further examples are elaborated in detail in the sections below.

1.7 Single cell RNA-Seq has been applied to the identification of novel cell types

Traditionally, cell type identification has been based on cellular morphology or molecular markers. In the absence of these, as explained previously, the discovery of novel cell types is prevented. Currently, our understanding of the cell types that are present in most multi-cellular organism is extremely limited. Even for human– a well studied organism – we do not yet know exact number of cell types within it. In primary school education material [25], there is indeed an estimate of about 200 cell types in a human body. However, this estimate is likely far from the truth. So, if a primary school pupil enquires more deeply and asks for a thorough catalogue of all human cell types, cell biologists can only respond that this question is already beyond the boundary of human knowledge. This lack of knowledge is surprising, given that we understand items as minuscule as an atom, and as far as distant galaxies; this lack of knowledge is troubling, given that it severely limits the progress of biological research and development of medical treatments. Thus, discovering novel cell types, which in part answers our inquisitive pupil's question, will provide tremendous benefit in advancing many aspects of biological research. Here we use one cell type – the *neuron*, out of the aforementioned "about 200 cell types" – as an example, to describe the severe lack of knowledge and over-simplification that currently prevail in cataloguing cell types. It is technically

correct to consider every cell that transmits nerve impulses as one single cell type: the *neuron*. However, it is agreed that there is tremendous diversity among them. It is well documented that different neurons exhibit striking differences in cell size, shape, axon length, dendrite count, impulse firing patterns, secretion and response of signalling molecules such as dopamine, serotonin, GABA, and glutamate. Studies have attempted to catalogue neurons into motor neurons, cortical neurons, interneurons, dopaminergic neurons, GABAergic neurons and increasing more refined classifications. However, very importantly, a recurring theme that most such studies emphasize is that much of the complex diversity among neurons is probably as yet unknown [26, 27]. Likewise, many studies also attempted to discover other novel cell types in the human body. However, the *a priori* discovery of novel cell types was impeded. When single cell RNA-Seq became available, we were able to overcome the difficulties as explained in Section 4 of this chapter. This approach has been successfully applied in several biological contexts, and has discovered many novel cell types that have never been described before, as I will describe in the examples below.

Mouse retina is a complex tissue containing many cell types. Decades of research based on morphology identified dozens of different cell types [28, 29]. However, not all cell types are distinguishable based on morphology or have a known molecular marker. As a result, it was possible that there were some cell types within the mouse retina that were not yet described. For the purpose of identifying novel cell types in mouse retina using an entirely objective approach, single cell RNA-Seq was applied to individual cells dissociated from it [30]. To

briefly describe the experiment: firstly, individual cells are encapsulated into individual nanoliter-scale aqueous compartments (i.e. droplets) that are isolated from each other in an oil solution. Secondly, a bead is injected into each droplet. The bead is connected to reverse transcription primers that encode a barcode sequence unique to each bead. Subsequently, in each isolated droplet, the cell is lysed and releases mRNA for reverse transcription to occur with the barcoded primer on the bead. As a result, each reverse transcribed cDNA contains the unique barcode of the beads within the same droplet. Thirdly, the droplets are then broken and merged together. All remaining steps of the experiment are performed on the merged solution. Because each reverse transcribed cDNA is barcoded, they can be traced back to a specific bead with which it co-localised within the same droplet during reverse transcription. Finally, all barcoded cDNA are amplified and fragmented into appropriate lengths for high throughput short read sequencing. The authors performed droplet based single cell RNA-Seq on a total of 49,300 single cells from seven different experimental batches. Using the previously described *a priori* cell type identification approach, this study identified 39 distinct cell types. For previously known cell types with known marker genes, the authors were able to examine the expression of these genes in the 39 cell types they identified. Of the 39 cell types, 12 matched a previously described cell type based on the expression of these cell type-specific marker genes. The ability to match up with known cell types, I argue, validates single cell RNA-Seq based *a priori* cell type identification approach. An important contribution of this study is that these known cell types now can be associated with many additional candidate molecular markers. A second important

contribution of this study is the discovery of novel cell types and their candidate markers. The authors also validated some of these markers using protein staining experiments. With cell type-specific molecular markers, the function of these novel cell types can be examined by further experiments. This study demonstrated the utility of single cell RNA-Seq in identifying novel cell types, and provided putative cell type-specific markers for many cell types in the mouse retina.

Single cell RNA-Seq was also applied to *in vitro* cultured cells, such as mouse Embryonic Stem Cells (mouse ESCs), to investigate their heterogeneity. Previously, mouse ESCs grown on serum medium were shown to consist of heterogeneous cells of diverse expression patterns [31, 32]. Because the NANOG protein is an important transcription factor of the core pluripotency network [33], it is very interesting that its expression is heterogeneous among mouse ESCs [34, 35]. The varying levels of NANOG protein suggest that mouse ESCs may consist of heterogenious population of cells in different cellular states despite being cultured in the same dish under the same condition. Further evidence shows that the varying levels of NANOG protein are correlated with varying levels of self-renewal efficiencies [33].. Additionally, studies have shown that some cells among mouse ESCs resemble cells in a two-cell-stage mouse embryo (i.e. 2C-like cells) [36]. The abovementioned heterogeneity of mouse ESCs was discovered by measuring a small number of relevant molecules. Further studies to re-examine this heterogeneity based on the complete transcriptome using single cell RNA-Seq could be more revealing.

Indeed, recently, Dr. Aleksandra A. Kolodziejczyk and her collaborators performed single cell RNA-Seq on mouse ESCs cultured in three different media: serum, 2i and alternative-2i [21]. They profiled a total of 703 single cells using the Fluidigm C1 system. Because each Fluidigm C1 experiment is only able to capture a maximum of 96 cells, the 703 cells were generated from 9 different experimental batches. In this study, using the *a priori* cell type identification methods described previously, the mouse ESCs in the serum medium were grouped into 3 clusters. The authors named the 3 clusters as: cells that are differentiated, cells that are permissive to differentiation, and cells that highly express pluripotent genes. Because these three clusters were assigned a cluster name subjectively based on the expression values of several known marker genes, one needs to refrain from over-interpreting these cluster names. This study nevertheless convincingly demonstrated that mouse ESCs cultured in serum exhibit heterogeneous transcriptomic states. When examining the three clusters, the authors discovered that the expression levels of *Nanog* are different among the 3 clusters of cells, recapitulating previous studies [35, 37-39]. Likewise, the heterogeneous expression pattern for key pluripotent genes, such as *Esrrb* [40] and *Zfp42* [41], are also recapitulated in this study. However, this study falls short from examining what proportions of the observed heterogeneity of *Nanog, Esrrb* and *Zfp42* is due to the technical noise of single cell RNA-Seq or due to underlying biological variation, because the observed heterogeneity is the sum of the two. After all, if any gene in any single cell RNA-Seq dataset is examined, its expression level will be heterogeneous regardless of whether biological heterogeneity is the underlying cause.

This study also identified that 10 of the 295 mouse ESCs cultured in 2i medium express the following genes highly: *Zscan4b, Zscan4c, Zscan4d, Zscan4e, Zscan4f, Tmem92, Pramef25, Zfp352, Usp17la, Usp17lb, Usp17lc* and *Usp17le*. The expression levels of these genes were shown to mark a rare population of cells (previously named as 2C-like cells) that are distinct from other mouse ESCs. This study validated previous results that mRNA abundances of these genes are heterogeneous among mouse ESCs. This study also demonstrated that these genes are co-expressed in a small number of cells. However, the PCA results obtained by the orginal authors showed evidences that contradict with this claim. As seen from the original figure in the published paper reproduced in Figure 1-1. We can see that all cultured cells (marked by the black ellipse), regardless of the culture medium, are more similar to *in vivo* mouse blastocyst cells (marked by the red circle), and are less similar to genuine 2-cell-stage cells (marked by the cyan ellipse). Assuming there are any *in vivo* "2C-like cells" among the cultured mouse ESCs, they should occupy a position that is near the *in vivo* 2-cell-stage cells. Perhaps, these so-named “2C-like cells” do not correspond to the *in vivo* 2C cells at all, despite a set of pre-selected genes are more highly expressed. In addition to the biological discoveries, the authors also proposed a novel analysis strategy to identify heterogeneously expressed and co-regulated genes in single cell RNA-Seq dataset. Based on gene-to-gene expression correlations, the authors discovered 2 groups of genes: one group known to be important for pluripotency, the other known to be important for differentiation. This study again demonstrated an ability to use single cell RNA-Seq followed by

bioinformatic analysis to uncover heterogeneous groups of cells without needing to know cell type specific marker genes beforehand.

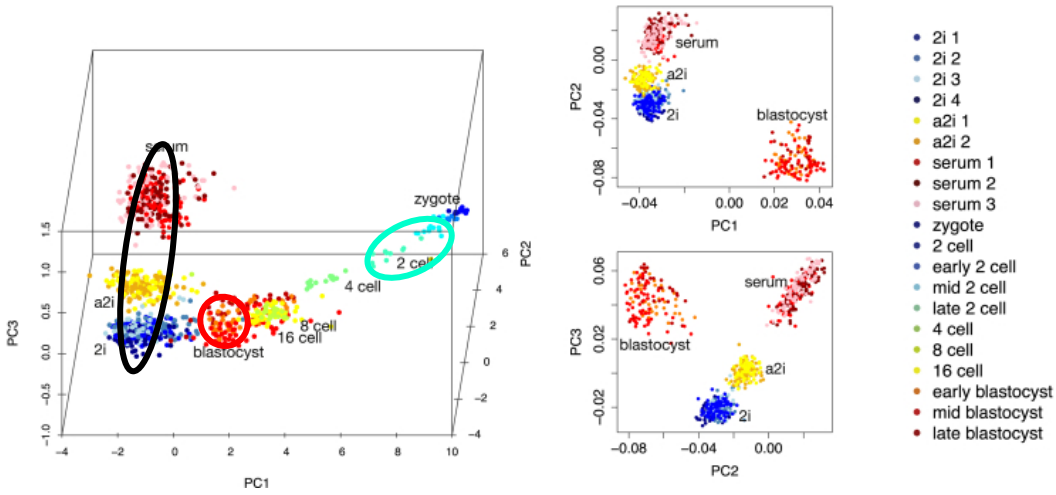


Figure 1-1: Principal Component Analysis by the original authors. [21] This figure shows the first three principal components of PCA analyses obtained by the original authors, which were used to indicate similarities of various *in vivo* cell types in an embryo to their cultured mouse ESCs. Figure reproduced from the original paper [21], which is published under the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/).

Other than applications to *in vivo* tissues and *in vitro* monolayer cell cultures, Single cell RNA-Seq may also be applied to *in vitro* organoid culture. Organoids are *in vitro* structures obtained using special culturing methods similar to those initially described in 2008 [42]. This culturing method is different from traditional monolayer *in vitro* cell culture. Because it better mimics the 3D spatial organisation of cells during culturing, the resultant organoids supposedly better resemble the corresponding *in vivo* organ. For the same reason, the resultant organoids supposedly contain all the constituent cell lineages of the corresponding *in vivo* organ, arranged in correct 3D relative positions. Recently, single cell RNA-Seq was performed on intestinal organoids [20]. This study set out to identify known and novel cell types located in the intestinal crypt organoid. There is an intrinsic difficulty in analyzing intestinal crypt organoid: among entrocyte, goblet, tuft, paneth and enteroendocrine cells in intestinal

crypt, the entrocyte cells are extremely abundant relative to other cell types. As a result, capturing cells pseudo-randomly from an un-purified population of cells dissociated from the intestinal crypt organoid will consist of mainly entrocyte cells and very few cells of other cell types. This makes single cell RNA-Seq experimental design and data analysis more challenging. For example, because enteroendocrine cells are relatively rare, a randomly captured population will contain few of them. This issue prevent a detailed investigation on enteroendocrine cells. The authors addressed this issue by taking the following iterative approach. They first performed one round of single-cell RNA-Seq without sorting cells based on any marker. Using this initial dataset, the authors identified that *Reg4* is potentially a good marker gene that marks all enteroendocrine cells. The authors then performed cell sorting by only retaining cells that express the REG4 protein. A second iteration of single cell RNA-Seq was then performed using the cells enriched for enteroendocrine based on REG4 expression. In total, 161 cells were randomly captured during the second iteration of single cell RNA-Seq experiment. Using the second iteration dataset, the authors determined that 60 of the 161 cells were enteroendocrine cells. Among the 60 enteroendocrine cells, the authors identified two subpopulations that were *Chga*- and *Chga*+ respectively. Within each of the *Chga*- and *Chga*+ subpopulations of enteroendocrine cells, the authors further identified multiple further refined subtypes, and their respective marker genes. The various subtypes and their marker genes were validated using transgenic mice and immuno-fluorescence experiments. Likewise, the authors examined heterogeneity among progeny of cells that had expressed LGR5 protein during

development. The experimental designs and analysis strategies taken in this study effectively allowed the experiment to iteratively zoom into the relatively rare population of enteroendocrine cells, instead of wasting resource on abundant cell types.

A common theme in the above examples is that cell type specific markers were not available to distinguish these cell types. Instead, the novel cell types and their marker genes were revealed from analysing single cell RNA-Seq dataset. These examples highlighted the capability of single cell RNA-Seq approaches for discovering novel cell types *a priori* (i.e. not requiring any biological knowledge of novel cell types beforehand). It is important to acknowledge that finding novel cell types in itself do not provide much biological insight, unless the properties and functions of the novel cell types can be discovered. As a result, these studies later draw support from existing knowledge and experimentation to obtain biological insights. For example, the anatomically localisation of the novel cell type, or the protein function its marker genes were examined to reveal the function of the novel cell types. These goals and approaches will also constitute a major part of the mouse midbrain dopaminergic neuron single cell RNA-Seq analysis project described in Chapter 4.

1.8 Single cell RNA-Seq is useful for quantification of biological variation

Other than discovering novel cell types, single cell RNA-Seq can be applied to uncover many other interesting aspects of biology. The expression levels of a gene in different cells usually vary, even among cells of the same cell type or the

same cell line. The heterogeneous expression of a gene in the same cell type could be due to differences in cell state (e.g. cell cycle state) or due to random activation of gene expression [43]. Although some of the heterogeneity in gene expression could be deleterious and thus is biologically buffered away [44], some could potentially play a functional role in cellular regulation [45].

Single cell RNA-Seq enables the expression of the same gene to be measured in multiple single cells. However, it is important to note that any measurement incurs measurement error. Consequently, heterogeneity in the expression levels observed in a single cell RNA-Seq experiment does not necessarily reflect heterogeneity in transcript abundances. To quantify heterogeneity of transcript abundance requires the ability to disentangle the perceived heterogeneity introduced by measurement error (i.e. technological noise) and the underlying biological heterogeneity. Innovative experimental designs have made good use of external spike-in mRNA molecules of known concentrations to address this issue. Because the spike-in molecules were added at exactly the same concentrations, any variance present in the estimated expression abundance of spike-in molecules is necessarily due to technical noise. As a result, it is possible to gauge the level of variation arise due to technical noise. By discounting the technical variation from the total observed variation, it is then possible to estimate the biological variation. This innovation enabled the identification of heterogeneously expressed genes caused by the underlying heterogeneity of transcript abundance, rather than caused by measurement errors prevalent among single cell RNA-Seq datasets [46].

It is important to note that a commonly used set of external spike-in mRNA molecules is the ERCC spike-in. It contains 92 different poly-adenylated RNAs generated largely from random unique sequences [47]. The different transcripts are present at different concentrations covering more than 6 orders of magnitudes [48]. The ERCC spike-in was initially intended for assessing the accuracy of micro-array measurements. Employing ERCC spike-ins for assessing technical noise present in single cell RNA-Seq protocol is an innovative repurposing. There are three main benefits of using ERCC as spike-ins for single cell RNA-Seq. (i) The ground truth of absolute molecule number is known for each mRNA species, which allowed relative transcript abundance of a cell measured by single cell RNA-Seq to be converted into absolute transcript abundance. (ii) The ground truth of molecule number is useful to assess the accuracy and bias of single cell RNA-Seq measurements. (iii) The wide range of cDNA concentration of different mRNA species is useful to assess the detection sensitivity at a wide range of concentrations including the extreme low-concentration transcripts. However, there are some disadvantages of using the ERCC spike-in, for example the lengths of poly-adenylated tails of ERCC transcripts are shorter than the average length of poly-adenylated tails of mammalian genes. Additionally, the transcript lengths of ERCC transcripts are also shorter than the average length of mammalian transcripts. As a result, it can be desirable to design a new set of RNA spike-in transcripts that are specifically engineered for assessing single cell RNA-Seq performances. The spike-ins shall be designed to assess technological variation, transcript coverage bias, detection sensitivity and other aspects that are important for single cell RNA-Seq. This

would enable the assessments of biological variations in single cell RNA-Seq dataset to be investigated in greater detail.

1.9 Single cell RNA-Seq is useful for assessing the effects of cell cycle and differentiation on the transcriptome

Transcriptomic variation across different stages of a cell cycle has been studied using bulk samples consisting of synchronised cells. Transcriptomic variation over cellular differentiation has been studied using bulk samples consisting of time series data generated during the course of differentiation. These studies are less precise because cells cannot be perfectly synchronised in a cell cycle or during differentiation. The ability to individually profile the transcriptome of a large number of cells allows these variations to be investigated at single cell precision. To this end, analysis tools have been developed recently that aim to disentangle various components of transcriptome variation. The components of variability include technical noise induced variability as explained in Section 1.8 above, and biological variability such as cell cycle difference induced variability and cell differentiation induced variability [23]. The method requires a set of genes to be pre-selected by the user according to existing literature (e.g. all genes that have been associated with cell cycle). Based on the single cell RNA-Seq data, the method identifies expression variation patterns that are shared among this pre-selected set of genes, and remove (i.e. regress out) this shared expression pattern. As an example, the set of pre-selected genes can be all cell cycle related genes, and variation patterns shared among cell cycle genes are thought to reflect variation induced by cell cycle. This

variation is then regressed out before downstream analyses. However, this method was criticized that among the variation being regressed out, the majority is due to data quality, library size and other confounding factors [49]. These methods and criticisms were published after the completion of my analyses. As a result, this method was not applied to my analyses.

A central claim of the above method is that the transcriptome varies tremendously at different stages of a cell cycle [23]. This naturally invites the further question of whether it is possible to predict the cell cycle stage based on a cell's transcriptome alone. One such approach works by: first, limiting analyses to cell cycle related genes annotated in cycle-base [50]; second, examining the similarity of this cell being investigated to a reference panel of cells of known cell cycle stage [22]; thirdly, assigning the cell cycle stage of the reference panel of cells having the highest similarity to the cell being investigated as its predicted cell cycle stage. The authors performed various tests and concluded that this method performs well only when the training and prediction phases are performed on cells of the same cell type [22]. Consequently, the authors concluded that cell cycle signatures can vary strongly between different cell types. This conclusion raises the following important question: does a universal cell cycle signature exist at all? Indeed, another report quantitatively affirmed this observation, by showing that different sets of genes were involved in the cell cycle of different cell types [51].

1.10 Introducing the single cell RNA-Seq protocols and platforms being used for my projects

There are many popular single cell capture and library preparation protocols for performing single cell RNA-Seq experiments, such as SMART-Seq1 [52], STRT-Seq [53], CEL-Seq [54], SMART-Seq2 [16], MARS-Seq [55], SCRB-Seq [56], In-Drop [19] and Drop-Seq [30]. A comparison and analyses of the various experimental protocols have been reviewed recently [57], and are not within the scope of this thesis. When my projects started, the most accessible protocol to the general research community users was the SMART-Seq1 protocol performed on the Fluidigm C1 system, because relevant reagent kits were then commercially available. For this reason, my laboratory collaborators for both the human ESC project (Chapter 3) and midbrain dopaminergic neuron project (Chapter 4) chose to use SMART-Seq1 to generate all single cell RNA-Seq data.

In addition to the choice of various protocols, there are also several microfluidic platforms available for performing single cell RNA-Seq on hundreds or thousands cells in parallel. At the time of performing the experiments, only the Fluidigm C1 system was commercially available and convenient to setup and operate. For this reason, the single cell RNA-Seq data for the projects described in Chapter 3 and Chapter 4 were generated using the Fluidigm C1 system. The Fluidigm C1 system captures single cells automatically and prepares cDNA from single cells as illustrated in Figure 1-2. The Fluidigm C1 system used in my projects can prepare the cDNA libraries for up to 96 separate cells in one experiment, which takes less than one day. This throughput influenced the total number of cells being sequenced in these two projects. After both my projects started, a newer version of the Fluidigm C1 microchip enabled the throughput to

be performed on 800 cells per experiment. However, in order to avoid a batch effect caused by switching to a different microchip, all projects continued to use the same protocol and same version of the microchip.

Below I will briefly introduce the experimental procedure of using the Fluidigm C1 system to generate cDNA. The first step is to provide several thousand dissociated cells in solution. The solution is then pumped to travel across an array of capture chambers connected in series. When a cell in solution enters a specific empty capture chamber, it may be captured, in which case the capture chamber is intended to lose the ability to capture a second cell. Alternatively, a cell in solution may also enter an empty capture chamber, and pass through without being captured, in which case the specific capture chamber remains empty and is available to capture the next incoming cell. Each capture chamber is designed to capture one cell, although, as we will discuss in Section 2.6 this is not always true. After the cell capture stage, the Fluidigm C1 system typically performs the SMART-Seq1 protocol [52] within the reaction chambers on the microchip and generates 96 separate cDNA samples. The 96 cDNA samples are then individually barcoded using 96 distinct Illumina sequencing barcodes. After barcoding, the cDNA samples are mixed together, amplified and prepared for short read high throughput sequencing by Illumina. The barcoded sequencing results are subsequently computationally de-multiplexed into 96 separate RNA-Seq datasets for bioinformatics analysis.



Figure 1-2: An illustration of the workflow using the Fluidigm C1 System. This system can be used to perform single cell RNA-Seq with the SMART-Seq1 protocol or other protocols[58]. The steps depicted here are the main steps performed by the Fluidigm C1 system: cell capture, wash and mRNA reverse transcription. The reverse transcribed libraries are then transferred manually to a plate, bar-coded and amplified for sequencing.

1.11 Structure of the thesis

This chapter has introduced the relevance of single cell measurements, single cell –omics measurements and single cell RNA-Seq for biological research. This chapter also presented several examples of applying single cell RNA-Seq to discover novel biological insights and introduced single cell RNA-Seq protocols and platforms that were employed for my projects. Chapter 2 discusses the challenges of sequencing experiments and single cell RNA-Seq experiments. Single cell RNA-Seq experiments are complex and their associated data analyses are challenging. The methods to analyze these datasets are a rapidly developing area of interdisciplinary research of statistics and molecular biology. This large

body of recent literature and developments of single cell RNA-Seq dataset analysis is thus covered systematically in this chapter. With the introduction of single cell RNA-Seq technology in Chapter 1, and the introduction of data analyses in Chapter 2, I have covered the background for all the research projects that will be described in my thesis. Subsequently, I described a single cell RNA-Seq analysis project analyzing human embryonic stem cells in Chapter 3. This is followed by describing a mouse dopaminergic neuron single cell RNA-Seq analyses project in Chapter 4. The materials and methods of each of the research project chapter are presented within each respective chapter. Finally, Chapter 5 summarizes the contribution of my research and suggests potential future research directions.

1.12 Summary

In this chapter, I described the disadvantages of bulk sample measurements and argued that it is important to measure abundances of biological molecules (i.e. proteins, nucleic acids metabolites etc.) in single cells. I then described non-omics measurement methods that require pre-selection of measurement targets to reveal their benefits and disadvantages when marker molecules are not known. Due to these disadvantages, I thus argued that it is important to perform –omics measurements in single cells for many biological studies: for example, identifying novel cell types in the absence of cell type specific markers. Based on these arguments, the utility of single cell –omics experiments thus became clear. Among various potential single cell –omics experiments, only single cell transcriptomics is feasible. Single cell RNA-Seq

technology recently became widely available to the research community because of several technological breakthroughs. I described several published applications of single cell RNA-Seq that were described to exemplify the utility of single cell RNA-Seq to discover novel biology. In these single cell RNA-Seq projects, although the biological system and objectives are not the same, the data analyses objectives are highly similar. The objectives are typically identifying novel cell types and heterogeneity among a population of cells and the cell type specific markers. I then introduced single cell RNA-Seq protocols and the microfluidic platform used in the projects of my thesis.

Chapter 2 Single Cell RNA-Seq Experiments and Associated Data Analyses are Challenging

In Chapter 1, I described the utility of single cell RNA-Seq in many biological research contexts, using a theoretical analysis and several published examples. Single cell RNA-Seq is also the main approach that was used to accomplish each of the research projects to be described in later chapters of this thesis. Single cell RNA-Seq is based on high throughput sequencing technology. This is a complex technology and its associated data analysis is challenging. Therefore, in this chapter I will introduce challenges of high throughput sequencing, single cell RNA-Seq and their data analyses. I will intentionally place an emphasis on the protocols and microfluidic platforms that were applied in my projects. It is worth noting that most single cell RNA-Seq datasets generated by other protocols and other platforms also share these characteristics and analytical challenges [57], thus the discussion in this chapter is also generally applicable to other single cell RNA-Seq datasets.

In this chapter, I start by providing an overview of high throughput sequencing, RNA-Seq and single cell RNA-Seq. These provide the background information necessary to understand the basic principles of these approaches. In the subsequent sections, I will briefly introduce the special characteristics, capabilities, limitations and challenges related to data analyses of high throughput sequencing, RNA-Seq and single cell RNA-Seq.

2.1 High throughput sequencing and its associated data analysis can be complex

High throughput sequencing is a relatively recent technology providing the ability to read the sequence (i.e. to sequence) DNA fragments in massive quantities. A popular example is the Illumina sequencing technology [59]. In one experiment, the Illumina technology typically is able to sequence millions to billions of DNA fragments simultaneously. Illumina sequencing requires fragments to be approximately 50-500 nucleotides long (not including adapters) and is able to read 36-150 nucleotides from either or both ends of the DNA fragments. Once the sequences of the short fragments are determined, they are saved in a file - typically in the *fastq* format. Because the sequencing procedure can introduce errors, each nucleotide in a fastq file is annotated with a quality score (i.e. a number indicating the confidence level of that nucleotide being correct). Other than Illumina sequencing technology, there are several other high throughput sequencing technologies, such as, Ion Torrent, SOLiD or Pacific Biosciences. More recently, Oxford Nanopore technology also became available. Oxford Nanopore has lower throughput, but the size of its sequencing device is only as big as two fingers. Although the underlying sequencing mechanisms of these technologies differ substantially, the outcomes of sequencing (i.e. the data files generated by their sequencing machines) are very similar. Therefore, the initial steps of processing, such as quality checking and short read mapping, can be achieved by using similar bioinformatics concepts and software tools.

With high throughput sequencing technologies, a large number of DNA fragments can be sequenced rapidly and economically. For example, Illumina

HiSeq 4000 is able to generate 250 million reads per lane. Sequencing takes about only 3 days and costs less than 2000 US dollars per lane. If any biological quantities (e.g. transcription factor binding, mRNA abundance) can be converted to DNA quantity, they can then be interrogated rapidly and economically using high throughput sequencing. This also extends the capability of an investigator, because once DNA fragments are generated, the remaining steps of sequencing experiments are nearly identical. As a result, many areas of biological research started to employ high throughput sequencing technology. For example, in the past, investigation of transcription factor binding onto DNA could be performed for only a handful of DNA loci due to the prohibitive cost and time required to repeat the same experiment manually for the entire genome. Chromatin immuno-precipitation followed by high throughput sequencing (i.e. ChIP-Seq) enabled transcription factor binding to be interrogated at genome-wide scale with reasonable effort and cost. In another example, RNA reverse transcription followed by high throughput sequencing (i.e. RNA-Seq) is designed to quantify almost the entire transcriptome of samples. It is achieved by first converting RNA to complementary DNA (i.e. cDNA), then by sequencing the cDNA. The abundance of the cDNA thus reflects the abundance of the corresponding RNA in the original sample. Although measuring the abundances of mRNA can be achieved by using microarrays with reasonable effort and cost, RNA-Seq has comparative advantage what was discussed in Chapter 1 Section 5 above.

Despite of these advantages, all experiments making use of short read high throughput sequencing suffer from an intrinsic disadvantage caused by short

read mapping ambiguities. All short read sequencing experiments requires a mapping step to infer the genomic coordinates of each sequenced short fragment. This mapping step is required because short reads do not retain the genomic coordinate information relating to where the DNA or RNA originated, the only information available in the short reads is the nucleotide sequence itself. This mapping step is accomplished by searching an existing reference genome (like searching a map) for a location that contains the (near) exact nucleotide sequence of the short fragment. Thus this process is frequently referred to as *mapping*. The first disadvantage of the mapping step is that it can be inaccurate. The inaccuracies arise when sequencing errors are present in the short read, or when the reference genome is an imperfect representation of the actual genome (e.g. due to incompleteness, assembly errors etc.). The second disadvantage of the mapping step relates to the two alleles of a diploid organism. The accurate diploid reference genome of the individual in question is usually not available. The publically available reference genome assembly is only one linear sequence (i.e. a haploid reference genome) that simplifies the diploid genome. As a result, for any given read, it is impossible to disambiguate whether the reads arose from the paternal or the maternal chromosome. The third disadvantage of the mapping step - aside from sequence errors, assembly imperfections and allelic ambiguity - is that many regions in the reference genome are identical to one another (i.e. repeated). Transposable elements, centomeres or telomeres are examples of such repeated regions. In such cases, the short fragment arise from any of these repeated regions can be mapped to every other instances. No information exists to disambiguate among the

alternatives. Some of these issues can be mitigated. The mitigating approaches include adopting a longer read length, such as sequencing using PacBio and Nanopore technologies, and adopting a more realistic reference genome, such as a graphical representation of the reference genome [60]. Improvements to high throughput sequencing data analysis are an active research area.

In summary, short read high throughput sequencing has provided significant cost reduction and flexibility for many molecular assays (i.e. ChIP-Seq, RNA-Seq, DNase-Seq etc). However, the erroneous or ambiguous mapping issues are not yet fully resolved.

2.2 RNA-Seq and its associated data analysis can be complex

RNA-Seq converts the quantity of interest – RNA – to cDNA. All subsequent steps of the experiment are performed according to standard high throughput sequencing. Therefore, its data analysis inherits all the complexities and issues of high throughput sequencing data analysis. Furthermore, the cDNA generated from RNA adds another set of complexities and issues that need to be considered during data analysis. Five aspects of RNA-Seq data analytical issues will be elaborated below: short read mapping, gene annotation set, transcript abundance estimation, experimental biases and losing strand information.

Short read mapping in any high throughput sequencing experiments is prone to errors. In addition to those issues common to high throughput sequencing, short read mapping for RNA-Seq faces several additional challenges and complexities. Firstly, because mature RNAs underwent splicing, they only

contain the concatenated exonic sequences, which are not necessarily the same as the contiguous sequences present in a linear reference genome. The simple tools developed for mapping genomic short reads onto the reference genome are no longer suitable for RNA-Seq datasets, because these tools will mistaken any fragment that underwent splicing as a genomic deletion (i.e. the intron is considered by the simple mapping tool as deleted from the reference genome). Secondly, many genes may have pseudogenes or paralogous genes that exhibit near-identical sequences. These regions may be mapped ambiguously, thereby causing inaccuracy of expression quantification. This phenomenon was studied in detail in [61] by performing simulations. Thirdly, In order to enhance results, many paired-end short read mapping algorithms need to estimate fragment size accurately. In the case of RNA-Seq, many genes have several isoforms. The observed fragment may originate from any of the possible isoforms. Therefore, estimating cDNA fragment size in RNA-Seq data is tremendously more complicated. To address the three issues related to mapping as mentioned above, a set of specialized mapping tools is employed to identify correct mapping locations for RNA-Seq reads. A few popular examples are: Tophat [62], MapSplice[63], PALMapper [64], GSNAP[65], STAR [66] and OSA [67]. Most of these offer two different modes of mapping: (i) unassisted mode and (ii) gene annotation set assisted mode. In the unassisted mode, the tools are designed to initially mapping a small part of the short read, followed by extending the mapped region while ensuring that splice sites are well tolerated during extension. In the gene annotation assisted mode, the mapping is typically designed to proceed in two separate stages. During the first stage, the algorithm

constructs a set of reference transcriptome sequences using the externally provided gene annotation set, and attempts to map all the short reads onto the reference transcriptome sequence. The reference transcriptome sequence is much smaller than the reference genome. As a result, mapping one short read using the annotation assisted mode is much faster than using the unassisted mode. During the second stage, any remaining reads that were not mapped during the first stage are now mapped to the reference genomic sequence as if using the unassisted mode. This is to ensure that un-annotated genes are not entirely dropped from the mapping. A detailed comparison and discussion of the existing methods and the remaining unresolved issues of RNA-Seq read mapping are reviewed elsewhere [68].

In the above paragraph about the complexities during short read mapping, I have alluded to a gene annotation set. Here, I discuss the utilities and complexities relating to gene annotation sets for RNA-Seq data analyses. The assisted mode of RNA-Seq read mapping utilizes the location of exonic sequences, splice junctions and isoforms in the gene annotation set. The provision of a gene annotation set is important for the mapping step, because studies have shown that without the annotation set, up to half of the reads that are mapped to splice junctions may be mapped entirely differently [69]. However, there are many different gene annotation sets available. The different annotation sets differ in several aspects such as information sources, annotation method, stringency and completeness. For human, available gene annotation sets include Ensembl [70], RefSeq [71], UCSC gene set [72], AceView [73],

Vega[74] and GENCODE [75]. The HANAVA gene set only include genes that are supported by experimental sequence evidence, and do not include genes predicted solely by *ab initio* methods [70]. The Ensembl gene set consists of two types of genes: HANAVA genes previously described and other genes that are predicted using the Ensembl gene building pipeline. The RefSeq gene set represents a collection of validated sequences based on the sequence data supplied by original authors in published literature [71]. The UCSC gene set is based on protein information from Swiss-Prot/TrEMBL (UniProt) and mRNA information from GenBank. The Vega gene set contains manually curated transcripts produced by the HAVANA group at the Wellcome Trust Sanger Institute. The Ensembl gene set contains all Vega genes. Since Ensembl 56, the GENCODE gene set is the same as the Ensembl genes set. AceView is a non-redundant curated representation of available human cDNA sequences. Different annotations can be selected for different analysis aims. For example, if the aim is to cover a stringent set of genes that have been validated experimentally, RefSeq or UCSC genes can be selected. If the aim is to investigate as many genes as possible, regardless of whether the gene sequence is validated experimentally, then the Ensembl genes set which contains both validated and predicted genes can be selected. For all the projects presented in this thesis, I selected the Ensembl gene set. This is because the purpose of identifying cell-to-cell heterogeneity, considering as many genes as possible will provide more information during downstream analyses. This argument is particularly true when the sample being investigated is not well studied, because less well-studied samples may express many genes that have not been experimentally

characterised. Therefore, choosing RefSeq in these cases would have discarded important information.

Once the short read mapping step with appropriate gene annotation is completed, the investigator can proceed to obtain the biological quantities of interest – mRNA transcript abundances. There are many ways of defining the expression abundances of a given gene. Popular examples include: (i) total read count overlapping all exons of the gene body, commonly referred to as “raw counts”, (ii) read count per million, commonly referred to as “CPM”, defined as the raw count divided by library size, then times 1 million, (iii) read count per kilo-bases per million reads, commonly referred to as “RPKM” [76], defined as the raw count divided by library size, then divided by transcript length, and then times 1 billion, (iv) transcripts per million, commonly referred to as “TPM” [77]. Among so many possible definitions, no single definition is suitable for all purposes. An appropriate definition shall be chosen according to the specific application. Three examples are provided below to illustrate the complexity of choices. (i) If we intend to apply certain analysis tools, its underlying statistical tests maybe designed to be only applicable for some definitions of expression but not for others. For example, it is important to provide the raw counts to DESeq2 [78] without applying any normalization beforehand. (ii) In an RNA-Seq experiment, for the same expression abundance, longer transcripts will generate more fragments than shorter transcripts. As a result, the raw counts are not suitable for comparisons between different transcripts within the same sample. In this case, it is important to scale the reads appropriately by the transcript

length, thus FPKM or TPM is more suitable. (iii) In an RNA-Seq experiment containing more than one sample, different RNA-Seq libraries would be of different sequencing depths. A more deeply sequenced library would exhibit greater counts than a shallowly sequenced library. In order to compare the expression of a same gene between different samples, it is important to scale the counts appropriately by the sequencing depth, so, CPM is more suitable. (iv) In a cross sample comparison study, since it is important to ensure that both within sample quantities and between sample quantities are comparable, FPKM or TPM is essential. In the projects described in this thesis, several of the above expression abundance measurements were employed appropriately for different analyses.

Experimental biases may affect the measurements obtained by RNA-Seq [79]. Thus it is important to account for these biases during data analyses [80]. It is acknowledged that many biases remain ill-defined. Below, I will briefly describe several kinds of better known biases. (i) In an RNA-Seq experiment, the measured cDNA was initially reverse transcribed from mRNA. For different transcripts, the efficiency of reverse transcription can be different. Such bias may be caused by different poly-adenylated tail length, RNA sequence composition or RNA secondary structure. As a result, the abundances of the cDNA initially generated from reverse transcription step are a biased representation of the abundances of the underlying RNAs. (ii) After cDNA fragments are generated, they are typically amplified using exponential amplification to reach sufficient DNA quantities required for high throughput sequencing. During exponential

amplification, a subtle preference in amplification is compounded multiplicatively as illustrated using a toy example below: a 10% amplification bias over 28-32 rounds of exponential amplification can result in a 14-21-fold apparent difference in abundance (1.1^{28} = 14.4; 1.1^{32} = 21.1), despite of the underlying abundances before amplification being equal. As a result, the amplified DNA library can be a highly biased representation of the corresponding initial cDNA. (iii) DNA sequencing platforms have their respective preferences over different DNA fragments [81, 82]. For example, Illumina sequencing is more efficient when sequencing fragments of about 200-500 nucleotides long [83]. The representations of fragments that are outside of this range are depleted – the further it is away from this range, the more severe is the depletion. Given these biases, of both known and unknown origin, it is important to assess whether RNA-Seq is able to provide a reasonable measurement of the underlying RNA quantity. One study demonstrated that the biases can be dramatic [79]. Therefore, conclusions drawn by RNA-Seq shall be validated using independent approaches such as qRT-PCR, Digital PCR, RNA *in situ* hybridisation and immuno-staining. For tackling the bioinformatics analyses challenges, methods can incorporate fragment size biases, sequence composition bias and transcript coverage bias to improve short read mapping [65, 84, 85].

A DNA double helix has two strands containing complementary sequences and RNA can be transcribed from either of these two strands. Before RNA is converted to double-stranded DNA, it is possible to determine the strand information. However, when RNA is converted to double stranded cDNA, two

different RNAs from different strands are converted to the same DNA molecule. Therefore, from the DNA sequence alone it is no longer possible to disambiguate which of the two possible RNAs provided the reverse transcription template. I would like to note that strand-specific RNA-Seq protocols are available. However, due to the additional complexity of applying stand-specific protocols, they are not commonly used in RNA-Seq experiments.

In summary, in any RNA-Seq experiment, the abundance of sequenced cDNA is not a perfect representation of the abundance of the RNAs from which they were reverse-transcribed. I described several sources of biases prevalent in RNA-Seq experiments, and highlighted the challenges to account for these biases during RNA-Seq data analyses.

It is interesting to mention that very recently, an *in situ* variant of RNA-Seq became available. This method is able to perform sequencing by synthesis, or sequencing by hybridisation, on fixed tissues *in situ*. This method provides tremendous benefits because it not only measures expression levels, but also reveals the 2D spatial relationships among the cells being measured [86]. Furthermore, RNA *in situ* hybridisation based technologies are also becoming available. They are able to estimate RNA abundance via successive rounds of bleaching and hybridisation using short pre-defined sequences [87, 88]. Although these recent developments are not directly relevant to my projects presented in this thesis, they are highly relevant to the field of single cell genomics research because they can reveal the positional information of single cells, in addition to their transcriptomic contents.

2.3 Single cell RNA-Seq datasets are high dimensional and require dimensionality reduction methods

Single cell RNA-Seq datasets are high dimensional, so dimensionality reduction tools are desirable to assist intuitive interpretations of single cell RNA-Seq data. A substantial amount of variation present in a single cell RNA-Seq dataset can be the result of technical variation and is not biologically relevant [89]. These technical variations are likely to be uncorrelated. On the other hand, if a group of genes function in the same pathway, the expression levels of these genes can be highly correlated. Correlated biological variation can be identified and described using a small number of summary attributes that are fewer than the original number of attributes (i.e. the number of genes). This procedure is called dimensionality reduction. The summary attributes would capture more of the correlated variation due to biological variations and less of the uncorrelated variation due to technical noises. As a result, it is preferable to perform dimensionality reduction to facilitate the understanding of single cell RNA-Seq data.

There are many dimensionality reduction methods available including principal component analysis, probabilistic principal component analysis, factor analysis, multi-dimensional scaling, and t-distributed stochastic neighbourhood embedding. Comparing and evaluating the theory and performance of dimensionality reduction methods is beyond the scope of this thesis. For comprehensive reviews on this topic, please refer to [90, 91]. As is shown in the

published single cell RNA-Seq literature to date, the most popular dimensionality reduction method for analysing single cell RNA-Seq data is principal component analyses (i.e. PCA), followed by t-distributed stochastic neighbourhood embedding (i.e. t-SNE). It is not easy to evaluate the performances of various dimensionality reduction methods in analysing single cell RNA-Seq data because the ground truth underlying single cell RNA-Seq datasets is generally not available. As a result, none of the published single cell RNA-Seq literature performed an evaluation of relative performance; none of them claim to have developed or applied the optimal dimensionality reduction method. Likewise, my thesis does not perform such evaluations or make similar claims. I do acknowledge that this is an important area for future research.

2.4 Single cell RNA-Seq expression values are distributed with a zero-inflated pattern and require special considerations

Previous observations of single cell RNA-Seq data suggest that the expression values of any given sample have excessive number of zeros (i.e. dropouts) than that are expected from the expression values of a traditional RNA-Seq experiment. This interesting phenomenon is universal across single cell RNA-Seq data generated from different samples, using different protocols and by different labs. Here, I will briefly describe three potential sources of these dropouts. (i) Previously, it has been shown that promoters are not constantly producing transcripts. Instead, they produce transcripts in bursts determined by the molecular properties of the on/off switching dynamics of the promoter

assembly, as described and modelled in [43]. This observation was later validated by a different research group using an independent experimental approach [44]. (ii) In a single cell RNA-Seq experiment, it is not the case that all mRNA molecules are reverse transcribed. This is due to inefficiencies of reverse transcription reactions described previously in Chapter 1 section 6. As a result, many transcripts that are present in the sample are not being transcribed into cDNA and thus their abundances are estimated as zeros (i.e. dropouts). (iii) cDNA fragments are present at low abundances, and thus need to be amplified by exponential amplification in order to be sequenced by high throughput sequencing. The exponential amplification may favour some cDNA fragments over others. If some cDNA molecules were not sufficiently amplified for sequencing, their final expression estimation will be zero (or nearly zero).

The excessive amount of zeros (i.e. dropouts) is a major challenge of single cell RNA-Seq data analysis. Facing this challenge, an innovative modelling method has been developed to account for this zero inflated distribution of expression values in single cell RNA-Seq datasets during dimensionality reduction analyses [92]. By mitigating the effect of these dropout events, the innovative modelling method is able to achieve more accurate results during dimensionality reduction analyses, and to better restore the underlying relationships between the data points. In this paper, simulated dataset that seeks to recapitulate the dropout characteristics of single cell RNA-Seq experiments was used to test the proposed approach. The authors demonstrated that in the presence of a high fraction of dropout events reminiscent of single cell RNA-Seq, classical

dimensionality reduction methods were unable to uncover the underlying biological variation. On the contrary, when the same simulated dataset was analyzed using their proposed method, the results were superior in recovering underlying biological variation. The improvement in results is apparent in Figure 2-1 (this figure is reproduced from the original publication [92]). Furthermore, a separate study also has confirmed that the number of zeros present in single cell RNA-Seq datasets can be a significant confounding factor for the identification of biological variability [93]. However, these excellent methods and studies only became available after analyses of all my projects were completed, so my projects were unable to make use of these insights. Fortunately, the biological variation can be discovered in my datasets even without applying these advanced analytical tools. This potentially indicate that data quality is sufficiently good and biological variation is sufficiently strong.

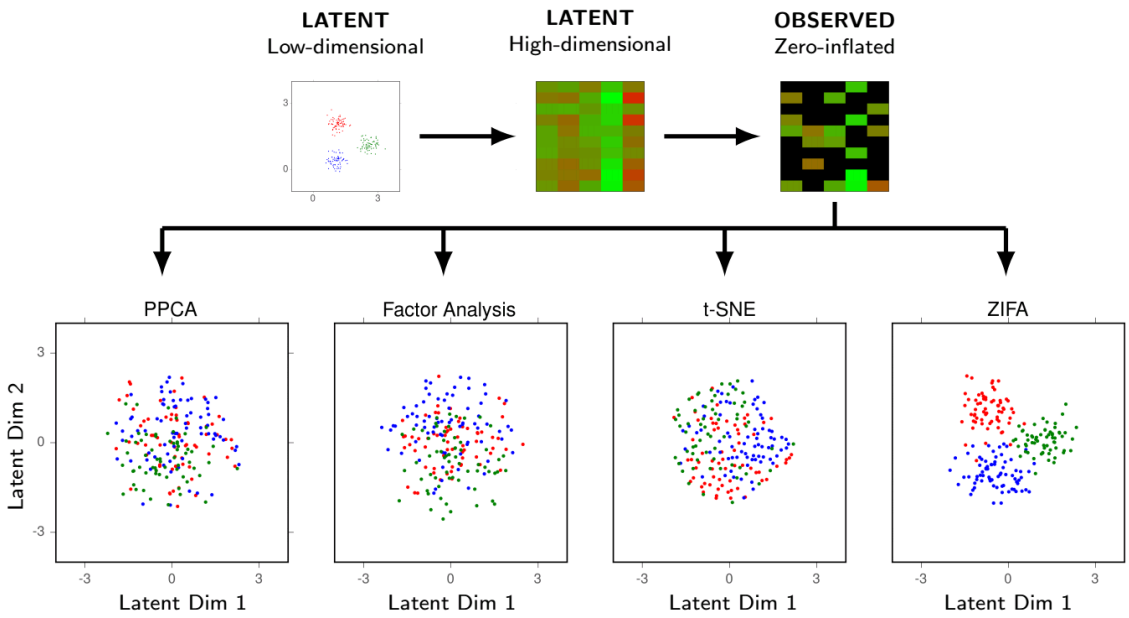


Figure 2-1: Comparison of the results obtained by ZIFA and other factor analysis algorithms. The upper half of the figure shows a flowchart describing the procedure to simulate a zero-inflated dataset. The ground truth of the dataset is that it contains three underlying clusters. The lower half of the figure shows the results from applying four different dimensionality reduction methods to the simulated dataset. We can see that the first three popular dimensionality reduction methods performed badly. This is potentially because they did not explicitly account for the zero inflated distribution. The fourth method "ZIFA" is the newly proposed method. This method is able to recapitulate the underlying ground truth of the method. This figure is reproduced from [92], which is published under the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/).

2.5 A large proportion of capture units fail to capture a healthy cell

Using the Fluidigm C1 platform for single cell RNA-Seq, often there is a high proportion of C1 capture sites that fail to generate RNA-Seq libraries that are useful for downstream analyses. This can be due to many reasons. Four

possibilities are discussed below. (i) When cell-containing solution flows across the capture sites in a Fluidigm C1 system, the capture sites can trap cells. However, this trapping mechanism has a low success rate. According to the Fluidigm C1 microfluidic chip design, about 2000 cells flow across the capture sites sequentially in one experiment. This effectively provides each capture site with about 2000 opportunities to capture a cell, hoping that it succeeds at least once. However, even with about 2000 cells, a 100% capture rate for all 96 capture sites on the Fluidigm C1 chip has been proved impossible. A substantial proportion of capture sites may not capture a cell to generate a useful single cell RNA-Seq library. Based on our collaborators' experience of different cell types and experimental conditions, capture efficiencies typically range between 20% and 95%. (ii) Capture sites are designed to capture cells of a given size range. If the optimal capture size of the Fluidigm C1 chip fails to match the actual size of a cell type, the capture efficiency will understandably be reduced. (iii) In most tissues, cells are tightly bound together by cell-to-cell junctions and are not easily separable. Individual cells within tissues can also be fragile to mechanical/chemical dissociation, isolation and sample handling. As a result, after dissociation, cellular debris will be present in large quantities together with intact cells in the solution. When the solution is pumped across the Fluidigm C1 capture sites, many sites occluded by cellular debris may not capture an intact cell. (iv) During dissociation, handling and single cell capture steps, many cells can become distressed and initiate apoptosis. As a result, many apparently intact cells captured by the capture sites may have most of their RNA already degraded. These apparently intact cells will have little RNA starting material to

generate a useful RNA-Seq library. In Chapter 4, I discuss the poor capture rate for dissociated midbrain dopamine neurons, which might be caused by a large amount of cellular debris arising from brain tissue dissociation.

The majority of empty capture sites can be identified with bright field or fluorescence microscopy which allows empty capture sites to be excluded from downstream library preparation and data analyses. All single cell RNA-Seq datasets generated using Fluidigm C1 system in this thesis have applied this procedure to eliminate empty capture sites. In addition to identifying physically-empty capture sites, it is also possible to identify occupied capture sites that contain dead cells. For example, a captured unhealthy cell can be identified with bright field microscope based on their abnormal morphology. A captured unhealthy cell can also be identified using cell live/dead staining and under fluorescence microscopes. One example of live/dead staining is the method available from Abcam. The method is based on the principle that live cells will have an intact cellular membrane which excludes red dye within their reagent; and will concurrently have an active intracellular esterase activity which generates green fluorescence when treated with their reagent. Those capture sites that were identified as containing dead cells then can be excluded from the set of cells being sequenced and analyzed. Despite the fact that microscopic observations is able to identify some problematic capture sites, it is not the case that all problematic sites can be identified with the methods described above. Some problematic sites will contribute to the final single cell RNA-Seq datasets. It is thus unsurprising that a sizable, but initially unknown, fraction of capture sites

in any single cell RNA-Seq experiment may not generate a high quality RNA-Seq library.

If these low quality RNA-Seq libraries are not excluded from the downstream analysis, they could strongly interfere with or completely obscure the biological signal that an investigator would be interested in. For example, in a clustering analysis, the un-annotated unhealthy cells could potentially form a separate cluster and erroneously present themselves as a new cell type. Hence, it is important to identify all possible capture sites that contain unhealthy cells or cell debris. However, wet lab methods were not able to identify all such problematic capture sites. This calls for the ability to identify such problematic capture sites based on bioinformatics analysis. An method based on support vector machine (SVM) was developed for this purpose [94]. To briefly illustrate this method, an SVM model was built with a set of about 30 input attributes and one output attribute. A few examples of the input attributes are: “total mappable reads”, “GC content”, “total reads mapped to apoptosis genes”, “total reads mapped to metabolic genes”, “total reads mapped to mitochondrial genes”, “total reads mapped to ribosomal genes”. The output attribute is the quality of the cell. The supervised training of the SVM was achieved by using approximately 1000 cells with known input attribute values and with known underlying truth of output attribute values. During supervised training, relevant information contained in the training dataset was learned by the model. After training, the model is then able to predict unseen cells as good or unhealthy. Although the parameters within an SVM model are not directly interpretable, it

is possible to infer some biological insights from them. For example, in a linear SVM, among the input attributes, if “reads mapped to apoptosis genes” is a heavily weighted attribute, we can infer that “reads mapped to apoptosis genes” is a good indicator of cell healthiness. Furthermore, the sign of the weight can also be useful for inferring whether less “reads mapped to apoptosis genes” is typically associated with a healthier cells. For another example, in the same linear SVM, if “GC content” is an attribute carrying no or negligible weight, we can infer that “GC content” is not a good indicator of the healthiness of the cell. Unfortunately, this method became available after the analyses of all my projects were completed. As a result, my projects did not make use of this tool during analyses. Instead, to achieve the same goal, I applied a simpler version of quality prediction steps based on several thresholds that were previously adopted by published single cell RNA-Seq research papers.

2.6 Single cell RNA-Seq dataset typically can also contain capture units with multiple cells

We refer to a single cell capture site in Fluidigm C1 device, or a single droplet in a Drop-Seq experiment, as a capture unit. In any single cell RNA-Seq experiment, although one capture unit is designed to capture one cell only, it may also capture more than one cell. In the past, little attention was paid to this issue. A white paper issued by Fluidigm described that with the older version of their C1 Auto-prep system, some capture units may contain more than one cell (i.e. doublets) that is not visible under simple microscopic inspection [95]. For simplicity, I will refer to these doublets, triplets, quadruplets etc. capture units

collectively as "doublets". This white paper reported that depending on the experimental parameters, approximately 30% of the capture sites of their initial version of the medium sized C1 chip (i.e. 10-17μm chip) could be capturing more than one cell, with a standard deviation of 10%. The reason of capturing excessive doublets is a flaw in the design of the microchip where a volume beneath the capture site also unintentionally traps additional cells. It is important to point out that the doublet issue is not limited to the initial version of the medium sized chips, because the initial versions of other sized chips also exhibit similar issues. Fluidigm redesigned all their C1 chips following the discovery of prevalent doublets.

Furthermore, other single cell sequencing platforms such as droplet-based methods [19, 30] also encounter a similar doublets issue albeit for an entirely different underlying technical reason. In droplet-based methods, a stream of barcode-containing droplets passes across a window, where individual cells are injected from solution into individual droplets randomly. Assuming that the cells are uniformly distributed in the solution, the number of cells injected into each droplet follows a Poisson distribution. According to this distribution, some droplets will contain two or more cells. The doublet issue poses a significant challenge to the utility of single cell RNA-Seq platforms unless its occurrence rate can be shown to be negligible. To assess the doublet rate, human and mouse cells are mixed together before generating single cell RNA-Seq libraries on the mixture. In this way, doublets can be identified when substantial sequences are mapped to both the human and mouse reference genomes. In this assessment

performed for Drop-Seq [30], the authors showed that the proportion of droplets that contained mRNA of more than one cell can be as low as 1.2% if the experimental conditions are chosen properly. This rate is lower than the rate of a Fluidigm C1 system, which is 4.2% as estimated using the same experimental design [30]. As a result, it is generally assumed that with an appropriate cell loading ratio, the doublet rate can be low in a Drop-Seq experiment and thus impact is potentially negligible [30].

One potential issue of having doublets in a single cell RNA-Seq dataset is illustrated in the hypothetical example described below. In the example, we assume that there are two underlying cell types, and assume that some capture sites captured doublets. During the cell type identification stage, applying *a priori* clustering on all cells would potentially identify three cell types: the two underlying cell types, plus a perceived "third cell type". This third cell type in fact represent capture sites that captured doublets - one from each cell type. The expression characteristics of this perceived "third cell type" would then be the average of the two underlying constituent cell types. It is worth noting that some capture sites will contain multiple cells of the same underlying cell type. In such scenarios, the doublet is indistinguishable from a singlet of its constituent cells. This is because RNA-Seq only provides an estimate of the relative abundances of the RNA molecules, but not the absolute abundances. Doublet capture site and of singlet capture site only differ by a multiplier, which is not discernible using RNA-Seq. I acknowledge that ERCC spike-ins allow estimation of absolute RNA

abundances in theory, the estimations are sufficiently accurate to discern a difference of only two times.

The risk of having doublets poses a potential challenge to many previously published findings that assumed each capture site only contained one cell. My projects were completed using the initial version of the Fluidigm C1 system without its redesigned chips. For the midbrain dopaminergic neuron project described in Chapter 4, our collaborators sought to assess the rate of doublets by taking microscopic images at multiple z-planes. Fortunately, our collaborators showed that midbrain dopaminergic neurons are not susceptible to a high frequency of doublets in this experiment. The same investigation has not been carried out for the project on human embryonic stem cells described in Chapter 3 owing to resource limitations. Fortunately, through a theoretical analyses, I concluded that doublets, if any, do not confound the results obtained in the project described in Chapter 3.

2.7 Single cell RNA-Seq dataset profiled cells are a biased representation of the cells present in the solution

Cells are of different shapes. Capture sites may not be able to capture cells of all shapes at the same efficiency. It is thus possible that cells of a certain shape (e.g. spherical) are more easily captured than other shapes (e.g. elongated). Likewise, many other properties such as cell sizes, cell surface adhesiveness, hydrophobicity can also bias capture efficiency. As a result, the cell capture process will not follow a simple random sampling procedure, thus not entail its nice statistical properties. Although the extent of bias has not been quantified by

any published study, it is important to be aware that the frequency of cells of each cell type profiled in a single cell RNA-Seq dataset is a biased representation of the frequency of cells in the dissociated cell solution. Hence, the frequency of cells captured in a Fluidigm C1 system cannot be used in isolation to support claims about frequency of cell type in dissociated cell solution.

2.8 Single Cell RNA-Seq exhibits strong amplification bias that can be accounted for using Unique Molecular Identifiers

In single cell RNA-Seq experiments, if the mRNA quantity obtained from one cell is low, the quantity of cDNA generated from reverse transcription will also be low. As a result, it is necessary to perform exponential amplification on the cDNA before the DNA concentration reaches nanogram levels required for high throughput sequencing (e.g. Illumina sequencing). For example, there are approximately 28 rounds of exponential amplification in the SMART-Seq2 protocol [16]. As explained previously in Chapter 2 Section 2, exponential amplification bias can dramatically bias the estimated abundances of mRNA molecules.

In order to address this issue caused by the amplification bias, a method was designed to introduce a random 'barcode' sequence to the primers used for reverse transcription [96]. The barcode sequence is thus attached at the beginning of the reverse transcribed cDNA molecule. As a result, each reverse transcription event is marked by a different barcode. If the same random barcode connecting to the same mRNA molecule is present multiply in a library, we can generally conclude that such duplicated observations are the result of

PCR amplification. Thus, we shall count the multiple observations due to PCR amplification as one single observation. A caveat of this approach is that the complexity of random barcodes need to be sufficiently high. Otherwise, two different reverse-transcription events occurred for the same gene may contain the same barcode sequence by random chance. In this case, the two reverse-transcription events will be indistinguishable from one single reverse-transcription event followed by PCR duplicates. As a result, the two reverse-transcription events will be erroneously counted as one single reverse-transcription event by downstream bioinformatics analyses.

Even with appropriate designs, this approach has two issues. Firstly, the unique barcode marks each reverse-transcription event, instead of a unique mRNA molecule. Consequently, an mRNA molecule being reverse transcribed twice into cDNA by different primers will be linked to two unique random barcodes. In current literature, this approach is termed as "Unique Molecular Identifier", but it is important to be aware that it reflects the number of unique mRNA reverse-transcription events, because the same mRNA molecule can be reverse transcribed multiple times by different primers. Secondly, recent studies have shown that there are sequence dependant biases in the libraries generated with the Unique Molecular Identifier approach [97]. While UMI was designed to address one type of bias (PCR amplification bias), it unintentionally introduced a second type of bias (sequence dependant bias).

2.9 Single cell RNA-Seq data analyses methods to identify cell types

With single cell RNA-Seq data, there are three possible approaches to define cell type for each cell: (i) Assemble a list of known marker genes for every possible cell type. Based on the expression levels of these marker genes, assign cell type identity. (ii) Assemble a list of known cell types and obtain their transcriptomes. Based on similarity to all known cell types, assign cell type identify. (iii) Without using any existing knowledge of known cell types or their markers, compute the pair-wise similarity between all possible pairs. Entirely relying on this pair-wise similarity, partition all cells of unknown identity into several groups, and assign cell identities based on such groupings. In Chapter 1 Section 4, I proposed that this is a major benefit of single cell RNA-Seq data.

In both Chapters 3 and 4, I did not take the first two approaches described above to identify cell types. This was because: (i) insufficient existing literature describing all cell types present in pre-implantation human blastocyst or in mouse midbrain was available, (ii) insufficient single cell RNA-Seq datasets providing accurately labelled cell types in a human blastocyst or in mouse midbrain were available, and (iii) I aimed to ensure that my analysis was not biased by the use of existing cell type classification systems.

Many recently published single cell RNA-Seq analyses have adopted this same approach. For example [30] and [98], they both serve to support the validity of my choice. Such an *a priori* clustering approach is powerful because it enables researchers to experiment on any tissue with multiple cell types, and use single cell RNA-Seq to identify its constituent cell types.

Using this *a priori* approach, and using the single cell RNA-Seq data of a random mixture of cells, I was able to identify potential cell types present in the mixture. This was followed by performing differential expression analysis between the identified cell types, in order to identify candidate marker genes of each cell type. This approach was extremely powerful, because it did not require prior knowledge of the cell types present in the mixture. Furthermore, once the clustering determined the cell types, a large number of cell type specific marker genes could be identified. These marker genes are useful for designing downstream experiments such as cell sorting, conditional gene knockouts or opto-genetic. Nevertheless, one issue of this approach is related to how coarse or how refined the cell types' definition is most appropriate. I will discuss this issue in the next section.

2.10 Determining cell type number from a pool of cells is a subjective issue

Previously, we presented three different potential approaches for classifying cell types, based on: (i) marker genes; (ii) transcriptomes of cells originating known cell types; and (iii) an *a priori* agnostic method. The three approaches have their own respective benefits and drawbacks. The first two are relatively definitive in the number of cell types identifiable and their respective identities. However, these methods require substantial amount of existing knowledge on all cell types present in the tissue being investigated, which unfortunately is not true for most tissues. The third approach, based on *a priori* clustering, does not require existing knowledge. However, it is usually not

conclusive on how many cell types are present in the tissue. This is because any clustering-based approach requires an input parameter that either directly dictates the number of clusters (e.g. the parameter K in K-means clustering), or indirectly influences the number of clusters (e.g. the cutting height in hierarchical clustering followed by dendrogram cutting).

Separating cells into a larger number of cell types is equivalent to a more refined classification, whereas separating them into a smaller number of cell types is equivalent to a broader classification. Philosophically, it is as equally correct to group all cells into one cell type as it is to consider each cell as one separate cell type; and the continuum of choices in-between the extremes are also equally correct. Thus, deciding on how many cell types are present in the mixture will always be a subjective choice and may attract endless debate. Such a debate lies outside of the scope of this thesis. Practically, the number of cell types is subjectively decided depending on a judgment call by the investigator, the biological context and the research question being investigated.

It is also important to note that, although the first two approaches appear more definitive on the number of cell types, it is not due to any superiority in the methodology. Rather, it is due to the contentious issue being shifted elsewhere. In other words, the burden of deciding how many cell types are present is shifted to previously published literature. For the first two approaches, the classification criteria used in previously published literature, regardless of their correctness, have directly dictated the cell types of the final outcome.

In Chapter 3, this issue will be addressed using the Dynamic Tree Cut algorithm and the expert judgement based on several well validated marker genes of known cell types. In Chapter 4, this issue will be addressed by iteratively splitting clusters until a point where splitting further would have prevented us from confidently identifying cell type specific marker genes.

2.11 Summary

In this chapter, I discussed the challenges of high throughput sequencing, RNA-Seq and single cell RNA-Seq. For each challenge, I also discussed the tools that are designed to address these issues. The challenges and their corresponding analytical methods included the following. (i) It is challenging to map short reads to splice sites. This challenge is addressed using splice site-aware mapping algorithms. (ii) It is challenging to estimate RNA abundance due to various biases introduced by RNA-Seq experimental steps. This challenge is addressed using tools that explicitly account for sequence bias, positional bias etc. during RNA abundance estimation. (iii) Single cell RNA-Seq data is present in a high dimensional space. It is very challenging for us human beings to interpret the relationship between cells in a high dimensional space. This challenge is addressed by applying methods designed to perform dimensionality reduction. (iv) The distribution of values contains more zeros than expected by negative binomial distribution, making clustering and other downstream analyses more challenging. This challenge is addressed using statistical models that are specifically designed to account for the excessive zeros. (v) Single cell RNA-Seq

dataset may result from many capture sites that did not contain a healthy cell. It is challenging to perform downstream analysis when large amount of libraries are of low quality. This challenge can be addressed with analytical approaches using a previously published method, or by taking a simpler, threshold-based, method. (vi) Using currently available single cell capture technologies, a capture unit is not guaranteed to capture one cell only. There are non-negligible probabilities that some capture units may contain more than one cell. This can be a major concern if these capture units are not eliminated and if downstream analysis depends on the assumption that every capture unit contains only one cell. Currently, in most published single cell RNA-Seq papers, this issue is not addressed. The only amelioration that I am aware of is the following: when an independent measurement determines that the prevalence of doublets is not too high, one can claim that the doublets may not distort results significantly. (vii) Cell dissociation and capture can be biased for certain cell types. As a result, the number of cells from each cell type being captured pseudo-randomly shall not be interpreted as a faithful representation of the number of cells in the tissue. (viii) Many rounds of PCR amplification are required by most single cell RNA-Seq protocols. As a result, expression quantification is significantly distorted by biases during PCR amplification. Experimental protocols that employ UMI allow downstream bioinformatics analyses to identify PCR duplicates and to account for this type of bias.

For my projects, all challenges discussed above, except (vii) and (viii), were addressed using the best method available at the time when I performed my

analyses. Challenge (vii) and (viii) cannot be addressed easily, because they require the experiment to generate a completely different single cell RNA-Seq dataset.

Chapter 3 ： Human Conceptoid Bodies Derived from Human ESCs, and Insights Obtained from Associated Single Cell RNA-Seq Data Analyses

3.1 Abstract

Human Embryonic Stem Cell (human ESC) is an important type of in vitro cultured cell derived from the Inner Cell Mass (ICM) cells of human pre-implantation blastocysts. It is important because it can propagate indefinitely in culture and can differentiate into various progeny cell types of all three germ layers in an early human embryo. This suggests that human ESC possess the potential to generate all cell types of a human body. Human ESCs show promise for many future medical uses, and thus have attracted tremendous interest in medical research. Despite this, there are several fundamental questions that have not yet been answered. (i) What is the *in vivo* equivalent cell type of human ESCs? (ii) What is the exact developmental potential of human ESCs? (iii) During early human development, from the totipotent blastomere, what are the earliest cell types that arise and what are their molecular markers? This study sought to provide important information and evidence to answer these fundamental questions.

My analyses on single cell RNA-Seq data generated from human blastocyst discovered that morula cells co-express both ICM and trophoblast markers. This co-expression state coincide with the fact that morula cells are bi-potent for ICM and trophoblast. Furthermore, by making use of *a priori* clustering analyses of single cell RNA-Seq data from human conceptoid bodies, I defined that there are

two constituent cell types in conceptoid bodies. More importantly, I managed to relate the two constituent cell types of conceptoid bodies to the two cell types present in a human pre-implantation blastocyst. The results of my analysis have four implications. (i) We showed that the developmental potential of human ESCs is similar to the ICM of human blastocyst - it not only gives rise to cells in the embryo proper, it also gives rise to extra-embryonic components of the embryo, for example, the trophoblast lineage. (ii) We showed that the conceptoid bodies are good *in vitro* models of human blastocyst, and can be adopted for research related to human blastocysts. (iii) Our results suggest that the *in vivo* counterpart of human ESCs is likely the ICM cells of human blastocyst, which is different from the present popular belief (i.e. human post-implantation epiblast). (iv) The results of my analysis also provided important novel molecular markers for the different cell types present in human blastocysts.

In this chapter, I start by introducing the biological background related to human pre-implantation blastocyst and human ESCs and some observations that inspired the conceptualization of the conceptoid body protocol. These are important background information closely related to the understanding of conceptoid bodies. Subsequently, I present the results of my bioinformatics analyses. This is followed by a discussion section that includes biological interpretation of analytical results and experimental results. In the final section, I present the analytical methods, procedures and parameters that I used to obtain the results.

3.2 Contributions and Publications

This chapter is based on the work described in my co-first author paper "In Vitro Derivation of Blastocyst-like Conceptoid Bodies Indicates Near Totipotency of Human Embryonic Stem Cells". At the time of completing this thesis, the manuscript is completely written, and is ready to be submitted to a peer reviewed journal. This is a collaborative project between three research groups led by: (i) Shyam Prabhakar from the Genome Institute of Singapore in Singapore; (ii) Chris Ponting and Wilfried Haerty from the Medical Research Council Functional Genomics Unit in UK; and (iii) Lili Sun and Paul Robson from the Jackson laboratories in the United States. Lili Sun and Paul Robson led the experimental aspects of the project. With advice from Chris Ponting, Shyam Prabhakar, and Wilfried Haerty, I designed and performed the bioinformatics aspects of the project.

Our collaborators from the Jackson laboratories, Lili Sun and Paul Robson, carried out the protocols to induce human ESCs to form fluid-filled cavities that closely resemble a pre-implantation blastocyst. They also dissociated Day 3 conceptoid bodies into single cells, and performed single cell RNA-Seq library preparation on them. They also performed single cell RNA-Seq library preparation on untreated human ESCs, as a control sample. This single cell RNA-Seq dataset is important for identifying whether there exist different cell types in the conceptoid body and more importantly, what their identities are.

After library preparation, high throughput sequencing was performed by the Genome Institute of Singapore Sequencing Core Facility. Subsequently, the

single cell RNA-Seq dataset was passed to me for bioinformatics analysis. With the help and guidance of Professor Chris Ponting, Dr. Shyam Prabhakar and Dr. Wilfried Haerty, I performed the computational analyses on several single cell RNA-Seq datasets. These included datasets from our collaborators and from public repositories. All parties in the collaboration contributed to the biological interpretation of my analytical results.

3.3 Introduction

Human conceptus (i.e. zygote or fertilized egg) develops into a multi-cellular structure called a blastocyst on approximately the 5th day after fertilization. The blastocyst consists of a single epithelial layer creating a hollow sphere. The sphere contains some blastocoel fluid, and a small number of non epithelial cells – the Inner Cell Mass (ICM) cells – within it. It is generally accepted that the first cell type differentiation occurs at this stage where a totipotent blastomere cell becomes either a trophoblast cell (i.e. the outer layer of the sphere), or an ICM cell (i.e. the non-epithelia cells attached to the interior of the sphere). The trophoblast cells belong to the trophectoderm lineage and will develop into the placenta and other extra-embryonic structures. The ICM cells will develop into the embryo proper and contribute to some other parts of extra-embryonic structures, such as the yolk sac.

Human ESCs are derived from the inner cell mass of the *pre*-implantation human blastocyst [99]. Human ESCs can propagate indefinitely in culture. Currently, there has been no published literature investigating whether the epiblast cells within human *post*-implantation blastocyst can be propagated in culture as human epiblast stem cells (human EpiSCs). This is partly because the human *post*-implantation blastocyst is not accessible for research due to ethical restrictions. Human ESCs have been shown to possess the ability to differentiate into cell types belonging to any of the three germ layers in the embryo proper. A non-exhaustive schematic of the differentiation potential of human ESCs is shown in Figure 3-1. This differentiation potential is of tremendous value. If its potential could be harnessed for drug testing, or tissue/organ regeneration, we

may begin to cure many currently incurable diseases. Owing to its potential value, it is important to understand the properties, characteristics and origin of human ESCs. As a result, a substantial amount of research and literature have been devoted to investigate the *in vivo* counterpart of human ESCs.

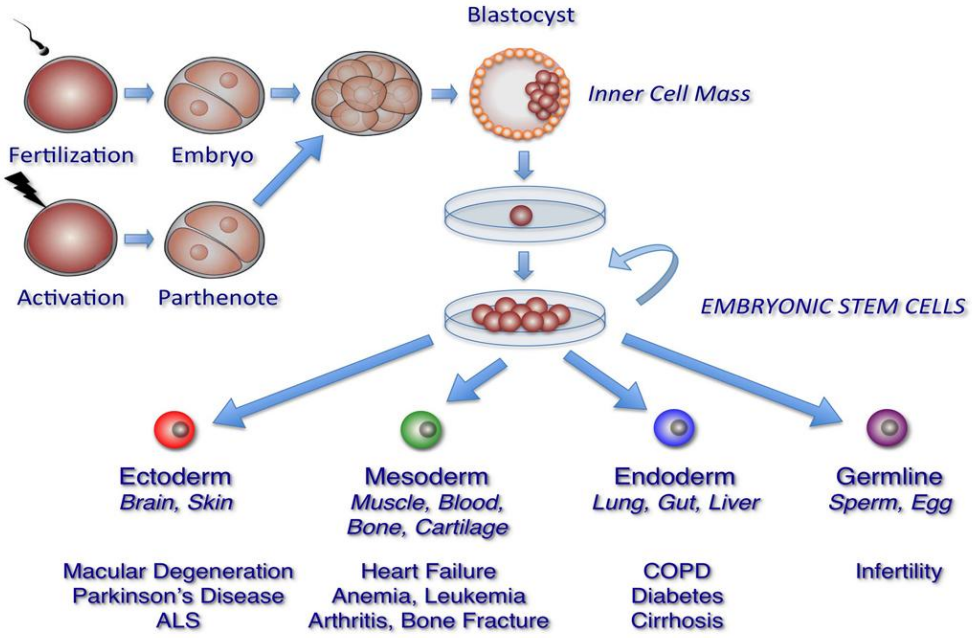


Figure 3-1: A schematic of human ESC derivation and its differentiation potential and their corresponding medical utility. Human ESCs are capable of differentiating into the three germ layers, the germ line and their downstream cell types. Figure reproduced from [100], which is published under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/).

Mouse Embryonic Stem Cells (mouse ESCs) are derived from the inner cell mass (ICM) or the epiblast of the mouse blastocyst *before* implantation. Mouse post-implantation Epiblast Stem Cells (mouse EpiSCs) are derived from the egg cylinder epiblast of the mouse blastocyst at a stage *after* implantation. Mouse ESCs are also referred to as "naive" cells; mouse EpiSCs are also referred to as "primed" cells [101]. The mouse ESCs are able to contribute to blastocyst

chimeras [102], a property not shared with mouse EpiSCs[103] . For mouse ESCs and mouse EpiSCs, the corresponding *in vivo* cell type can be identified molecularly [104, 105], because the *in vivo* cell types are accessible for experiments. Current molecular information argues that mouse ESCs are equivalent to the epiblast within pre-implantation blastocyst [106], and mouse EpiSCs are equivalent to the epiblast within post-implantation blastocyst [107].

When mouse EpiSC was first derived [108, 109], the authors argued that human ESC gene expression pattern are more similar to mouse EpiSCs than to mouse ESCs, although it is important to acknowledge that corss-species expression comparison is very challenging, if possible at all. Furthermore, the authors pointed out that there are some degree of similarity between the culture conditions of mouse EpiSC and the culture conditions of human ESC. Consequently, the stem cell research community began to hypothesize that human ESC is the mouse version of mouse EpiSC, and thus is the counterpart of the post-implantation epiblast (i.e. the “human ESC = human post implantation epiblast” hypothesis). Because human post-implantation blastocysts are not accessible for research due to ethical restrictions and medical reasons, there is no way to test the aforementioned “human ESC = human post implantation epiblast” hypothesis molecularly. Over the years, this hypothesis has never been tested at the molecular and transcriptomic level. In the absence of molecular evidence, peripheral and anecdotal information are often construed as evidence. For example, culture conditions, morphology, metabolic activity, gene expression and epigenetic status [108, 109]. It is important to note that the flaw in any of

these construed evidences are the inability to access the human post-implantation epiblast. How could someone argue that human ESC equals to human post-implantation epiblast when she/he had absolutely no data on human post-implantation epiblast?

Strongly contradicting these unsupported arguments is the ability of human ESCs to give rise to the trophoblast lineage [110-112]; sometimes even occur spontaneously in culture. Mouse EpiSC has never been shown to spontaneously give rise to trophoblast lineage. If the human ESCs were post-implantation epiblast, they have already segregated away from the trophoblast lineage during pre-implantation period, and thus should not be capable of generating cells from the trophoblast lineage. However, if human ESC indeed give rise to trophoblast cells, this could be explained by human ESCs having not yet fully segregated from trophoblast cells. This suggests that human ESCs are in fact more like human ICM. Given the fact that human ESCs are derived from the ICM of pre-implantation blastocyst, and not derived from the post-implantation epiblast, the more parsimonious and thus more likely explanation is that human ESCs are representative of the ICM, thus are more potent, and are not yet fully segregated from trophoblast. On the other hand, trying to argue that human ESCs are derived from ICM, differentiate into post-implantation epiblast, sometimes spontaneously revert back to ICM, and then evolve into trophoblast appears not to be the most parsimonious explanation.

This hypothesis could have been tested if we could examine *in vivo* samples of post-implantation human blastocyst. However, due to ethical restrictions,

research examination of any peri-implantation and post-implantation sample is not possible in human. It is known that there are drastic differences of pre-implantation and early post-implantation development between the mouse and human. The earliest stages of embryonic development are far more variable across species than expected from their evolutionary relationships [113]. (i) The primitive endoderm (i.e. PrE) – the third cell type to emerge from the developing conceptus – is dependent on FGF signalling in the mouse. On the other hand, in human, PrE is not dependent on FGF signalling [114]. (ii) The fourth human cell type to emerge during development, the primitive syncytium at the polar end of the implanting blastocyst [115], does not have an identifiable mouse blastocyst counterpart. (iii) Mouse pregnancy is under the control of the oestrous cycle where the animal reabsorbs the endometrium if implantation does not occur during that cycle. On the contrary, human pregnancy is under the control of the simian-specific menstrual cycle [116], where spontaneously decidualized maternal endometrium is shed through menstruation if implantation does not occur. (iv) The early post-implantation epiblast has a flat, disc-like morphology in the human (and most other mammals), whereas in the mouse it is cup-shaped. In summary, due to a substantial divergence, it is not surprising that mouse and human ESCs differ substantially in developmental potentials, growth requirements, gene expression levels, and regulatory controls [99, 117, 118]. This will be elaborated in more detail below.

Human ESCs have been shown to be capable of differentiating into cells of the trophoblast lineage [110-112]. Hence, we sought to test two hypotheses. (i)

Whether differentiation from human ESCs into trophoblast cells occurs through a transient intermediate state, where each individual cell is “blastomere-like”, and capable of generating both ICM cells and trophoblast cells? (ii) Could we harness the potential of this “blastomere-like” transient cell state to create a blastocyst organoid? In this study, we show that before hESCs fully differentiate to trophoblast, each cell attains a transient state where both hESC and trophoblast lineage-restricted marker genes are highly expressed, a transient state reminiscent of progenitor cells. Furthermore, we showed that using cells in this transient state, a blastocyst-like organoid can be derived.

We then sought to investigate the properties of this blastocyst-like organoid. It was previously known that human blastocyst contained at least two cell types, the outer epithelial layer that forms a fluid-filled cavity and some cells contained within the fluid-filled cavity. We thus ask the question whether the organoid contains distinct cell types, similar to those of the blastocyst? To answer this question, our collaborators performed single cell RNA-Seq and sequenced up to 96 cells from day 3 blastocyst-like organoids. I performed the bioinformatics analysis and identified two distinct cell types present in the day 3 blastocyst-like organoid. Furthermore, my analysis also demonstrated that the two cell types are equivalent to the two cell types present in human blastocysts, which are the ICM cells and the trophoblast cells. Consequently, we named this blastocyst-like organoid the “conceptoid body”, in relation to embryoid body, because it not only contained the progenitors of germ layers, but also contained cells of the blastocyst lineage. We then performed marker protein staining and

other functional validations to demonstrate that these novel structures are functionally similar to human blastocysts in many critical aspects.

3.4 Results

Protocol to induce trophoblast from human ESCs

When human ESCs were treated with a previously reported trophoblast induction protocol [119] (referred to as the SU/BMP protocol in this thesis) for several days, it was shown that their morphology became similar to trophoblast cells [119, 120]. We sought to validate these observations and to compare the gene expression dynamics of human ESC-derived trophoblast with those of the *in vivo* formation of the trophoblast lineage. To do this, our collaborators Lili Sun and Paul Robson treated human ESCs using the trophoblast induction protocol and performed bulk RNA-Seq at 2 days interval over a time course of 8 days. The untreated human ESCs are denoted as Day0. Global transcriptome comparisons to a panel of human tissues [121, 122] indicate that the differentiating human ESCs progress toward a trophoblast phenotype over 8 days of differentiation (Figure 3-2, Materials and Methods M8).

We next sought to determine whether the transcriptome of these *in vitro* human ESC-derived trophoblast cells was similar to that of the earliest *in vivo* trophoblast cells that first form in the pre-implantation blastocyst (i.e. the blastocystic trophoblast). Ideally, gene expression profiling of human blastocyst shall be carried out by our collaborator using the protocol that was applied to generate the human ESC derived trophoblast dataset. However, due to practical

constraints, our collaborators were not able to perform these experiments. In order to achieve this analysis, we compared previously published *in vivo* blastocyst gene expression [123] to in vitro human ESC-derived Day 8 blastocystic trophoblast gene expression. As seen in Figure 3-3, genes differentially expressed between human ESCs and these trophoblasts showed a significant tendency to be regulated in the same direction, regardless of whether in vivo or in vitro trophoblast cells were examined ($p<1.3x10^{-14}$, Materials and Methods M9).

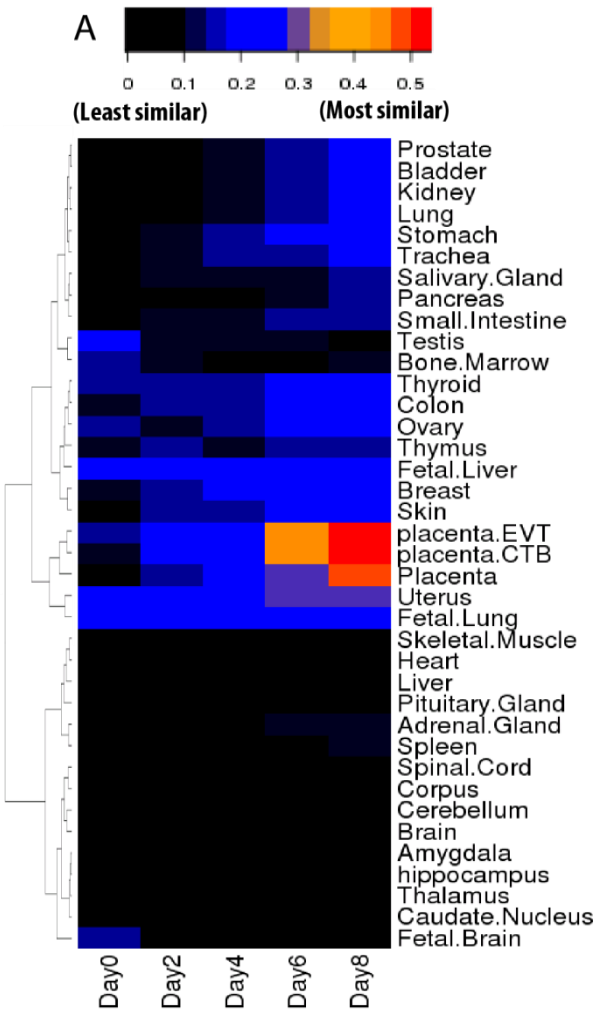


Figure 3-2: The similarity of treated human ESC cells to a panel of 38 normal human tissues. The figure indicates that human placental tissues, not other tissues types, are most similar to the treated human ESCs. Each row represents a normal human tissue type. There are 38 tissue types covering tissues derived from all three germ layers. Each column represents a time point (Day2, Day4, Day6 and Day8) when hESC is undergoing SU/BMP treatment to generate trophoblast cells. The un-treated hESC is represented as Day0. The colours of the heatmap represent the similarity (Pearson correlation coefficient computed using the expression matrix) of the hESC-derived sample to the normal tissues. We can see that over the time course of hESCs differentiation to trophoblast in vitro, the cells are progressively becoming more similar to the placental tissues.

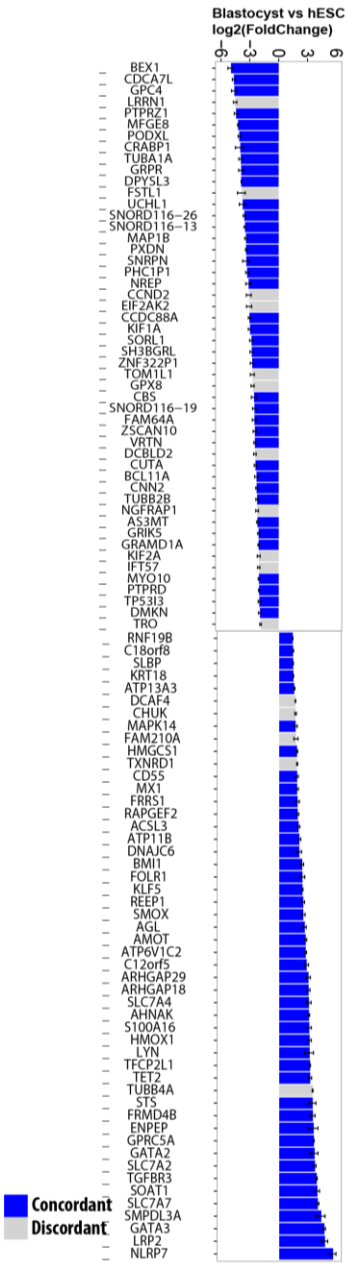


Figure 3-3: Differential gene expression analysis between hESC and in vivo blastocystic trophoblast. The bar length indicates the fold changes calculated by dividing the expression value in blastocystic trophoblast with the expression value in hESCs [123] and taking log2. The division is performed between expression values obtained from identical platforms. The error bars indicate the standard deviation by considering all the blastocyst biological replicates in [123]. The bars are coloured blue (indicating concordance) only if the same direction of regulation is also observed when we compared the in vitro hESC-derived blastocystic trophoblast with the un-stimulated hESC. Vice versa, the bars are coloured grey to indicate discordance (Materials and Methods M9).

Discovery of a transient bi-potent state co-expressing both ICM and trophoblast markers

We were intrigued by the observation that hESC cells progressively become trophoblast-like under SU/BMP treatment, and wished to understand what occurs at the early stages of SU/BMP treatment. In addition, based on our collaborator Lili Sun's and Paul Robson's extensive experience in culturing these hESC under SU/BMP treatment, they observed that hESC cells treated with SU/BMP for brief periods are able to revert back to hESC phenotype when treatment is withdrawn, whereas hESC cells treated with SU/BMP for several days are unable to revert back to the hESC phenotype. Our collaborators hypothesized that when cells are transitioning from hESC cells to trophoblast cells, they go through an intermediate state that lies in-between the two cell types. Furthermore, because progenitor cells are known to co-express lineage restricted markers of downstream cell types [124-127], our collaborator hypothesized that the transcriptomic profile of individual cells at the intermediate state is similar to that of the progenitor cell state.

It is known that human ESCs express pluripotency marker genes such as *POU5F1 (OCT4)*, *ZSCAN10* and *NANOG* and that trophoblast express lineage restricted marker genes such as *GATA3*, *GATA2* and *MSX2*. To investigate the above mentioned hypotheses, single cell qRT-PCR was performed on cells at various time points during the transition to measure the expression of the abovementioned lineage restricted marker genes. 5 time points were sampled: before applying the differentiation stimulus (also referred to as Day 0), and Day 1,2,3,4 after applying the differentiation stimulus. This experiment is able to

elucidate how the transcriptome evolves during the differentiation of human ESCs to trophoblast cells. In order to visually identify the transition of individual cells, for each cell the expression values of the six genes were displayed as a pie chart (Figure 3-4; Materials and Methods M16). Each circle represents one cell, and the sizes of the six pie slices of varying colours represent the normalized expression level (Materials and Methods M16) of the six different genes. The human ESCs initially abundantly expressed pluripotent markers and lowly expressed trophoblast markers. On Day 1 after the SU/BMP stimulus was applied, each cell co-expressed both pluripotent markers and trophoblast makers. This co-expression state lasted from Day 1 to Day 2. However, at Day 3, cells are no longer in the co-expression state. Instead, each cell expressed trophoblast markers exclusively. These results show that the cells converting their fate from human ESCs to trophoblasts went through a transient stage when cells expressed the markers of both lineages. This co-expression state lasted for about two days. The time when cells exit the co-expression state coincide with the time when cells are committed to the trophoblast fate and no longer spontaneously revert back to a human ESC fate.

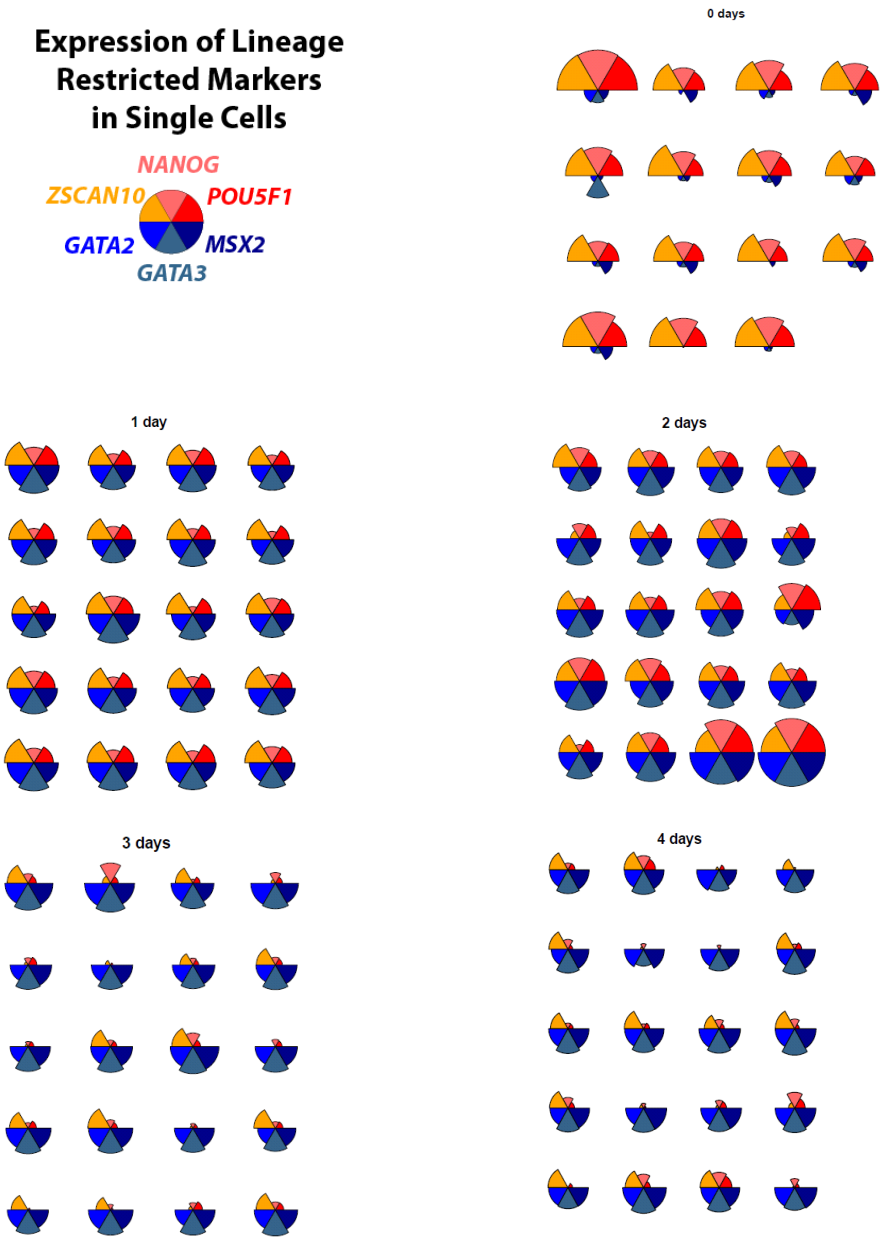


Figure 3-4: Expression of marker genes examined at single cell resolution using qRT-PCR. The single cell expression of *ZSCAN10, NANOG, POU5F1(OCT4), GATA2, GATA3* and *MSX2*were denoted in a pie chart consisting of six slices, with each slice denoting one gene, colored according to the legend on the top left corner of the figure. The expression abundance at detected by qRT-PCR is denoted by the radius of the pie chart. The expreesion values were then normalised linearly so that all output values lie in the range of [0,1] (Materials and Methods M16).

We demonstrated that cells undergoing differentiation from hESC to trophoblast transiently enter a transient stage for 1-2 days, during which they expressed markers of both lineages. I hypothesized that this *in vitro* derived

transient stage represents a bi-potent stage when cells are capable of becoming either cell type. More generally, I hypothesized that during *in vivo* cell differentiation, when one cell type gives rise to two cell types, the founding cell type also concurrently expresses the markers of both of its daughter cell lines. To test this hypothesis, I used a previously published single cell RNA-Seq dataset profiling gene expression during human pre-implantation development [128] (from oocyte to blastocyst, including morula stage). During human pre-implantation development, the morula develops into the blastocyst. At the morula stage of development, the cells are called blastomeres; they are believed to be totipotent – meaning they are capable of generating all cell types of the embryonic and extra-embryonic cell types. At the transition from morula to blastocyst, the totipotent blastomere becomes two different cell types: the trophoblast cells, located at the periphery of the blastocyst. This epithelium layer of trophoblast cells creates a fluid-filled cavity that is called the blastocoel. Attached to the inner surface of the trophoblast is a small number of cells called the Inner Cell Mass (ICM), which is pluripotent and will develop into the embryo proper. This transition from morula to blastocyst is perfect for studying the transcriptomic transition from one founding cell type, to two daughter cell types. In this dataset, a total of 16 cells originating from 2 morula stage embryos were profiled. 13 of the 16 cells passed the same quality filter applied throughout this study (Materials and Methods M2). In Figure 3-5, expression abundances of lineage restricted genes are plotted for both ICM cells (the columns indicated by the orange top bar), trophoblast cells (the columns indicated by the blue top bar) and blastomere cells (the columns indicated by the black top bar). The cell type

identities of ICM cells and trophoblast cells were determined according to Materials and Methods M6. The cell type identity of blastomere cells were determined according to the developmental timing noted by the original authors. We can see that for many of the genes that are shown to be restricted in the daughter cell types (the ICMs or the trophoblasts), they already begin to be expressed in the founding cell type (the blastomeres). A similar phenomenon was previously observed in mouse pre-implantation development [126]. After our analysis was completed, another paper was published proposing that some of these principles can be generalised to other cell types, such as intestinal stem cells, hematopoietic stem cells and human adult pancreatic cells [129]. Our analysis, together with others' recent observations, confirm that this co-expression state is generally applicable to many progenitor cells. In other words, being in this co-expression state suggests that the cell is capable of giving rise to multiple daughter lineages. We further hypothesize that if a cell can be artificially induced to enter this expression state, this cell is also capable of giving rise to cells of both daughter lineages under suitable conditions.

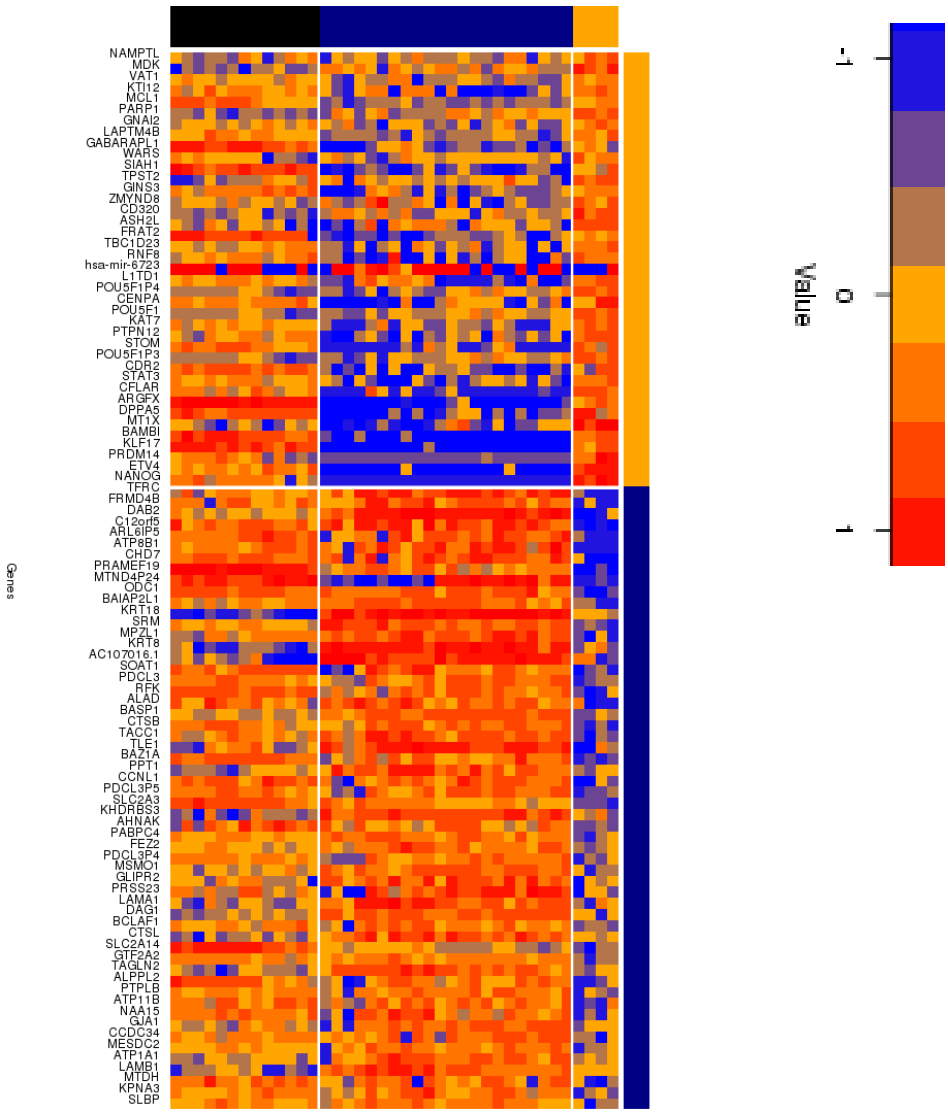


Figure 3-5: Expression of ICM marker genes and trophoblast marker genes among blastocyst and morula cells. Rows represent genes differentially expressed between ICM and trophoblast (cell types defined as Materials and Methods M6). The orange vertical bar marks genes that are more abundantly expressed in ICM cells, and the blue vertical bar marks genes that are more abundantly expressed in trophoblast cells. Each column represents a single cell. The black horizontal bar indicates columns that represent the founding cell type: blastomere cells in the morula. The blue horizontal bar indicates columns that represent a daughter cell type: the trophoblast cells. The orange horizontal bar indicates columns that represent a second daughter cell type: the ICM cells.

Protocol to exploit the transient bi-potent state to create an organoid: a conceptoid body

Recent research has shown that by providing suitable stimuli, 3D environment and cell-to-cell interactions, the formation of an organoid can be induced in vitro that recapitulates at greater detail of tissue/organ function than does a monolayer culture. For example, it has been shown that an intestinal crypt organoid can be induced in culture [130, 131], likewise for a liver organoid [132], and many more examples reviewed elsewhere [133]. Furthermore, research has also shown that in mouse, individual blastomeres are totipotent and are able to contribute to any part of a blastocyst [127], likewise in human [134].

Recall that: (i) our collaborators generated cells in this transient co-expression state that behave like blastomere cells, (ii) blastomere cells can give rise to any cell type in the blastocyst, and (iii) organoids could form given suitable conditions and starting cell population. Further experiments were carried out by our collaborators in an attempt to integrate these three properties. These experiments sought to use the blastomere-like cells to create a blastocyst organoid. To do this, our collaborators use clumps of about 10-50 cells that reached the blastomere-like transient state (24 hours after trophoblast induction) and cultured the clusters in suspension to allow the cell clumps to self-organize for another 48 hours (i.e. 3 days in total). The medium used for culturing the cell clumps was the medium used for human blastocyst culture in IVF clinics. Surprisingly, a large number of cell clumps formed a fluid-filled cavity that resembles the blastocyst as shown in Figure 3-6. These structures contain an

epithelial outer layer, similar to the blastocyst's layer of trophoblast cells, and a small group of cells attached to the interior of the epithelial layer. These conceptoid bodies are the key samples on which most analyses described in this chapter are based. We named these structures "conceptoid bodies", to distinguish them from "embryoid bodies" because they further contained the extra-embryonic components of the embryo.

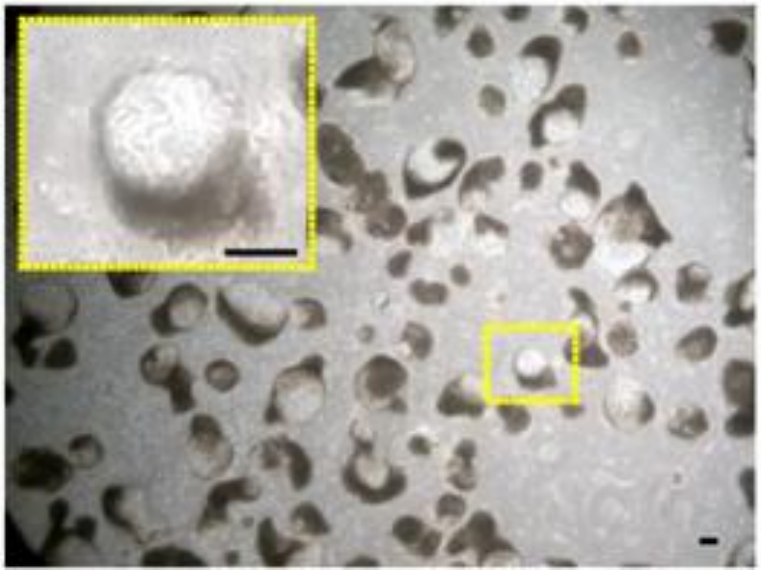


Figure 3-6: Formation of conceptoid bodies. The image is captured and provided by our collaborators: Lili Sun and Paul Robson. Bright-field image showing the formation of structures (i.e. conceptoid bodies) containing fluid-filled cavities that morphologically resemble blastocysts. Scale bars represent 50 microns. Our collaborators did not retain a higher resolution version of this image, unfortunately.

Bioinformatics analyses on single cell RNA-Seq data reveal two distinct cell types in conceptoid body

The formation of this conceptoid body was intriguing. However, besides the morphology of conceptoid body, at this point no molecular information was available. The conceptoid body could consist of any cell types: either human

ESCs, a descendant cell type of human ESC that has been characterized (e.g. trophoblast or ICM), or a descendant cell type of human ESC that has not been characterized previously. The conceptoid body could consist of a single type or a mixture of several different cell types.

Due to the morphological similarity to cavitated blastocyst, we hypothesized that the conceptoid body consist of cell types present in a human blastocyst. To investigate this hypothesis, it was important to investigate single cells of the conceptoid body. Because if investigated as a whole, the mixture of all cells would have prevented the characterization of the cell types present in the conceptoid body. To achieve this, our collaborators dissociated individual cells from Day 3 conceptoid bodies, pseudo-randomly captured up to 96 cells using the Fluidigm C1 microfluidic device for single cell RNA-Seq. It is important to note that the positional information of where the cells originated was completely lost during the dissociation step.

The single cell RNA-Seq dataset was generated by our collaborators using the Fluidigm C1 system and the raw sequencing data generated by Illumina sequencer (Materials and Methods M1) was passed to me for analysis. There was no prior knowledge regarding the total number of cell types present in the mixture, and the markers of different cell types. There were at least three major challenges in the bioinformatics analysis of this study: (i) The number of cell types was not known, yet many clustering algorithms (e.g. k-means clustering) require as an input parameter the number of clusters to be identified. (ii) Just like any previously published single-cell RNA-Seq dataset, the biological and

technological variation is high [46]. (iii) Because the relative concentrations of amplified cDNA differ across different cells prior to mixing together for sequencing, the raw reads yielded by each cell were expected to vary greatly. Below, I describe the results I obtained from analysing the Day 3 conceptoid body single cell RNA-Seq data.

Because the number of cell types and their identities were unknown, as a first step, I performed *a priori* clustering (Materials and Methods M4), and identified two major cell types among the cells dissociated from Day 3 conceptoid bodies. However, we are uncertain about the identity of these two cell types. If the Day 3 conceptoid bodies were similar to Day 6 pre-hatching pre-implantation blastocyst, it would contain the following two cell types: trophoblast cells and Inner Cell Mass (ICM) cells. Before making use of in vivo blastocyst dataset, we sought to investigate this hypothesis based on previously marker genes. I first performed differential expression analysis (Materials and Methods M5, M12) between the two cell types with conceptoid bodies, to identify the genes specifically expressed in each of the two cell types. I identified 804 differentially expressed genes (Materials and Methods M12). From Figure 3-7 it can be seen that *IFITM2, ESRG, IFITM1, GDF3, NANOG, NODAL, ZIC2* and *TDGF1* are among the genes more abundantly expressed in the first cluster of cells: these are well known classical marker genes for pluripotent and ICM cells [135]. On the other hand *KRT7, DLX5, GATA3, EFNA1, ANXA4, TFAP2A, SLC7A2, HOXA1, KRT19, GRH2L, SLC2A3, ANXA2, SLC19A2* and *TJP1* are among the genes more abundantly expressed in the second cluster of cells: these are classical

marker genes for trophoblast cells [135]. Without relying on in vivo blastocyst data, I argue it is plausible that the two cell types present in Day 3 conceptoid bodies are ICM and trophoblast cells, respectively. However, a more reliable analysis would be to compare this conceptoid dataset with a dataset generated from *in vivo* blastocysts.

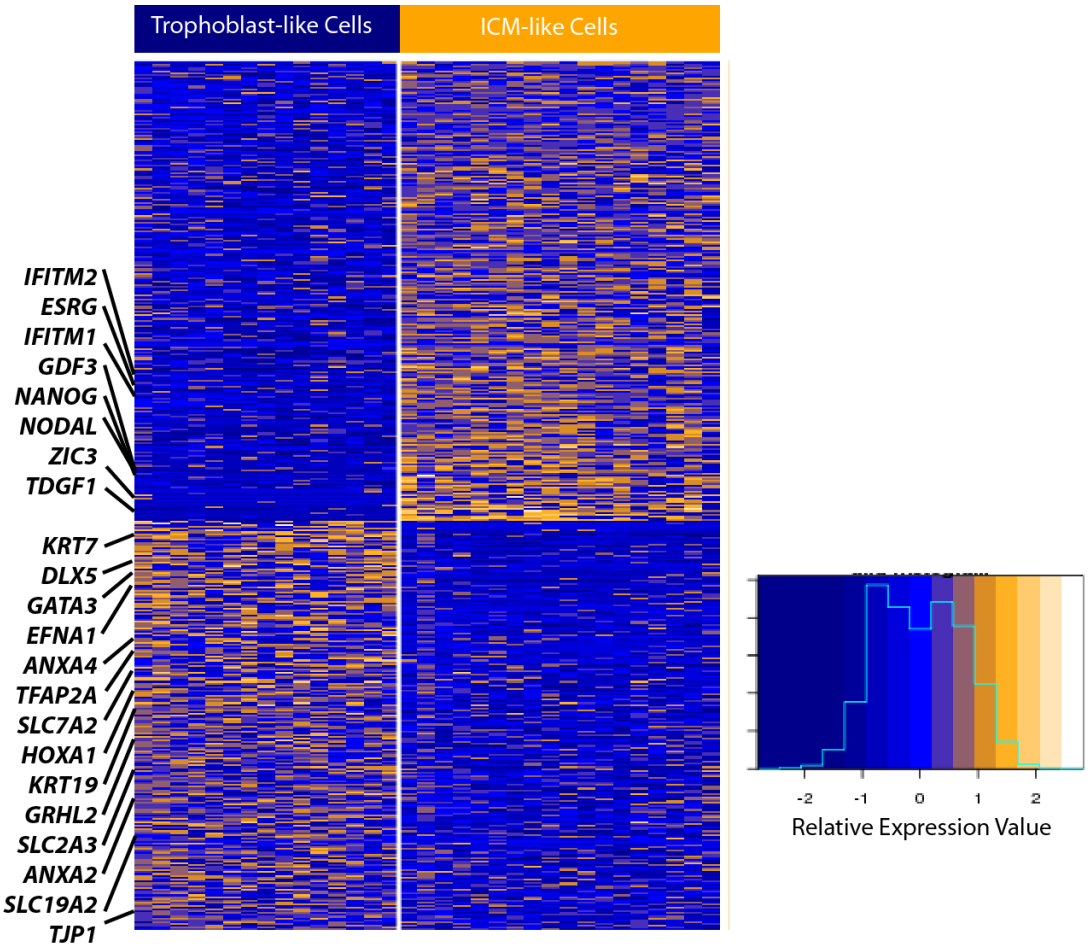


Figure 3-7: Expression of ICM and trophoblast marker genes among single cells dissociated from conceptoid bodies. ICM cells and trophoblast cells inferred from single cells. Genes differentially expressed between the ICM cells and trophoblast cells from single cell RNA-Seq data generated from Day 3 conceptoid bodies. Each row represents one gene, each column represents one cell. Due to space limitations, only a handful of marker genes are annotated in the figure. Expression is in the units of log10 (FPKM), where a minimum value of 1 is applied to FPKM values to prevent log(0) to generate negative infinity values. The log10 (FPKM) values are centred and scaled to obtain the relative expression values displayed in the heatmap.

So far, we identified two cell types present in Day 3 conceptoid bodies, and we tentatively proposed their respective cell type identity. However, we have not obtained information regarding the relative position of the two cell types in a conceptoid body. This was because the positional information was completely lost when the conceptoid bodies were dissociated into single cells. It is possible that among the collection of all conceptoid bodies, some entirely consist of ICM-like cells and others entirely consist of trophoblast-like cells. Alternatively, it is possible that every conceptoid body consists of both ICM-like cells and trophoblast-like cells. In the alternative scenario, the relative localizations of the two cell types also need to be determined. We hypothesized that the conceptoid bodies may resemble a blastocyst with the trophoblast-like cells located at the periphery and the ICM-like cells located at the middle of the sphere. We thus sought to investigate the relative position of the two constituent cell types of the conceptoid bodies. Based on the list of marker genes that I identified previously by performing differential expression analysis (Materials and Methods M12), our collaborators selected two markers for trophoblast cells based on antibody availability and their cellular localisation, and used immuno-fluorescence to perform co-staining in conceptoid bodies in order to observe the relative position of each cell type. As shown in Figure 3-8A,B, the trophoblast-like cells, marked by staining of SLC2A3 (GLUT3) and GATA3, always lie at the periphery of conceptoid bodies, with a small number of non-trophoblast cells attached to its interior (probably ICM-like cells, as marked by the presence of DAPI blue staining

and absence of trophoblast marker staining). Furthermore, our collaborators also performed co-staining of trophoblast and ICM markers concurrently on the same conceptoid body, as shown in Figure 3-8C, the trophoblast-like cells, as marked by CLDN10 in green, are at the periphery of conceptoid bodies, whereas the ICM-like cells, as marked by NANOG in red, are at the centre of conceptoid bodies and are attached to the interior of the trophoblast. Based on these observations, we conclude that the conceptoid bodies contain two cell types: trophoblast-like cells and ICM-like cells. The two cell types are arranged exactly as they are in an *in vivo* blastocyst.

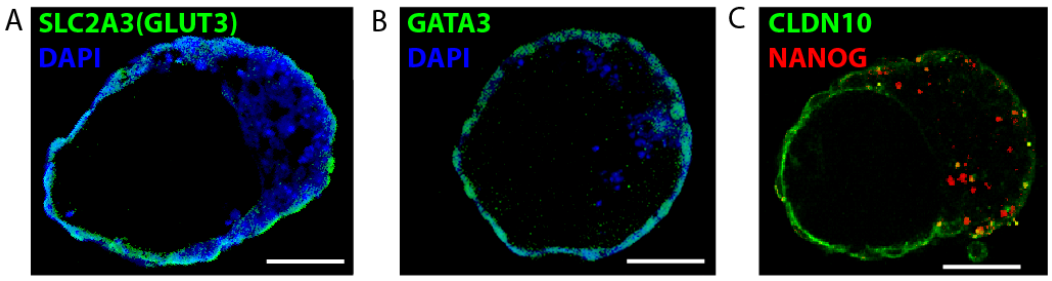


Figure 3-8: Protein staining of several markers in conceptoid bodies. (A) Immuno-fluorescence assay showing expression of SLC2A3 (GLUT3) at the periphery of the conceptoid body (green), but not in the inner cells (blue, DAPI). (B) Same assay, showing GATA3 expressed preferentially at the periphery. (C) CLDN10 is expressed at the periphery of the conceptoid body, while NANOG is expressed in interior cells. Scale bars: 50 microns. Our collaborators Lili Sun and Paul Robson performed these experiments and captured the raw images of all the panels in this figure. After receiving the raw images from our collaborators, I added the annotations, scale bars onto the raw images and created all panels in this figure.

Transcriptomic demonstration of that cell types in conceptoid bodies are similar to those in blastocysts

Thus far, I demonstrated that there are two major cell types present in conceptoid bodies: ICM-like and the trophoblast-like cells. Our collaborators

demonstrated that their relative positions are similar to those in an early pre-implantation blastocyst containing ICM and trophoblast. However, this was based on a handful of known marker genes of ICM and trophoblast. I thus further sought to demonstrate the identity of the two cell types present in conceptoid bodies, without relying on known marker genes, rather, based entirely on the transcriptome of *in vivo* blastocysts.

Ideally, this shall be achieved by generating transcriptomic data from single cells of *in vivo* blastocysts using the same protocol applied to conceptoid bodies, so that differences arising from different experimental protocols are minimized. However, due to limited resources and other practical constraints, our wet lab collaborators are unable to perform the *in vivo* blastocyst experiments. In the absence of the ideal *in vivo* blastocyst dataset, I turned to publicly available datasets. Single cell transcriptome data recently generated from various stages of human pre-implantation embryos spanning from oocyte to blastocyst [128] provided us with an opportunity to compare the conceptoid body single cell transcriptome to that of their *in vivo* counterparts. This dataset profiled individual cells of three Day 6 pre-implantation blastocysts. The original authors classified the cells into three known cell types based on older analytical methods. Our re-analysis of this dataset identified a slightly different classification and discovered that these embryos contain mainly ICM cells and trophoblast cells (Materials and Methods M6).

In order to assess the cell identities of the two cell types present in conceptoid bodies based on global transcriptomic profiles, I sought to compare

the two cell types obtained in our dataset to the ICM and trophoblast cells present in these human embryos. Furthermore, because both datasets contained human ESCs, these were used as a positive control cell type allowing us to benchmark the validity of our analysis. Principal Component Analysis (PCA; Materials and Methods M15) as shown in Figure 3-9 indicated that the three cell types - human ESC, ICM and trophoblast cells - are clearly identifiable and well separated from each other based on the first two principal components. Furthermore, the control cell types - human ESCs – both from our own dataset and from the downloaded dataset [128], are closely overlapping. This demonstrates the validity of this analysis method, because cells closely located in the PCA space can be identified as being of the same cell type. When we examined the trophoblast-like cells from conceptoid bodies, it became clear that the trophoblast cells from conceptoid bodies are located adjacent to the trophoblast cells from pre-implantation blastocyst (Figure 3-9A), indicating the close similarity of the trophoblast cells from the two sources. The same is true for the ICM cells, except that there is a slightly higher heterogeneity among the ICM cells. The heterogeneity among ICM cells is evident both for ICM cells from the blastocyst as indicated by the relative far distance among the four ICM cells, and for the ICM cells from conceptoid bodies. The higher degree of heterogeneity is not surprising, because the ICM cells in a blastocyst are in a transient stage of rapid development, and will differentiate into epiblast, primitive endoderm and potentially many uncharacterized cell types within a very short time frame. It is important to note that this analysis does not rely on markers mentioned in the existing literature. Instead, this analysis is entirely

based on the global transcriptomes of single cells from genuine *in vivo* human blastocysts. This *a priori* analysis strongly validated the cell type identities assigned previously. The strength of my analyses is particularly highlighted by the close proximity of the control cell type - human ESCs - profiled by the two different protocols (points shown in grey/black in Figure 3-9A).

It is worth mentioning that the four ICM cells derived from the *in vivo* human blastocyst and all ICM-like cells within the conceptoid bodies appear to be highly diverse, as can be seen in Figure 3-9A. This is an interesting phenomenon and can potentially be attributed to several factors. Firstly, the ICM within a blastocyst will give rise to many cell types, such as the hypoblast, yolk sac, germ layers and several extra-embryonic tissue types, as a result, the greater spread of these cells may indicate that the ICM cells in fact represent a population of heterogeneous cells that are more heterogeneous than human trophoblasts or untreated human ESCs. Secondly, the ICM cell is an intermediate cell type that is potentially rapidly changing over time. ICM only appears ephemerally, within a very short development window, and disappears. As a result, the ICM cells captured among the conceptoid bodies and among the *in vivo* blastocysts may represent a continuum along the temporal axis of ICM maturation. To thoroughly investigate this hypothesis is not possible at this time, because there are only 4 ICM cells among the *in vivo* blastocyst dataset. However, it would be interesting to examine whether the ICM cells from blastocyst and the ICM-like cell from conceptoid bodies show any systematic variation for other principal components. I thus examined two further principal components, PC3

and PC4, shown in Figure 3-9B. I observed that there are no obvious patterns identifiable in any remaining principal components.

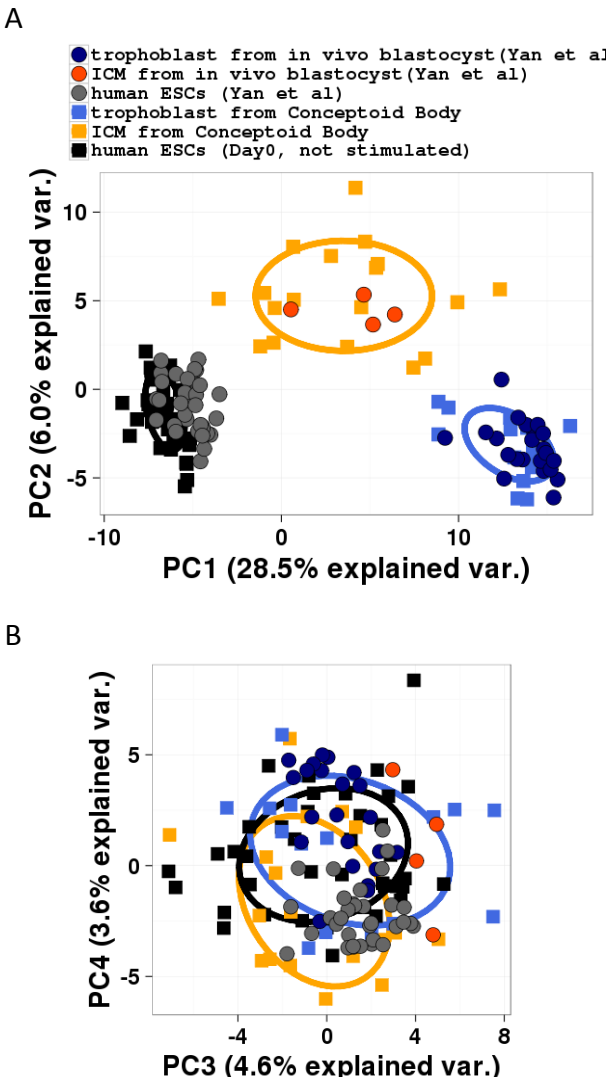


Figure 3-9: Results obtained using PCA on all six different cell types. The two cell types present in conceptoid bodies: trophoblast and ICM, plus protocol matched human ESCs, are represented by squares of varying colours; The two cell types present in human pre-implantation blastocysts: trophoblast and ICM, plus protocol match human ESCs, are represented by circles of varying colours. In Panel A, the first two principal components are shown. The figure shows that inner cell mass (ICM), Trophoblast (TE) and experimental protocol matched hESC occupy three distinct locations in the cell type space. The majority of their embryonic in situ counterpart cell types are located approximately in the same location of the cell type space. Elliptical contours represent the 60% probability density region under the fitted multivariate normal distribution [136]. In Panel B, PC3, PC4 are without any obviously identifiable patterns. Furthermore, no obviously identifiable pattern can be discovered when other principal components were examined (not shown).

Examine the expression of individual genes in the cell types in conceptoid bodies and in blastocyst

Thus far, I showed that for the global transcriptome, the two cell types in conceptoid bodies are similar to the two cell types in human blastocyst. I further sought to investigate this observation at the resolution of individual genes and individual cells. For the conceptoid bodies, I performed differential gene expression analysis between its ICM cells and its trophoblast cells, to obtain a list of differentially expressed genes (Materials and Methods M12). For the *in vivo* blastocysts, I performed the same differential gene expression analysis and obtained a second list of genes (Materials and Methods M11). In the ideal scenario, these two lists and their expression values should be equivalent. However, the single cell RNA-Seq protocol employed in [128] differs from our single cell RNA-Seq protocol in three key aspects. Firstly, the method for first strand synthesis is different, with [128] using the method described in [13], whereas our dataset is based on the template-switching method described in the SMART-Seq1 protocol [137]. Secondly, the cDNA fragmentation steps also differ: [128] is based on shearing the DNA using Covaris and our data is based on the transposon tagmentation method using the Illumina Nextera Kit. Finally, the sequencing platform and read length of the two datasets are different. As a result, it is not surprising that some genes are detectable and accurately measurable by the SMART-Seq1 protocol we used, but are entirely undetectable or less accurately measurable in Tang's protocol [13], and vice versa (Materials and Methods M10). It was thus necessary to intersect the two lists of genes to avoid those that are absent or noisy due to technical reasons. The intersected gene list contained 53 genes. In other words, for a gene to be among the 53

genes, it is required to be detectable in both the SMART-Seq1 [137] and in Tang's protocol [13], and the noise level does not obscure it from being identified as differential expressed. It is important to note that it is not a requirement that the genes are regulated in the direction that is concordant between conceptoid bodies and blastocysts. In fact, among the list of 53 genes identified *a priori*, there are indeed a small number of genes that are regulated in discordant directions between the conceptoid bodies and the blastocysts (Figure 3-10). Despite the stochastic variation in expression expected in single cell transcriptomes [46], these genes showed generally concordant expression changes within conceptoid bodies and within blastocysts. Of the 53 differentially expressed genes, 49 were concordant within conceptoid bodies and within blastocysts ($p<7.0x10^{-11}$; Materials and Methods M7). This concordancy is highly significant.

Overall, these results I obtained suggest that the two cell types present in the conceptoid bodies are equivalent to the two cell types present in Day 6 pre-implantation blastocyst before hatching. It is important to note that the argument is not purely based on a handful of marker genes, but rather based on global transcriptomic expression pattern.

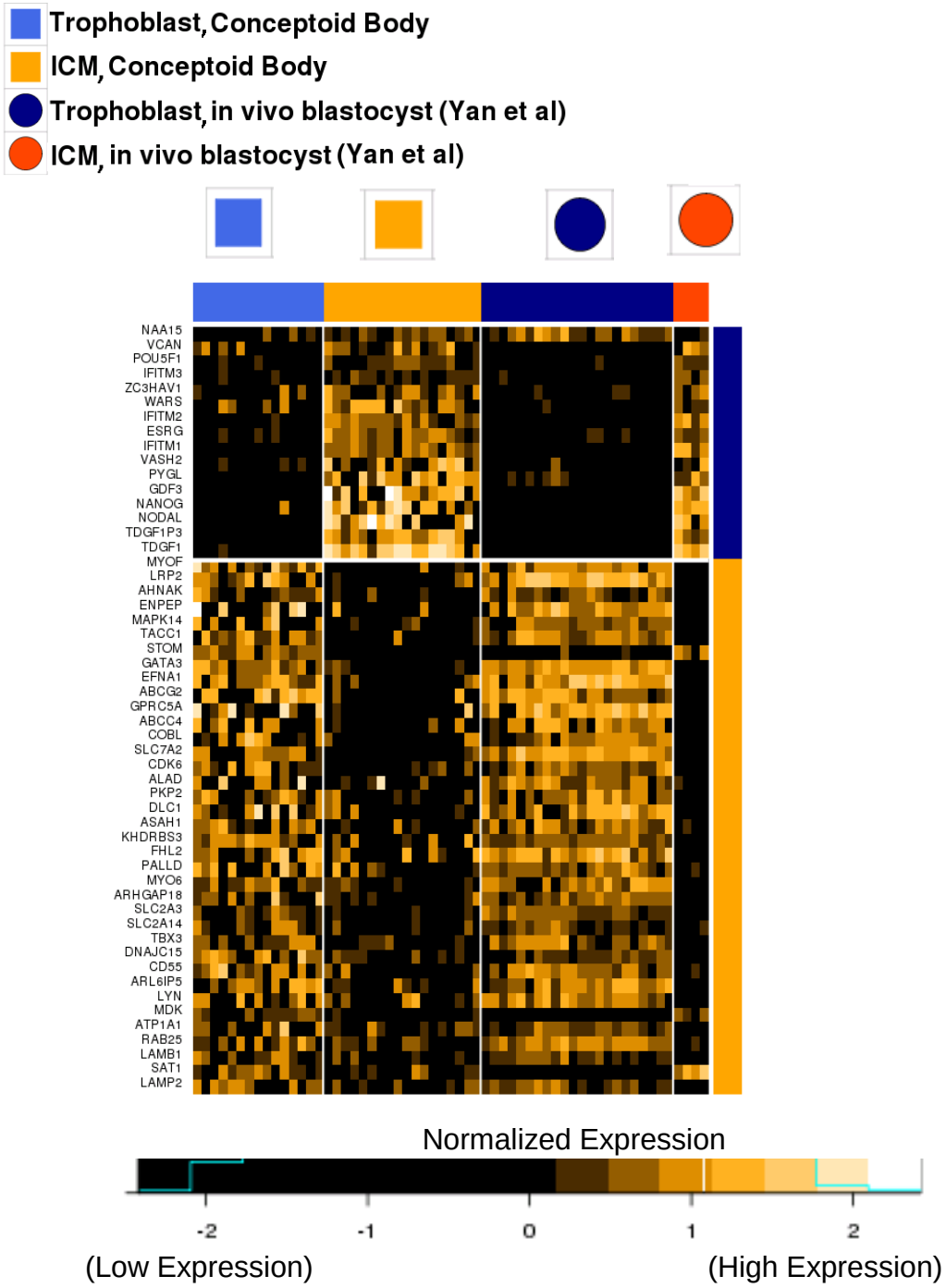


Figure 3-10: Expression of ICM and trophoblast markers genes are similar in conceptoid bodies and in blastocysts. Each row represents one gene, each column represents one cell. The list of genes shown is obtained from a data-driven a priori approach described in Materials and Methods M7 without considering the direction of gene regulation. As a result, it is surprising that only 4 genes are regulated in an opposing direction between conceptoid bodies and human blastocysts whereas the remaining 49 are regulated in the same direction between the ICM and trophoblast cells. The level of concordance is highly significant (p<7.0x10^{-11}, Materials and Methods M7).

Experimental validation of conceptoid bodies with respect to function

After I demonstrated using bioinformatics methods that the cell types present in the conceptoid bodies are equivalent to the cell types present in Day 6 pre-implantation blastocyst before hatching, our collaborators sought to perform functional validations on the conceptoid bodies. Specifically, they sought to demonstrate that the day 3 conceptoid bodies were functionally equivalent to the human pre-implantation blastocyst in the manner in which they interact with the endometrium. At this stage of development, the pre-implantation blastocyst interacts with the uterus and performs several key functions: to attach to the endometrium, to invade and to migrate into the endometrium for a successful implantation. For studying mouse development, the gold standard assay would be to allow implantation in the female reproductive tract. However, for studying human development, such assays are strictly forbidden due to ethical restrictions.

In the absence of such a gold standard assay, they designed assays to separately examine the following key functional interactions with the endometrium: (i) attachment to the endometrium surface, (ii) invasion into the layer of endometrium stroma and, (iii) cell mobility for their eventual successful migration into the endometrium.

For attachment, our collaborators used a previously published attachment assay [138] that measures the number of times the spheroids remain attached to gelatine and to a human endometrial cell line RL95-2, under the condition of centrifugal force. Conceptoid bodies displayed a significantly higher attachment rate than embryoid bodies on gelatine (Figure 3-11A, Materials and Methods M17). My analyses showed that this was of a similar level when compared to

spheroids formed from a choriocarcinoma cell line (i.e. positive control; JAR cell line [138]). The same phenomenon was observed again when using the RL95-2 cells as the attachment substrate, which better mimics the endometrium (Figure 3-11B, Materials and Methods M17).

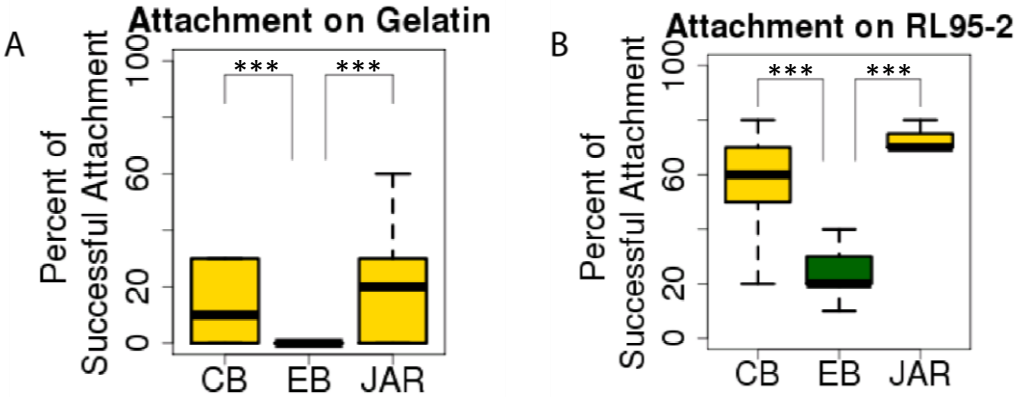


Figure 3-11: Attachment of conceptoid bodies on to various substrates. (A) Percent of structures remaining attached to gelatine (component of endometrial extracellular matrix) under centrifugal force. CB: conceptoid body; EB: embryoid body; JAR: cell clump derived from choriocarcinoma cell line (positive control). Statistically significant differences are indicated by *** (p<0.001; Materials and Methods M17). (B) Similar assay, attachment to cultured endometrial cells (RL95-2 cell line). These attachment assays were performed by our collaborators Lili Sun and Paul Robson.

For invasion, our collaborators placed the conceptoid bodies on a layer of endometrium cells; to act as controls, similarly sized embryoid bodies were used in the same experiment. After co-culturing, immuno-fluorescence staining on the cell nucleus (DAPI), the cell membrane and for the trophoblast marker KRT7 were carried out. The results are shown in Figure 3-12. We can see that for the embryoid bodies, there was a clear boundary separating them from the endometrium stroma cells (Figure 3-12A, boundary indicated by green arrows).

Moreover, the embryoid bodies remained negative for the trophoblast marker KRT7 (Figure 3-12B). In contrast, cells from the conceptoid body spread extensively into the endometrium stroma and expressed KRT7 (Figure 3-12C). At closer inspection, the conceptoid body-derived KRT7-positive cells were spindle shaped with the morphology of motile cells (Figure 3-12D, cells indicated by red arrows), while the embryoid body-derived cells were rounded with a much greater nuclear-to-cytoplasmic ratio representative of a non-motile cell morphology similar in appearance to undifferentiated human ESCs (Figure 3-12A, cells indicated by red arrows).

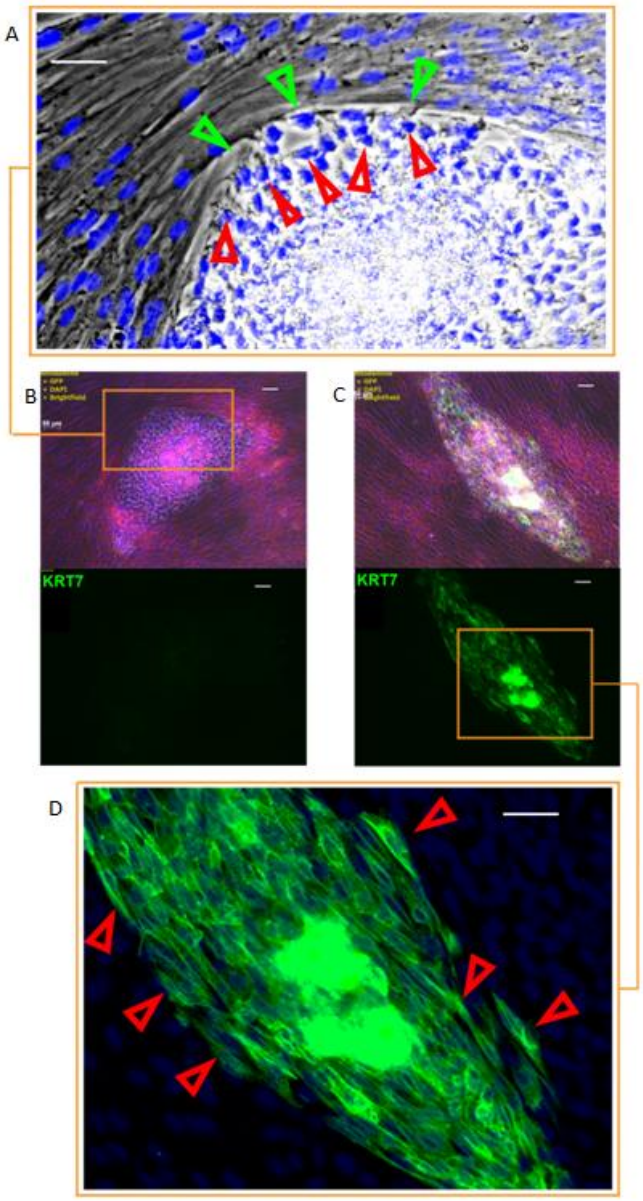


Figure 3-12: Experiments showing the interaction between conceptoid bodies and cultured endometrium cells. (A) Zoom-in of (B), showing clear separation between embryoid body cells and endometrial stroma cells. (B) Behaviour of embryoid body on a layer of endometrial stroma. The upper half shows a composite of a bright-field image with immuno-fluorescence profiling of the locations of KRT7 (green), cell nuclei (blue, DAPI) and cell membranes (red). The lower half shows KRT7 alone. (C) Behaviour of conceptoid body on a bed of endometrial stroma cells, similar to as in (B). (D) Zoom-in of (C), showing invasive (elongated and spindle-shaped) morphology of outer cells from the conceptoid body. (A-D) Our collaborators performed the experiments and captured the raw images of all the panels in this figure. After receiving the raw images from our collaborators, I added the illustrations onto the raw images and created all panels in this figure.

Finally, our collaborators also tested the migratory properties of hESC-derived blastocystic trophoblast monolayer culture. To do this, our collaborators used a trans-well migration assay, with undifferentiated human ESCs as the control cell type. After 2 days, on average 195 blastocystic trophoblast cells migrated across the membrane, compared to approximately 2 human ESCs, as show in Figure 3-13. The difference of the success rate between the two cell types is significant ($p < 1.3 \times 10^{-7}$, Materials and Methods M17).

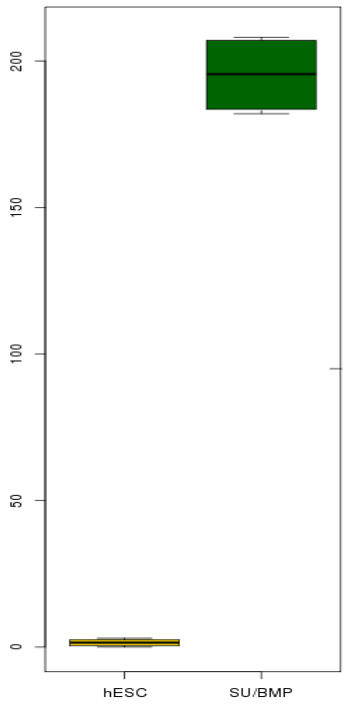


Figure 3-13: comparing of cell migration rate. The Y axis represent the number of cells migrated through the pores in this trans-well migration assay. The two box plots show the comparison between undifferentiated hESCs and trophoblast cells derived from treating hESCs with the SU/BMP protocol.

The results of these functional validation assays, in addition to their ability to self-organize into blastocyst-like structures (a functional validation in itself), indicate that conceptoid bodies and the human ESC-derived blastocystic trophoblast recapitulate the functional properties of both attachment and migration that a genuine human blastocyst possesses.

3.5 Discussion

The developmental position and potential of human ESCs

Model organisms such as the mouse have assisted in the discovery of many insights into human biology. However, one should be cautious about to what extent knowledge can be directly inferred across mouse and human, because mouse has diverged from human for about 96 million years [139]. Evidence is clear that drastic adaptive changes have occurred in their foetal-maternal interactions. Thus one should not automatically take findings in mouse development, or mouse stem cells, and assume that these should be perfectly recapitulated in human development. There are thus potential limitations in inferring the properties of human ESCs by relating these to either mouse ESCs, or to mouse EpiSCs.

Mouse EpiSCs are derived from post implantation epiblast and are already segregated from the trophoblast lineage. Consequently, they are not capable of giving rise to cells of the trophoblast lineage or other extra-embryonic lineages. As expected, mouse EpiSCs do not contribute to the trophoblast when they are artificially aggregated to the mouse blastocyst [108]. Human ESCs have long been argued to be similar to the mouse EpiSCs[108]. Downstream corollaries of this argument are that human ESCs are not "naïve", not toti-potent, and not able to differentiate into cells of the trophoblast lineage.

Because human ESCs were believed to be not "naïve", many researchers sought to derive their “naïve” counterparts. There exist many versions of proposed “naïve” human ESCs culture conditions as reviewed in [140]. To argue

that the resultant cells are naive ESCs requires an ability to examine whether a cell is indeed a naive ESC. For mouse, this can be achieved in part by examining the expression of genes that were known to be more highly expressed in mouse ESCs relative to mouse EpiSC. For human, this is not possible because thus far there has been no study comparing *in vivo* ICM cells against human ESCs. In the absence of true marker genes, several genes were selected based on knowledge obtained from mouse. It is often construed, without supporting evidence, that these genes will also mark naive pluripotency when comparing naive human ESC with regular human ESCs. As a result, the elevated expression of these genes determined from mouse studies have often been used to argue for "naive" human ESC among these publications.

For example, in [141], the elevated expression of *OCT4, NANOG, TBX3, REX1, STELLA, TFCP2L1, KLF2, KLF4, GBX2* and *SALL4* were used to argue that the resultant cells are naive human ESC. Likewise, in [142], the elevated expression of *DPPA5, STELLA, DPPA2, REX1, KLF4, KLF5, TFCP2L1* and *NANOG* were used, and in [143], the elevated expression of *NANOG, KLF4, DPPA3, FLK5, DNMT3L, TBX3, ZNF600* and *LAMB1* were used. However, because human ICM cells were profiled in [128], we can now examine whether these genes are indeed more highly expressed among naive human ICM cells relative to regular human ESCs. To achieve this, in Figure 3-14, I plotted the expression pattern of these genes among single cells from *in vivo* blastocysts. We can see that out of 16 genes that were supposedly markers for naive ICM cells, 15 do not show consistently elevated expression between the 4 truly naive human ICMs relative to regular

human ESCs. This supported the concern raised repeatedly in my thesis that observations and expression patterns of mouse ESCs should not, without further evidence, be directly transferred to human ESCs. Not surprisingly, these postulated naive human ESCs have been argued by others to be *in vitro* culture artefacts rather than naturally occurring cell types [144].

These culture conditions add to our understanding of human ESCs because a large set of signalling molecules and pathways were perturbed to reveal their effect on human ESCs. The different chemical cocktail applied in each experimental condition was summarized in Figure 3-15. We can see that the derivation of conceptoid bodies perturbed several of the pathways that were implicated in human ESCs, although the molecules and effects were different in each culture condition.

In comparison to the direct reprogramming protocols that were applied among the previously published culture conditions, our culture conditions entail two distinct steps: the first step applies the chemical stimuli and the second step withdraws all chemical stimuli and grows in clumps in suspension. Our two-step culture condition is unique because it rely more on the innate abilities of the trophoblast cells to provide an optimal signal and environment to enable the human ESC to become ICM-like cells. This represents a method that is entirely orthogonal to the other five human ESC derivation methods published previously and will potentially inspire a series of culture conditions that rely on the mutual interaction of multiple cell types, rather than chemical cocktails. Our results provided additional understanding of human ESCs to the existing collection of

human ESC literature by showing that: (1) with the appropriate treatment, human ESCs can give rise to cells that are similar to trophoblast cells within a Day 6 pre-implantation blastocysts; and, (2) with the appropriate treatment and appropriate environment, human ESCs can give rise to cells similar to the Inner Cell Mass of a Day 6 pre-implantation blastocyst. Because primed EpiSCs shall be giving rise to later cell types, such as the germ layers or the yolk sac, instead of giving rise to the early cell types, such as the trophoblast and Inner Cell Mass. Our results suggest human ESCs do not represent primed EpiSCs.

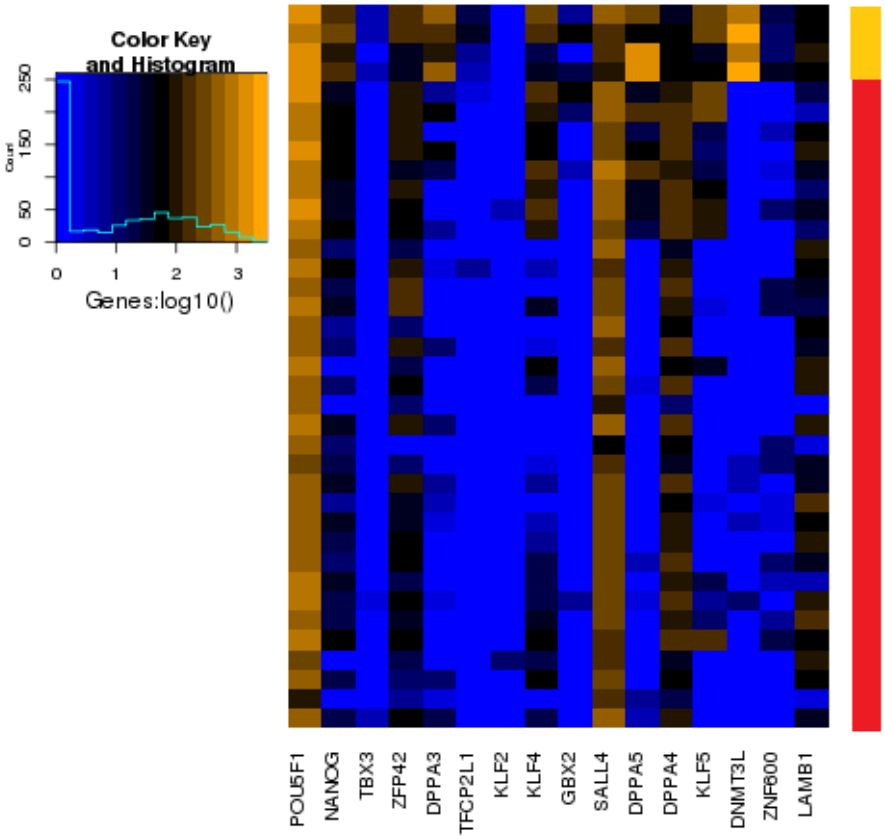


Figure 3-14: Expression value of naive mouse ESC marker genes among human ICMs and human ESCs. Each row is a cell and each column is a gene. The *in vivo* ICM cells are marked by the yellow side bar. Regular human ESCs are marked by the red side bar. Expression abundance of the naive pluripotency marker genes were plotted in each *in vivo* ICM cells and in each human ESCs profiled in [128].

	Gafni 2013	Chan 2013	Valamehr 2014	Ware 2014	Theunissen 2014	Takashima 2014	Concep-toid (this thesis)
MEK-i	MEK-i	MEK-i	MEK-i	MEK-i	MEK-i	MEK-i	
GSK3-i	GSK3-i	GSK3-i	GSK3-i	GSK3-i	GSK3-i	GSK3-i	
JNK-i	JNK-i						
P38-i	P38-i						
PKC-i	PKC-i					PKC-i (Go6983)	
ROCK-i	ROCK-i		ROCK-i		ROCK-i	ROCK-i	
BMP-i		BMP-i					
BRAF-i					BRAF-i		
SRC-i					SRC-i		
FGF-i							FGF-i
BMP4							BMP4
FGF2	FGF2		FGF2	FGF2			
TGFb					Activin		
hLIF	hLIF	hLIF	hLIF		hLIF	hLIF	

Figure 3-15: A summary of various stimuli applied to generate naive human ESCs. The signalling pathways that were stimulated in several published naive human ESC assays, and the initial stimuli applied during the derivation of conceptoid body derivation process.

The conceptoid body is a useful model for studying human early development

The conceptoid body structure that we derived provides a powerful model with which to study the molecular events in human development that are generally difficult or impossible by other means. To date, there has been virtually no molecular description of cell types in the early post-implantation human conceptus. Considering the extensive variation between species early in development, we would argue that the best way to define these cells and their function is through the conceptoid bodies described here rather than using other

organisms. Insights drawn from this model should have clinical bearing on human placental pathologies, assisted-reproductive technologies, as well as should provide a better understanding of the embryonic origins of birth defects and later onset adult diseases associated with alterations in fetal-maternal interaction. In addition, pathways co-opted through mutation in cancer cells may have their functional origins in early development, as Warburg suggested in his observation related to aerobic glycolysis in cancer and in embryonic development [145]. Thus the utility of the conceptoid body structures likely extends into mechanistic studies of pathways active in cancer cells.

3D environment and cell-to-cell contact are important aspects of cell differentiation

In this project, we showed that conceptoid bodies that contain multiple cell types can form in a suspension culture. I argue that the 3D environment is the most important factor that enabled the derivation of the relevant cell types, and enabled the cell types to arrange in a manner that resembles a human blastocyst.

One of the unexpected phenomena observed in this study is that without additional intervention or extrinsic direction, the two relevant cell types (trophoblast and ICM) are obtained in a single culture. I hypothesize that the main enabling factor is that we restored cell-to-cell contact, and other paracrine/ autocrine signalling in the cell-clump suspension culture. To support this hypothesis, below I will cite evidence presented in existing literature. In [134], the authors demonstrated that in human blastocysts, early trophoblast cells are

not yet fully committed to a trophoblast fate. They extracted trophoblast cells of several blastocysts and aggregated these together to form a reconstituted blastocyst that only contained trophoblast cells. They found that, despite no ICM cells being aggregated into the reconstituted blastocyst, ICM cells started to appear after a period of culturing. This observation suggests that the trophoblast cells previously located in the outer layer of a blastocyst can alter their fate if they are now located in a different position in the reconstituted blastocyst. I interpret their results as follows: that a different cell-to-cell contact signal is the enabling factor that converts a small number of trophoblast cells into ICM cells. Likewise, in conceptoid body culture, the cell-to-cell contact that was re-established by the suspension culture enabled the trophoblast cells and ICM cells to form. Other studies also demonstrated that in certain scenarios, cell fate decisions can entirely depend on the surface properties sensed by the extracellular matrix [146].

Another notable phenomenon is that the two different cell types in conceptoid bodies are arranged according to the physiologically-correct spatial arrangements. The ICM cells are located at the centre of conceptoid bodies and the trophoblast cells forms a hollow sphere - the exact same arrangement of a blastocyst. I hypothesize that if a cell is placed in a suitable biological context, it can interpret the biological context and acts accordingly. The cellular action may include: cell migration, cell type differentiation or apoptosis. This hypothesis can draw support from a previous study. In [134], the newly emerged ICM cells are always located at the correct relative position of the reconstituted blastocyst.

This study [134], together with our results, suggest that the eventual ICM cells were able to respond to the context in which they were placed within a reconstituted blastocyst and migrate when necessary.

Based on the observations above, it is clear that restoring the 3D spatial contact plays an important role for cell differentiation, cell migration, or both. Unfortunately, we did not identify the underlying mechanism and pathway that control this process. These can potentially be studied by repeating the same conceptoid body derivations after perturbing relevant cell-to-cell contact protein-coding genes.

Studying single cell RNA-Seq data from blastocyst requires *a priori* clustering

In the human embryonic development dataset [128], among the 30 cells dissected from human blastocyst, a very small subset of them (eight cells) were identified as trophoblast cells during physical dissection because of their relative location in the blastocyst is clearly distinguishable. However, a human blastocyst consists of at least two cell types, and the identity of the other 22 cells from the blastocyst can either be trophoblast or ICM. In order to discover the cell types present in a blastocyst, it is necessary to classify all 30 cells into different cell types before any subsequent analyses can proceed. After removing 3 cells that failed quality check, I used the clustering method described in Materials and Methods M6, and identified two cell types present in blastocyst. Subsequently, because the identities of a small subset of trophoblast cells (i.e. the previously mentioned eight cells) are already known during dissection, this information is

utilized to identify all remaining trophoblast cells that are grouped together with the subset of eight trophoblast cells. Subsequently, the other cluster are assigned to the ICM cell identity. After cell type identities were assigned, the expression levels of known marker genes in ICM cells and trophoblast cells were inspected to confirm their expected expression patterns.

Potential doublet issue does not confound the results we obtained so far

Among all cells that we sequenced, we discovered two cell types represented by the blue square and the orange square shown in Figure 3-10. One cell type express known trophoblast specific genes; the other cell type express known ICM specific genes. As discussed in Chapter 2 Section 6. The older version of Fluidigm C1 microchips are susceptible of capturing doublets that cannot be excluded by microscopic inspection. Because this project use the older version of Fluidigm C1 microchips, this issue could potentially be a major confounding factor to our results presented in this chapter. For example, a cell type we identified could be doublets. Below, I formulate a hypothesis that my results can be explained by doublets, and show that the hypothesis has no merit and cannot be true. If I assume there is one underlying cell type, having doublets would not be able to explain the observed two cell types that each known marker genes of trophoblast and ICM cells. If I assume that there are two underlying cell types: A and B. Having doublets would cause the experiment to observe three perceived cell types: A, B, C (where C represent doublets consisting of A and B). The expression profile of C thus will be the average expression profile of A and B. In Figure 3-10, we can see that two observations would contradict the doublets

hypothesis: (i) there are two cell types instead of the three that doublets will predict, and (ii) both the two cell types being identified express several cell type-specific genes that are not expressed in the other cell type. Through this theoretical analysis, I conclude that the cell type we identified cannot be explained by the presence of doublets.

Comparing data from two different labs is challenging and requires special attention

In this experiment, we needed to compare single cells from conceptoid body with single cells from *in vivo* blastocyst. It would be ideal if the protocol used for sequencing the conceptoid body was applied to sequence *in vivo* human blastocysts, so that the single cells present in the organoid could be compared directly to the single cells in a human blastocyst without being distorted by batch effect or protocol differences. However, human blastocysts available for research are extremely scarce. Furthermore, because all the wet-lab experiments were carried out by our collaborators and the circumstances of our collaborators do not enable this protocol-matched experiment to be carried out. As a result, a sub-optimal strategy was adopted by comparing our dataset with a publicly available *in vivo* human blastocyst single cell RNA-Seq dataset. As a result, it was important to ensure that differences between protocols did not confound the analysis conclusions drawn from the datasets obtained using two different protocols.

In most biological experiments, if data is collected during several experimental batches, there will be a batch-specific effect which may potentially confound experimental results. When possible, this should be mitigated by stratifying samples across several different batches. This stratification ensures that the factor of interest (e.g. treatment versus control) is not confounded by experimental batch. However, a perfectly stratified experiment is not always possible in many scenarios, for example when some parts of the dataset have already been generated before realizing the need for stratification. For another example, if the data collection has to be performed in a manner that is incompatible with sample stratification, especially clinically obtained fresh patient tissue samples.

In this project, one of our aims was to match human conceptoid body samples with *in vivo* derived human blastocyst samples. In this aspect of the project, we encountered a batch effect that could not be mitigated because the *in vivo* derived human blastocyst single cell RNA-Seq dataset is a public dataset [128] that was generated by other labs using a single cell RNA-Seq protocol described in [13] and with manual handling. On the other hand, the conceptoid body single cell RNA-Seq dataset was generated in our lab using the SMART-Seq1 protocol on the Fluidigm C1 system. The two protocols are very different as explained in detail in Materials and Methods M10. As a result, even for the same sample, the outcome will be very different due to the batch effect. Given this potentially severe batch effect, it is a major challenge to match the cells present in them.

Although the batch effect is severe, it is fortunate that it is not confounding. This is both because all three different biological conditions of interest (human ESCs, *in vivo* ICM cells and *in vivo* trophoblast cells) are present in the [128] dataset and because all three different biological conditions of interest (human ESCs, conceptoid ICM-like cells, and conceptoid trophoblast-like cells) are present in our dataset. As a result, I was able to salvage the biological signal of interest from the batch effect. The analytical approach described in the Materials and Methods M15 was designed specifically for the purpose of minimizing the batch effect among these two protocols and focus only on the biological signal of interest.

It is also important to compare the developmental stages of the two different datasets. Conceptoid bodies were derived from human ESCs which no longer carry the usual notion of developmental stage; this implies that conceptoid bodies do not carry with them a precise post-fertilization time mark with which to assign developmental stage. A practical method for determining its developmental stage is based on morphology. Based on the opinion of our collaborators, the morphology suggests that the conceptoid bodies are potentially day 4-5 because this is the time when embryo cavitation occurs. Comparatively, the single cells being profiled in the [128] dataset were obtained from Day 6 pre-hatching blastocysts, a time which is slightly different than the inferred developmental stage of conceptoid bodies. This difference in development stage represents another major challenge of combining two datasets generated from two different labs.

3.6 Materials and Methods

M1 Single cell RNA-Seq assays

Cells were captured, lysed, and full length cDNA libraries constructed individually using the microfluidic C1 Auto Prep System (Fluidigm) and the SMART-Seq protocol (Clontech) as previously described in [147]. Illumina HiSeq sequencing libraries were generated using the Nextera XT kit (Illumina). Each library was barcoded individually before being multiplexed together for sequencing using the single-end 50 base pair (bp) format on HiSeq2000 (Illumina) and de-multiplexed and base-called using the Illumina software CASAVA1.8.

M2 RNA-Seq reads mapping and quality filtering on each cell

50 bp non-stranded single-end reads were mapped using GSNAP [148] to the human genome assembly (hg38) using the Ensembl75 gene annotation set [149]. Annotations not based on hg38 were lifted over to hg38 using the UCSC browser. In total, 9.8 x 10^8 reads were successfully mapped to the human genome assembly hg38; details of mapped reads for each cell are provided in Figure 3-16. We identified failed cDNA libraries as having very low fractions of exonic sequences, very few expressed genes, or very few successfully mapped reads. These libraries were most likely generated from apoptosed cells, cells in a bad overall condition, or cellular debris captured in the capture chambers of the C1 microfluidic device, where the starting mRNA is either highly degraded or absent. To prevent low quality libraries affecting the analysis and conclusions, cells were disregarded if they failed one or more of the following three quality filters. (i) Among all mapped reads, fewer than 35% are mapped to annotated

exons. (ii) Total number of expressed genes <8000 for day 3 conceptoid body (expressed genes defined as FPKM>1) and total number of expressed genes <10000 for human ESCs. This is approximately equivalent to the bottom quartile of each experiment because human ESCs have slightly more genes expressed in each cell. (iii) Total number of mapped reads < 1 x 10^5. In total, 48 cells obtained from day 3 conceptoid bodies and 39 unstimulated human ESCs passed these filters and were retained for subsequent analyses.

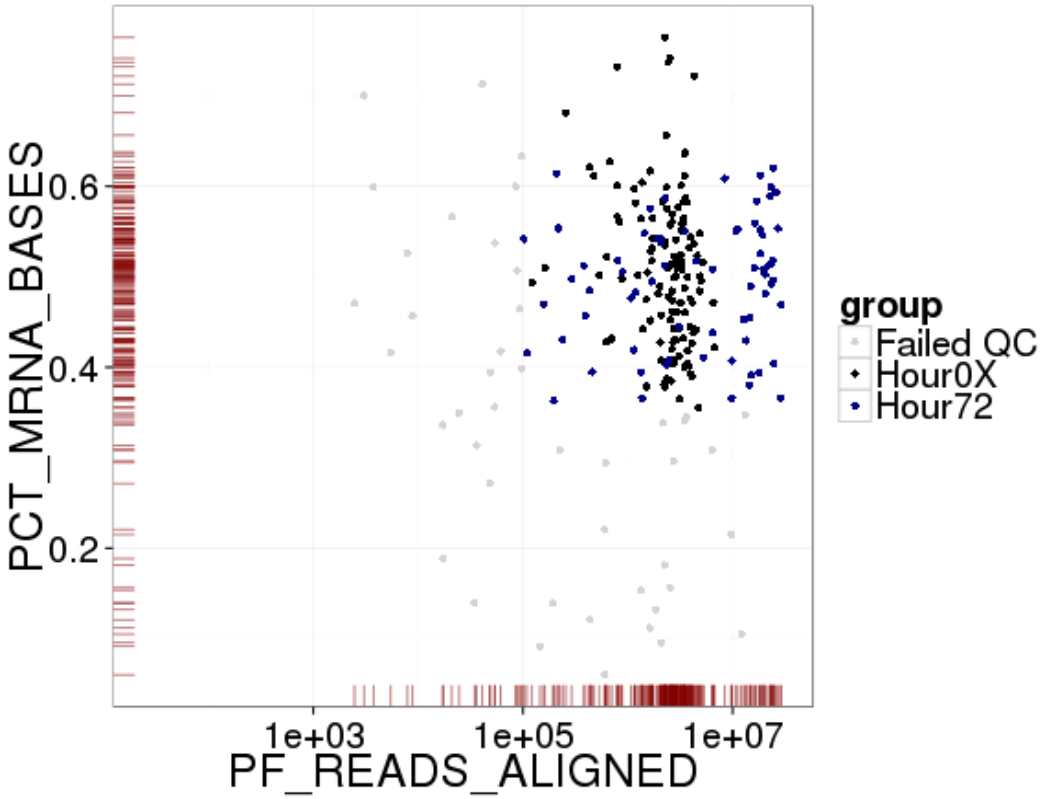


Figure 3-16: Summary of short read mapping statistics. Each dot represents a single cell. The Y-axis represent the fraction of nucleotides aligned to annotated mRNA exonic regions out of total number of nucleotides aligned in each cell. The X-axis represent the number of pass filter reads aligned to the reference genome hg38. The dots colored in grey represent cells failed either or both of the quality filters.

M3 Genes expression quantification

Gene expression abundance estimates were obtained using cufflinks version 2.0.2 [150]. The following command line options were used to obtain the abundance (FPKM): cufflinks --label reference –GTF reference.gtf --num-threads 25 --frag-bias-correct hg38.fa --library-type fr-unstranded --compatible-hits-norm --multi-read-correct library_name.bam >& reference_library_name.cufflinks. TPM values were obtained as previously described in [151]. Normalized read counts were obtained by normalization against the effect of library sizes using the estimateSizeFactors sub function in DESeq Version 1.12.0 [152].

M4 Cell type identification in hESC-derived conceptoid bodies

Due to the morphological characteristic of forming a liquid filled-cavity closely resembling the blastocoels in a human blastocyst, it is plausible that the conceptoid bodies would consist of both cell types present in the human blastocyst (i.e. ICM cells and trophoblast cells). We performed *a priori* hierarchical clustering based analysis to identify potential cell types among all cells passing quality control filters in the cavitated human ESC clumps (i.e. conceptoid bodies). Bioinformatics clustering was performed using the hclust function in R with the option method=”average”. The pair-wise weighted correlation coefficient [153] was used as the similarity measure, and sqrt (1 - similarity) as the dissimilarity measure [154, 155]. Subsequently, Dynamic Tree Cut [156] was used to identify the clusters and outliers from the dendrogram. We sought to investigate the cell types present in Day 3 conceptoid bodies based on clustering. We identified 2 clusters of cells, which are marked by the horizontal blue bar and orange bar respectively in Figure 3-17. The remainder of

the cells marked by grey lie far from both clusters and thus do not belong to either of the two main clusters as determined by Dynamic Tree Cut [156]. These cells marked by grey are considered as outliers and not analyzed further because they are potentially unhealthy cells with most mRNA degraded.

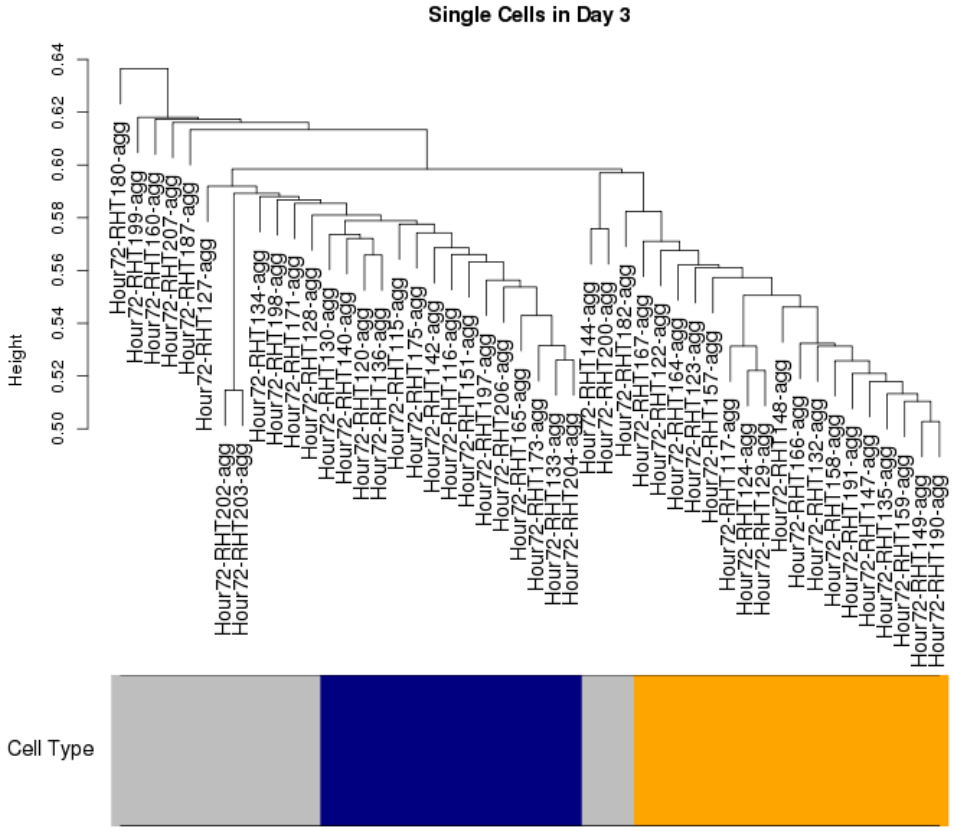


Figure 3-17: Clustering single cells obtained from conceptoid bodies. Upper panel: Dendrogram of hierarchical clustering of cells from Day 3 conceptoid bodies that passed quality filtering. Lower panel: cluster label assignment by Dynamic Tree Cut [156]. There are 2 clusters identified by Dynamic Tree Cut. The two distinct cell types are marked using blue and orange horizontal bars respectively. The cells marked by grey are the cells that do not belong to either cluster. There is a very tight cluster of 2 cells RHT202 and RHT203. These two samples are adjacent to one another on the Fluidigm C1 chip and on the 96 well place, which potentially represents an event of cross-well contamination caused manual pipetting error.

M5 Differential expression analysis

Differential expression analysis between any two clusters was performed using SCDE [153] on non-zero genes. A gene is deemed as differentially expressed if it satisfies the following criteria: (i) FDR<20%, (ii) Fold Change < 1/3 or Fold Change >3 across the different cell types, and (iii) the maximum likelihood estimate of the expression in the upregulated cell types is have FPKM > 100. The same cut-off criteria were used throughout the study. The maximum likelihood estimate of the expression in one cell type is computed as follows: for each gene, SCDE computes the joint posterior probability distribution function of expression value of all cells of the cell type and the mode of the joint posterior probability distribution was used to obtain the maximum likelihood estimate of expression value for that gene in the corresponding cell type.

M6 Quality Check and Cell type identification in IVF-derived blastocyst

After downloading the raw data from [128], I performed a quality assessment among the 30 cells dissociated from blastocyst for single cell RNA-Seq. 28 of the 30 cells pass the quality check filter described in Materials and Methods M2; the two cells that failed quality check were discarded. Because single cell RNA-Seq data analytical methods have progressed rapidly over the two years following the publication of this dataset [128], I decided to reanalyze the data using more contemporaneous methods that were applied in my thesis. Firstly, my *a priori* hierarchical clustering based analysis was applied to identify potential cell types among all the pass QC cells in the IVF-derived blastocyst. Clustering was performed using the hclust function in R with the option method="average". The pair-wise weighted correlation coefficient [153] was

used as the similarity measure, and sqrt (1 - similarity) as the dissimilarity measure [154, 155]. Subsequently, Dynamic Tree Cut [156] were used to identify the clusters and outliers from the dendrogram. We identified 2 clusters among all the pass QC cells in the IVF-derived blastocyst (Figure 3-18). Based on the expression of known pluripotency marker genes (Figure 3-10): *NANOG, TDGF1, GDF3* [135]*, NODAL* [157] and *IFITM1* [158], it is thus clear that the first cluster with 4 cells (2 cells from blastocyst BN2, 2 cells from blastocyst BN3) represent ICM cells. It is important to highlight that the results obtained using more contemporary methods differ from the results described in the original publication. I attribute these differences in results to the expression cut-off, distribution modelling and clustering methods between the analytical procedures. I believe that using more contemporary methods provides more reliable results. I thus chose to discard the analytical results described in the original publication. I also used the experimental evidence presented in the original paper to confirm my analytical results. The blue shaded cells are confirmed trophoblast cells as defined by physical location during dissection of each blastocyst according to the experimental evidence. These confirmed trophoblast cells are valuable, because they allow us to identify the rest of cells that belong to the same cluster with them as trophoblasts. Our cell type definition also matches well with a recent reanalysis of the same blastocyst dataset [135]. Furthermore, these cells also more abundantly express known trophoblast maker genes (Figure 3-10): *LRP2, SLC7A2, ABCG2, EFNA1, ENPEP* and *GATA3* [135].

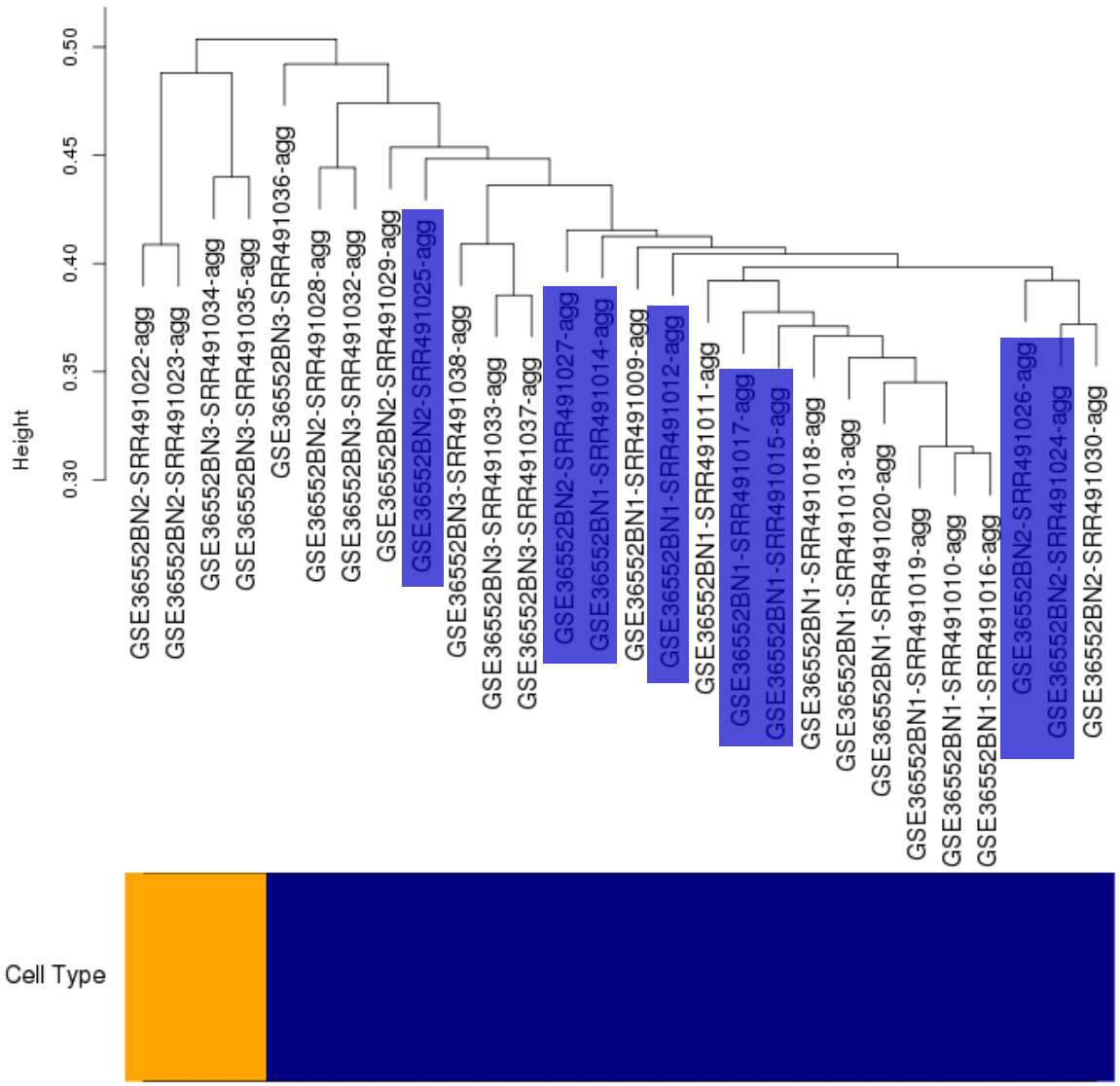


Figure 3-18: Clustering of single cells from human *in vivo* blastocyst. The clustering was performed *a priori* using all pass QC human blastocyst cells profiled by [128]. Upper panel: Dendrogram of hierarchical clustering of 28 cells from human blastocyst that passed quality filtering. Lower panel: cluster labels assigned automatically by Dynamic Tree Cut [156]. There are 2 major clusters – the ICM cluster and the trophoblast cluster. The last three clusters correspond to trophoblast cells derived from three different blastocysts. The three biological replicates (i.e. three independent separate blastocysts) are encoded in the library name as: BN1, BN2 and BN3. The blue shaded cells are confirmed trophoblast cells as defined by physical location during dissection of each blastocyst. These confirmed trophoblast cells are valuable because they assist the cell type classification of the rest of cells that belong to the same cluster with them.

M7 Data-driven analysis to examine the ICM and the trophoblast within conceptoid bodies as well as the ICM and the trophoblast within human blastocyst

The cell types in the conceptoid bodies are defined based on the *a priori* method described in section M4 above. After identifying the two cell types, I performed differential expression analysis between the two cell types as described in section M12. Likewise, the two cell types in blastocyst were identified *a priori* using the method described in M6 above. The second list of significantly differentially expressed genes was obtained as described in section M11. I intersected the two lists of significantly differentially expressed genes regardless of whether the direction of regulation was discordant or concordant in order to obtain the final list of 53 genes that were significantly differentially expressed (Figure 3-10). Of the list of 53 significantly differentially expressed genes, 49 are concordantly regulated (i.e. are regulated in the same direction for the *in vivo* human blastocyst and for the *in vitro* conceptoid bodies). I used the sign test to assess the significance level of concordance, with the null hypothesis stating that discordant and concordant directions of regulation are equally likely. In this case, the sign test rejects the null hypothesis and provides evidence for significance ($p<7.0\text{x}10^{-11}$), obtained using the R command binom.test (49, 53, p = 0.5). This level of consistency is highly significant.

M8 Expression similarity to a panel of human tissues

Bulk reference transcriptomes were downloaded from GSE2361 (36 normal human tissue samples) and GSE9773 (2 placental tissue samples, each containing 5-6 technical replicates). Genes probed in both datasets were chosen for full

quantile normalization. For GSE9773, all technical replicates of extra villous trophoblast (EVT) were averaged to form one single EVT profile, likewise for cytotrophoblast (CTB). Subsequently, the 36 normal human tissue samples, one EVT sample and one CTB sample were combined to form the final reference panel of tissues. Highly variable genes within this final reference panel of tissues were selected based on their fold-change over the median expression (log2(fold - change) > 3). The pair-wise similarity between each human ESC-derived trophoblast and each reference tissue was computed using the Pearson correlation coefficient for the highly variable genes.

M9 Similarity of *in vivo* and *in vitro* blastocystic trophoblast

We compared *in vivo* blastocyst gene expression to *in vitro* human ESC-derived Day 8 blastocystic trophoblast gene expression using protocol matched samples from [123]. The whole-blastocyst transcriptome is highly representative of blastocystic trophoblast, because over 90% of cells in the human blastocyst belong to the trophoblast lineage [159]. The transcriptomes of human blastocyst and human ESCs profiled in [123] were obtained using the Affymetrix Human Gene 1.0 ST Array. Its associated data was obtained from GEO under the dataset accession GSE29397. For differential expression analysis, I used the eBayes function in the LIMMA [160] package to perform differential expression analysis contrasting whole pre-implantation embryo (3 replicates) and human ESCs (4 replicate) profiled in [123], and obtained the list of top 50 genes upregulated in human ESCs using the function topTable in the "LIMMA" package in R. Likewise, I obtained a list of the top 50 genes upregulated in human blastocyst. This list of

100 genes was used to assess the concordance of gene expression in our study's monolayer cultured blastocystic trophoblast. If a gene (e.g. *TUBB4A* in Figure 3-3) is up-regulated in human ESCs in the *in vivo* [123] dataset, and is down-regulated in the human ESCs in our dataset, this gene is deemed to be discordant, and is indicated in grey. Likewise, if a gene is regulated towards the same direction in both datasets, it is considered as a concordant gene, indicated in blue.

Among the top 100 differentially expressed genes discovered using the *in vivo* dataset, 87 are regulated in the same direction *in vitro* Figure 3-3. I used the sign test to assess the significance level of concordance, with the null hypothesis stating that discordant and concordant directions of regulation are equally likely. In this case the sign test provides strong evident to reject the null hypothesis ($p < 1.3 \times 10^{-14}$), obtained using R command binom.test (87, 100, p = 0.5). This level of consistency is highly significant.

Where necessary, Affymetrix Human Gene 1.0 ST Array probe sets were mapped to Ensembl IDs using AltAnalyze probe-set to Ensembl mapping downloaded from GEO under the platform accession GPL18401. Mapping between different Ensembl versions was performed using Ensembl BiomaRt.

M10 Protocol differences

The single cell RNA-Seq protocol employed in [128] differs from our single cell RNA-Seq protocol in two key respects. First, the method for first strand synthesis is different, with the *in vivo* blastocyst data [128] generated using the method described in [13], whereas our dataset is based on the template-

switching method described in the SMART-Seq protocol [137]. Second, the cDNA fragmentation steps also differ, where [128] is based on shearing the DNA using Covaris and our data is based on the transposon tagmentation method provided by the Illumina Nextera Kit. This is further compounded by the fact that our data was generated using tiny volume reactions in the Fluidigm C1 microfluidic devices, whereas the regular reaction volumes were used in [128]. Such substantial differences between the two protocols could result in substantial differences between the resultant datasets. For example, some genes might be reliably detectable by one protocol but not by the other protocol. Consequently, only genes that are reliably measureable in both protocols can be used for downstream analysis. Furthermore, the relative expression values that are centred at the weighted mean of the expression values for each protocol being the most meaningful value to be used for all analysis that mitigates this protocol differences. The expression values of individual cells are then divided by the weighted mean to obtain the relative fold change, which is then taken log2 to obtain the normalized express value.

M11 Differential expression analysis between blastocystic trophoblast and ICM cells derived from *in vivo* blastocyst

Differential expression analysis was performed by contrasting human blastocystic trophoblast cells from *in vivo* blastocyst and human ICM cells from *in vivo* blastocyst according to the differential expression analysis method described in section M5 and followed by the normalisation described in section

M10 above. The same thresholds and parameters described in section M5 were applied.

M12 Differential expression analysis between blastocystic trophoblast and ICM cells derived from *in vitro* conceptoid bodies

Likewise, an independent differential expression analysis was performed by contrasting blastocystic trophoblast cells from *in vitro* conceptoid bodies and the ICM cells from *in vitro* conceptoid bodies.

M13 Differentially expressed genes between human ICM cells from in vivo blastocyst and protocol-matched human ESCs

Differential expression analysis was performed by contrasting human ICM cells from *in vivo* blastocyst and protocol-matched human ESCs from the same experiment according to the differential expression analysis described in section M5 above and followed by the normalisation described in section M10 above. The same thresholds and parameters described in section M5 were applied.

M14 Differentially expressed genes between human ICM cells from in vitro conceptoid bodies and protocol-matched human ESCs

Likewise, an independent differential expression analysis was performed by contrasting ICM cells from *in vitro* conceptoid bodies and protocol-matched human ESCs from the same experiment.

M15 Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of possibly correlated variables into a set of uncorrelated variables called principal components. It is useful in identifying the major patterns of variation in a high dimension dataset and display it in a lower dimension which can be more intuitively visualized and interpreted by human beings. PCA were performed using all pass filter cells sequenced in our experiment: 39 human ESC cells, 15 trophoblast-like cells and 18 ICM-like cells from conceptoid bodies. The first two principal components capture the largest fractions of the total variance (28.5% and 6% respectively), thus the first two Principal Components (PCs) were displayed in Figure 3-9A. All blastocyst cells and human ESCs (passage 10 hESCs) profiled in [128] and passing filters were projected into the same PCA space. All cells generated by our experiment are displayed as squares and all cells generated by [128] are displayed as circles. The cell type identities were inferred by the clustering method described previously in sections M4 and M6. The list of genes used for the principal component analyses was obtained by joining the two lists of gene expression values obtained from section M11 and M13 above. We only consider the *in vivo* blastocyst dataset to define the list of biologically relevant genes for the principal component analysis, because each of these cells has a clearly defined *in vivo* cell type identity.

M16 Analysis of single cell expression detected by qRT-PCR

At Days 1, 2, 3 and 4 during the monolayer trophoblast differentiation and un-stimulated human ESCs (named as Day 0), single cell expression of *ZSCAN10,*

NANOG, POU5F1(OCT4), GATA2, GATA3, MSX2, ACTB and *GAPDH* was measured using qRT-PCR on approximately 20 – 30 cells at each time point to obtain raw Ct values. The average expression of *ACTB* and *GAPDH* were used as normalizing constant for the other six cell type-specific marker genes: *ZSCAN10, NANOG, POU5F1 (OCT4), GATA2, GATA3, MSX2*. Ct values exceeding 28 were considered undetectable, and were set to 28. Subsequently, the Ct values were inverted using the formula: (28 – Ct), so that larger values represent high expression and lower values represent lower expression. The values were then further normalised linearly according the formula [Ct - Min(Ct)] / [Max(Ct) - Min(Ct)], so that all output values lie in the range of [0,1]. The normalized expression values were plotted as the size of the pie chart, using the "stars" function in R.

M17 Statistical analysis of attachment assay and cell migration assay

JAR: Cell clumps are derived from a choriocarcinoma cell line (i.e. JAR cell line [138]; positive control) and were significantly more successful in the attachment assay when compared to embryoid bodies (EB). Statistical test was performed using the categorical label (i.e. JAR, EB) as the independent variable, and the attachment success rate as the response variable. The null hypothesis was that the response variable was not influenced by the categorical label. The test was performed using the R function: summary (lm (as.vector(success_rate) ~ as.vector(categorical_label))).

For examining cell attachment experiment performed on gelatine as substrate, the significance levels are the following: JAR vs. EB, the null hypothesis

was rejected with $p < 2.3 \times 10^{-4}$. Conceptoid Bodies vs. EB, the null hypothesis was rejected with $p < 1.2 \times 10^{-3}$. Furthermore, as expected, JAR vs. Conceptoid Bodies comparison was not significantly different ($p > 0.1$), suggesting the behaviour of conceptoid bodies are not distinguishable from that of the positive controls.

For examining cell attachment experiment performed using the endometrium cell line (RL95-2 cell line) as attachment substrate, the significance levels are the following: JAR vs. EB, the null hypothesis was rejected with $p < 2.5 \times 10^{-8}$. Conceptoid Bodies vs. EB, the null hypothesis was rejected with $p < 3.1 \times 10^{-5}$. Furthermore, as expected, JAR vs. Conceptoid Bodies comparison was not significantly different ($p > 0.1$), suggesting the behaviour of conceptoid bodies are not distinguishable from that of the positive controls.

For the cell migration experiment, the significance level of applying the above described statistical test is the following: the null hypothesis was rejected ($p < 1.3 \times 10^{-7}$).

Chapter 4 Identification of Different Classes of Dopaminergic Neurons in the Mouse Midbrain

4.1 Abstract

Dopaminergic neurons in the mouse midbrain consist of a heterogeneous population of cells. Most existing investigations of the molecular heterogeneity of these neurons are based on dissecting the midbrain into multiple regions. The resolution of these studies is limited to distinguishing neurons that occupy distinct anatomical regions of the midbrain. Our study, using single cell RNA-Seq technology, improved the resolution by which different classes of midbrain dopaminergic neurons can be identified. We identified up to 10 potential classes of midbrain dopaminergic neurons using an entirely data-driven approach. Two classes, the *Aldh1a1* positive class and the *Aldh1a1* negative class within the ventral tegmental area, have been validated by experimental approaches in other studies [3], thereby confirming the validity of our single cell RNA-Seq dataset and my bioinformatics analysis methods. In addition to the two known classes, other novel classes of potential midbrain dopaminergic neurons were also identified. For example, the *Lix1* positive class within the dorsal substantia nigra pars compacta, and the *Crhbp* positive class within the ventral tegmental area and within the ventral midline area. I also identified the marker genes for each predicted class. Using RNA *in situ* hybridisation on their respective markers, the various predicted classes of midbrain dopaminergic neurons were shown to be localised to different regions of the midbrain by my collaborators.

4.2 Publication and Contributions

This Chapter is based on the work described in "Identification and characterization of midbrain dopaminergic neuron subtypes" (Mary Green, Tim Xiaoming Hu, et al.). This manuscript is being prepared and will be submitted to a peer reviewed journal. This is a collaborative project between the Francis Crick Institute (formally the National Institute of Medical Research, MRC, London, UK) and the MRC Functional Genomics Unit (Oxford, UK). Dr. Mary Green, Dr. Siew-Lan Ang, Prof. Chris Ponting, Dr. Wilfried Haerty and I designed the study together. The sample preparation, micro-dissection, Fluorescence Activated Cell Sorting (FACS), single cell RNA-Seq and experimental validation experiments were performed by Dr. Mary Green - the other co-first author. Operation of the Fluidigm C1 Autoprep system and the sequencing by Illumina HiSeq were carried out by the single cell core facility of the Francis Crick Institute. I performed the bioinformatics analyses under the close guidance of Prof. Chris Ponting and Dr. Wilfried Haerty. All parties in the collaboration actively participated in the interpretation of analytical results.

4.3 Introduction

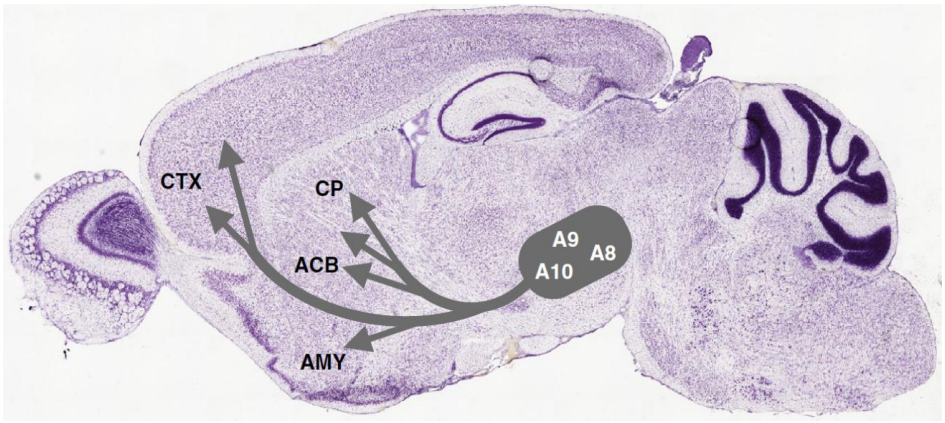
Motivation

Dopaminergic neurons are an important group of neurons located in many regions of the brain and have complex projections toward many other brain regions (reviewed in detail by [161]). Of all dopaminergic neurons, about 75% [162] are located within the midbrain and play vital roles in the regulation of complex behaviours, such as voluntary movement, emotion, addiction [163], food intake [164] and risk/reward behaviour [165]. Abnormalities of midbrain dopaminergic neurons have been associated with neurological diseases, such as depression [166], Parkinson's Disease [167] and schizophrenia [168]. Most notably, implantation of human foetal-derived midbrain dopaminergic neurons into Parkinson's disease patients has been shown to relieve some symptoms [169]. There would be tremendous medical potential if appropriate midbrain dopaminergic neurons could be differentiated from stem cells *in vitro*, followed by their implantation into the appropriate brain region to treat Parkinson's Disease. It is known that only specific classes of dopaminergic neuron primarily degenerate in Parkinson's disease [170]. This suggests that dopaminergic neurons are heterogeneous. Clinical applications heavily depend on a comprehensive understanding on the signalling and molecular controls during development and differentiation of various midbrain dopaminergic neuron classes [171]. But, unfortunately, we know little about different classes of mature midbrain dopaminergic neurons, their molecular markers and functional properties. A thorough understanding of different classes of midbrain dopaminergic neurons is essential to understand regulation of specific

behaviours and related diseases. This knowledge ultimately could contribute to the management and treatment of these behaviours and related diseases.

Ideally, to maximize medical utility, knowledge of midbrain dopaminergic neurons should be acquired directly from human brain samples. However, it is extremely difficult to obtain fresh human brain samples donated for medical research. Even frozen human brain samples donated for medical research are rare. Furthermore, validation experiments and further functional investigations are generally invasive. In general, such invasive investigations using live human brains are ethically unacceptable. Due to these sample limitations, most research on the brain that requires fresh brain tissue samples choose instead to experiment on the mouse brain, as in this study. Major benefits of working with the mouse are manifold. (i) Many experimental validations such as RNA *in situ* hybridisation, immuno-staining and *in vivo* perturbation studies are possible. (ii) The ability to make transgenic mice for experimental use, especially the ability to purchase various transgenic mice that are readily available from mouse strain services (e.g. Jackson Laboratory) has made many genetic experiments more convenient and rapid. (iii) Being a highly popular model organism, there is a vast amount of literature related to mouse midbrain dopaminergic neurons, describing their axon projections, functions and molecular signatures of midbrain dopaminergic neurons from various brain regions ([161, 172, 173] and references therein). For a pictorial description regarding the locations and axon projections of mouse midbrain dopaminergic neurons, please refer to Figure 4-1A. For a coronal section of the midbrain, please refer to Figure 4-1B.

A



B

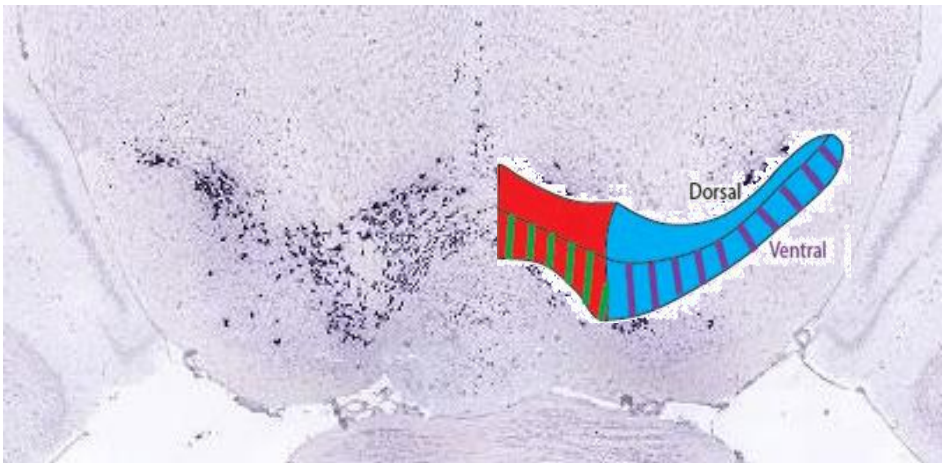


Figure 4-1: Stained sections of mouse brain showing the midbrain dopaminergic neurons. Panel A: A sagittal section of the mouse brain. The schematic showing midbrain dopaminergic neurons in adult mouse and their axon projections to other regions of the mouse brain. The naming of A8 to A10 regions was first introduced by Dahlstroem and Fuxe in 1964 [174], and remains in use in the current literature. The A9 region is the substantia nigra pars compacta (SNc), the A10 region is the ventral tegmental area (VTA). We refer to the dopaminergic neurons in these two regions collectively as the midbrain dopaminergic neurons. Arrows illustrate major axon projections of dopaminergic neurons to other parts of the mouse brain: amygdala (AMY), caudate putamen (CP), accumbens (ACB) and cortex (CTX). Image reproduced from [173], which is published under Elsevier User License that permits reuse of the figure for non-commercial purposes (https://www.elsevier.com/about/company-information/policies/open-access-licenses/elsevier-user-license). Panel B: A coronal section of mouse midbrain provided by my collaborator Dr. Mary Green. The staining of TH protein in the mouse midbrain marks the dopaminergic neurons. A cartoon illustration superimposed on the right half of the midbrain is provided by my collaborator Dr. Mary Green. The cartoon depicted the VTA

region in red and the SNc region in blue. The VTA region is divided into two tiers: the ventral tier (marked by green vertical bars) and the dorsal tier (not marked by vertical bars). Likewise, the SNc region is divided into two tiers: the ventral tier (marked by blue vertical bars) and the dorsal tier (not marked by vertical bars).

Much evidence demonstrates that mouse midbrain dopaminergic neurons are not a homogeneous group of cells [3]. Several recent reviews on the currently known molecularly defined classes of midbrain dopaminergic neurons are available [173, 175] . So far, it is often assumed that there are two major classes of midbrain dopaminergic neurons: the substantia nigra pars compacta (SNpc; or SNc) dopaminergic neurons - projecting to the striatum; and the ventral tegmental area (VTA) dopaminergic neurons - projecting to limbic and cortical areas. In the past, investigations have been carried out attempting to reveal heterogeneity among midbrain dopaminergic neurons. This is achieved by first dissecting out the SNc region and the VTA region, followed by using microarrays to measure expression profile for both samples. By contrasting expression profiles of the two samples, this study was able to identify the molecular markers for VTA and SNc respectively [176]. These markers included *Sncg* and *Sox6* that are specifically expressed in the dopaminergic neurons in the SNc region; and *Grp, Lpl* and *Otx2* that are specifically expressed in the dopaminergic neurons in the VTA region.

However, these dissection based approaches suffer from two important problems: firstly, assuming that some neuronal classes are anatomically segregated, then without molecular markers it is not possible to delineate exact

boundaries; secondly, some classes of midbrain dopaminergic neurons can be evenly distributed within an anatomical region, making them difficult to purify without a molecular marker. As a result, most existing studies have relied on coarse anatomical delineation of different classes of dopaminergic neurons, such as VTA and SNc. Considerable evidence suggests that current coarse anatomical delineation represents an oversimplification. A few examples are discussed below. (i) It has been shown that dopaminergic neurons of the SNc not only project to the striatum, but also to the limbic and cortical areas. [177-179]. (ii) Dopaminergic neurons of the VTA also project to the ventral striatum and the ventro-medial part of the caudate-putamen in rodents (the primate equivalent is the nucleus caudatus) [177]. (iii) Dopaminergic neurons located in the dorsal aspect of VTA and SNc are round or fusiform in shape. They also express high levels of *Calb1* (calbindin) and low levels of *Slc6a3* (dopamine transporter) [161]. (iv) Dopaminergic neurons in the ventral tier are more densely packed in a sheet and angular in shape [161]. (v) Studies of Parkinson's disease have shown that one class of dopaminergic neurons located in the SNc region of the midbrain is more susceptible to toxins and other stressors, and that these neurons die early in patients with Parkinson's disease ([180, 181] and references therein). (vi) Studies have also shown that positive and negative motivation signals are conveyed via different types of dopaminergic neurons [182]. Regarding the molecular events that occur during the maturation of midbrain dopaminergic neurons, limited information on the prenatal development of midbrain dopaminergic neurons is available through one-by-one gene knockout or gene over-expression experiments (reviewed in [172, 175, 183]). Overall, despite of

multiple evidence that suggests the existence of several classes of midbrain dopaminergic neurons [3, 184], a comprehensive molecular description of different classes of midbrain dopaminergic neurons remains lacking. Non-omics experimental approaches based on known markers lack the ability to identify new marker genes or novel classes. The absence of molecular markers has limited investigators' abilities to perform many types of molecular experiments, such as the localization of different classes of dopaminergic neurons in the midbrain using immuno-staining; performing conditional knock-out, knock-in experiments; or performing other functional characterizations such as opto-genetics. Thus there is a considerable need to identify and to molecularly characterize different classes of midbrain dopaminergic neurons.

Recently, efforts have been made to characterize different classes of midbrain dopaminergic neurons using single cell qRT-PCR [3]. SLC6A3 protein is a dopamine transporter and it is assumed to be expressed in all mature midbrain dopaminergic neurons. The authors used the transgenic mouse with Slc6a3::cre::Ai9 which has been previously described to mark the majority of midbrain dopaminergic neurons [185, 186]. This allowed the authors to enrich for midbrain dopaminergic neurons using FACS by assessing the levels of dTomato protein. Expression was assessed using a microfluidic system capable of measuring 96 genes in 96 samples concurrently to perform 9,216 single cell qRT-PCR experiments. In total, 274 neurons were profiled over 3 batches. However, there is a major drawback of this single cell study. As explained in Chapter 1 Section 2 of the thesis, non-omics approaches require all genes for the

experiment to be subjectively selected beforehand. To select the set of genes, the authors first included several housekeeping genes (*Gapdh, Hprt* and *Actb*), and several genes essential to carry out dopaminergic neuron function (*Ddc, Th, Slc6a3* and *Slc18a2*). The housekeeping genes were measured so that the quality of each cell could be assessed, and only high quality cells were retained for downstream analysis. The four dopaminergic neuron essential genes were measured to ensure that only genuine dopaminergic neurons were retained for downstream analysis. The authors then relied on prior literature and experience to select a set of genes in the hope that some these genes would be good markers of different subpopulations of midbrain dopaminergic neurons. For this reason, the authors selected genes that were known to be differentially expressed between VTA and SNc dopaminergic neurons based on the literature and the author's preference, and genes shown to be related to Parkinson's disease: *Atp13a2, Lrrk2, Park2, Park7, Pink1* and *Snca*. Their results showed that, among these genes, some are indeed heterogeneously expressed among the captured neurons. Based on the PCA analysis presented in the original paper, I argue that, of the 96 genes examined, only four: *Aldh1a1, Slc32a1, Calb1, Vip* mark a specific class of midbrain dopaminergic neurons. For example, the results of the PCA analysis performed by the original authors are reproduced in Figure 4-2. The loadings plot on the right suggests that Class 1A indicated in dark-red is heavily weighted by *Aldh1a1* on PC3, and not by other gene. The small number of cell type specific genes identifiable by this approach is due to the aforementioned drawback that only a limited number of pre-selected genes could be considered (96 genes in this case). Furthermore, many novel cell types

may not be distinguishable if their markers are not selected to be among the 96 genes. Had the authors been able to measure the entire transcriptome, they would be able to identify far more marker genes for each of the classes they identified; they would also be able to identify other novel cell types.

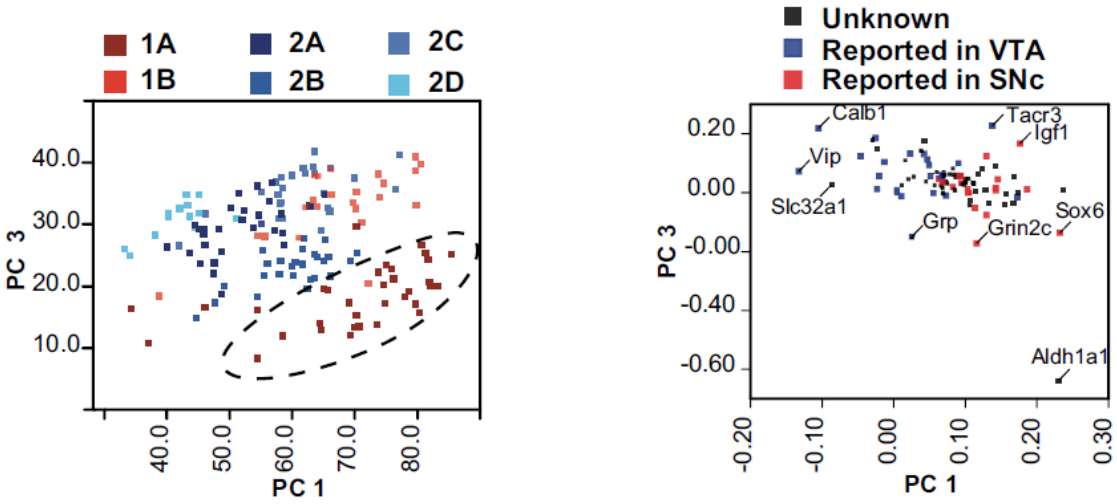


Figure 4-2: The results obtained using PCA by the original authors. It shows the results of principal component analysis (PCA) performed by the authors in [3]. The left panel shows the positions of cells in the PCA space on dimension PC1 and PC3. Based on the left panel, it can be seen that the cells marked by dark-red colour, named as Class 1A dopaminergic neurons are positioned far from other cells. The right panel shows the loadings of the genes that contributed to PC1 and PC3. We can see that the major distinctions of Class 1A dopaminergic neurons can be attributed mostly to the expression of *Aldh1a1*. This figure is reproduced from [3] which was published with the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/).

Recently, a dramatic improvement in technology has enabled RNA-Seq to be performed on individual cells. Thus, for the first time, we have been able to interrogate almost the entire poly-adenylated transcriptome of individual midbrain dopaminergic neurons. This has the potential to identify different classes of midbrain dopaminergic neurons that neither have any known markers

nor are separable by anatomical dissection. For a hypothetical example, if VTA were to contain two classes of dopaminergic neurons, and the two classes are evenly distributed within the same physical location, a more traditional study using dissection would fail to distinguish the two different classes present in the VTA. With single cell RNA-Seq, both classes of dopaminergic neurons within the VTA would be captured and profiled as individual cells, allowing them to be separated into two distinct classes bioinformatically.

We thus sought to apply single cell RNA-Seq on midbrain dopaminergic neurons to generate a transcriptomic description of them at postnatal day 4. We hypothesised that this would enable us to catalogue the different classes of midbrain dopaminergic neurons that are present during this developmental time point.

Experimental Design

The developmental stage for all single cell RNA-Seq experiment was selected to be postnatal Day 4 for two reasons. Firstly, due to the two opposing requirements explained below: due to their complex axon and dendrite connections, the more mature neurons are more difficult to dissociate into single cells while maintaining their viability and integrity. Consequently, an earlier time point was preferred. On the contrary, although midbrain dopaminergic neurons are post-mitotic from embryonic Day 15 onwards, they may remain immature over several days. Immature dopaminergic neurons might not yet have differentiated into different classes, and could appear to be a homogeneous

population of cells. Thus, a later time point was preferred. Between these two opposing requirements, the interval between embryonic day 15 to postal natal Day 4 is acceptable, because within this interval, neurons remain capable of being dissociated into single cells and yet are reasonably mature. Secondly, due to existing independent dataset published previously. Subtypes of dopaminergic neurons have previously been investigated using single cell qRT-PCR [3]. This study choose to carry out their single cell investigations using postnatal Day 4 mouse. Having an independent dataset generated at the same time point potentially could facilitate cross-validation and comparisons between the two independent datasets. As a result, postnatal Day 4 were selected as the time point to carry out all investigations for this experiment.

The mouse brain is a complex tissue consisting of many cell types: a recent survey of the mouse cortex revealed that there are at least 47 molecularly distinct subclasses of cells [98]. Although a full survey of cell types present in the mouse midbrain has not yet been carried out, it is plausible that the complexity of the midbrain is similar to that of the cortex, and midbrain dopaminergic neurons represent only a small fraction of the total number of cells present in the mouse midbrain. When we pseudo-randomly capture single cells from the midbrain to perform single cell RNA-Seq, we will inevitably capture midbrain dopaminergic neurons and cells of many other types. Depending on the relative abundances of each cell type, if in our experiment we did not enrich for midbrain dopaminergic neurons, we might have had to profile thousands of cells to obtain a handful of midbrain dopaminergic neurons which would be extremely

inefficient and costly. As a result, we applied a method to enrich for the midbrain dopaminergic neurons. Our collaborator, Dr. Mary Green in the Francis Crick Institute chose to use a mouse strain that replaced one allele of *Pitx3* with Green Fluorescence Protein (*GFP*) gene. This was because *Pitx3* gene expression has previously been shown to specifically mark dopaminergic neurons in the mouse midbrain [187-189]. With GFP protein marking midbrain dopaminergic neurons, Dr. Mary Green was then able to dissect out the relevant regions with the help of a fluorescence microscope. Furthermore, again with GFP distinguishing individual midbrain dopaminergic neurons, after she dissociated the dissected mouse midbrain, Dr. Mary Green used FACS to enrich specifically for GFP protein positive midbrain dopaminergic neurons from the cell suspension. Thereafter, the sorted population was re-suspended and single cells in the sorted population were pseudo-randomly captured by the Fluidigm C1 system and profiled using SMART-Seq1 protocol on the Fluidigm C1 system by the core facility of the Francis Crick Institute. The sequencing dataset was passed to me for analyses.

During the cell capture stage, it is common for some capture sites to capture two or more cells (i.e. doublets), for example, in one of our experiment 4 out of 96 capture sites contained two cells. These doublet capture sites, if visible, were identified and excluded from sequencing. However, not every doublet capture sites are visible, especially if the doublet cell is captured at a region beneath the typical viewing z-plane. After all data were generated, we became aware that the version of the Fluidigm C1 microchip we used is prone to capture cells beneath the typical view z-plane, thus some doublets may not have been

excluded. In order to validate that this issue was not confounding for our experiment, our collaborator Dr. Mary Green examined multiple z-plane views of a fully loaded Fluidigm C1 microchip using a microscope and showed no doublets were captured at a different z-plane. We conclude that doublets, if any, were already excluded by visual inspection. Thus this issue does not affect our study.

I performed all bioinformatics analyses whose goals were three-fold. First, to confirm that a substantial fraction of cells profiled in our single cell RNA-Seq were midbrain dopaminergic neurons. Second, to verify that these captured midbrain dopaminergic neurons are mature, and thus have already differentiated into distinct classes of midbrain dopaminergic neurons, especially that the previously known two classes of midbrain dopaminergic neurons (namely VTA and SNc specific dopaminergic neurons) [176] can be recapitulated in our single cell RNA-Seq data. Last, I sought to discover new classes of midbrain dopaminergic neurons and novel molecular markers of each class of midbrain dopaminergic neurons.

To achieve these analytical goals, different analysis methods were required. For the first goal, the expression of markers of dopaminergic neurons and all other cell types that are known to be present in the brain were examined, in order to separate dopaminergic neurons and non-dopaminergic neurons in our dataset. For the second goal, I clustered dopaminergic neurons into several groups based on the transcriptome, and plotted the expression of genes that are known to mark subpopulations of midbrain dopaminergic neurons in order to examine whether these known markers were enriched in some groups identified

using my clustering. For the final goal, it is important to note that by making use of existing knowledge to limit genes that participate in analyses can potentially misguide the analyses to only discover cell types that have been studied and obscure cell types that are currently unknown. As a result, novel classes of midbrain dopaminergic neurons shall be identified using data-driven approach based on all expressed genes. I then could identify their marker genes by performing differential expression analysis that uncovered genes that are specifically expressed in each class of midbrain dopaminergic neurons.

4.4 Results

Overview of all subpopulations of cells that will be discussed in this chapter

My analyses discovered that among all GFP+ cells being captured in our study, 172 passed quality filtering and could be separated into two major groups. As shown by the second row of Figure 4-3, one group contained 62 cells (marked by the light to dark grey bar; referred to as grey-scale bar thereafter) that did not highly express *Th, Ddc, Slc6a3* and *Slc18a2,* whose expression is essential for of dopaminergic neurons. The other group contained 110 cells (marked by the red to sky-blue multi-colored bar; referred to as red-to-skyblue bar thereafter) that highly express these four genes, which we consider to be mature dopaminergic neurons. As shown by the third row of Figure 4-3, among the 110 mature dopaminergic neurons, my data-driven analyses identified two major known classes: VTA and SNc dopaminergic neurons; among the 62 other GFP+ cells, my data-driven analyses identified two major cell types. The successful identification of known classes (i.e. VTA and SNc dopaminergic neurons) validated my data-driven analysis approaches. Other subpopulations were experimentally validated by our collaborators Dr. Mary Green and Dr. Siew-Lan Ang. Figure 4-3 shows the global overview of all seven validated subpopulations as the seven leaf nodes. This overview will be referred to throughout this chapter. Furthermore, the structure of the tree hierarchy presented in Figure 4-3 provides the framework of how the results will be presented. As a result, this figure will be repeated in later sections with different highlighted aspects in order to indicate specific groups of cells that are discussed.

It is worth noting that additional putative subpopulations were identifiable among the leaf nodes in this figure based on my analyses. However, due to space constraints, the leaf nodes in this figure were not split further to indicate these putative subpopulations.

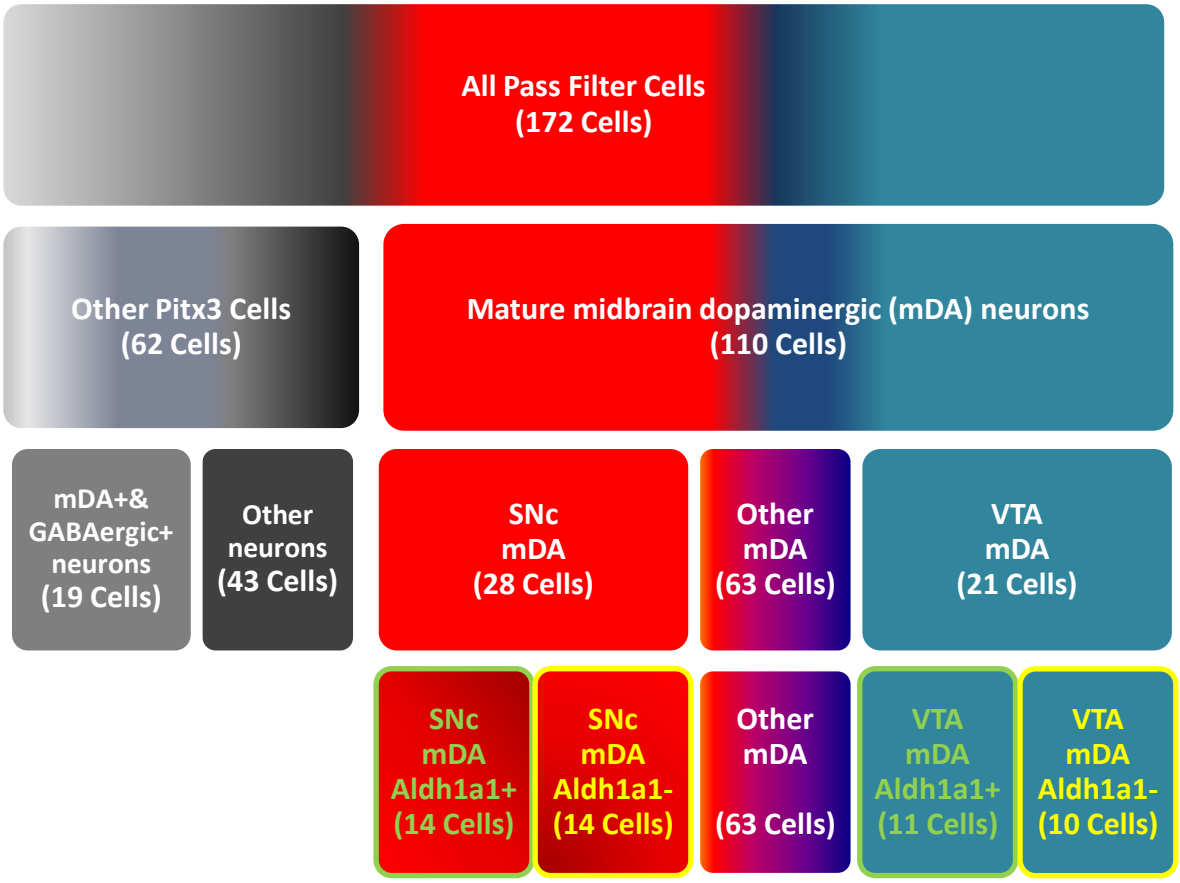


Figure 4-3: Overview of all cell types discovered in this project. The illustration showing the global hierarchical clustering structure and assigned cluster names for all subpopulations that were identified and validated in this project.

The first iteration of analyses split all 172 pass QC cells into two major groups: 110 dopaminergic neurons and 62 other GFP+ cells

Although the *Pitx3* gene marks dopaminergic neurons in the developing midbrain, it is not known whether other cells near the midbrain region also express *Pitx3* at postnatal Day 4. As a result, it was important to determine whether all cells, or only a subset of cells, captured in this experiment are dopaminergic neurons. To achieve this using an entirely data-driven approach, I first sought to determine the major cell types present in our dataset generated at this time point, followed by examining several known markers of dopaminergic neurons.

All 172 cells that passed quality filters (Materials and Methods D2) were used for downstream analysis. This quality filter was a general filter that only assessed the quality of the cells without considering midbrain dopaminergic neuronal markers. As a result, it was expected that among the 172 cells that passed QC, some may not be dopaminergic neurons. The *a priori* clustering of these cells was achieved using the method proposed in [190] (Materials and Methods D5). The method requires a set of GO terms as input, in our case downloaded from biomaRt [191]. The resulting dendrogram of applying this method is shown in Figure 4-4A. After the cells were organised into this dendrogram, I first sought to verify that the clustering of cells was not due to confounding factors such as the number of genes detected in each cell or experimental batches. This was achieved by colouring cells according to these potential confounding factors. It can be seen in the Figure 4-4A,B that there is a random distribution of experimental batches and a random distribution of genes

detected that are spread evenly among the clustering dendrogram. This indicates that these potential confounding factors are not the main factors driving the formation of the clustering dendrogram. The number of genes detected by the two major clusters were not significantly different when examined using Student's t test ($p > 0.25$).

Subsequently, we sought to understand the cell types represented in this dendrogram. Because the experimental design was to investigate dopaminergic neurons, and sorting on GFP protein was known to be able to enrich for dopaminergic neurons, we first inspected the expression levels of the four essential components for dopamine transmission: *Th, Ddc, Slc6a3(Dat)* and *Slc18a2(Vmat)*. As seen in Figure 4-4C, all four genes are highly expressed across the cells marked by the red-to-skyblue bar, suggesting that these cells are dopaminergic neurons and have strong dopaminergic activity. On the other hand, the cells marked by the grey-scale horizontal bar exhibit systematically reduced expression of these four genes. This expression pattern corresponds nearly perfectly to the clustering dendrogram if we split the dendrogram according to the first major branching point. The clustering dendrogram and the expression pattern of these four genes suggest that the first split of the clustering dendrogram precisely separates dopaminergic neurons from other GFP+ cells. I thus proposed that the 110 cells marked by the red-to-skyblue bar horizontal bar are putative dopaminergic neurons, and that the 62 cells marked by the grey-scale horizontal bar are other GFP+ cells.

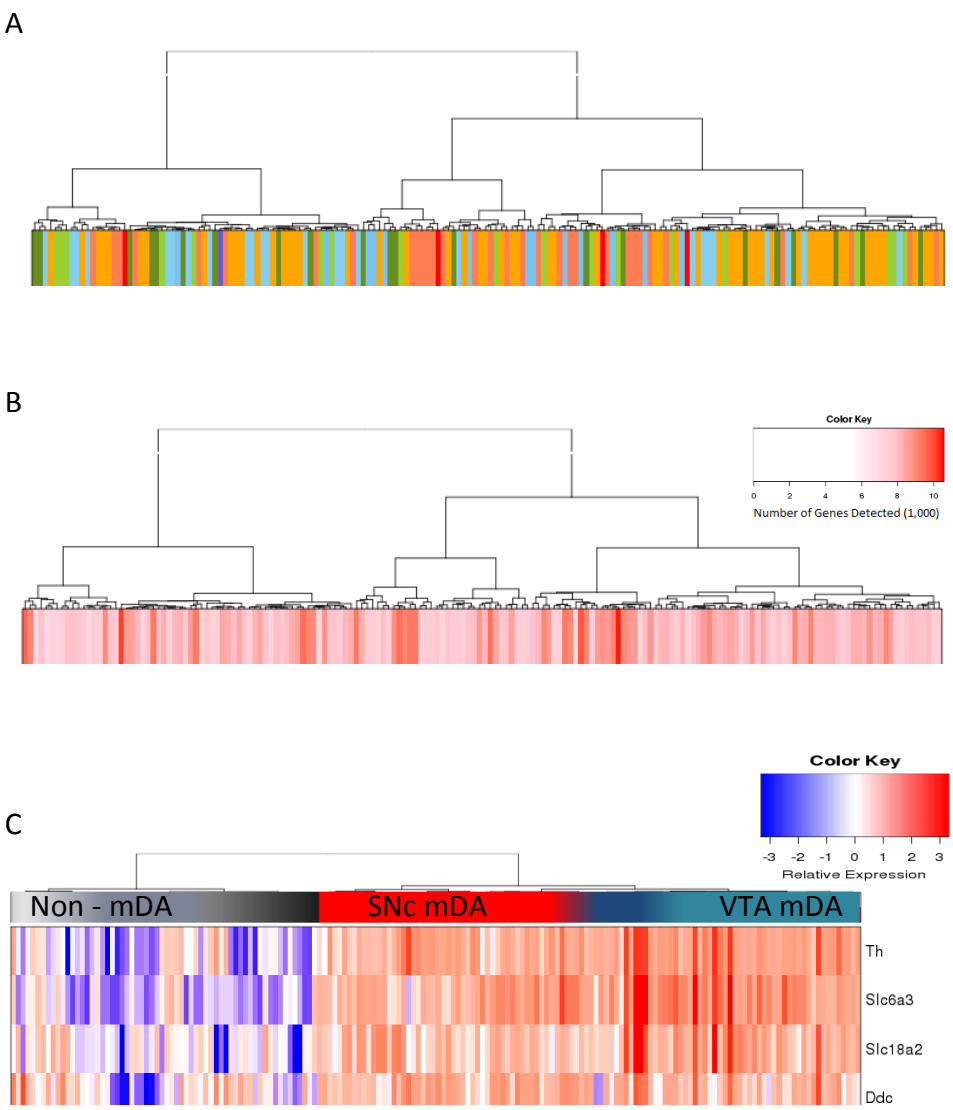


Figure 4-4: Clustering of all cells sequenced in the experiment that passed quality check, together with batch, number of genes detected and dopaminergic marker genes. Panel A: Dendrogram determined using method described in [190]. The colour bars beneath the dendrogram represent the experimental batch. Panel B: The dendrogram is the same as Panel A. The colour of the bars denote the number of genes detected in each cell. Panel C: Heatmap showing the relative expression levels of 4 genes, *Th, Slc6a3, Slc18a2* and *Ddc*, that are known to be essential for the dopamine synthesis, re-uptake and secretion as discussed in [192]. Blue indicates relatively low expression, red represents relatively high expression as indicated by the colour scale key, which is used for all heatmaps throughout this thesis. The clustering dendrogram in all three panels are the same, which was determined using the expression of all expressed genes according to Materials and Methods D5, using the 172 cells presented in the figure.

By examining these four well-known dopaminergic neuron markers that were subjectively selected, it is highly suggestive that the cells marked by the horizontal bar from red to sky-blue are dopaminergic neurons. However, this subjective approach is potentially biased. I thus sought to take an unbiased approach to identify all genes that are characteristic for this group of putative dopaminergic neurons. To achieve this, I performed an *a priori* differential gene expression analysis (Materials and Methods D6) by contrasting the group of dopaminergic neurons with the group of other GFP+ cells, and obtained 61 genes that are more abundantly expressed among the putative dopaminergic neurons (shown in Figure 4-5). The five most differentially expressed genes are *Slc6a3, Sox6, Drd2, Gm27031* and *Th. Slc6a3* and *Th* are well-known markers of midbrain dopaminergic neurons, and *Sox6* marks dopaminergic neurons in the SNc region of the midbrain. *Drd2* encodes the dopaminergic receptor which plays an important auto-regulatory role in dopaminergic neurons [193]. *Gm27031* gene body coordinates overlaps with the gene body coordinates of the eighth most differentially expressed gene *Sv2c*; and that locus was re-assigned as the 3'UTR of *Sv2c* in the most recent Ensembl gene annotation set. Looking beyond the top 5 genes, *Slc18a2* (ranked 40 in this list) is another essential marker of dopaminergic neurons. *Ddc* is also among the *a priori* list of genes discovered using this approach when less stringent thresholds were applied (data not shown). Furthermore, these *Th+* cells more abundantly express *Oprk1* and *Tacr3*, which were regarded as markers of mature dopaminergic neurons [194], and were previously associated with dopaminergic neuronal functions [194, 195]. In summary, this *a priori* identified list of genes strongly supports the hypothesis

that these cells are indeed mature midbrain dopaminergic neurons. We next focused on this group of mature midbrain dopaminergic neurons to identify further classes of dopaminergic neuron. A detailed investigation of this group of 110 mature midbrain dopaminergic neurons will be presented in the next section.

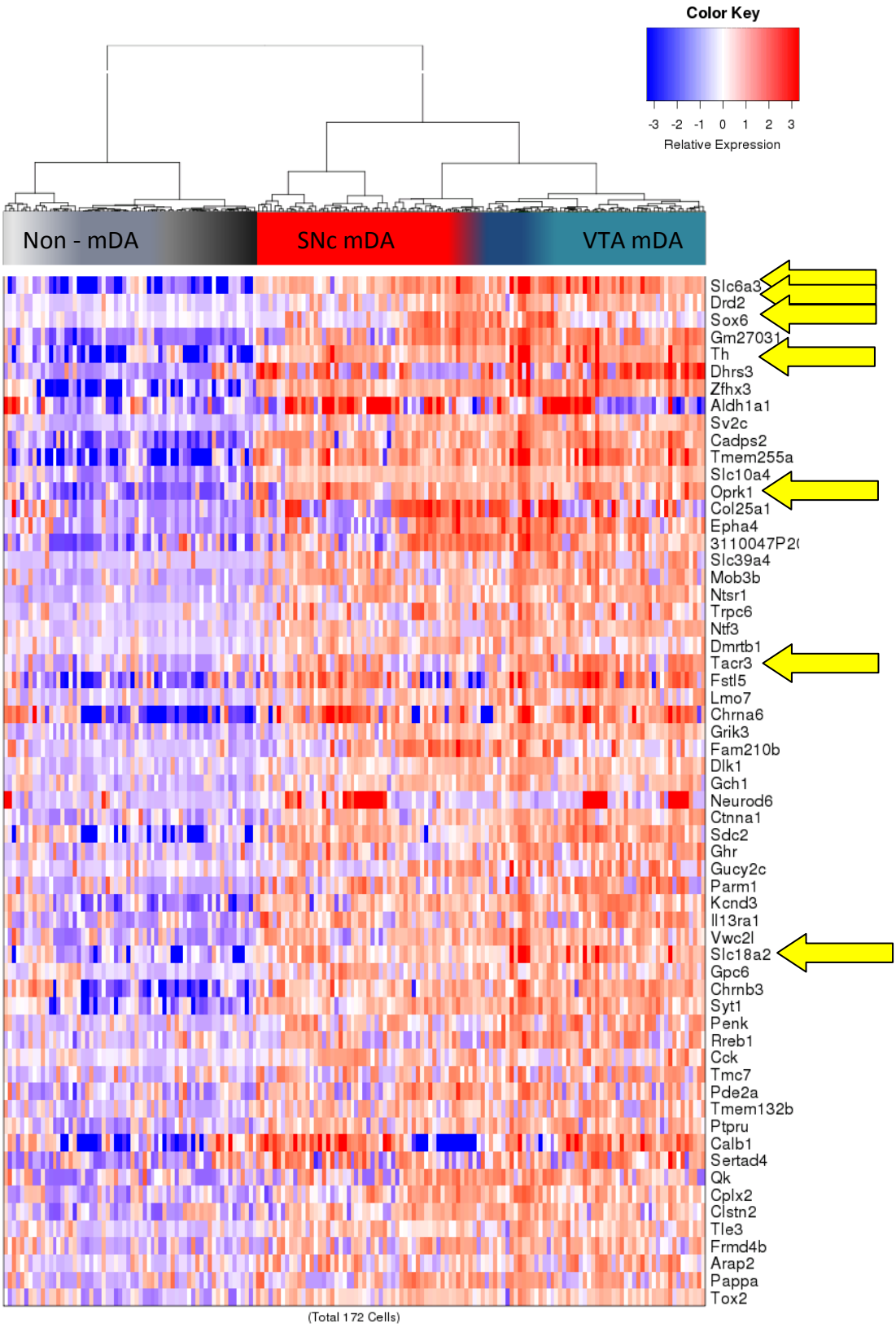


Figure 4-5: Heatmap showing genes that are more abundantly expressed in the group of cell that are marked by the red-to-skyblue bar. Differential gene expression is performed according to Materials and Methods D6, FDR < 20%, FPKM in the expressed group > 50, fold change > 4. The genes are ordered according to the fold change between the two groups. The clustering

dendrogram on top of the heatmap was determined using the expression of all expressed genes according to Materials and Methods D5, using the 172 cells presented in the heatmap.

Thus far, 110 dopaminergic neurons have been classified, yet I also sought to understand the remaining 62 GFP+ cells. Other than knowing that these cells are probably not dopaminergic neurons, I did not have any expectation or knowledge regarding the classification of these cells. As a result, I next chose to examine the expression patterns of marker genes of all major cell types in the brain. According to [196], there are 10 major cell types in the brain: oligodendrocyte, microglia, astrocyte, cholinergic neuron, GABAergic neuron, serotonergic neuron, dopaminergic neuron, glutamatergic neuron, radial glia, and neuroepithelia. In addition, I included known marker genes of mature neurons and immature neurons. I plotted the expression of these cell type specific markers [196] shown in Figure 4-6. It can be seen that dopaminergic marker genes are highly expressed among the 110 mature dopaminergic neurons. On the other hand, among the 62 other GFP+ cells, some express *Neurod1*, a marker for immature neurons; some express *Gad1, Gad2* and *Slc6a1*, indicating that some of these cells are capable of transmitting GABA. If we examine the expression of *Gad2* and *Slc6a1* in further details, we can see that there is further heterogeneity among the group of 62 cells. For example, the cells that express *Gad2* do not express *Gad1* and *Slc6a1*, and vice versa. The heterogeneity among this group of cells will be discussed in a later section with the title "among the group of 62 other GFP+ cells, two different classes of neurons were identifiable".

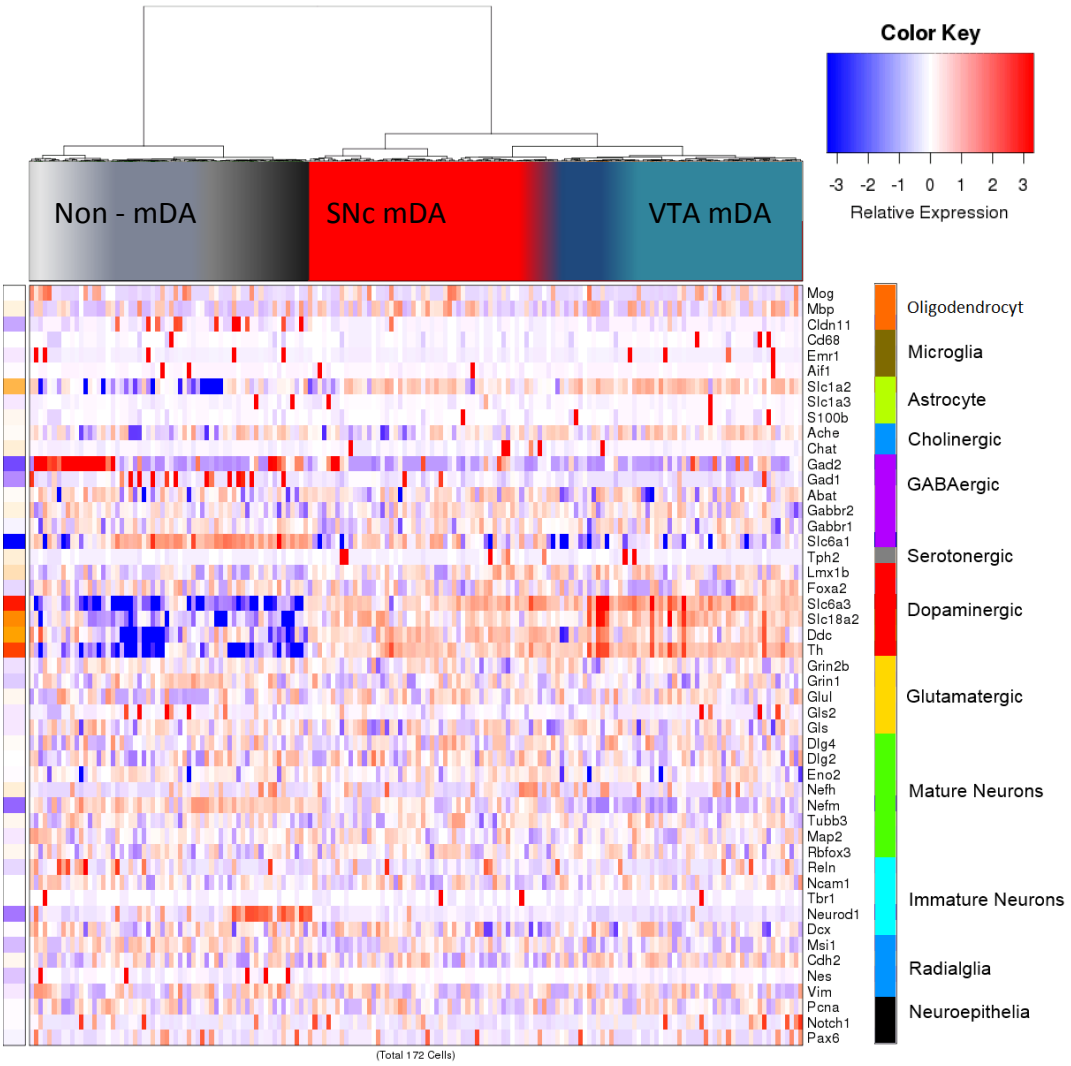


Figure 4-6: Heatmap showing the expression of known marker genes of 10 major cell types in the brain. Furthermore, I added the known marker genes of mature neurons and immature neurons. The clustering dendrogram on the top of the heatmap was determined using the expression of all expressed genes according to Materials and Methods D5 using the 62 cells presented in the heatmap.

The above approach using known makers requires a list of pre-selected genes based on prior knowledge. However there are disadvantages to this approach. For example the supplied list of genes is likely not to be sufficiently comprehensive to mark all cell types in the midbrain. In addition, the list of

marker genes that were identified in other organisms or at other developmental time points might not necessarily be applicable to the mouse midbrain at postnatal Day 4. I reasoned that this group of other GFP+ cells may originate from one of the three sources below. (i) Due to the imperfections of FACS sorting for example, when a doublet of GFP+ and GFP- cell pass though the FACS at the same time, although some cells do not express the GFP protein, these GFP protein negative cells could have been erroneously included. (ii) Although *Pitx3* is a good marker for midbrain dopaminergic neurons, some non-dopaminergic neurons in the mouse midbrain may also activate *Pitx3* promoter. As a result, these cells are positive for GFP and are included. (iii) Although it is known that all dopaminergic neurons are post-mitotic at postnatal Day 4, it is not known whether all of them are equally mature. It is plausible that some dopaminergic neurons are not sufficiently mature, and thus may lack the expression pattern of a typical mature dopaminergic neuron. In theory, the abundance of GFP protein in these cells could be further examined by estimating the abundance of *GFP* mRNA transcript in these cells. However, due to the complex workflow of mapping the entire dataset again onto the reference genome, this additional analysis is not performed. If later it becomes important to evaluate whether some cells fail to express the *GFP* transcript, a retrospective re-ananlysis can be performed to estimate the abundance of *GFP* mRNA transcripts in each cell.

Because of the potentially highly heterogeneous sources for this group of cells as explained above, we did not expect these cells to originate from a single cell type. As a result, I reasoned that it would be more informative to also take

the *a priori* differentiation expression analyses (Materials and Methods D6) approach to identify genes that are more abundantly expressed among these cells. Among 84 genes that are more abundantly expressed in these cells (Figure 4-7), three stand out: *Ebf2, Ebf3* and *Slc17a6* (indicated by yellow arrows). It has been suggested that proteins of the EBF family are markers for immature neurons [197]. In particular, *Ebf2* was shown to be required for development of dopamine neurons in the midbrain [198]. Furthermore, although *Slc17a6* was traditionally considered as a marker of Glutamatergic/GABAergic neurons, it was recently hypothesized to be indicative of those immature neurons that have yet to decide their final repertoire of neurotransmitters [199]. *Lmx1a*, a marker of maturation and differentiation of dopaminergic neurons [172, 194], is also among the list of genes more abundant among this group of cells when less stringent thresholds was applied (data not shown), and our collaborators have previously shown that *Lmx1a* transcripts decreases with age in mice [200]. Although there might be substantial heterogeneity among this group of GFP+ cells, based on the evidence discussed above, I conclude that at least some cells among this group of GFP+ cells are immature neurons.

Importantly, I believe that there can be more heterogeneity within them, because for example, 2 of the 84 genes in Figure 4-7, *Gad2* and *Slc6a1* (indicated by green arrows) are classical markers of GABAergic neurons. The expression of these two genes suggests that some other cells among this group of GFP+ cells are capable of synthesizing and transmitting GABA. However, because these *Gad2*+&*Slc6a1*+ cells and *Gad1*+&*Slc6a1*+ express a non-negligible amount of *Th,*

Ddc, Slc6a3 and *Slc18a2* (Figure 4-6), it is not clear whether they only transmit GABA, or alternatively, co-transmit GABA together with dopamine.

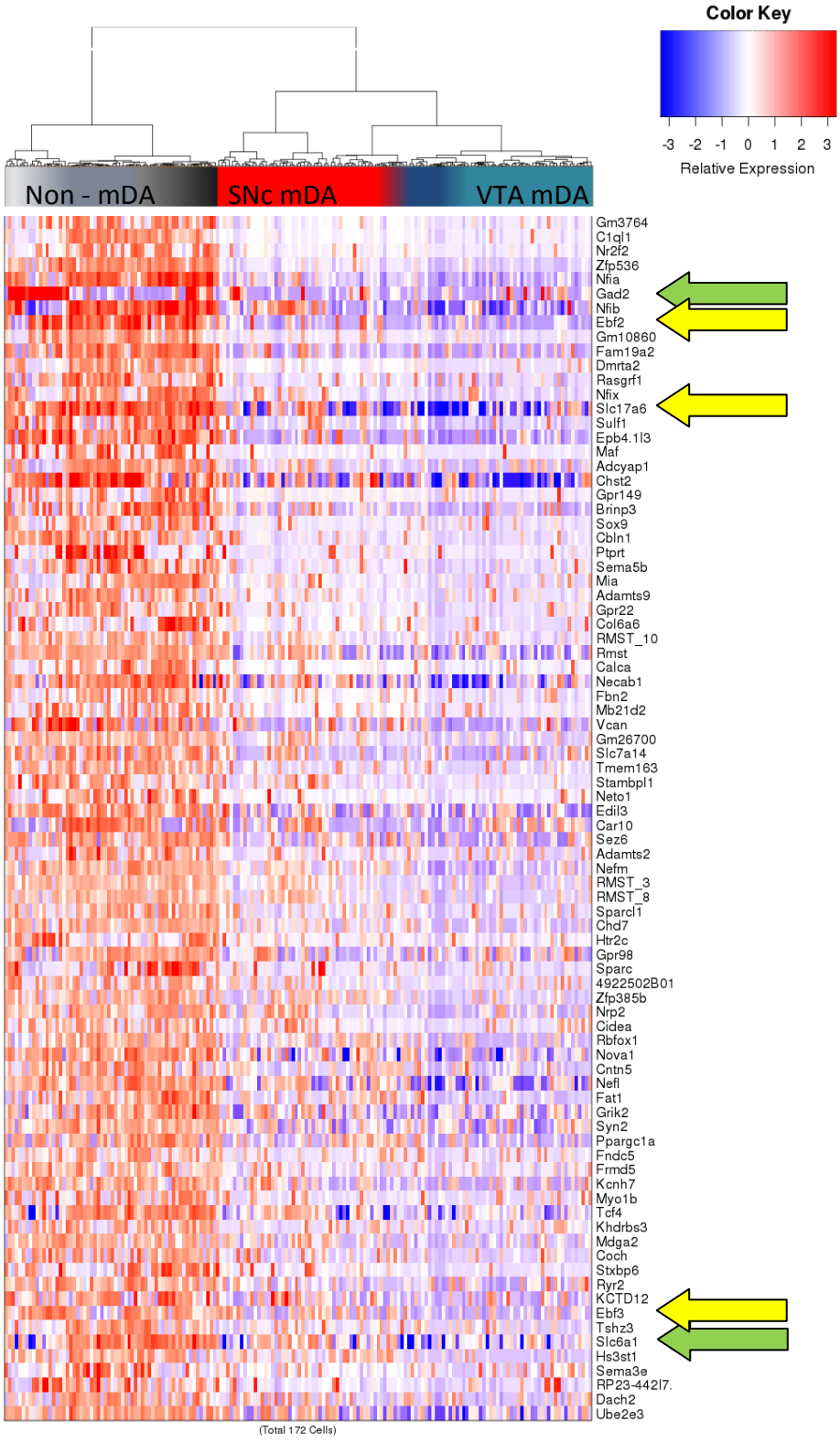


Figure 4-7: Heatmap showing genes that are more abundantly expressed in the cells marked by the grey-scale colored bar. Differential gene expression analysis is performed according to Materials and Methods D6, FDR < 20%, FPKM in the expressed group > 50, fold change > 4. The yellow arrows indicate immature neuron marker genes. The green arrows indicate known GABAergic neuron markers. The clustering dendrogram on the top of the heatmap was determined using the expression of all expressed genes according to Materials and Methods D5 using the 62 cells presented in the heatmap.

Analysis of the group of 110 mature midbrain dopaminergic neurons recapitulated known classes: VTA and SNc dopaminergic neurons

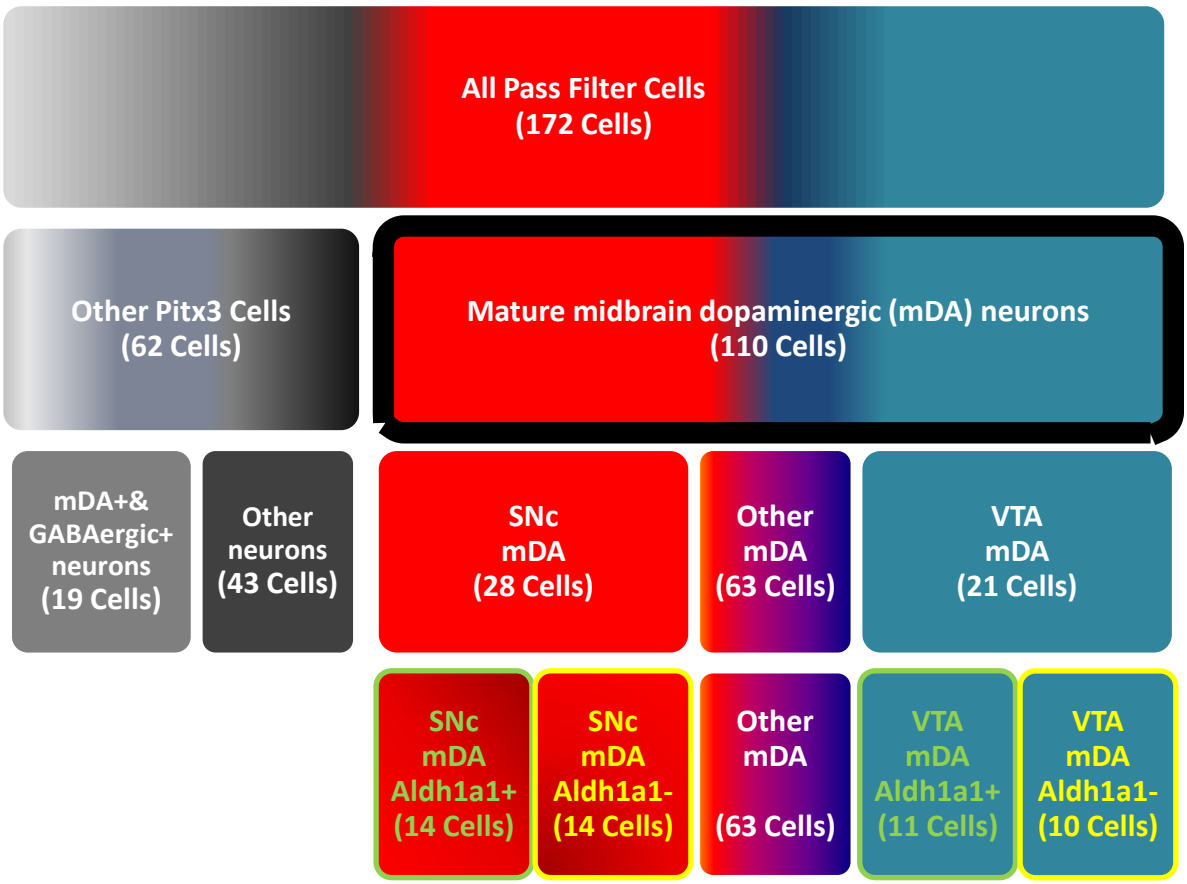


Figure 4-8: Illustration showing the global hierarchical clustering structure and assigned cluster names for all subpopulations that were identified and validated in this project. The 110 cells highlighted by the dashed box are considered to be mature dopaminergic neurons which will be analyzed in details in the sections below.

In the analyses presented above, I considered all 172 pass filter cells (Materials and Methods D2). I predicted that these contained two major groups: the 110 mature midbrain dopaminergic neurons and the other 62 GFP+ cells. In this section, I focus on the 110 mature midbrain dopaminergic neurons. I seek to identify distinct classes of midbrain dopaminergic neurons among them. To achieve this, I performed *a priori* clustering (Materials and Methods D5), and obtained the clustering dendrogram shown in Figure 4-9. Again, I verified that the clustering is not driven by two potential confounders: experimental batch and number of genes detected in the cell, which is achieved by inspecting Figure 4-9A,B. The number of genes detected by the two major clusters were not significantly different when examined using Student's t test ($p > 0.1$).

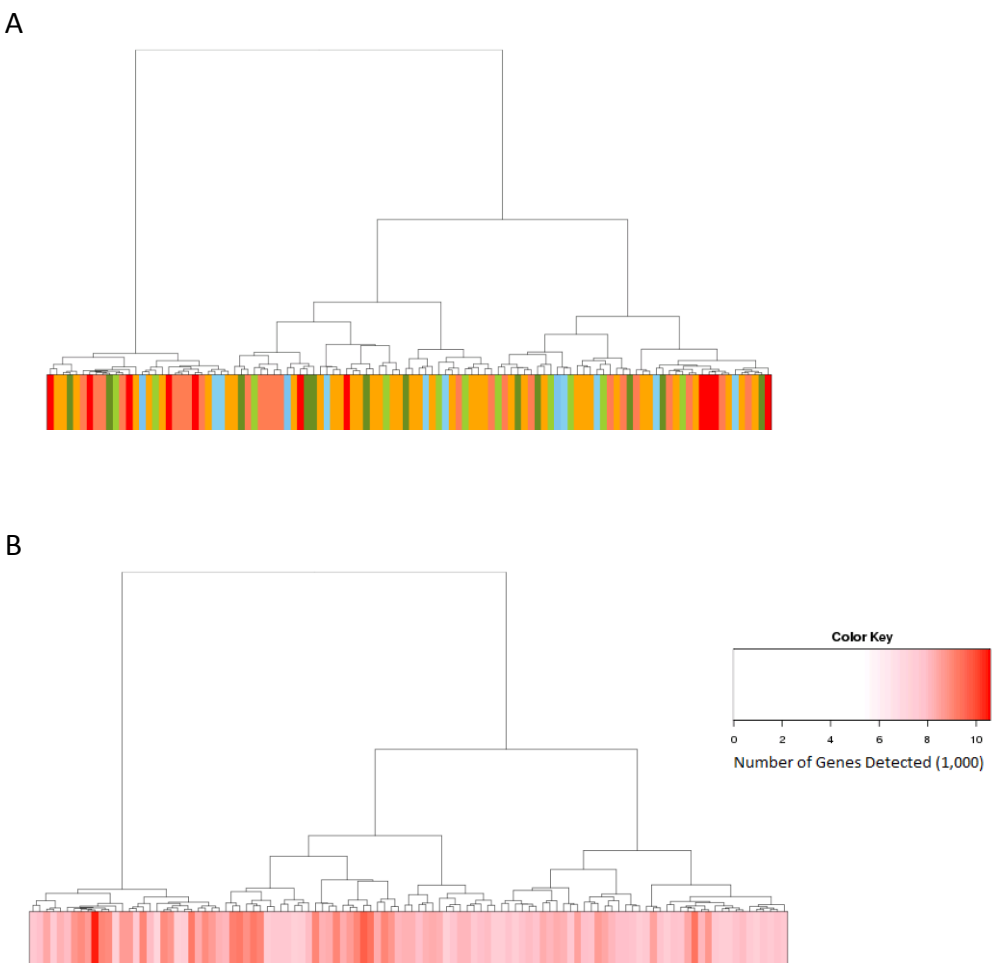


Figure 4-9: Clustering of all mature midbrain dopaminergic neurons sequenced in the experiment, showing the batch and number of genes detected. Panel A: Dendrogram determined using the method described in [190]. The colour bars with 9 different colours beneath the dendrogram represent the 9 different experimental batch. Panel B: The dendrogram is the same as Panel A. The colour of the bars denotes the number of genes detected in each cell. The clustering dendrogram in both panels are the same, which was determined using the expression of all expressed genes according to Materials and Methods D5, using the 110 cells presented in the figure.

Subsequently, I sought to recapitulate several classes of midbrain dopaminergic neurons known to be present in the adult mouse midbrain. It was known that dopaminergic neurons in the VTA or the SNc regions express several different marker genes [3, 176]. These known classes of dopaminergic neurons

were previously determined based on their anatomical locations in the midbrain. On the other hand, my clustering results presented in Figure 4-9 were based on transcriptomic similarity among single cells in an entirely data-driven approach and were not based on any anatomical information. I hypothesized that if the VTA and SNc regions contain different classes of dopaminergic neurons that are discernible by their transcriptomes, then the genes previously discovered by contrasting VTA and SNc should be heterogeneously expressed among the different classes of midbrain dopaminergic neurons I discovered using my entirely data-driven approach.

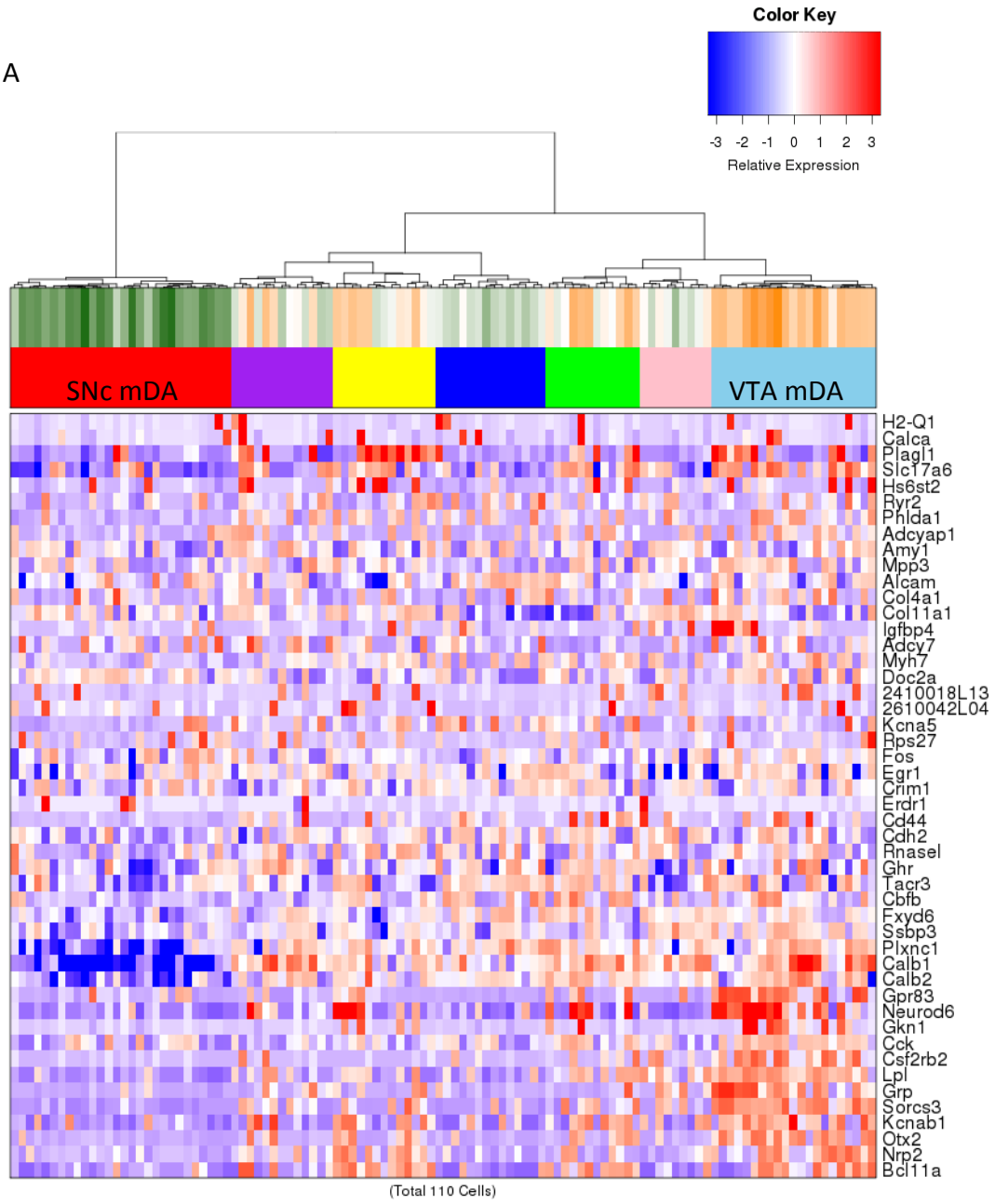
I first sought to test the above hypothesis for VTA dopaminergic neurons. I foresee that the previously identified VTA markers based on anatomy could potentially exhibit a heterogeneous expression pattern that varies systematically for cells within some sub-branches of my clustering dendrogram. This was achieved by plotting the expression of 52 VTA dopaminergic neuron marker genes previously reported from adult mice [176] (the exact age was not documented in the original manuscript), and organising the cells according to my previously determined clustering dendrogram. (Figure 4-10 A) The seven coloured bars beneath the dendrogram indicate putative classes of dopaminergic neurons indicated by the clustering dendrogram. The coloured bars that contain various shades of orange and green indicate the systematic variation summarising the expression of all VTA specific genes based on the method described in [190]. In this case, a deep orange shade indicates that the corresponding cell is expressing several VTA specific genes highly. Due to a limit

of space, if more than 50 genes were supplied, only the top 50 genes that contributed towards a systematic variation among all cells are displayed [190]. It is clear that the orange colour shades are deepest among the cells indicated by the sky-blue horizontal bar, and some cells indicated by the yellow and green bar.

It is important to note that these marker genes were discovered in adult mice [176]. Although it is preferable to determine the cell-type specific genes using adult mice because the cell type identities are more distinguishable, it is also valuable to consider earlier time points considering that our data was generated at postnatal Day 4. As a result, I also repeated this analyses using a second list of VTA specific genes that were determined at a closer developmental time point, namely postnatal Day 1 [201] (Figure 4-10 B). This second list of genes were obtained by re-analyzing data from supplementary table S1 in [201] to identify all genes that are upregulated in non-mutant VTA in comparison to non-mutant SNc. Again, I observed the same expression pattern indicating that the cells that are marked by the sky-blue bar are expressing VTA specific markers more highly than other cells. Note that these previous studies [176, 201] were carried out using mice at developmental time points different than that for our single cell RNA-Seq dataset. Because postnatal Day 4 lies in between postnatal Day 1 and adult, it is reasonably safe to assume that major cell migrations do not cause a spatial organisation to be present at postnatal Day 1 and adult, but absent at postnatal Day 4. Thus it is reasonably safe to interpolate postnatal Day 4, which is between postnatal Day 1 and adult. Based on the results presented above by connecting my results to two different previous studies, I conclude that

the cells indicated by the sky-blue bar are most likely to be dopaminergic neurons located within the VTA region of the midbrain.

A



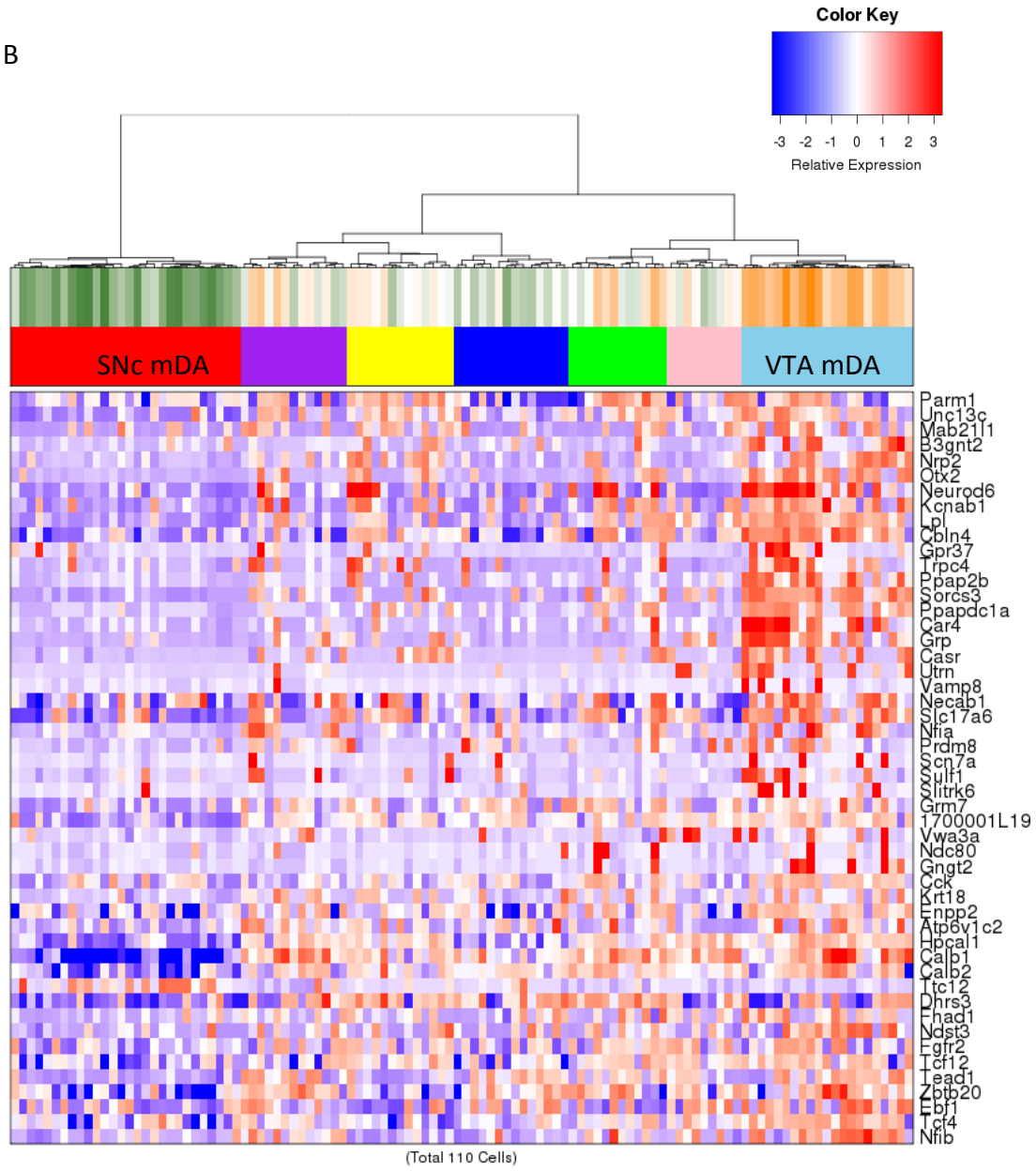


Figure 4-10: Heatmap showing the expression of VTA specific genes. Panel A, the VTA specific genes obtained from a performed in adult mice [176]. Panel B. similar to Panel A, except that the genes were identified in a different study and at a different developmental time point: postnatal Day 1 [201]. Using the method in [190], the coloured bars on top showing various shades of orange and green summarize the systematic variation of all VTA specific genes. In this case, a strong orange shade indicates that the corresponding cell is expressing most of the VTA specific genes highly. In both Panel A and B, if more than 50 genes were supplied, only the top 50 genes that contributed towards a systematic variation assessed based on [190] were plotted. The clustering dendrograms on the top of the heatmaps was determined using the expression of all expressed genes according to Materials and Methods D5 using the 100 cells presented in the heatmaps.

After successfully demonstrating that the aforementioned hypothesis for the VTA dopaminergic neurons is supported by our data, I further sought to examine whether the same would be true for the SNc dopaminergic neurons. This was achieved using the same approach, by plotting the expression of SNc dopaminergic neuron marker genes reported in adult mice [176] (Figure 4-11A). I used as input the expression of all genes upregulated in SNc relative to VTA in adult mice [176]. It is clear that the SNc dopaminergic neuron marker genes are expressed most strongly among the cells in the cluster marked by the red horizontal bar, and weakly for some cells marked by purple and blue horizontal bars. Again, I repeated the same analyses using a second list of VTA specific genes that were determined at postnatal Day 1 [201] (Figure 4-11B). This second list of genes were obtained by re-analyzing data from supplementary table S1 in [201] to identify all genes that are upregulated in non-mutant SNc in comparison to non-mutant VTA. Again, I observed the same expression pattern indicating that the cells that are marked by the red bar are expressing SNc specific markers more highly than other cells. Based on the results presented above by connecting my results to two different previous studies, and again assuming cell migration cannot be a confounding factor, I conclude that the cells indicated by the red bar are most likely to be dopaminergic neurons located within the SNc region of the midbrain.

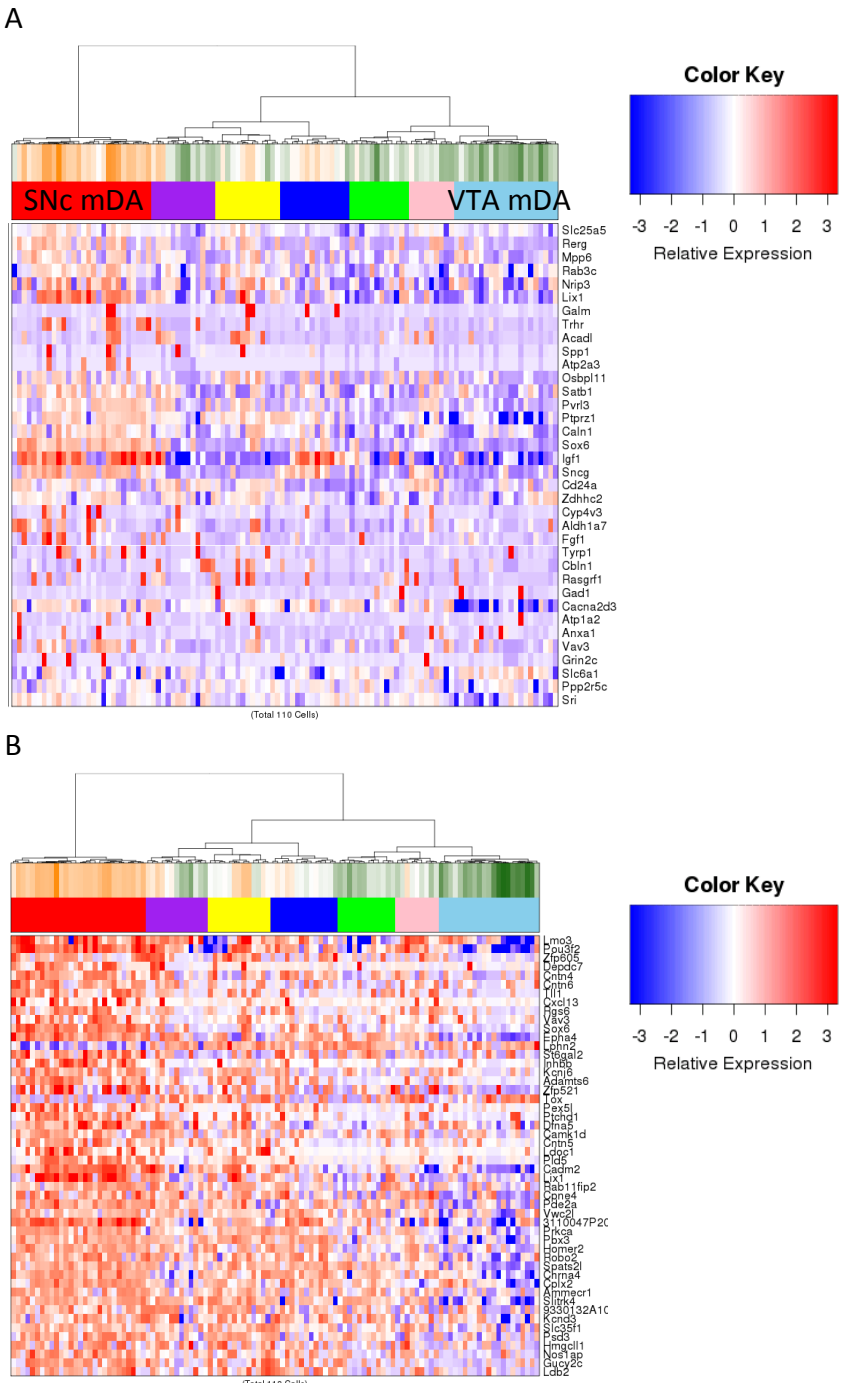


Figure 4-11: Heatmap showing the expression of SNc specific genes. Panel A. The SNc specific genes obtained from a study performed in adult mice [176]. Panel B is similar to Panel A, except that the genes were identified in a different study and at a different developmental time point: postnatal Day 1 [201]. Using the method in [190], the coloured bars on top showing various shades of orange and green summarize the systematic variation of all SNc specific genes. In this case, a strong orange shade indicates that the corresponding cell is expressing most of the SNc specific genes highly. In both Panel A and B, if more than 50

genes were supplied, only the top 50 genes that contributed towards a systematic variation assessed based on [190] were plotted. The clustering dendrograms on the top of the heatmaps was determined using the expression of all expressed genes according to Materials and Methods D5 using the 110 cells presented in the heatmaps.

It is important to highlight that when we compared our results with the two previously published results [176, 201], the developmental time point of the experiments of their samples were different than our samples. This difference may jeopardize the concordance between our results and the published results. Despite this jeopardy, the successful matching of a subset of our results, namely pairing the red and sky-blue clusters to previous results, has provided substantial confidence that our single cell RNA-Seq datasets are of high quality and is capturing true cell identities that are biologically relevant. Furthermore, the ability to use only the transcriptome and hierarchical clustering to recapitulate known classes of dopaminergic neurons has validated my entirely data-driven approach to discover further novel cell types.

Throughout Figure 4-10A, 4-10B, 4-11A, 4-11B, the cells indicated by purple, yellow, blue, green and pink horizontal bars are not strongly showing the expression pattern of any of the above two anatomically defined cell types based on gene lists defined from adult mice and postnatal Day 1 mice. This is not surprising given that the two previously known classes of midbrain dopaminergic neuron markers were obtained entirely based on approximate anatomical locations, a method that is potentially inferior in its ability to distinguish cell types that are interspersed. We consider these cells as other midbrain dopaminergic neurons that were not distinguishable based on previous studies.

Differential expression analysis identified novel markers for SNc dopaminergic neurons at postnatal Day 4

Knowing that the cells marked by the red bar are SNc dopaminergic neurons I next sought to identify novel markers for this group of cells at postnatal Day 4. To achieve this, I contrasted the two groups of cells: (i) the 28 SNc dopaminergic neurons marked by the red bar, and (ii) the remaining 82 dopaminergic neurons (not marked by the red bar). Differential expression analyses (Materials and Methods D6) identified the genes that are specifically expressed in SNc dopaminergic neurons and the genes that are depleted in SNc dopaminergic neurons (Figure 4-12).

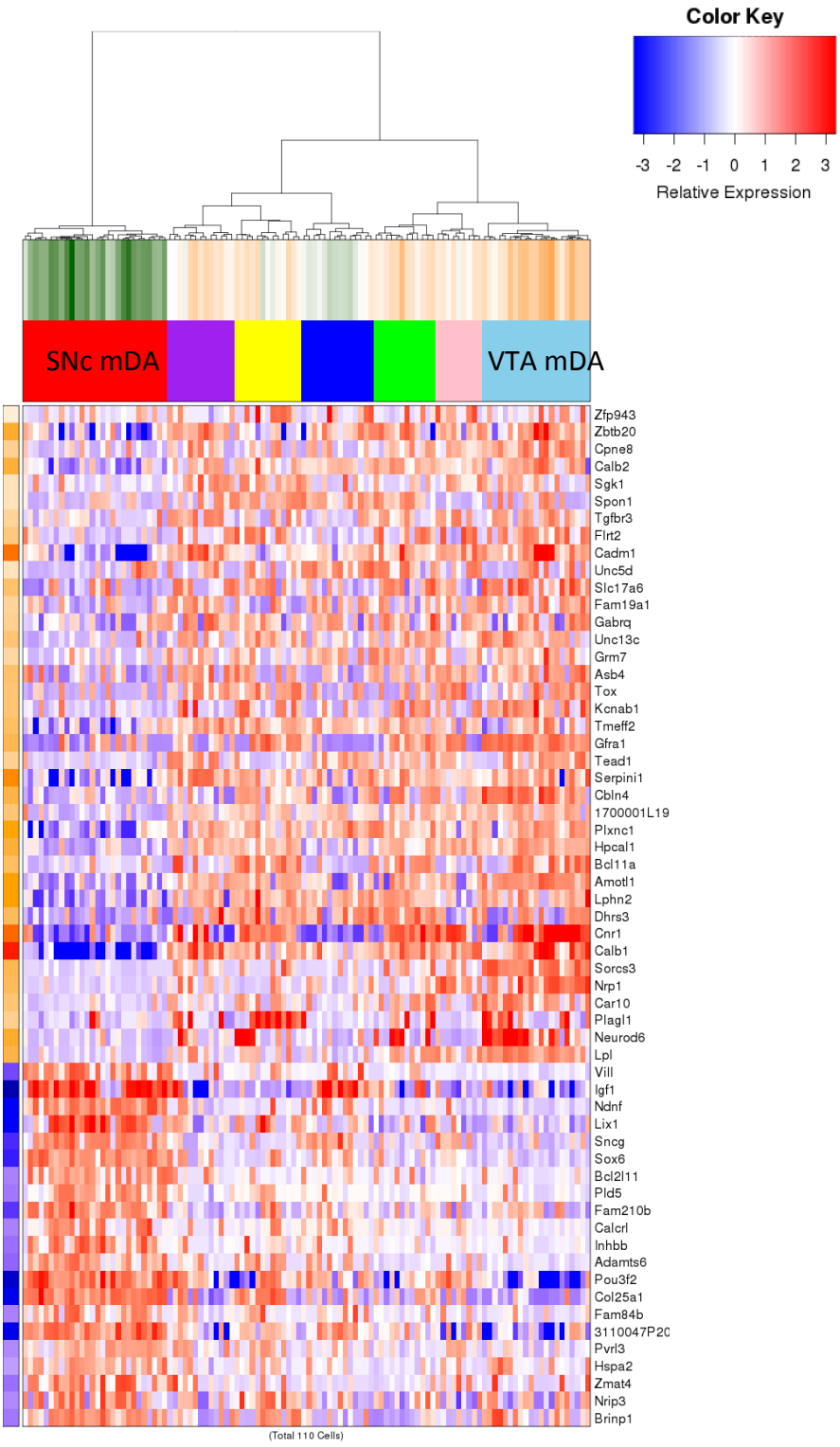


Figure 4-12: Heatmap showing the expression of genes identified by differential expression analysis. The upper half of the heatmap where the rows are indicated by the orange side bar represent the genes that are depleted among SNc dopaminergic neurons. The lower half of the heatmap where the rows are indicated by the blue side bar represent the genes that are enriched among the SNc dopaminergic neurons.

Differential expression analysis identified novel markers for VTA dopaminergic neurons at postnatal Day 4

I next sought to identify novel markers that distinguish the 21 VTA dopaminergic neurons (cells marked by sky-blue) using the same method. Figure 4-13 top (bottom) section illustrates genes whose expression is lower (higher) in the VTA DA neurons.

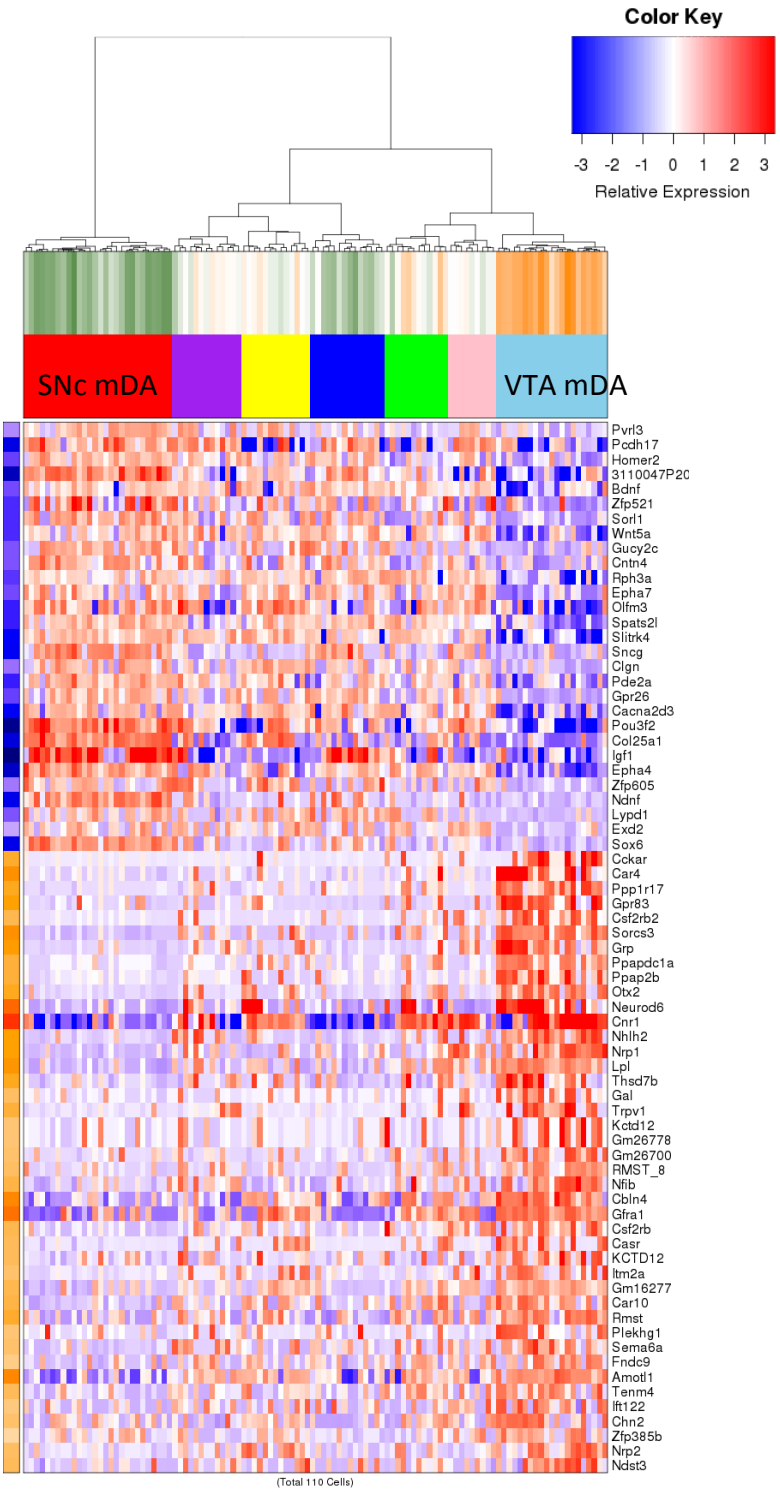


Figure 4-13: Heatmap showing the expression of genes identified by differential expression analysis. The upper half of the heatmap where the rows are indicated by the orange side bar represent genes that are depleted among the VTA dopaminergic neurons. The lower half of the heatmap where the rows are indicated by the blue side bar represent genes that are enriched among the VTA dopaminergic neurons.

Differential expression analysis on the group of 110 dopaminergic neurons identified novel classes of midbrain dopaminergic neurons

In the previous analyses, I demonstrated that the marker genes can be identified for the two known classes of midbrain dopaminergic neurons: VTA and SNc. Yet recall that the goal of the project was to identify novel classes of midbrain dopaminergic neurons. Thus for each of the remaining five clusters (out of a total of seven clusters), I also performed the same differential expression analyses to identify genes that could potentially distinguish them from other midbrain dopaminergic neurons. The results for the yellow group is shown in Figure 4-14. The top two distinguishing marker gene for this group is *Smoc2* and *Plagl1* indicated by yellow arrows. The function of these two genes in neurons has not been elucidated.

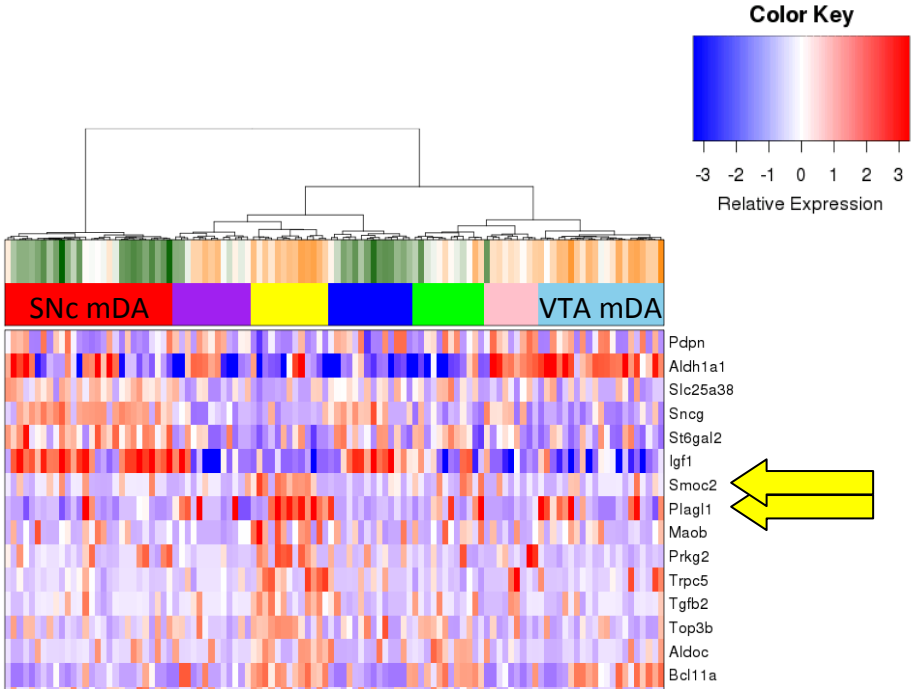


Figure 4-14: Heatmap showing the expression of genes identified by differential expression analysis. The first 6 genes are those that are depleted among the yellow group of dopaminergic neurons. The remaining genes on the heatmap represent genes with expression enriched among the yellow group of dopaminergic neurons. The yellow arrows indicate the top two genes with the highest fold change.

The results for the pink group are shown in Figure 4-15. The best distinguishing marker genes for this group are *Wdr52* and *Zhx2* as indicated by the yellow arrows.

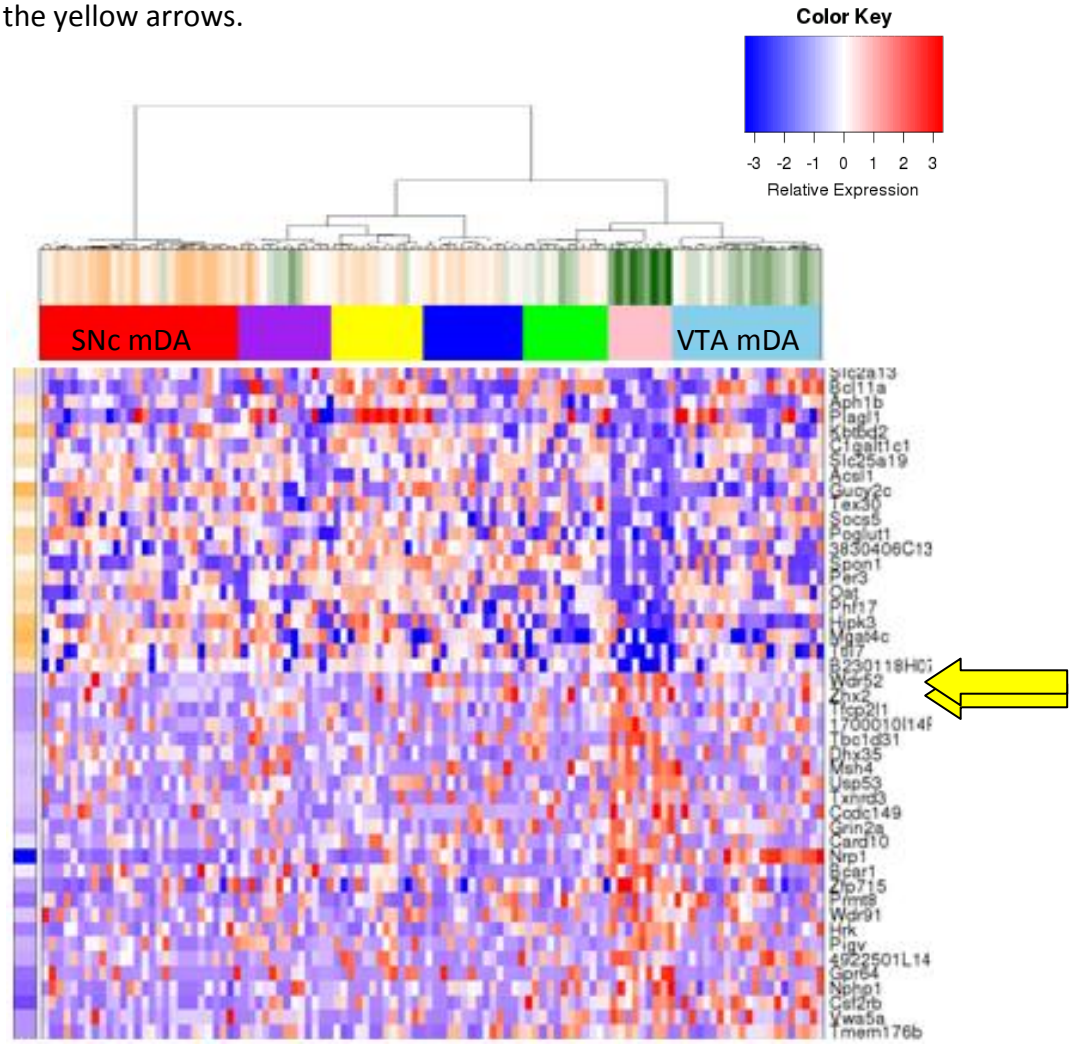


Figure 4-15: Heatmap showing the expression of genes identified by differential expression analysis. The upper half of the heatmap where the rows are indicated by the orange side bar represent genes that are depleted among the pink group of dopaminergic neurons. The lower half of the heatmap where the rows are indicated by the blue side bar on the left represent genes that are enriched among the pink group of dopaminergic neurons. The yellow arrows indicate the top two genes with the highest fold change.

The results for the green group are shown in Figure 4-16. Due to the small number of differentially expressed genes that are confidently identifiable from this group, it is not clear whether this is a genuine cluster of midbrain dopaminergic neurons or is a spurious cluster. A greater number of cells of this cluster undergoing single cell RNA-Seq may be required to confidently determine the marker genes for this cluster of cells.

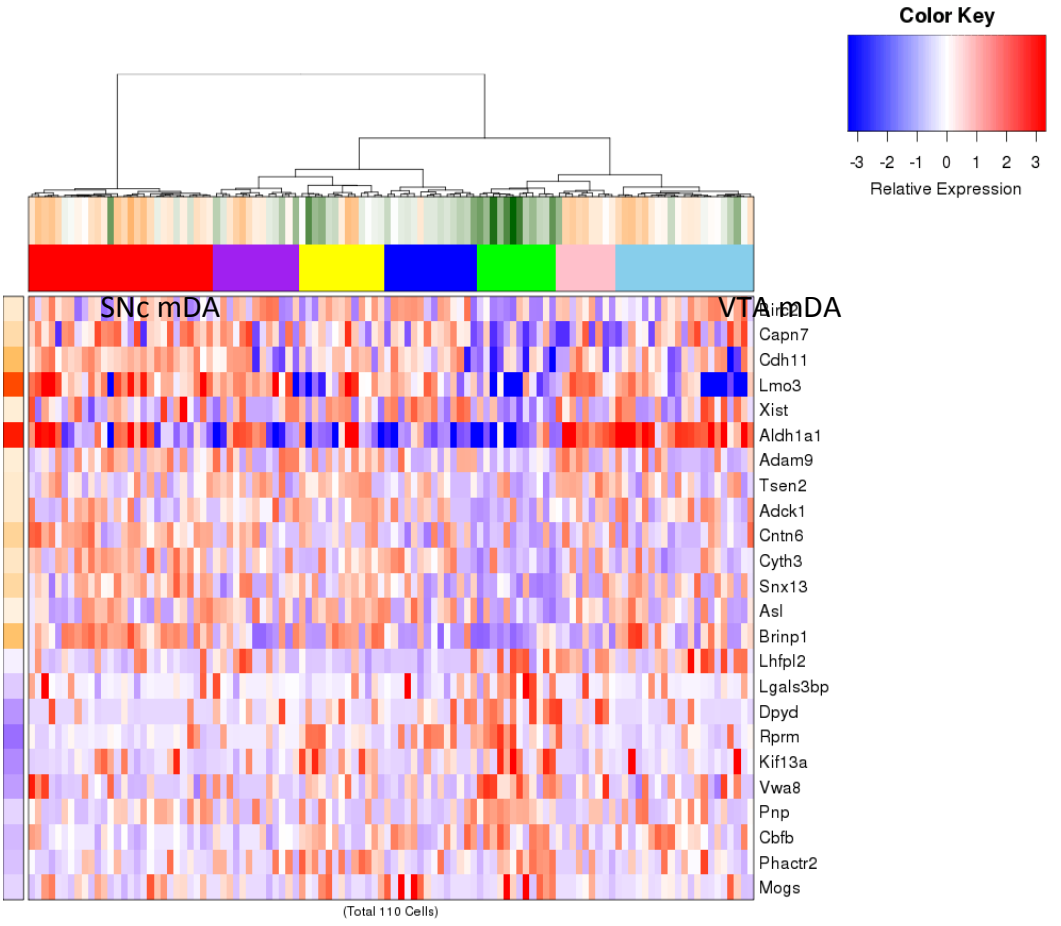


Figure 4-16: Heatmap showing the expression of genes identified by differential expression analysis. The upper half of the heatmap where the rows are indicated by the orange side bar represent genes that are depleted among the green group of dopaminergic neurons. The lower half of the heatmap where the rows are indicated by the blue side bar represent genes that are enriched among the green group of dopaminergic neurons.

The results for the blue group are shown in Figure 4-17. Due to the small number of differentially expressed genes that are confidently identifiable from this group, it is not clear whether this is a genuine cluster of midbrain dopaminergic neurons or is a spurious cluster. A greater number of cells of this cluster undergoing single cell RNA-Seq may be required to confidently determine the marker genes for this cluster of cells.

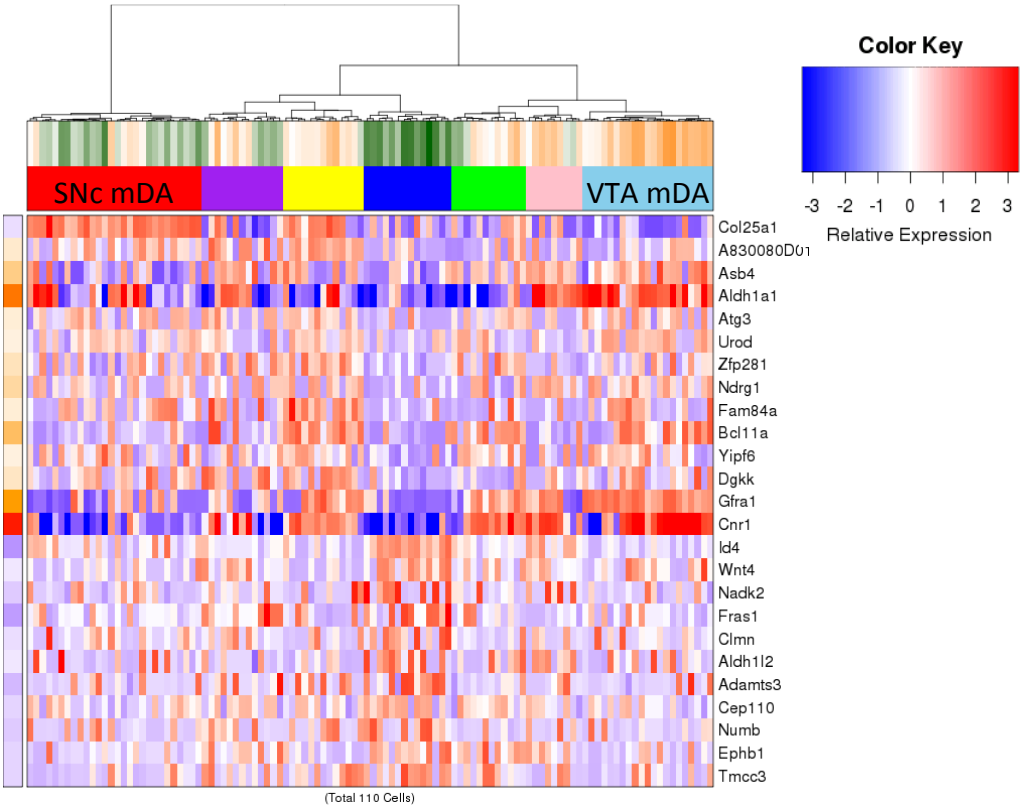


Figure 4-17: Heatmap showing the expression of genes identified by differential expression analysis. The upper half of the heatmap where the rows are indicated by the orange side bar represent genes that are depleted among the blue group of dopaminergic neurons. The lower half of the heatmap where the rows are indicated by the blue side bar represent genes that are enriched among the blue group of dopaminergic neurons.

The results for the purple group are shown in Figure 4-18. Due to the small number of differentially expressed genes that are confidently identifiable from this group, it is not clear whether this is a genuine cluster of midbrain dopaminergic neurons or is a spurious cluster. A greater number of cells of this cluster undergoing single cell RNA-Seq may be required to confidently determine the marker genes for this cluster of cells.

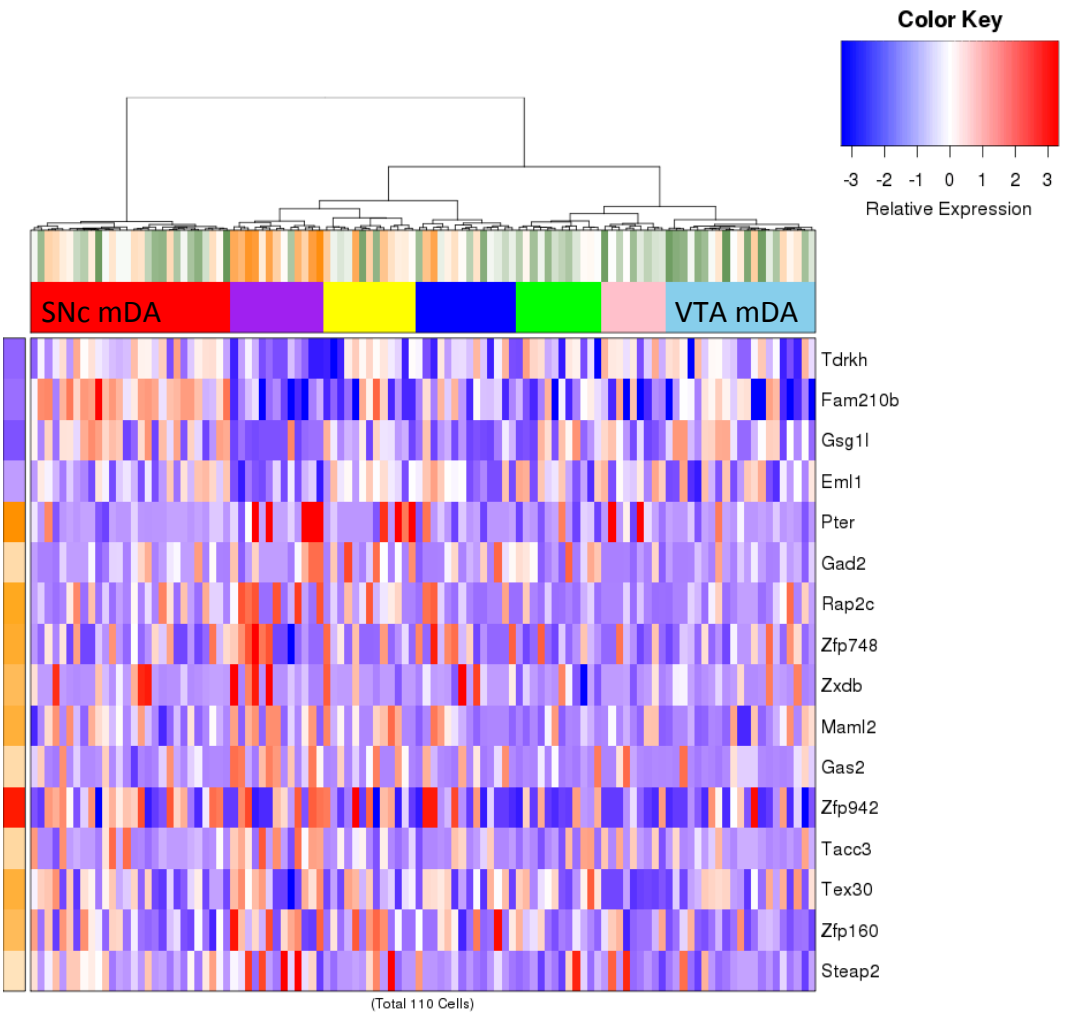
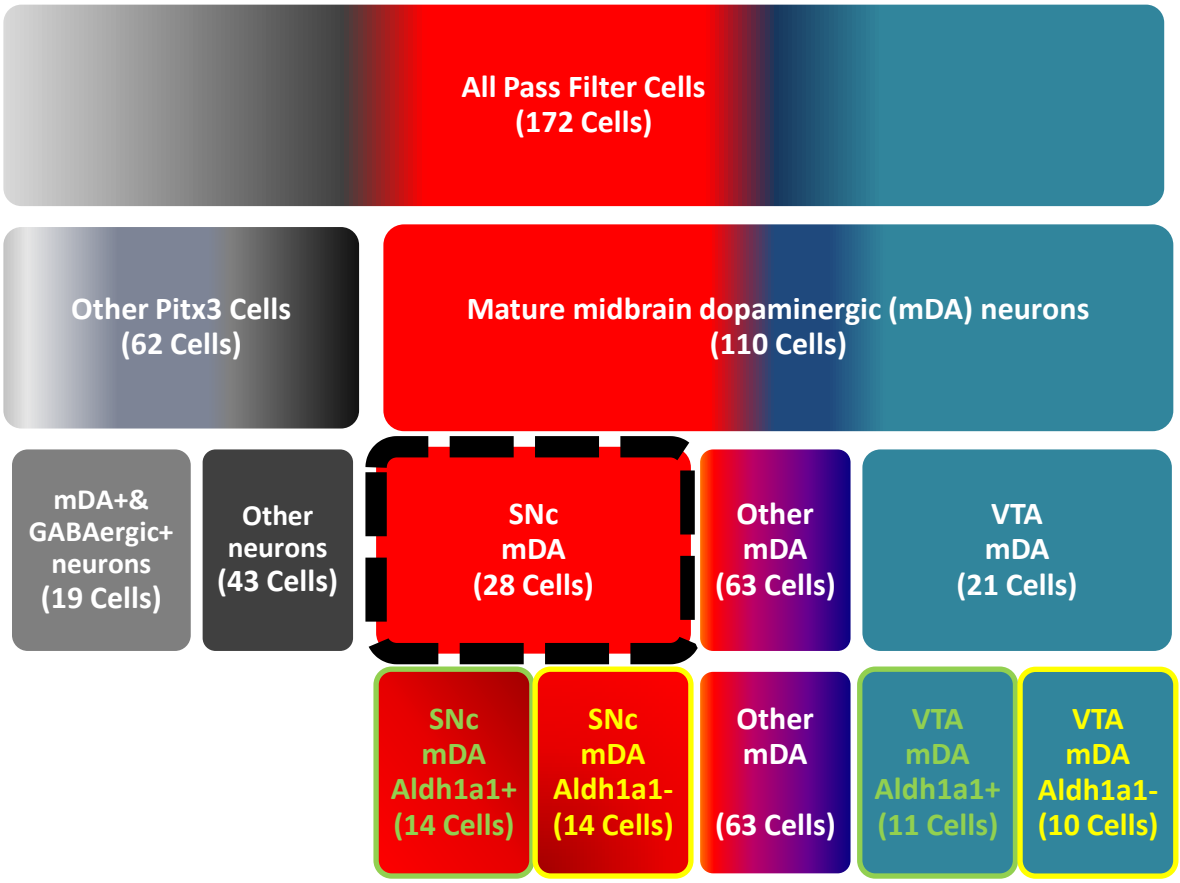


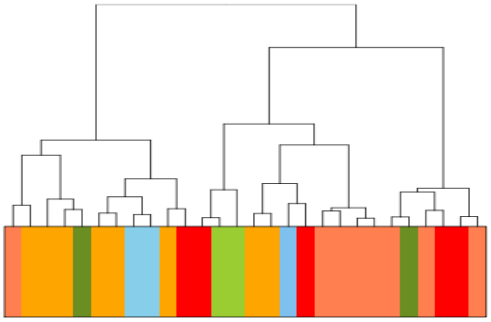
Figure 4-18: Heatmap showing the expression of genes identified by differential expression analysis. The upper half of the heatmap where the rows are indicated by the orange side bar represent genes that are depleted among the purple group of dopaminergic neurons. The lower half of the heatmap where the rows are indicated by the blue side bar represent genes that are enriched among the purple group of dopaminergic neurons.

Among the 28 SNc dopaminergic neurons, clustering revealed two sub-classes of SNc dopaminergic neurons that can be validated



Previously, we identified that a cluster of 28 cells among all mature midbrain dopaminergic neurons express marker genes of SNc dopaminergic neurons. I decided that this number of cells, although relatively few, may still be sufficient to identify further classes of SNc dopaminergic neurons among them. I repeated the same *a priori* clustering and splitting procedure on these 28 neurons, as shown in Figure 4-19. As before, I determined that the experimental batch and number of expressed genes did not drive the clustering. The number of genes detected by the two major clusters were not significantly different when examined using Student's t test ($p > 0.34$).

A



B

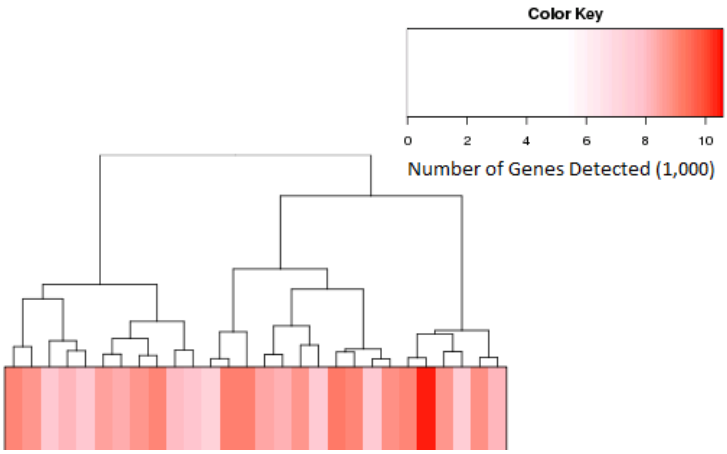


Figure 4-19: Clustering dendrogram of 28 SNc dopaminergic neurons. Panel A: The different colours in the top dendrogram indicate the experimental batch. Panel B: The varying shades of red colour in the bottom dendrogram indicate the number of genes expressed. The clustering dendrogram in both panels are the same, which was determined using the expression of all expressed genes according to Materials and Methods D5, using the 28 SNc Neurons presented in the figure.

After splitting the 28 SNc dopaminergic neurons into two subpopulations, I used differential expression analysis to identify their respective marker genes (Figure 4-20). *Aldh1a1* is among the top 3 genes that show the highest fold change (the 11 cells marked by the green box versus the remaining 17 cell). The authors in [3] showed that there are there are at least two subpopulations of SNc neurons, one subpopulation express *Aldh1a1* and the other subpopulation lacks *Aldh1a1* expression. Other studies also supported these subpopulations [202, 203]. I conclude that the two subpopulations of SNc dopaminergic neurons identified in our data-driven approach correspond well with the two subpopulations of SNc dopaminergic neurons that were distinguished by *Aldh1a1* expression in [3]. I consider this as a validation that the entirely data-driven approach using single cell RNA-Seq data was able to discover physiologically relevant subpopulations of SNc dopaminergic neurons. The protein localisation of ALDH1A1 in [3] is reproduced in Figure 4-21A.

It is interesting to note that among the 17 dopaminergic neurons marked by the yellow shaded box, there are 3 cells (marked by the inner green shaded box) that exhibit expression patterns that are similar to both the 11 cells to the left, and are also similar to the 14 cell to the right. I am unsure whether these represent capture sites with doublet cells, or represent some cells that exhibit dual properties of both the *Aldh1a1+* subpopulation and the *Aldh1a1-* subpopulation. For now, they are considered as part of the *Aldh1a1+* subpopulation. In summary, we identified two major subpopulations of SNc

dopaminergic neurons: the *Aldh1a1+* subpopulation with 14 cells, and the *Aldh1a1-* subpopulation with 14 cells.

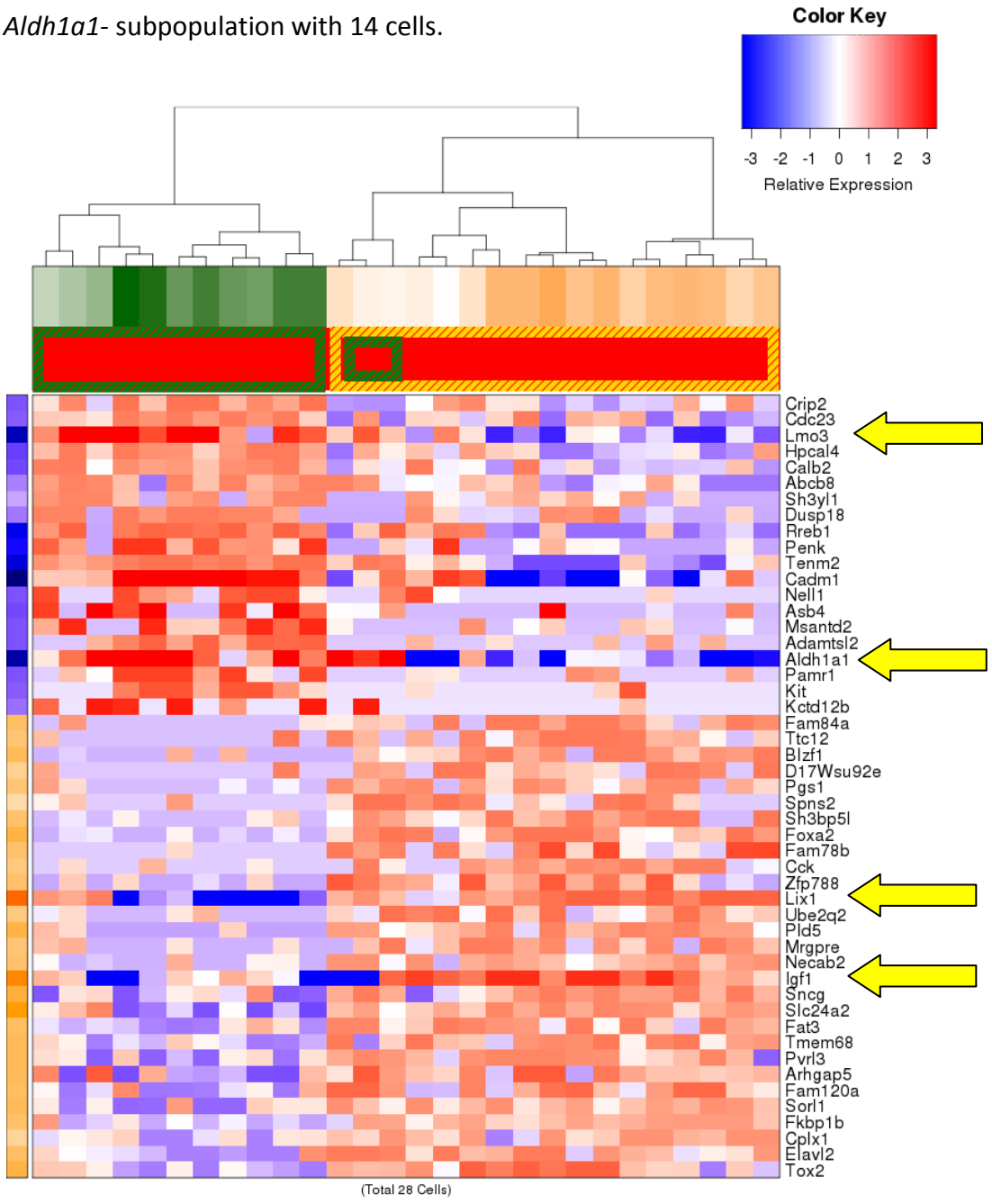


Figure 4-20: Heatmap showing the expression of genes identified by differential expression analysis. The upper half of the heatmap where the rows are indicated by the orange side bar represent genes that are depleted among 11 dopaminergic neurons marked by the green shaded box. The lower half of the heatmap where the rows are indicated by the blue side bar represent genes that are enriched among the 17 dopaminergic neurons marked by the yellow shaded box. The yellow arrows indicate the genes that are selected for validation by protein or mRNA staining experiments.

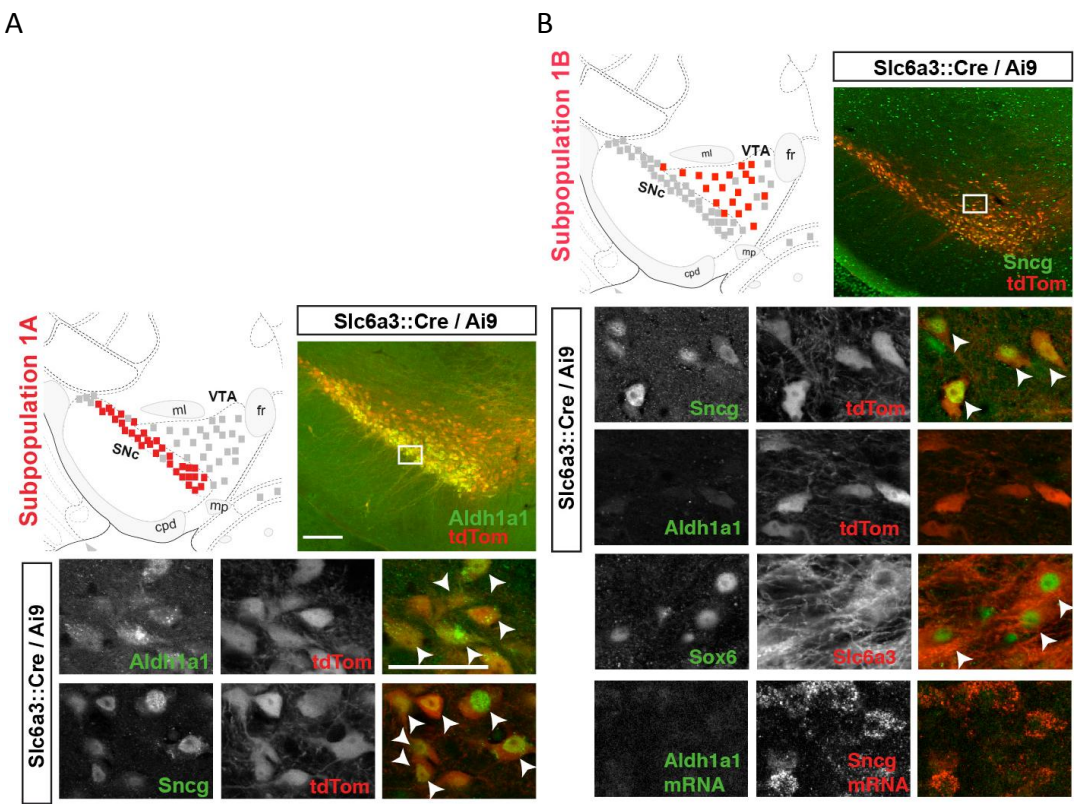


Figure 4-21: The staining results showing that in the SNc region, there are two populations of cells: ALDH1A1+ and ALDH1A1-. Panel A. ALDH1A1 and SNCG protein staining in the SNc region demonstrates that the ALDH1A1+ subpopulation of SNc neurons are located in the SNc. Panel B. *Aldh1a1* mRNA, SNCG protein and SOX6 protein staining in the SNc region reveals that there is a subpopulation of SNc neurons that are absent of *Aldh1a1* mRNA expression, the same cells are positive for *Sncg* mRNA, SOX6 protein and SLC6A3 protein suggesting they are indeed SNc dopaminergic neurons. The high magnification image for VTA was not discussed in the manuscript, thus it is unclear whether the same phenomenon can also be observed in VTA. Scale bar in low magnification = 200 microns, scale bar in high magnification = 50 microns. Figures reproduced from [3], which is published under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/3.0/).

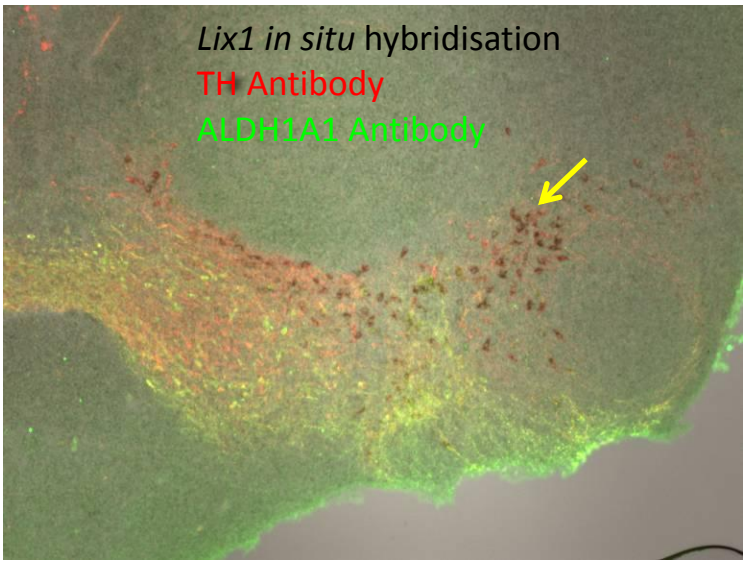
Although the study in [3] was able to identify the *Aldh1a1*+ subpopulation of SNc dopaminergic neurons, it was not able to identify any additional marker genes for this population. This is because in a single cell qRT-PCR based approach as in [3], the genes examined were selected beforehand based on existing literature. By contrast, here we identified several dozens of candidate marker genes for this subpopulation of SNc dopaminergic neurons.

Furthermore, the study in [3] was only able to distinguish the other subpopulation of SNc dopaminergic neurons by relying on the absence of *Aldh1a1*, without providing a specific positive marker of this *Aldh1a1*- subpopulation. As a result, the only validation was showing that some SNc dopaminergic neurons lacks the expression of ALDH1A1 protein (Figure 4-21B). On the contrary, using our -omics approach based on single cell RNA-Seq, we measured the expression values of almost the entire transcriptome. Thus our study has been able to identify novel marker genes for each of the subpopulations. For the full list of marker genes, please refer to the heatmap in Figure 4-20. Among those, my collaborators Dr. Mary Green and Dr. Siew-Lan Ang selected *Lix1* and *Igf1* as marker genes for the *Aldh1a1*- SNc subpopulation for further experiments. The experiments seek to use RNA *in situ* hybridisation and antibody staining to demonstrate that the *Lix1+&Igf1+&Aldh1a1*- SNc subpopulation is distinct from the *Lix1-&Igf1-&Aldh1a1*+ SNc subpopulation and that both subpopulations are located in the SNc region. Our collaborators' staining results are shown in Figure 4-22A,B. We can see that among the TH protein expressing SNc dopaminergic neurons, there are two types: (i) one type

that expresses the ALDH1A1 protein (yellow in colour) and is located generally at the ventral tier of SNc; (ii) the other type that is absent of the ALDH1A1 protein (red in colour) and is located generally at the dorsal tier of SNc. Only the dorsal SNc dopaminergic neurons absent for ALDH1A1 are expressing *Lix1* mRNA (Figure 4-22A) and *Igf1* mRNA (Figure 4-22B). This result based on othogonal methods again suggest that it is possible that my discovery based on single cell RNA-Seq data is correct: the SNc cells can potentially be separated into two disjoint subpopulations: the *Lix1+&Igf1+&Aldh1a1-* subpopulation and the *Lix1-&Igf1-&Aldh1a1+* subpopulation.

In the Figure 4-22B, if we turn our attention to VTA, It is interesting to observe that some dorsal VTA cells also express *Lix1* and *Igf1* mRNA. These cells also express SOX6 protein (data not shown), suggesting their molecular identity is closer to SNc dopaminergic neurons, despite being located in the VTA. This demonstrates that the SNc-VTA anatomically distinction that was used in literature is unable to delineate with SNc dopaminergic neurons with high precision. As a result, it highlights the necessity of performing measurements at the resolution of individual cells.

A



B

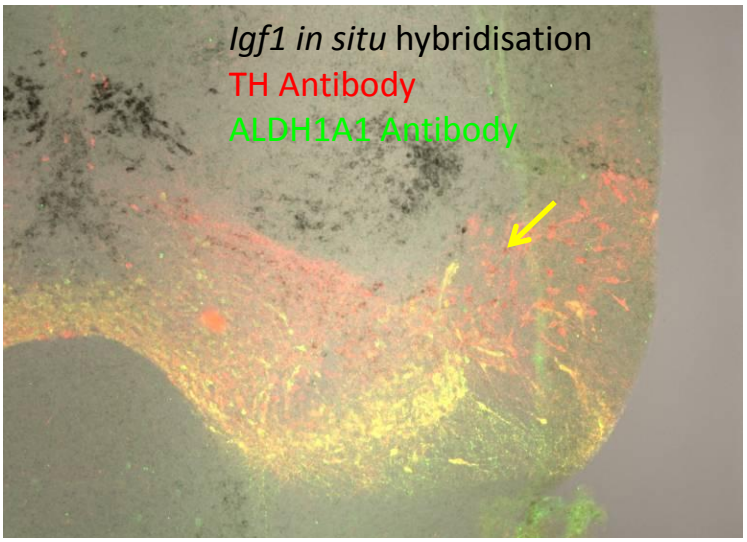


Figure 4-22: Staining showing the localisation of Lix1+ subpopulation and Igf1+ subpopulation. The experiments are performed by collaborator Dr. Mary Green. Panel A: RNA *in situ* hybridisation of *Lix1* mRNA in the midbrain. Antibody staining of TH protein (represented by red) and ALDH1A1 protein (represented by green) of the same sample. The ALDH1A1+ cells are observed at the ventral tier of both SNc and VTA. The *Lix1*+ cells are observed at the dorsal tier of both SNc and VTA (indicated by the yellow arrow). The *Lix1*+ cells are distinct from the ALDH1A1+ cells. Panel B: Same as Panel A, except that the RNA *in situ* hybridisation is for *Igf1* mRNA. The *Igf1*+ cells (indicated by the yellow arrow) are distinct from the ALDH1A1+ cells.

Explore further heterogeneities among the 28 SNc dopaminergic neurons

Thus far, I identified two classes of SNc dopaminergic neurons: the 11 neurons that are generally *Aldh1a1* positive and *Lix1* negative and the 17 neurons that are generally *Aldh1a1* negative and *Lix1* positive that are experimentally supported. However, in Figure 4-20, the three cells that are marked by both the yellow shaded box and the small green shaded box suggest that there might be further heterogeneity. I sought to examine whether there exist additional more refined subpopulations among these 28 SNc dopaminergic neurons. In theory, I could split the dendrogram to any number of clusters. However, a very small number of starting cells may prevent the identification of reliable clustering results and cell-type specific markers. I will use the following toy example to elaborate the issue with defining subpopulation with very few cells. Assume we randomly selected 2 cells originating from 2 cell types among a total of 50 cells and decided to use differential expression to identify the marker genes for the 2 randomly selected cells. Although the 2 cells originate from 2 different cell types, due to stochastic expression, by random chance we will be able to identify many genes that are specifically expressed in the 2 randomly selected cells and not expressed in the remaining 48 cells. This list of genes, despite appears to specifically mark the 2 cells, arise purely due to random stochastic expression, because the 2 cells originate from 2 different cell types. If we had selected a lot more cells to perform the same differential expression experiment, e.g. randomly selecting 20 cells out of 50, the chances of finding many genes that are specifically expressed among 20 randomly selected cells originate from different cell types due to stochastic expression is extremely rare.

As a result, when the number of cells within a putative subpopulation is small, it is difficult to examine whether the identified cell clusters and their corresponding marker genes are due to stochastic expression, or due to genuine biological processes. Thus, given this small number of starting cells, I decided to split the dendrogram only into four clusters. This is effectively splitting the yellow shaded cluster of 17 cells described in the previous section in Figure 4-20 into three additional clusters while maintaining the green shaded cluster of 11 cells unchanged. For each subpopulation, I identified its marker genes using differential expression analysis by contrasting the chosen subpopulation against the remaining cells of the 28 SNc dopaminergic neurons. The results for each subpopulation are shown in Figure 4-23 A,B,C,D (limited to the top 25 genes of either direction due to space limit). Please note that the green shaded cluster in Figure 4-20 is the same as the green shaded cluster in Figure 4-23D, thus the differential expression result is the same, except that Figure 4-23D shows less number of genes (up to 25 genes of either direction).

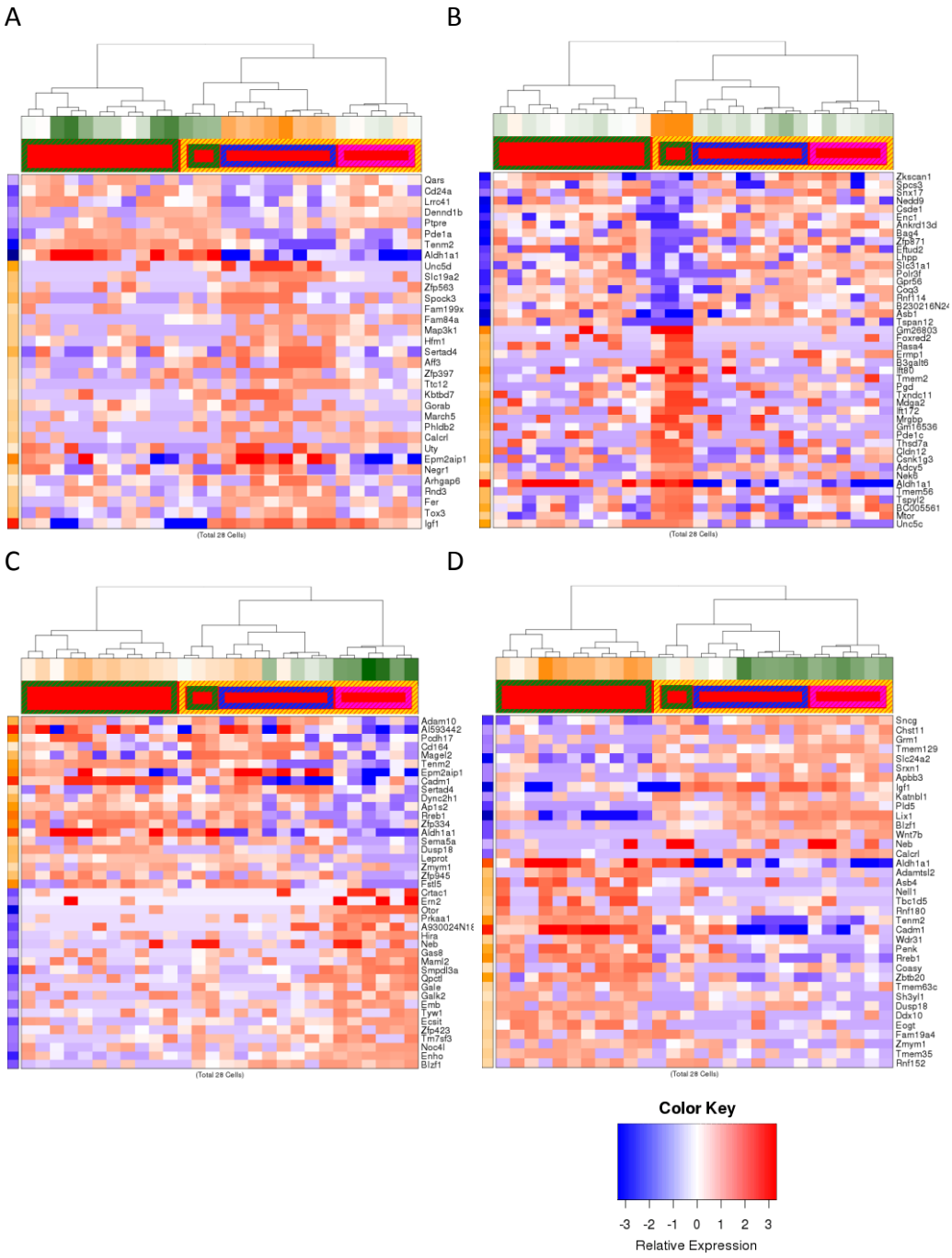


Figure 4-23: Heatmap showing the expression of genes identified by differential expression analysis by contrasting the chosen group with the remaining SNc dopaminergic neurons. Panel A, the subpopulation marked by blue inner shaded box. Panel B, the subpopulation marked by green inner shaded box. Panel C, the subpopulation marked by pink inner shaded box. Panel D: the same subpopulation marked by a single green shaded box.

The above approach depends on splitting the clustering dendrogram. However, this clustering dendrogram may not be entirely accurate. I sought to directly identify the list of genes that are highly variable among the 28 SNc dopaminergic neurons without relying on the dendrogram (Materials and Methods D7). The reason that highly variable genes are of interest is because cell type specific marker genes - by definition those that are highly expressed in some cells and absent from others - are highly variable among the full population of cells. This list of genes can be useful for FISH or immuno-staining to examine whether they mark neurons that are located in certain anatomical regions of the SNc or mark different sets dopaminergic neurons. The top 100 most variable genes among the 28 SNc dopaminergic neurons were also used to perform hierarchical clustering. The results are shown in Figure 4-24.

In summary, my exploratory analyses putatively suggested that there could be additional subpopulations identifiable among the 28 SNc dopaminergic neurons. By sequencing additional cells from the same source and determining that these subtypes re-occur, we could validate that these putative subpopulations are biologically reproducible. However, at the time of writing, we did not analyse additional single cells. We could also seek to demonstrate that these subpopulations are biologically relevant by performing staining experiments to reveal their distinct anatomical locations, or by performing other studies such as axon tracing and electrophysiological recordings. These follow up experimental studies are currently being discussed and carried out by my

collaborators. At the time of writing, the results of these experimental studies are not yet available.

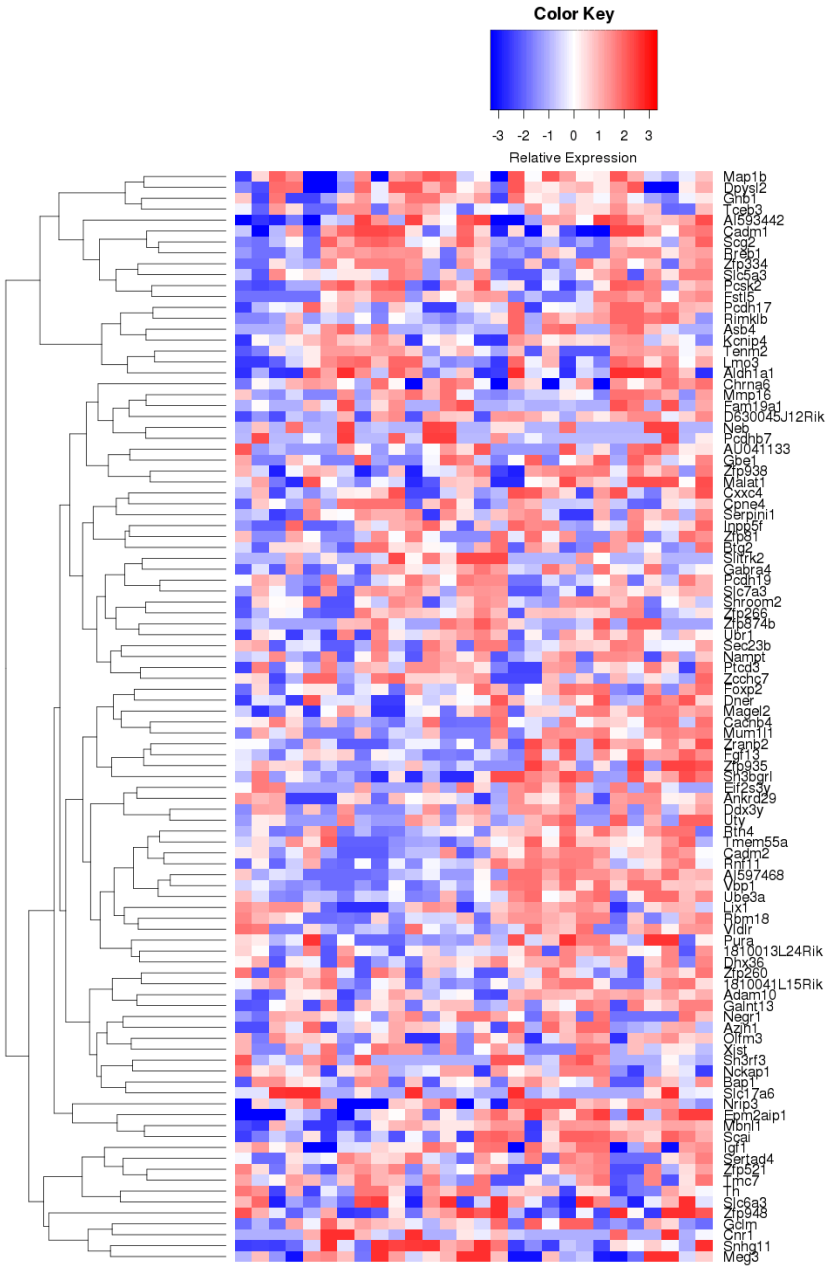
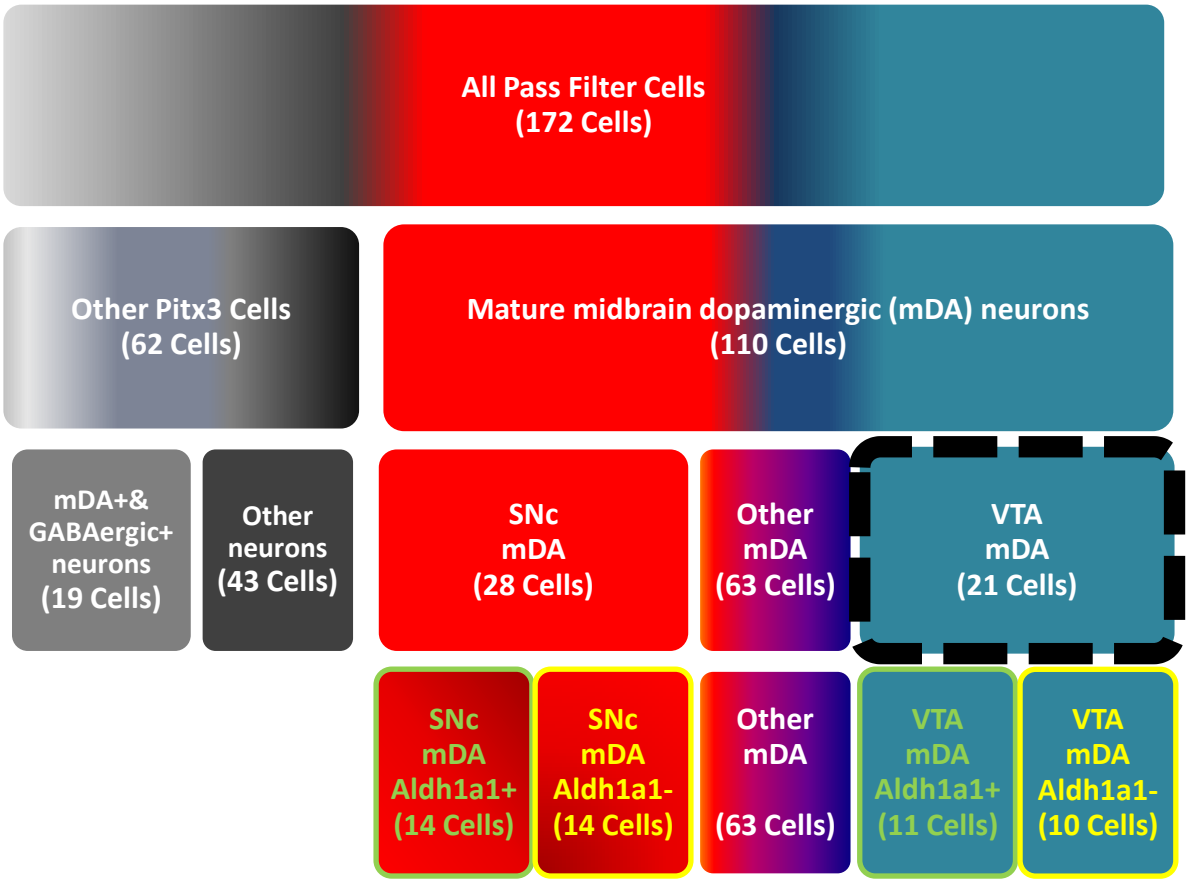


Figure 4-24: Heatmap showing the expression of genes identified by selecting the top 100 most variable genes. These genes are potentially subpopulation-specific genes because their expression values are highly divergent between different cell groups.

Among the 21 VTA dopaminergic neurons, clustering revealed two sub-classes of VTA dopaminergic neurons that can be validated



Previously, from 110 mature midbrain dopaminergic neurons, I identified a cluster of 21 cells that express marker genes of VTA dopaminergic neurons. Despite being a very small number of cells, I considered 21 cells might be sufficient to identify further subpopulations among them. Just as for SNc dopaminergic neurons, I sought to examine whether I could identify more refined subpopulations of dopaminergic neurons by repeating the same *a priori* clustering and splitting procedure on these 21 neurons. I also verified that potential confounding factors such as experimental batch and number of expressed genes were not driving the clustering (data not shown). The clustering dendrogram divided the 21 VTA dopaminergic neurons into two major groups. I

sought to identify the genes that are specifically expressed in either group of cells as shown in Figure 4-25.

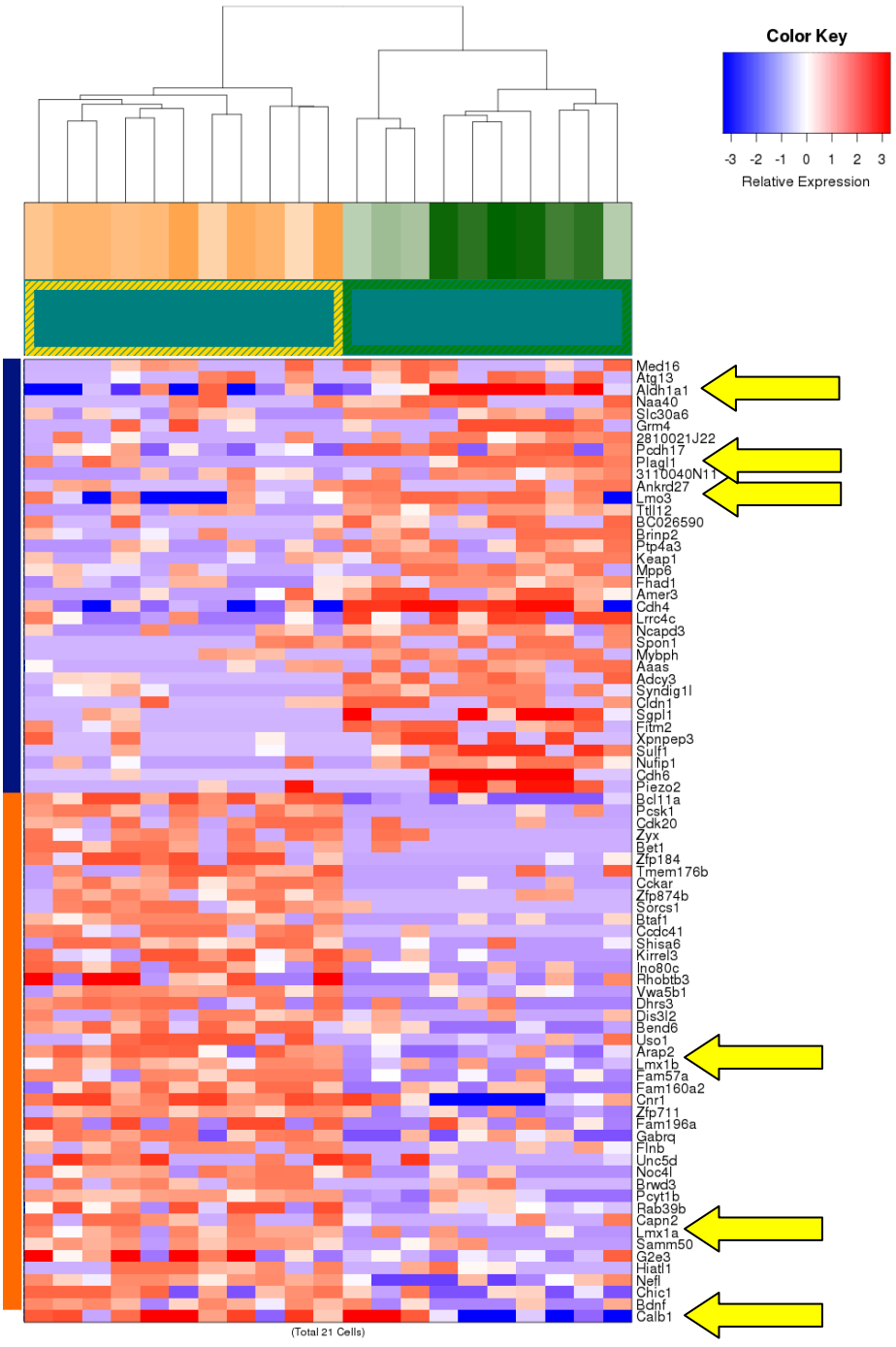


Figure 4-25: Heatmap showing the expression of genes identified by differential expression analysis. The upper half of the heatmap where the rows are indicated by the blue side bar represent genes that are depleted among 11 dopaminergic neurons marked by the yellow shaded box. The lower half of the heatmap where the rows are indicated by the orange side bar represent genes that are enriched among the 10 dopaminergic neurons marked by the green shaded box. *Calb1* has been added manually to the list of genes to examine its expression pattern. Arrow indicates the genes that were known to mark subpopulations of VTA dopaminergic neurons.

Previously in the SNc, *Aldh1a1* was shown to be specifically expressed in one subpopulation of dopaminergic neurons. It is thus interesting to note that here in VTA, *Aldh1a1* was again identified as a gene that is specifically expressed in one subpopulation of VTA dopaminergic neurons. Recall that in the above section, the subpopulation of *Aldh1a1*+ SNc cells match the expression pattern of group DA^{1A} neurons identified by [3]. Thus, I sought to investigate whether the currently identified *Aldh1a1* positive subpopulation of VTA dopaminergic neurons could be matched to one of the populations identified in [3] by re-examining the staining images contained in [3]. The staining images from the original paper are reproduced in Figure 4-26. We can see that the DA^{2B} subpopulation (in VTA) defined by [3] express ALDH1A1 protein. In the same staining image, we can also see that there is another population of VTA dopaminergic neurons in which ALDH1A1 protein staining is absent. The staining images confirmed that the *Aldh1a1*+ and *Aldh1a1*- groups are *bona fide* subclasses of dopaminergic neurons in the VTA region. Previously, the authors in [3] identified that *Calb1* and *Aldh1a1* are co-expressed among the DA^{2B} population. As a result, *Calb1* has been added manually to Figure 4-25 to examine its expression pattern. We can see that *Calb1* is only expressed among the VTA cells in which *Aldh1a1* expression is absent, which contradicts with [3]. This contradiction should be interrogated by co-staining *Aldh1a1* and *Calb1* in the same sample.

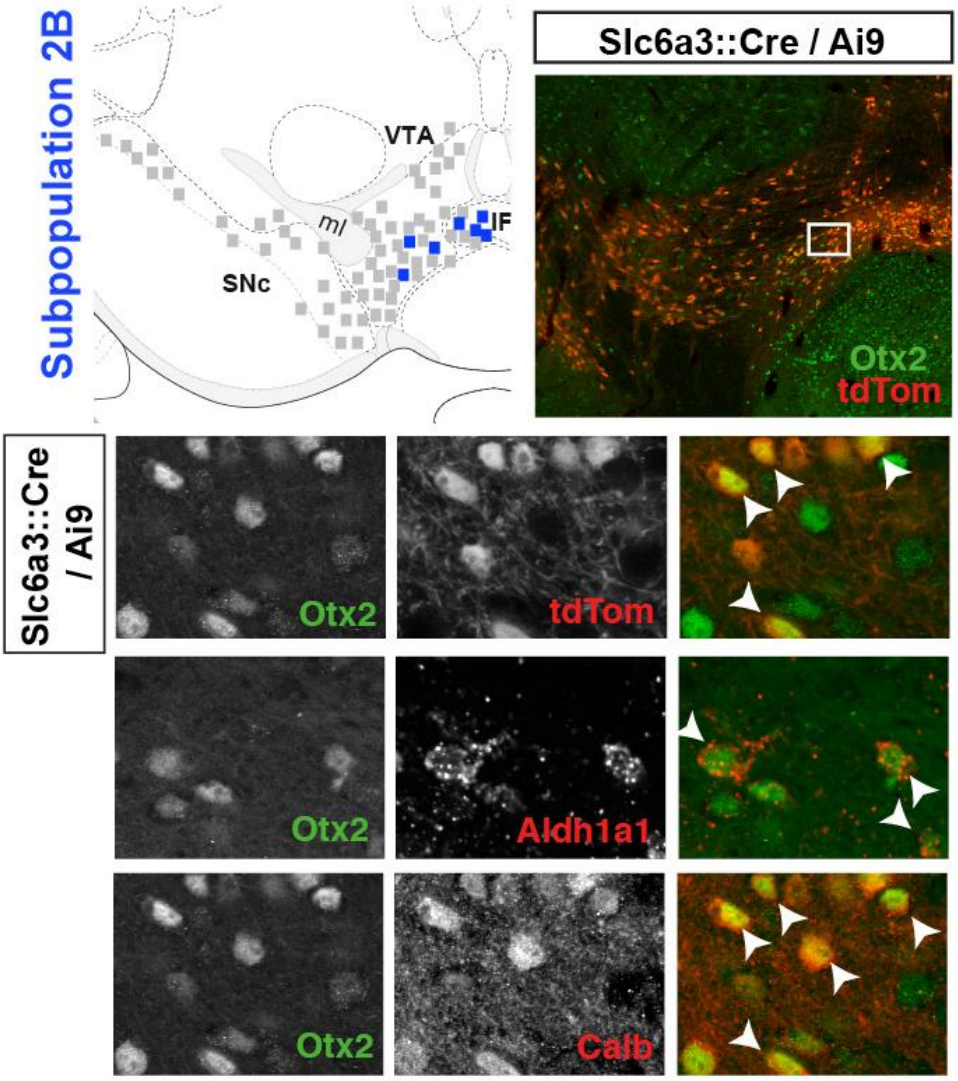


Figure 4-26: ALDH1A1, OTX2 and CALB1 protein staining in the VTA region. These staining results reveal that there is a subpopulation of VTA dopaminergic neurons that express OTX2 and SLC6A3 and lacks ALDH1A1 expression. The authors claimed that the same population (referred to by them as population DA^{2B}) also express CALB1. However, a co-staining of CALB1 and ALDH1A1 was not available from [3]. Image scale is identical to that of Figure 4-21 above. Figures reproduced from [3], which is published under the under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/3.0/).

Explore further heterogeneities among the 21 VTA dopaminergic neurons

Again, just as for SNc dopaminergic neurons, I sought to directly identify top 100 genes that are highly variable among the 21 VTA dopaminergic neurons without using the dendrogram (Materials and Methods D7). This list of genes could be useful for FISH or immuno-staining to examine whether they mark neurons that are located in certain anatomical regions of the VTA or mark non-overlapping dopaminergic neurons. The top 100 most variable genes among the 21 VTA dopaminergic neurons were used to perform hierarchical clustering (Figure 4-27). My exploratory analyses suggest there could be additional subpopulations present among the 21 VTA dopaminergic neurons. Similarly, due to the small number of cells analyzed here, further studies on these potential subpopulations are required to demonstrate that these are *bona fide* subpopulations.

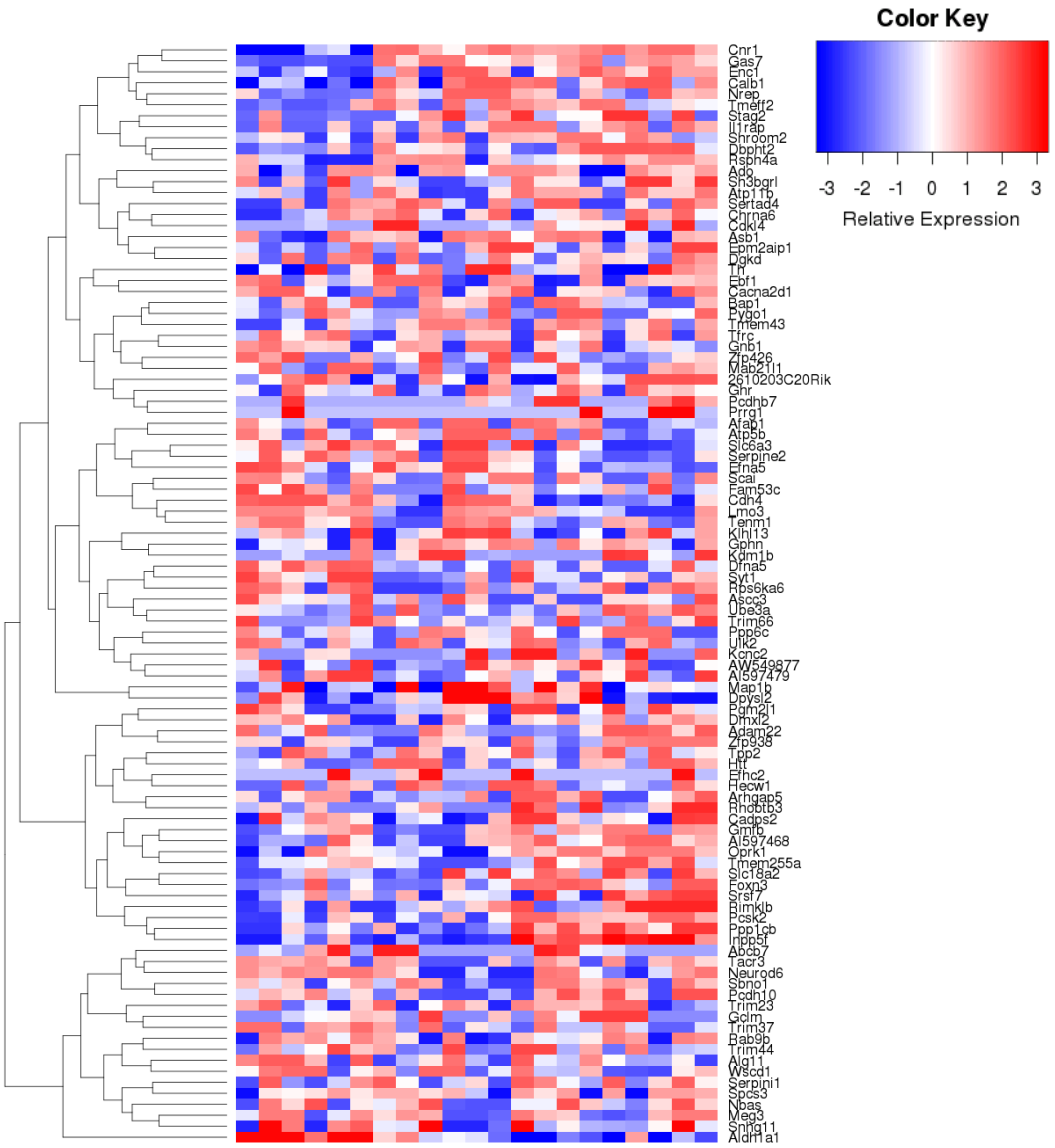
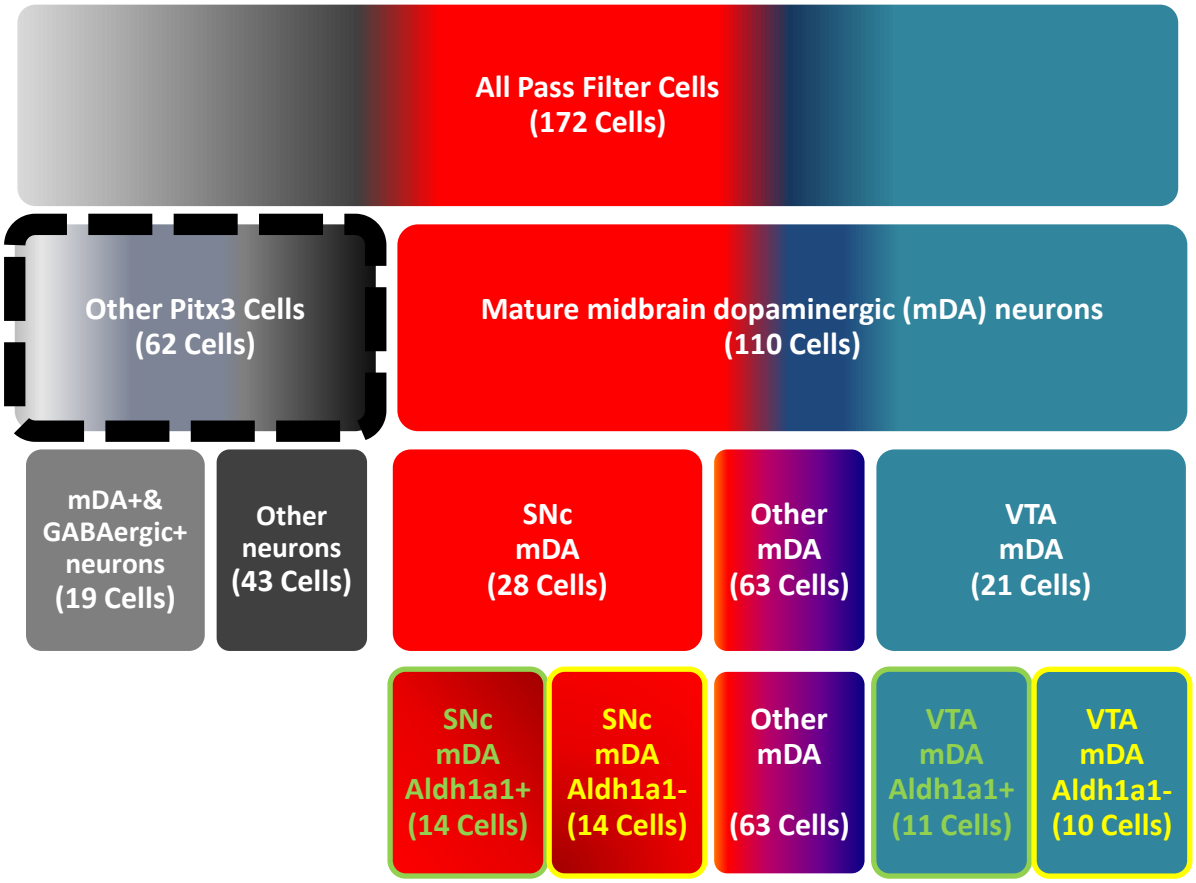


Figure 4-27: Heatmap showing the expression of genes identified by selecting the top 100 most variable genes. These genes are potentially subpopulation-specific genes because their expression values are highly divergent between different cell groups.

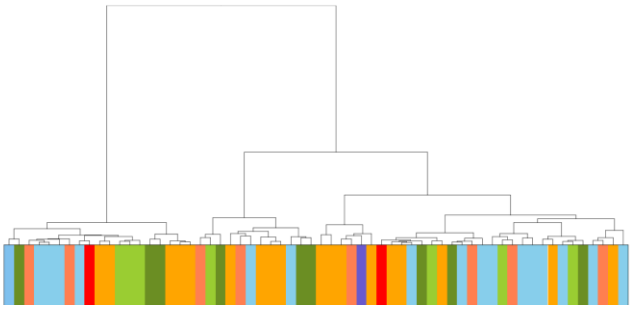
Among the group of 62 other GFP+ cells, two different classes of neurons were identifiable



Recall that all 172 cells were split into two major groups: the group of 110 mature midbrain dopaminergic neurons that highly express the core dopamine machinery (*Th, Ddc, Slc6a3* and *Slc18a2*), and the other group of the remaining 62 GFP+ cells that have intermediate or low expression of the four genes. We clustered these remaining GFP+ cells using the same *a priori* hierarchical clustering based approach (Materials and Methods D5). I sought to identify distinct cell types among them. To achieve this, I performed *a priori* clustering, and obtained the clustering dendrogram shown in Figure 4-28. I first verified that the clustering was not driven by confounders such as experimental batch. The number of genes detected by the two major clusters were significantly different when examined using Student's t test ($p < 6.5 \times 10^{-3}$), suggesting the 2 clusters

may have different number of genes expressed. The average number of genes detected in the two clusters are 7729, 8264 respectively, with The 95 percent confidence interval of the difference in mean being (160 ~ 909), suggesting in general, there is slightly less number of genes detected in the first cluster.

A



B

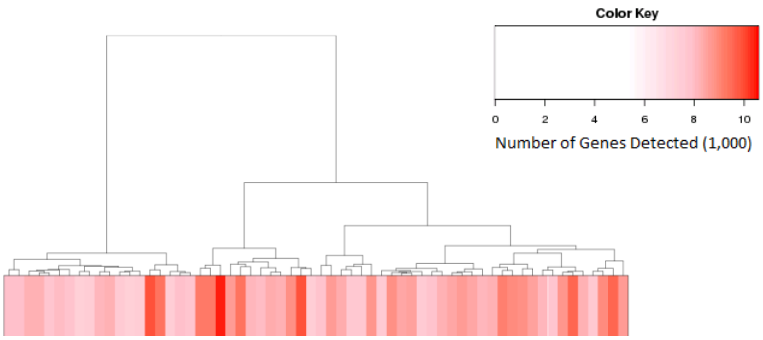
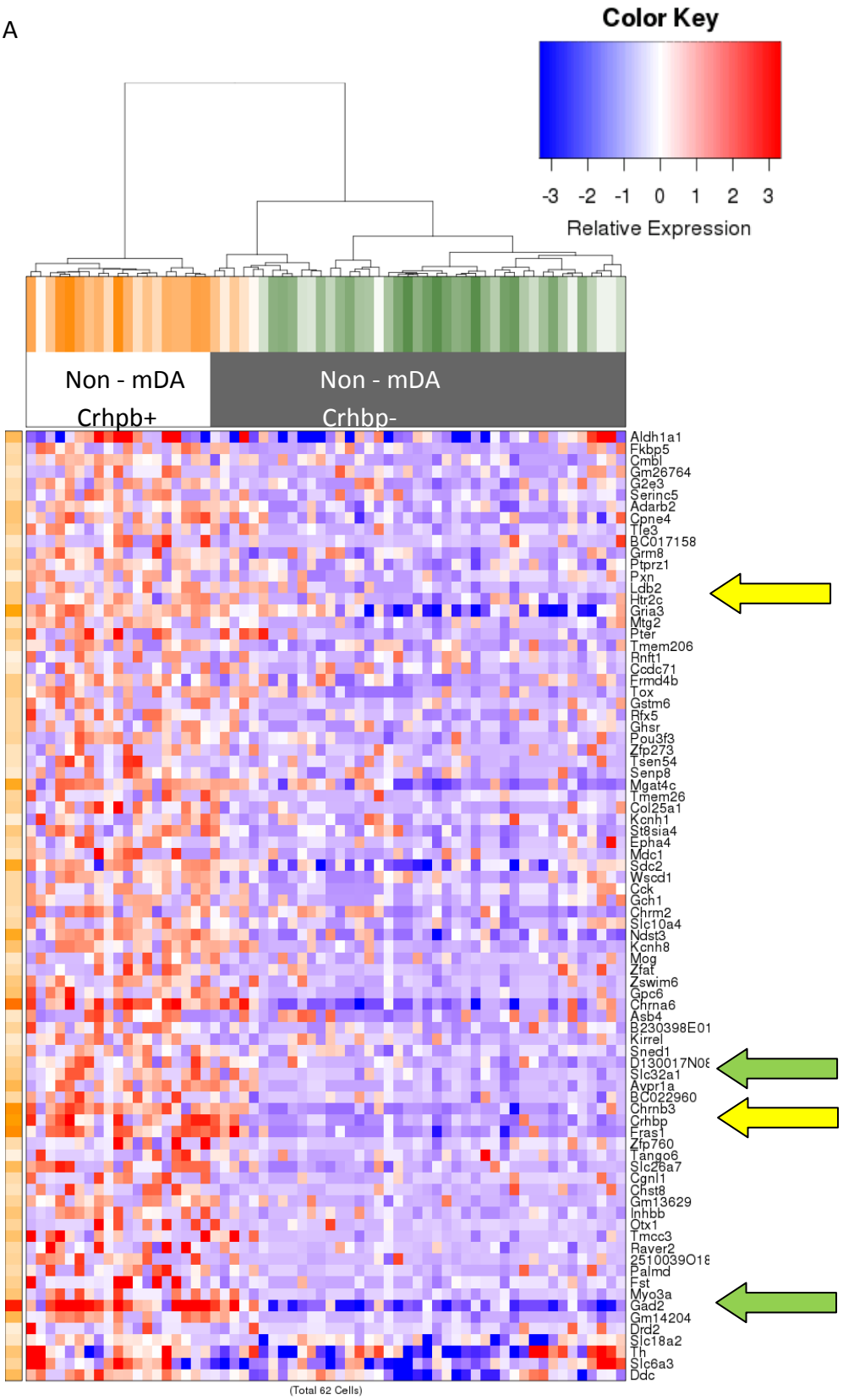


Figure 4-28: Clustering of all mature midbrain dopaminergic neurons sequenced in the experiment. Panel A: Dendrogram determined using method described in [190]. The colour bars beneath the dendrogram represent the experimental batch. Panel B: The dendrogram is the same as Panel A. The colour of the bars denote the number of genes detected in each cell.

I split the 62 cells into two subpopulations according to the dendrogram. The two subpopulations are marked by the white and dark-grey horizontal bars in Figure 4-29A. To understand the cells marked by the white horizontal bar, I performed differential expression analyses (Materials and Methods D9) between the two groups. I identified the list of genes that are more highly expressed by the white subpopulation relative to the dark grey population as shown in Figure 4-29. We can see that the subpopulation of cells indicated by white strongly express *Crhbp, Htr2c, Slc32a1* and *Gad2*. We will name this population the *Crhbp+* subpopulation. Previously, we have seen that some of these 62 other GFP+ cells express non-negligible amount of dopaminergic machinery, albeit their expression levels are lower than that of 110 mature dopaminergic neurons. In order to investigate whether the dopaminergic marker expressing cells are enriched in either of the two clusters, I manually added the five dopaminergic neuron marker genes: *Ddc, Slc6a3, Slc18a2, Drd2* and *Th* to Figure 4-29A. It can be seen that a large fraction of the cells in the white clusters express the dopaminergic neuron marker, suggesting they are potentially capable of transmitting dopamine. The localisation of this group of *Crhbp+* cells were investigated by our collaborators and will be presented in the next section.

Likewise, I also identified the list of genes that are more highly expressed by the dark-grey subpopulation relative to the white subpopulation as shown in Figure 4-29B. The function and identity of this subpopulation of cells are not understood. However, because they do not express marker genes of dopaminergic neurons, I did not wish to investigate further.

A


Color Key
-3 -2 -1 0 1 2 3
Relative Expression
Non - mDA
Crhpb+
Non - mDA
Crhbp-
Aldh1a1
Fkbp5
Cmbl
Gm26764
G2e3
Serinc5
Adarb2
Cpne4
Tle3
BC017158
Grm8
Ptprz1
Pxn
Ldb2
Htr2c
Gria3
Mtg2
Pter
Tmem206
Rnft1
Ccdc71
Frmd4b
Tox
Gstm6
Rfx5
Ghsr
Pou3f3
Zfp273
Tsen54
Senp8
Mgat4c
Tmem26
Col25a1
Kcnh1
St8sia4
Epha4
Mdc1
Sdc2
Wscd1
Cck
Gch1
Chrm2
Slc10a4
Ndst3
Kcnh8
Mog
Zfat
Zswim6
Gpc6
Chrna6
Asb4
B230398E01
Kirrel
Sned1
D130017N0
Slc32a1
Avpr1a
BC022960
Chrnb3
Crhbp
Fras1
Zfp760
Tango6
Slc26a7
Cgnl1
Chst8
Gm13629
Inhbb
Otx1
Tmcc3
Raver2
2510039O18
Palmd
Fst
Myo3a
Gad2
Gm14204
Drd2
Slc18a2
Th
Slc6a3
Ddc
(Total 62 Cells)

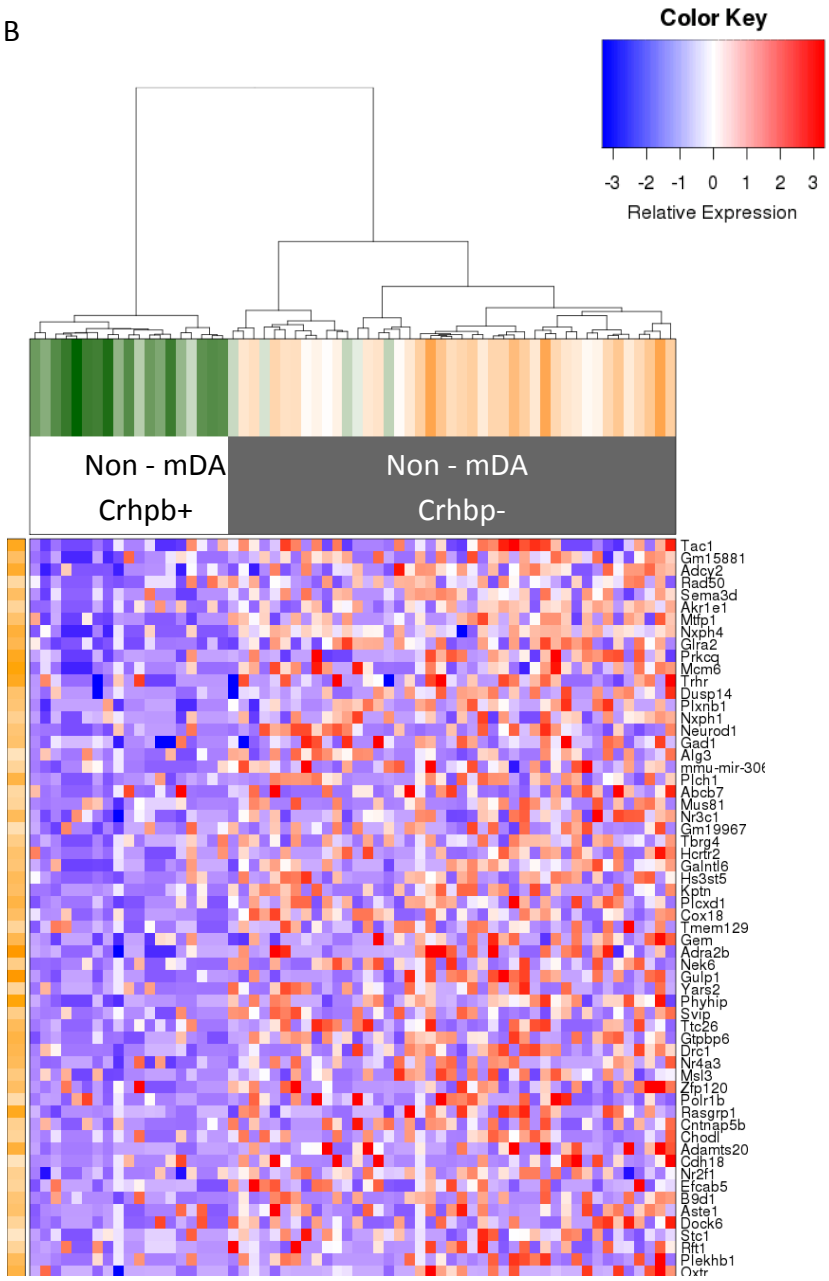


Figure 4-29. Heatmap showing the differentially expressed genes. Panel A: Heatmap showing genes that are more highly expressed in the cells indicated by white bar obtained by performing differentially expression analyses contrasting the cells indicated by white and the cells indicated by dark-grey. The five dopaminergic neuron marker genes: *Ddc, Slc6a3, Slc18a2, Drd2* and *Th* are manually added to the end of the heatmap obtained from differential expression analysis. Genes indicated by the yellow arrows indicate the genes that were selected for experimental validations. Genes indicated by the green arrows indicate the two genes that have a known function in the *Slc32a1*+ subpopulation: *Gad2* for GABA synthesis, and *Slc32a1* for non-canonical GABA release. Panel B: Same as panel A, except that showing genes more highly expressed in the dark-grey group.

The group of *Crhbp*+ cells are located at two anatomical regions: the ventral midline region and scattered at the VTA of mouse midbrain

This group of cells is potentially the same group of cells that were identified by [3] as the *Slc32a1*+ dopaminergic neurons within the VTA. The original staining for *Slc32a1* [3] is reproduced in Figure 4-30. We can see that the *Slc32a1*+ cells are located both in the ventral-midline and in the VTA. It is interesting that among the four cells expressing *Slc32a1* in the high magnification image (i.e. the four red nuclei), three of these express the Th protein, and one do not. This staining result is consistent with our single cell RNA-Seq results shown in Figure 4-29, where some *Slc32a1*+ cells express *Th* mRNA and the other four dopaminergic neuron marker genes; and some other *Slc32a1*+ cells do not.

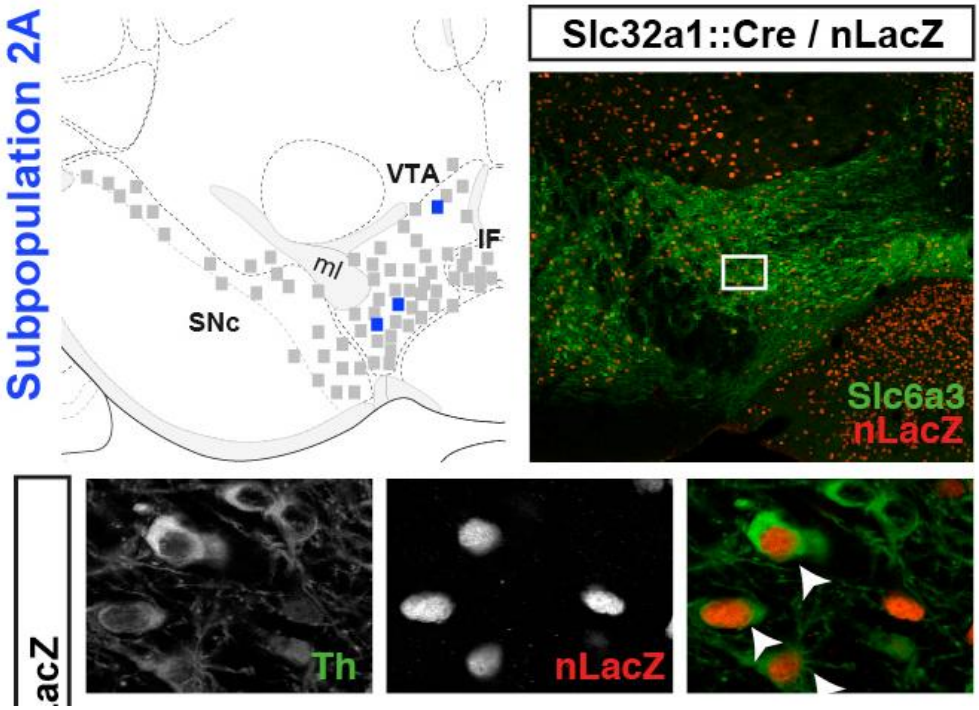


Figure 4-30: Staining showing the subpopulation of dopaminergic neurons marked by Slc32a1. TH protein was used as a dopaminergic neuron marker. We can see that there are two anatomic regions contained cells that express *Slc32a1*: the ventral midline and the VTA. Within the ventral midline, it is not clear whether the *Slc32a1*-expressing cells also express TH protein. Within the VTA, we can see that some cells co-express TH protein and *Slc32a1*: specifically, from the high magnification image, we can see that 3 out of 4 *Slc32a1*-expressing cells located in the VTA co-express TH protein. Figures reproduced from [3], which is published under the under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/3.0/).

Among the candidate marker genes for the *Crhbp*+ group of cells (marked by the white horizontal bar in Figure 4-29), we selected two abundantly expressed marker genes for validation: *Crhbp* and *Htr2c*. The chosen genes were used to perform RNA *in situ* hybridisation experiments by our collaborator Dr. Mary Green. Experimental results indicate that these genes mark a group of cells located within the VTA, and a second group of cells lie near the ventral midline

region, as shown in Figure 4-31. It appears that these two populations matched well with the localisation of *Slc32a1*+ cells in Figure 4-30.

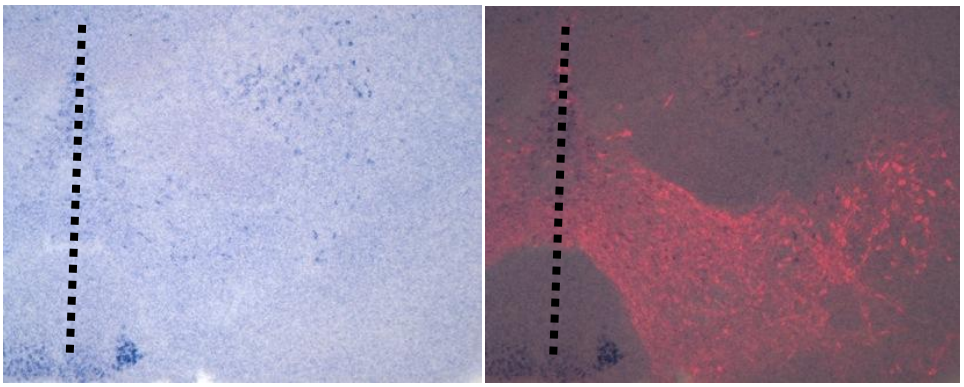


Figure 4-31: In situ hybridisation showing the relative localisation of Crhbp. The experiments were performed by collaborator Dr. Mary Green. Left panel: in situ hybridisation of Crhbp in the midbrain (the Caudal slice). Right panel: antibody staining of TH protein in the same slice. Dotted line represents the mid line between the left and right half of the brain. The Crhbp+ cells are observed in two distinct anatomical locations: (i) the ventral midline region and (ii) scattered within the VTA.

Heterogeneous expression of ephrins and ephrin receptors among 110 dopaminergic neurons

Ephrin receptor and ephrin ligand families of proteins form a complex cell-to-cell signalling system. Ephrin receptor is the largest subfamily of receptor tyrosine kinases. This signalling system is unique receptor tyrosine kinases system in the following aspects: (i) upon ephrin ligand binding onto ephrin receptor, both the cell expressing ephrin receptors receives a signal (forward signalling) and the cell expressing ephrin ligands receives a signal (reverse signalling) (ii) it acts via direct cell-to-cell contact rather than soluble ligands; (iii) the interaction involves either cleavage of the membrane bound ligand-receptor,

or endocytosis of the entire interacting ligand-receptor complex; (iv) they require higher-order clustering of the ligand-receptor complex to be effective; (v) *cis* interactions of ephrin ligands and receptors co-expressed in the same cell strongly influence the sensitivity of signalling; and (vi) varying concentrations of ligand and receptor can significantly alter the signalling outcome. These special features make it an ideal system for axon path-finding for developing neurons. I hypothesize that varying concentration of different ephrin ligands and receptors is a key component instructing midbrain dopaminergic neurons to project to other brain regions correctly. As a result, I investigated the expression levels of ephrin ligands and receptors across different classes of midbrain dopaminergic neurons. Other proteins participating in the ephrin signalling pathway were retrieved from Reactome [204]. The expression is shown in Figure 4-32 A. We can see the expression of *Epha4, Epha7* and *Nck2* are highly correlated. Furthermore, they are expressed in almost all dopaminergic neurons except the 21 VTA dopaminergic neurons. The expression of *Efna5* is highly anti-correlated with *Epha4, Epha7* and *Nck2*, suggesting divergent axon projections can potentially be configured by these ephrin genes *in vivo*. Furthermore, we can see that the expression of *Efnb3* is depleted from SNc dopaminergic neurons. Although the expression of *Efnb3* is weakly similar to *Efna5*, there is much more heterogeneity observed for the expression of *Efna5*. I also examined the expression of genes involved in the semaphorin pathway as shown in Figure 4-32 B. We can see that the divergence between SNc and VTA can easily be observed among this semaphorin signalling pathway, suggesting that these pathways are used in concert during dopaminergic neuron development at postnatal Day 4.

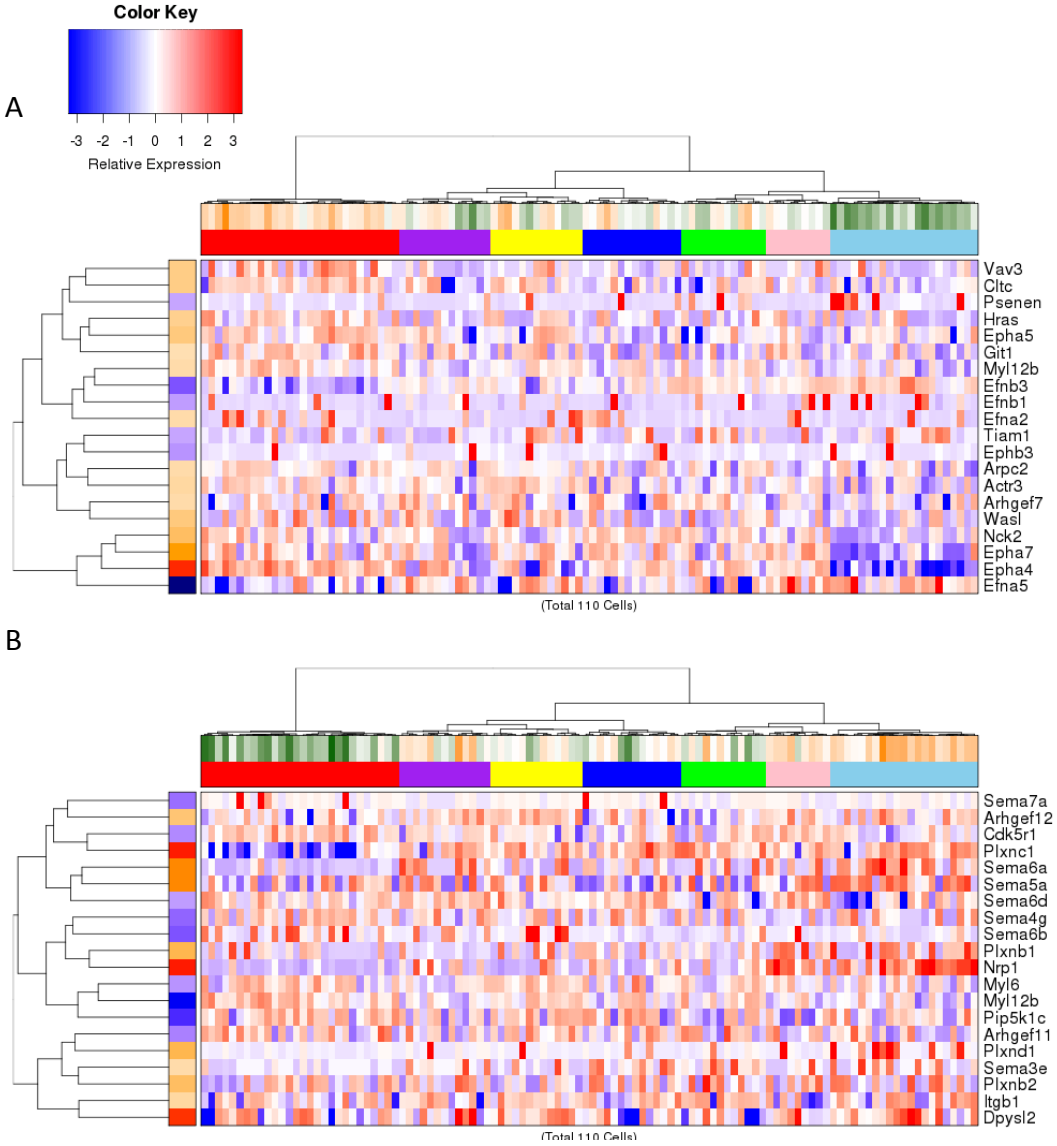


Figure 4-32: showing the expression of various genes that have been implicated in axon guidance. Panel A: Heatmap showing the expression of ephrin ligands, receptors and other genes that participate in this pathway. Panel B: Heatmap showing the expression of genes involved in the semaphorin pathway.

Above are several examples illustrating how cell-to-cell contact and other signalling events are used by midbrain dopaminergic neurons to construct their complex axon projection patterns. There are several other pathways, such as Ncam, Netrin, Robo, L1cam, and Ret signalling pathways [204], that may play important roles in constructing axonal projections during dopaminergic neuron development. These proteins are highly indicative for the projection patterns and cellular location of a neuron because they are essential contributors to the complex brain structure that formed during development. Although, I hypothesize that the expression of these proteins is not necessarily subpopulation specific. This is because different subpopulations can potentially project to the same brain regions and thus could employ similar axon guidance cues; and the same subpopulation can potentially project to different brain regions and thus could exhibit different axon guidance cues.

4.5 Discussion

Rationale and importance of this research

This study profiled almost 500 single cells at postnatal Day 4 in the mouse midbrain, where midbrain dopaminergic neurons are enriched by sorting the GFP protein levels in a mouse model where GFP protein level is correlated with Pitx-3 protein level [3]. This dataset is thus far the only dataset that has profiled the transcriptomes of hundreds of midbrain dopaminergic neurons. Thus it is the first comprehensive transcriptomic description of midbrain dopaminergic neurons at single cell resolution. I argue that this dataset is potentially able to reveal the different classes of midbrain dopaminergic neurons, surpassing previous dissection-based molecular profiling of different classes of midbrain dopaminergic neurons. Dissection-based approaches can be cross-contaminated due to potential inaccuracies during dissection. Worse, different classes of midbrain dopaminergic neurons might be evenly distributed and intermingled with each other, and such diversity could never be revealed purely by dissection.

An accurate and finely-resolved molecular description of midbrain dopaminergic neurons is an important step towards a more thorough understanding of these neurons. This includes a molecular description of the diversity of midbrain dopaminergic neurons, their respective markers, functions and vulnerabilities. Our study provided a molecular map at unprecedented single cell resolution. Using this dataset, we identified multiple novel classes of midbrain dopaminergic neurons and their respective marker genes. We validated these cell types and marker genes using experimental approaches. Given that

midbrain dopaminergic neurons control a large number of complex human behaviours such as movement control, risk taking, addiction etc, our results could greatly enhance the understanding of the complex roles of midbrain dopaminergic neurons. Our results may also assist us to understand the pathology in Parkinson's disease patients conveyed by subpopulations of midbrain dopaminergic neurons. By providing the molecular catalogue of the different classes and their respective marker genes, our studies could provoke a plethora of future experiments that target different classes of midbrain dopaminergic neurons.

Four validated subpopulations of dopaminergic neurons configured by two orthogonal parameters

In Figure 4-33 A, we can see that the 28 SNc dopaminergic neurons (cells marked by the red horizontal bar) were collectively marked by high expression of *Sox6, Pou3f2* and *Sncg*, and are depleted of *Otx2, Lpl* and *Grp*. On the other hand, the 21 VTA dopaminergic neuron (cells marked by the sky-blue horizontal bar) were exactly the opposite. In Figure 4-33 B, we can see that among the 28 SNc dopaminergic neurons, there are two major subpopulations. One subpopulation of SNc dopaminergic neurons is marked by the expression of *Aldh1a1* and *Lmo3*, the other subpopulation of SNc dopaminergic neurons does not express these two genes. It is interesting, or perhaps unsurprising, that In Figure 4-33 C, we can see that the major two subpopulations among the 21 VTA dopaminergic neurons are also divergent in their expression of *Aldh1a1* and *Lmo3.* The four classes of dopaminergic neurons mentioned above are determined by a 2 by 2 (2^2=4)

matrix as shown in Figure 4-34. I interpret the underlying configuration being laid out by two parameters that are orthogonal, where the row in Figure 4-33 represent one parameter (*Aldh1a1*+ vs *Aldh1a1*-) and the column in Figure 4-34 represent a second parameter (VTA vs SNc). Two parameters can configure four identities, and three parameters can configure eight identities. The exponential growth potentially empowers the gene regulatory circuitry to configure complexity that might be required in the midbrain. In fact, this is the most effective way to implement tremendous diversity. I think that if we had more cells to identify further subpopulations, it is possible to identify one additional orthogonal parameter in addition to the aforementioned two.

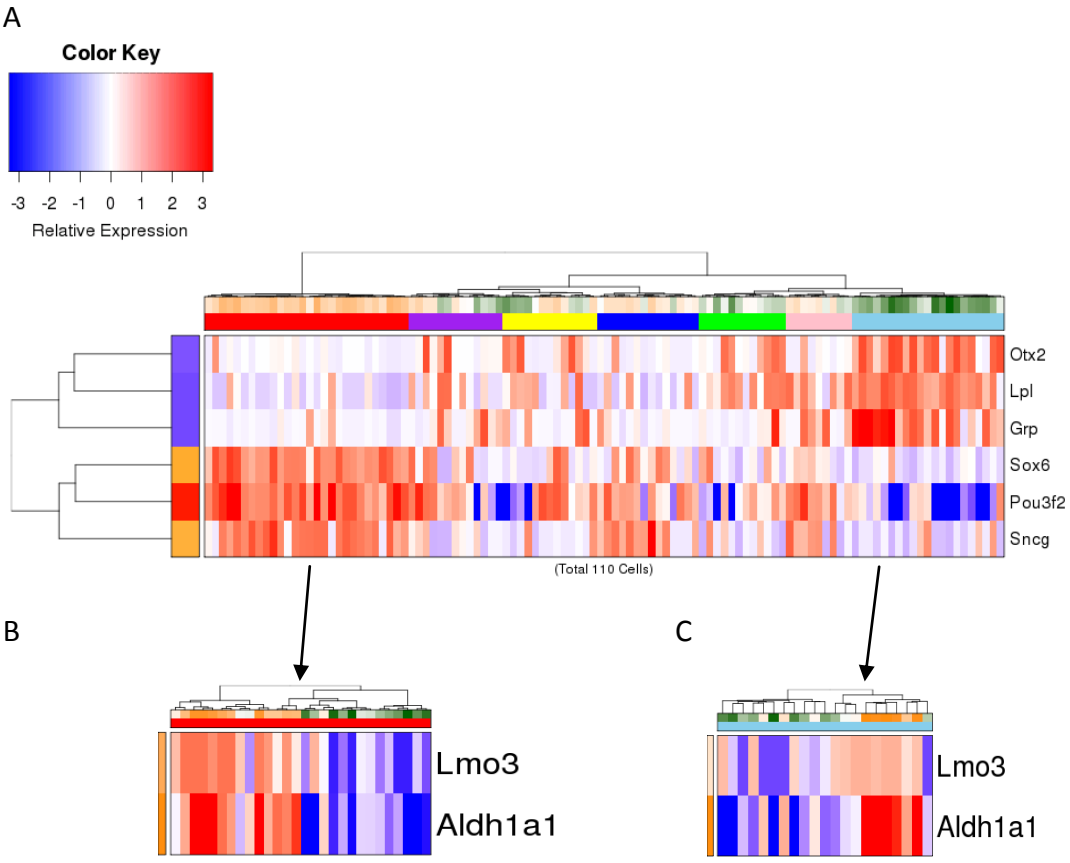


Figure 4-33: Lmo3 and Aldh1a1 defines two subpopulation of cells in both the VTA region, and the SNc region. Upper Panel: the expression of SNc marker genes: *Sox6, Pou3f2* and *Sncg* is higher among the 28 SNc dopaminergic neurons indicated by the red horizontal bar. The expression of VTA marker genes: *Otx2, Lpl* and *Grp* is higher among the 21 VTA dopaminergic neurons indicated by the sky-blue horizontal bar. Lower panels, the 28 SNc dopaminergic neurons are further divided into two major subpopulations by the expression of *Aldh1a1* and *Lmo3*; the 21 VTA dopaminergic neurons are further divided into two major subpopulations by the expression of *Aldh1a1* and *Lmo3*.

	SNc	VTA	Total
Aldh1a1+	14 cells	10 cells	24 *Aldh1a1*+ cells
Aldh1a1-	14 cells	11 cells	25 *ALDH1A1*- cells
Total	28 SNc cells	21 VTA cells	

Figure 4-34: Two orthogonal parameters configures four classes of dopaminergic neurons. One parameter being the SNc: VTA axis; the second parameter being the *Aldh1a1*+: *Aldh1a1*- axis.

Subpopulations of dopaminergic neurons that can potentially co-transmit glutamate and GABA

Dopaminergic neurons have been shown to release glutamate as a second neurotransmitter [205, 206]. Early studies showed that dopaminergic neurons express *Slc17a6* (*Vglut2*), and are lacking in *Slc17a7* (*Vglut1*) or *Slc17a8* (*Vglut3*) expression *in vitro* and *in vivo* [207, 208]. Studies have shown that *Slc17a6* is more broadly expressed among dopaminergic neurons during early development, and becomes more restricted in the adult [199, 209]. Our data generally supports this phenomenon. We can see from Figure 4-35 that *Slc17a6* (*Vglut2*) is heterogeneously expressed among mature midbrain dopaminergic neurons. The gradual restriction of *Slc17a6* expression is more obvious when its expression in mature dopaminergic neurons is compared with in immature neurons. Furthermore, *Slc17a7* (*Vglut1*) is not expressed in any cell, supporting previous observations [207, 208]. However, in contrast to the previously reported absence of *Slc17a8* (*Vglut3*), our data shows that *Slc17a8* (*Vglut3*) is expressed in almost every midbrain dopaminergic neuron in our study obtained at postnatal Day 4.

This incongruent observation of *Slc17a8* (*Vglut3*) could be due to the time-point of our study which differs from those of previous studies. There are, however, some doubts over whether these cells co-expressing *Th* and *Slc7a6* mRNA are *bona fide* dopaminergic neurons, because it was shown that the synthesis of the TH protein is suppressed in these neurons [209].

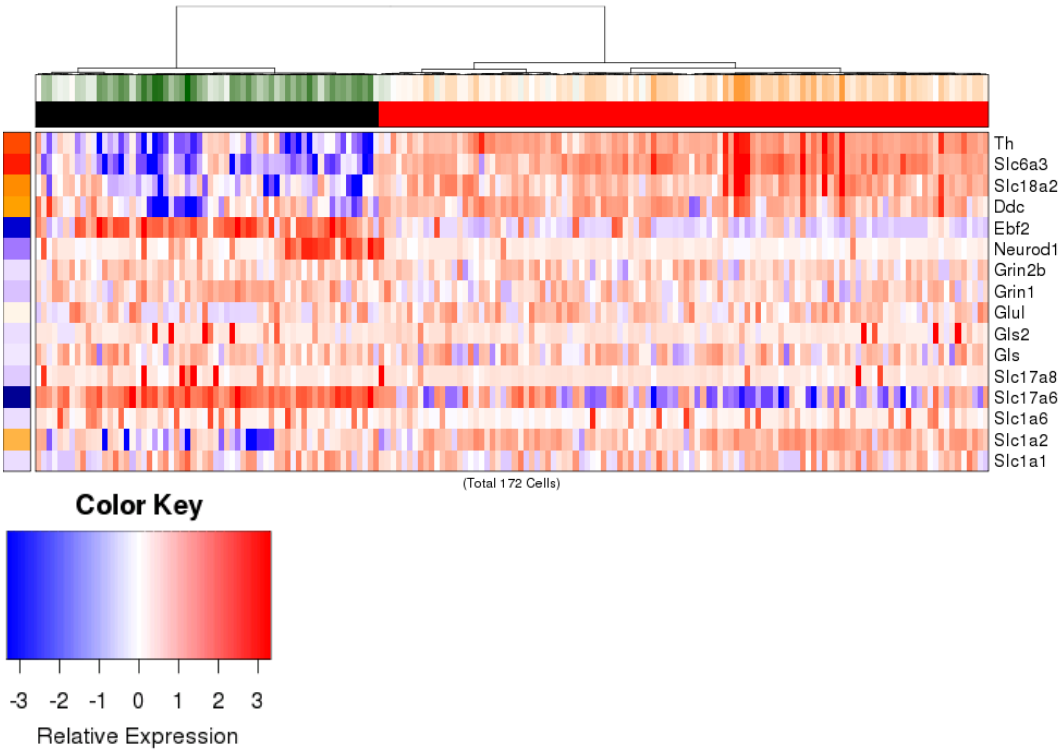


Figure 4-35: Heatmap showing the expression of several known markers. Dopaminergic neuron markers: *Th, Slc6a3, Slc18a2* and *Ddc*. Immature neuron markers: *Ebf2, Neurod1*. Essential subunits of glutamate receptor: *Grin2b, Grin1*. Genes for glutamate synthesise *Gls* (k-glutaminase) and *Gls2* (l-glutaminase). Genes for glutamine synthesise *Glul*. Genes for glutamate transport *Slc17a8* (*Vglut3*), *Slc17a6* (*Vglut2*), *Slc1a1* (Excitatory Amino Acid Carrier 1), *Slc1a2* (Excitatory Amino Acid Carrier 2) and *Slc1a6* (Excitatory Amino Acid Carrier 4).

Not only do they co-transmit glutamate, it has also been shown that midbrain dopaminergic neurons also co-transmit GABA. For example,

nigrostriatal midbrain dopaminergic neurons were shown to release GABA independent of *Slc32a1* (*Vgat;* GABA Vesicular Transporter). Instead, the GABA release of these cells was shown to be dependent on the presence of *Slc18a2* (*Vmat2*; Vesicular Monoamine Transporter) [210]. Furthermore, surprisingly, the midbrain dopaminergic neurons that co-release GABA do not express enzymes that synthesize GABA (i.e. *Gad1, Gad2*) [211]. The authors showed that these midbrain dopaminergic neurons recycle GABA from their target sites using *Slc6a1* (GABA Transporter 1) and *Slc6a11* (GABA Transporter 2) rather than synthesizing GABA *de novo* [211]. I examined the expression patterns of the genes that are important for canonical GABA synthesis (*Gad1, Gad2*), transport (*Slc6a1* and *Slc6a11*) and release (*Slc32a1*), as well as non-canonical release (*Slc18a2*), as shown in Figure 4-36. We can see that our data accord perfectly with the observation that midbrain dopaminergic neurons do not express the canonical GABA synthesis enzymes. Instead, they express *Slc6a11* for GABA re-uptake from their surroundings and express *Slc18a2* for GABA non-canonical release. Furthermore, It was recently shown that dopaminergic neurons can also synthesize GABA using an alternative pathway different from the canonical GABA synthesis pathway (*Gad1* and *Gad2*), instead, using *Aldh1a1* [212]. Although the properties of these alternative pathways are not well understood yet, It is possible that dopaminergic neurons operates a variety of GABA synthesis, re-uptake and release pathways that are different from that of canonical GABAergic neurons.

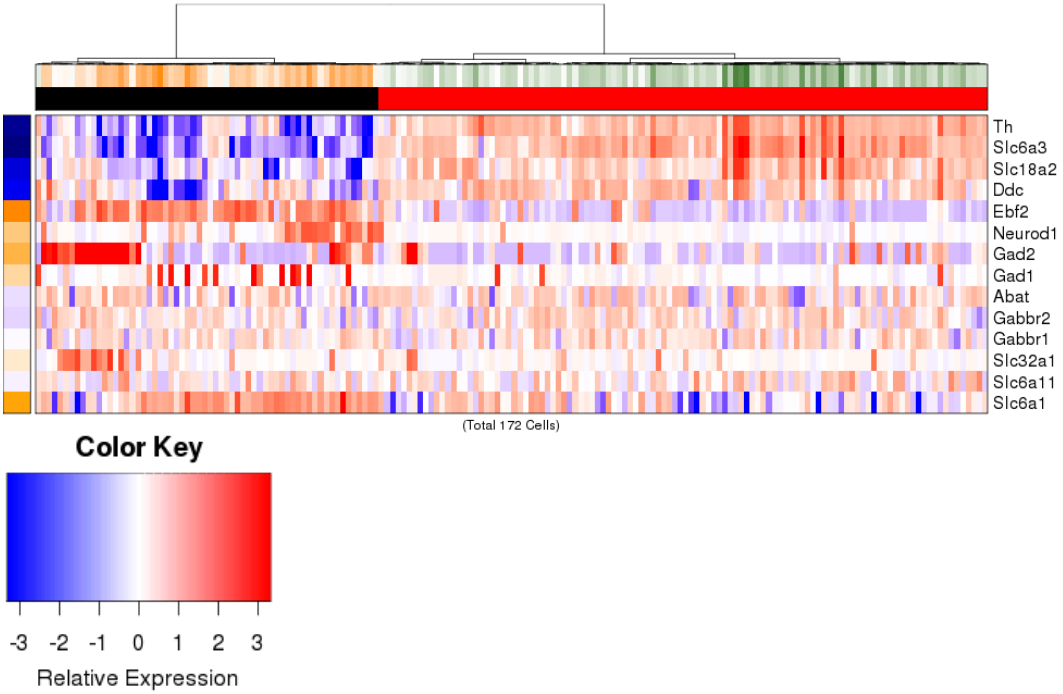


Figure 4-36: Heatmap showing the expression of several known markers. Dopaminergic neuron markers: *Th, Slc6a3, Slc18a2* and *Ddc*. Immature neuron markers: *Ebf2, Neurod1*. The GABA synthesis enzymes: *Gad1, Gad2*. GABA transaminase: *Abat*. GABA receptor: *Gabbr1, Gabbr2*. Genes essential for GABA re-uptake: *Slc6a1, Slc6a11*. And the gene essential for the non-canonical release of GABA: *Slc18a2* [210].

Overall, our data based on mRNA expression levels supports previous observations that some dopaminergic neurons can be capable of co-transmitting dopamine together with GABA, and some other dopaminergic neurons can be capable of co-transmitting dopamine together with glutamate.

The *Crhbp+* class of neurons are potentially capable of co-transmitting dopamine and GABA

The *Crhbp+*&*Gad2+* neurons identified from the group of 62 remaining GFP+ cells were uncharacterized to the best of my knowledge. Surprisingly, they weakly express genes that are essential constituents of both dopaminergic

neurons and GABAergic neurons as was seen in Figure 4-29. Our experiments using FISH revealed that the *Crhbp+*&*Gad2+* neurons can be found at two different locations of the midbrain: (i) scattered within the VTA; and (ii) within the rostral linear nucleus of Raphe (i.e. at the ventral midline adjacent to the VTA). The function of this group of neurons is unknown. With their specific markers identified and validated in this study, further studies could be designed, for example to verify its ability to transmit dopamine and GABA, to investigate their electro-physiological function using patch clamp recording, or to perform opto-genetic studies to study their function by *in situ* perturbations.

The benefits of entirely data-driven approaches for analysing single cell RNA-Seq dataset

Importantly, we did not rely on existing knowledge for the *a priori* classification of midbrain dopaminergic neurons. In other words, the classification analysis was entirely data-driven and the dataset itself informed us about the various classes of dopaminergic neurons that are present in the experiment. This approach has the benefit of not being affected by previous cell type classifications or by existing knowledge on midbrain dopaminergic neurons, obtained at potentially more inferior resolutions. In our approach, after the classification was completed, we use existing literature only to ensure that we name and interpret the putative cell types in a way that is consistent with current nomenclature.

The superior resolution when identifying subpopulations of midbrain dopaminergic neurons and the small number of cells required to identify their marker genes

It is important to note that the previously identified subpopulations of midbrain dopaminergic neurons and their markers could be recapitulated in our single cell RNA-Seq by my data-driven analyses. Our study, however, provide a superior resolution relative to existing studies in that we are no longer relying on anatomically defined subpopulations and instead are moving towards molecularly defined subpopulation which is of much higher resolution in delineating subpopulations. Although single cell qRT-PCR based methods have been successfully applied to interrogate subpopulations of midbrain dopaminergic neurons [3], these methods fail to detect the subpopulations that did not have a known marker gene.

It is also important to note that the number of cells present in our dataset for the experimentally validated classes of midbrain neurons is as small as 10-17 cells, suggesting that for successfully identifying a class of midbrain dopaminergic neurons in a single cell RNA-Seq dataset, only this small number of neurons of this class are required to be represented in the total captured population. My results provide important information regarding the number of cells that are required to be captured and sequenced in order to correctly identify the cell type, and to identify their respective markers.

Limitation of the clustering based classification method

In order to classify all midbrain neurons profiled in the experiment, we have undertaken an approach of hierarchical clustering followed by cutting the

dendrogram. I needed to subjectively select a input parameter during the dendrogram cutting step: the height at which to cut the dendrogram. It is clear that the number of classes is directly dependent on the choice of this parameter. Currently, there exists no mathematically ideal method by which to decide on this parameter elegantly and optimally. As a result, this input parameter is subjectively decided, to yield a reasonable and manageable number of clusters, and to ensure the clusters are biologically relevant. Technically speaking, it is correct to classify all midbrain dopaminergic neurons as a single class, and it is equally correct to classify every single midbrain dopaminergic neuron as one separate class because every neuron's transcriptome is different. The fundamental question of how many classes exist is a theoretical exercise relates to a different research domain, thus does not lie within the scope of this thesis. In the scope of this thesis, we approach the challenges of clustering from a biological perspective rather than from a theoretical perspective of defining clusters optimally. We thus focus on examining whether different cells in the same cluster share transcriptomic similarity, anatomical locations, and other important physiological properties, rather than the theoretically optimality of deciding the number of clusters.

Several methods have been proposed to circumvent the subjective selection of the cutting height. One popular method uses an approach to mimic statistical bootstrap, using a method titled "clustering with p-values" [213]. It is important to highlight that this approach is not actually a bootstrapping procedure, and thus does not carry any of the statistical interpretation more

usually associated with bootstrapping. Furthermore, this approach did not remove the requirement for subjectively selecting a cutting parameter: this approach requires as input a subjectively selected threshold to determine what level of repetitive discovery is considered as being acceptable. This approach merely shifts the subjective choice of the cutting height parameter to the subjective choice of repetitive discovery parameter. A false sense of statistic foundation was caused by the misleading title "clustering with p-values" of this approach, although there is no underlying statistic foundation to support this title. Another group of such methods, represented by Bayesian Information Criteria, Akaike Information Criteria etc, alters the likelihood score by adding a weighted term to penalise a large number of non-zero parameters being used at the selected model. This group of approaches suffers the same disadvantages of subjectively assigning certain weight to the penalty of each additional parameter. If the weight is too low, the model selection will favour the larger number of parameters (i.e. equivalent to cut into more clusters), and *vice versa*. This again shifts the subjective choice of the cutting height to the subjective choice of the penalty weight. My understanding is that these approaches fail to resolve the subjective choice of parameters rigorously. Instead, mathematical proof has shown that, under currently accepted definitions of clustering, theoretically rigorous clustering is an impossible task [214]. As a result, before mathematical studies is able to rigorously resolve the fundamental definition of clustering, a rigorous mathematical approach to split the clusters would not exist. Despite that, pragmatic approaches are desirable for identifying groups of cells that share a set of physiological properties, for example, groups of cells that exhibit

some common electrophysiological properties, that are located at a specific anatomical region, or that regulates several genes in a coordinated manner. As a result, we have taken the pragmatic approaches that first identify putative clusters, then identify genes that are specifically expressed among the putative clusters, and finally investigate the anatomical location and other biological properties of these putative clusters.

Limitation of the pseudo-random capture process

The dissected midbrain tissue was dissociated in solution. The dissociated cells were then captured by the Fluidigm C1 system. During each Fluidigm C1 experiment, the system pseudo-randomly captures up to 96 cells out of approximately 2000 cells, and discards all remaining cells. The large fraction of cells wasted during this step dictates that, in this experiment, we would not be able to obtain all midbrain dopaminergic neurons of a mouse. The set of cells that contributed to the single cell RNA-Seq dataset thus were a small subset pseudo-randomly captured. This pseudo-random capture step has three major drawbacks. First, using the Fluidigm C1 system, we will never be able to assess the total number of midbrain dopaminergic neurons in a mouse. Second, due to the same incomplete capturing process, the classes of dopaminergic neurons that we are able to investigate in this study will always be a subset of all classes present in a mouse midbrain, mostly the classes that are relatively abundant. As a result, classes of rare midbrain dopaminergic neurons may not be investigated in our experiment. For example, the *Vip*+ subpopulation of dopaminergic neurons discovered in [3] is not discovered by our study. Thirdly, the capture

step can be biased towards cells with certain properties, for example, cells within a certain size range, or cells with a certain surface property. In summary, the cell capture step does not entail the nice statistical properties of a simple random sampling. As a result, the relative abundance of the classes captured in a study cannot be considered to be an accurate representation of the relative abundance of the different classes of cells present in the mouse midbrain.

Limitation of molecular information based classification

There are many ways to classify neurons: based on their morphology, ion channels, neural transmitters, synaptic connectivity, spiking frequency, number of dendrites, lengths of axon or sizes of cell body. In this study, we chose to use molecular information as the only quantity of measurement, more specifically, the abundance of messenger RNA. This choice is for convenience and is by no means a perfect choice. A consequence is that our classification does not take any other aspect of neurons' properties into consideration. It is possible that many of the properties of a neuronal function rest entirely in its anatomy, connectivity or morphology, which will not be measurable in an RNA-Seq study. As a result, a substantial amount of follow-up experimental work remains required.

4.6 Materials and Methods

D1 Single cell RNA-Seq assays

Cells were captured, lysed, and full length cDNA libraries constructed individually using the microfluidic C1 Auto Prep System (Fluidigm) and the SMART-Seq protocol (Clontech) as previously described in [147]. Illumina HiSeq sequencing libraries were generated using the Nextera XT kit (Illumina). Excluding the wells marked as debris or as empty, there were 492 cells in total being sequenced. Each library was barcoded individually before being multiplexed together for sequencing using the paired-end 50 base pair (bp) format on HiSeq2000 (Illumina) and de-multiplexed and base-called using the software CASAVA provided by Illumina.

D2 RNA-Seq reads mapping and quality filtering on each cell

50 bp non-stranded paired-end reads were mapped using GSNAP [148] to the mouse genome assembly (mm10) using the Ensembl75 gene annotation set [149]. In total, 1.3 x 10^9 reads were successfully mapped to the mouse genome assembly mm10.

Neurons are highly entangled with each other with a complex layout of axons and dendrites. The dissociation step inevitably tears off most of the axons and dendrites which are damaging to the cells. We identified failed cDNA libraries as having very low fractions of exonic sequences, very few expressed genes, or very few successfully mapped reads. These libraries were most likely generated from apoptosed cells, cells in a bad overall condition, or cellular debris

being captured by the C1 microfluidic device, where the starting mRNA is either highly degraded or absent. To prevent low quality libraries affecting the analysis and conclusions, cells were disregarded if they failed one or more of the following three quality filters. (i) Among all mapped reads, fewer than 5% are mapped to annotated exons. (ii) Total number of expressed genes <6000. (iii) Total number of mapped reads < 1 x 10^6. Of 492 cells in total, 172 cells passed the quality filtering and were retained for subsequent analyses.

D3 Expression estimation for genes

Gene expression abundance estimates were obtained using cufflinks version 2.0.2 [150]. The following command line options were used to obtain the abundance (FPKM): cufflinks --label reference –GTF reference.gtf --num-threads 25 --frag-bias-correct mm10.fa --library-type fr-unstranded --compatible-hits-norm --multi-read-correct library_name.bam >& reference_library_name.cufflinks. TPM values, if required, were obtained as previously described in [151]. Normalized read counts, if required, were obtained by normalization against the effect of library sizes using the estimateSizeFactors sub function in DESeq Version 1.12.0 [152].

D4 Assess whether any batch failed completely

Each Fluidigm C1 run only captures up to 96 cells. As a result, in order to collect about 500 cells, the experiment requires several Fluidigm C1 runs. Due to the difficulty of obtaining sufficient dissociated midbrain dopaminergic neurons

from the tiny brain region dissected, about 10 mice were sacrificed for each Fluidigm C1 run. Thus, for logistical reasons, the dataset was generated across nine separate Fluidigm C1 runs. We first investigated the number of genes having FPKM>1 per cell in each batch, in order to assess whether any of the 9 batches failed completely. From Figure 4-37, we can see that all 9 batches included several cells with a few thousand genes having FPKM>1, and batches are not visually too different from other batches. Consequently, we considered that none of the 9 batches failed completely.

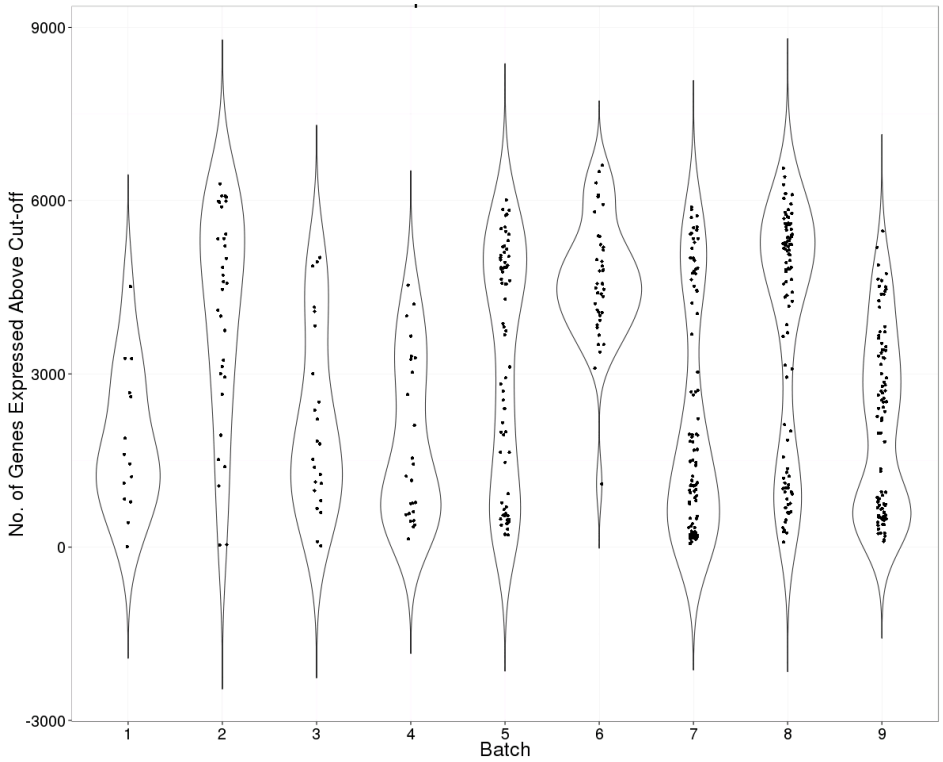


Figure 4-37: Number of genes detected in each cell. Each dot represents a cell. The Y axis of the dot represents the number of genes detected in the cell at a threshold of FPKM>1. The X axis of the dots represent the batch number that the cells were sequenced in. Every batch contained several failed cells (few genes having FPKM>1) , and several successful cells (many genes having FPKM>1). This is a common phenomenon in single cell RNA-Seq dataset generated using the Fluidigm C1 system.

D5 Clustering of cells *a priori*

There is existing literature describing different classes of midbrain dopaminergic neurons and their respective markers [194]. However, making use of such information could potentially bias our findings towards previously described classes or towards previously applied classification criteria. For example, if we were to limit the clustering analysis to the marker genes previously discovered by contrasting SNc and VTA regions of the midbrain [176], we would likely only discover the differences of molecular signatures between SNc and VTA, but not other novel cell types. In my analysis, I sought to investigate the classes of midbrain dopaminergic neurons independent of existing literature and to allow our single cell RNA-Seq data to dictate the classification results. To achieve this, firstly, between each pair of cells, I used all genes to compute the pair wise correlation coefficient [153]. Subsequently, I applied square root (1 – pair wise correlation coefficient) to convert correlation into a distance measure, and obtained a square distance matrix. With this, I was able to perform hierarchical clustering using the hclust function in R on all cells and to obtain a clustering dendrogram

This experiment sought to discover different class of midbrain dopaminergic neurons. Thus, after obtaining the clustering dendrogram, I split the dendrogram to separate cells into multiple classes. I then repeated this procedure iteratively to split the resultant clusters yet further. Specifically, in each iteration, I split the clustering dendrogram into two to three resultant clusters depending on relevant biological context and putative cell types in

question. I then repeated the same approach for the resultant clusters separately to further discover cell types within each resultant cluster that are populated with many cells.

D6 Differential expression analysis

Differential expression analysis between any two clusters was performed using SCDE [153] on non-zero genes. A gene is deemed to be differentially expressed if it satisfies the following criteria: (i) FDR<20%, (ii) Fold Change < 1/4 or Fold Change >4 across the two different groups of cells, and (iii) The maximum likelihood estimate of the expression in the up-regulated cell types is FPKM > 50. The same cut-off criteria were used throughout the study, except when this threshold returned fewer than 5 differentially expressed genes.

D7 Iterative splitting strategy and highly variable gene strategy to identify novel cell groups and their respective markers

Most strategies for identifying novel cell types by clustering require an input parameter to determine the number of clusters. Determining this input parameter is an issue of philosophical debate and is not within the scope of the thesis as explained in Chapter 2 Section 10. However, for the purpose of performing biological investigations, it is not helpful to argue that it is theoretically correct to partition cells into any number of clusters. Biological investigations seek to discover common characteristics of cells that are recurrent and performs a relevant biological function. As a result, there is a need to identify putative novel cell types that are recurrent, and their putative cell type-

specific markers. In this regard, I propose two different potential strategies described below.

The first strategy contains four step. (Step 1) For n in (2 ... 9), partition all cells into n groups, named as X_1 ... X_n. (Step 2) For X_i in (X_1 ... X_n), perform a differential expression test for all cells in the X_i group against all cells not in the X_i group to identify genes that are more highly expressed among cells in the X_i group. (Step 3) Repeat the same procedure for all possible choices of n and X_i and concatenate the list of genes that are being identified. This strategy will enable the concatenated list of genes to contain genes that are specifically expressed in both large groups of cells, and in smaller groups of cells. (Step 4) Cluster the cells and cells that have similar expression patterns using the previously concatenated list of genes and plot the heatmap to examine whether the cells and genes form recurrent expression patterns.

The second strategy seek to identify all genes that are heterogeneously expressed among a population of cells using the method described in [46]. Followed by clustering all cells using this group of highly variable genes. This has the benefit of not relying on any input clustering results because the input clustering results are not necessary correct. After ranking all genes according to its variability among all cells, the most variable genes (usually the top 100 genes) are selected. Clustering was then performed based on the selected genes to identify potential novel cell types.

Chapter 5 Conclusion

In this final chapter, I briefly summarize the main findings of each research project. This is followed by a summary of the power of single cell RNA-Seq data analyses that was demonstrated both by my projects and by other published single cell RNA-Seq data analysis projects. Finally, I will point out several interesting directions of single cell RNA-Seq data analyses that require further research.

5.1 Summary of the results presented in my thesis

In Chapter 2, through literature review and investigating the underlying principles of high throughput sequencing, I argued that the analyses of high throughput sequencing datasets are complex and require many special considerations. I also argued that there are many caveats when interpreting results obtained from high throughput sequencing experiments. RNA-Seq and single cell RNA-Seq experiments are one kind of high throughput sequencing experiments that interrogate the transcriptome. I argued that these are even more challenging to perform experimentally, and their associated datasets can be more challenging to analyze and interpret. Currently the Fluidigm C1 system is a popular commercial platform for performing single cell RNA-Seq experiments worldwide. A large number of published research projects have been based on this platform. Based on datasets generated from this platform, I highlighted the following major challenges associated with single cell RNA-Seq data analyses. Firstly, the high dimensionality (i.e. number of genes) and high number of data points (i.e. number of cells) and the severe degree of technical noise together

demand appropriate dimensionality reduction to be performed before much insight from single cell RNA-Seq datasets can be obtained. Secondly, expression values in single cell RNA-Seq datasets often contained many zero entries, even for transcripts that were known to be present in the sample. This phenomenon is called “dropout”. This required special statistical methods to take this into consideration during the modelling of single cell RNA-Seq datasets. Thirdly, many cell capture units may not capture a healthy cell to generate useful single cell RNA-Seq library. Thus there is a strong need to identify and discard data from these capture units prior to downstream analyses. This can be done both experimentally using microscopic examination, and bioinformatically using a previously published method. Fourthly, using the current single cell capturing technologies available, a capture unit is not guaranteed to capture one cell only. There are non-negligible probabilities that some capture units may contain more than one cell. This can be a major concern if downstream analysis depends on the assumption that every capture unit contains only one cell. Currently, in most single cell RNA-Seq papers published to date, this issue has not been addressed. The only amelioration I am aware of is that if an independent measurement can be used to determine the prevalence of doublets is not too high, one can claim that the doublets may not distort results significantly. Sixthly, cell dissociation and capture can be biased for certain cell types. As a result, the number of cells from each cell type being captured shall not be interpreted as a faithful representation of the number of cells in the tissue. Seventhly, many rounds of PCR amplification are required by most single cell RNA-Seq protocols. As a result, expression quantification is significantly distorted by biases during PCR

amplification. Experimental protocols that employ UMI will allow downstream bioinformatics analyses to identify PCR duplicates and correct against this type of bias.

In Chapter 3, our collaborators designed an experimental protocol to induce human embryonic stem cells to form a multi-cellular structure (i.e. conceptoid bodies) that contains different cell types. After the multi-cellular structure was induced, single cell RNA-Seq was performed on them. I performed bioinformatics analyses to investigate whether multiple cell types were present in the induced structure. As explained in Section 1.7 of the thesis, single cell RNA-Seq is particularly suited for the identification of cell types present in a random mixture of cells when no prior information or cell type specific markers are known. Applying *a priori* cell type identification strategies, I identified two cell types are present in conceptoid bodies, and their marker genes. My identified markers of the two cell types in conceptoid bodies were validated using protein-level experimental approaches. Furthermore, based on the similarities between their transcriptomes, I matched the two cell types in the conceptoid body with the two cell types (ICM and trophoblast) in Day 6 pre-hatching human blastocysts. Doing so allowed us to assign the putative *in vivo* identities of the two cell types in conceptoid bodies. The first major finding of this study was that human ESC induction protocol was successful in creating a conceptoid body that resembles the human blastocyst and contains the two cell types present in a human blastocyst. The second major finding of this study was that human ESC is able to form trophoblast under the treatment of BMP/SU. This suggest that after

being treated with appropriate stimuli and under an appropriate environment, human ESCs can become cells that are similar to ICM cells, which are totipotent. More importantly, the methods I designed identifying cell types *a priori* followed by identifying their respective markers were successful and are widely applicable.

In Chapter 4, we set out to investigate whether there are different classes of dopaminergic neurons present in the mouse midbrain. Few studies exist that have performed an unbiased survey of different classes of dopaminergic neurons in mouse midbrain. As explained in Chapter 1 Section 7, single cell RNA-Seq was particularly well suited to identifying cell types *a priori* when little information on the novel cell types and their markers was available. Midbrain samples were dissected, dissociated from post-natal day 4 mice, and dopaminergic neurons were enriched from the dissociated cells by performing FACS on dopaminergic neurons markers. Individual cells were randomly captured and single cell RNA-Seq was performed. Again, I applied the strategy described in Chapter 1 Section 7 to *a priori* identify clusters of cells among the randomly captured cells that passed quality filter and identified two major clusters. Cells in one cluster do not express the core machinery of dopamine function, and thus may not be dopaminergic neurons; cells in the other cluster highly express genes encoding the core machinery of dopamine function, and thus are likely to be genuine dopaminergic neurons. I then iteratively applied the same *a priori* cell type identification method to the cluster of dopaminergic neurons and identified several putative classes of dopaminergic neurons and their respective markers.

Some markers were experimentally validated by our collaborators. The first major finding in this study was that dopaminergic neurons can be separated into two major classes using entirely unbiased method based on their transcriptomes: one class expresses known marker genes of VTA dopaminergic neurons and one class expresses known marker genes of SNc dopaminergic neurons. Although the two types of dopaminergic neurons were known before, our result is important. This is because contrary to relying on anatomical information and biased by previous belief, our study is the first unbiased study that identified these two major classes of midbrain dopaminergic neurons. The second major finding of our study is that we identified two subpopulations of SNc dopaminergic neurons and their respective marker genes. Their respective markers were shown to be expressed in different locations in the SNc region using RNA *in situ* hybridisation and protein staining. The third major finding of the study is that we also identified two subpopulations of dopaminergic neurons in the VTA region and their respective marker genes. These findings were confirmed by experimental methods in an independent study. More importantly, my analyses' methods successfully identified novel classes of cells, and their respective markers. Moreover, my methods have the power to iteratively zoom into major classes to reveal the more refined subclasses within.

5.2 Summary of the importance of my thesis

I argued that for answering many biological research questions, performing measurements on bulk samples can be inadequate. Even single cell measurements can be inadequate if the measurements are not –omics

measurement. For example, to discover novel cell types that cannot be purified. I discussed the benefits of single cell –omics experiments. Very recently, the first single cell omics experiment – single cell RNA-Seq – became technologically feasible and can be performed on hundreds to thousands of cells in parallel. The tremendous challenges of analyzing and interpreting single cell RNA-Seq data was discussed based on the existing literature. Despite the tremendous challenges, the application of single cell RNA-Seq has enabled many major discoveries that are demonstrated by published examples. Using a similar strategy, in the projects presented in this thesis, I demonstrated that single cell RNA-Seq can be applied to discover cell types in a mixture of cells for *in vitro* and match the cell types to the corresponding known cell types *in vivo*. Likewise, I demonstrated that single cell RNA-Seq can be applied to discover novel classes of cells for *in vivo* biological context, and can be applied iteratively to reveal more refined classes of cells. To sum up, published studies and my results collectively demonstrate that single cell –omics measurements are important for many biological discoveries.

5.3 Suggestions for future research directions

An ideal single cell –omics experiment seeks to measure all molecules within a single cell. Single cell RNA-Seq only fulfils part of this aim by measuring the transcriptome. There is much information not available in the transcriptome and it would be interesting to collectively measure several aspects of a single cell. Experimental protocols have been developed to measure the epigenome

and the transcriptome of the same single cell [215]. Furthermore, experimental protocols to measure the transcriptome and the abundance of several proteins in the same single cell are currently being developed (personal communications). Thus, it is potentially interesting to design analytical methods that are able to combine multiple data types (e.g. combine methylome and transcriptome) to perform integrated analyses.

Clustering single cells based on single cell RNA-Seq measurements is a task that was performed in all published single cell RNA-Seq papers to date. The method chosen is typically hierarchical clustering followed by tree-cutting at a subjectively chosen tree height. Despite that there are many subjective parameter choices that can alter clustering and tree cutting outcome, these approach has been successful to some extent. Further research is required to perform cell clustering based on more suitable approaches. There are already some tentative approaches that have been proposed to address this issue [216].

References

1. White, M.D., et al., *Long-Lived Binding of Sox2 to DNA Predicts Cell Fate in the Four-Cell Mouse Embryo.* Cell, 2016. **165**(1): p. 75-87.
2. Torres-Padilla, M.E. and I. Chambers, *Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage.* Development, 2014. **141**(11): p. 2173-81.
3. Poulin, J.F., et al., *Defining midbrain dopaminergic neuron diversity by single-cell gene expression profiling.* Cell Rep, 2014. **9**(3): p. 930-43.
4. Maier, T., M. Guell, and L. Serrano, *Correlation of mRNA and protein in complex biological samples.* FEBS Lett, 2009. **583**(24): p. 3966-73.
5. Roberts, A., et al., *Identification of novel transcripts in annotated genomes using RNA-Seq.* Bioinformatics, 2011. **27**(17): p. 2325-9.
6. Mironov, A.A., J.W. Fickett, and M.S. Gelfand, *Frequent alternative splicing of human genes.* Genome Res, 1999. **9**(12): p. 1288-93.
7. Li, B., et al., *Evaluation of de novo transcriptome assemblies from RNA-Seq data.* Genome Biol, 2014. **15**(12): p. 553.
8. Laumer, C.E., A. Hejnol, and G. Giribet, *Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation.* Elife, 2015. **4**.
9. Zhao, S., et al., *Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells.* PLoS One, 2014. **9**(1): p. e78644.
10. Wang, Y. and N.E. Navin, *Advances and applications of single-cell sequencing technologies.* Mol Cell, 2015. **58**(4): p. 598-609.
11. Macaulay, I.C. and T. Voet, *Single cell genomics: advances and future perspectives.* PLoS Genet, 2014. **10**(1): p. e1004126.
12. Lao, K.Q., et al., *mRNA-sequencing whole transcriptome analysis of a single cell on the SOLiD system.* J Biomol Tech, 2009. **20**(5): p. 266-71.
13. Tang, F., et al., *mRNA-Seq whole-transcriptome analysis of a single cell.* Nat Methods, 2009. **6**(5): p. 377-82.
14. Kurimoto, K., et al., *Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis.* Nat Protoc, 2007. **2**(3): p. 739-52.
15. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nat Protoc, 2014. **9**(1): p. 171-81.
16. Picelli, S., et al., *Smart-seq2 for sensitive full-length transcriptome profiling in single cells.* Nat Methods, 2013.
17. Adey, A., et al., *Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition.* Genome Biol, 2010. **11**(12): p. R119.
18. Araci, I.E. and S.R. Quake, *Microfluidic very large scale integration (mVLSI) with integrated micromechanical valves.* Lab Chip, 2012. **12**(16): p. 2803-6.
19. Klein, A.M., et al., *Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.* Cell, 2015. **161**(5): p. 1187-201.
20. Grun, D., et al., *Single-cell messenger RNA sequencing reveals rare intestinal cell types.* Nature, 2015. **525**(7568): p. 251-5.
21. Kolodziejczyk, A.A., et al., *Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation.* Cell Stem Cell, 2015. **17**(4): p. 471-85.
22. Scialdone, A., et al., *Computational assignment of cell-cycle stage from single-cell transcriptome data.* Methods, 2015. **85**: p. 54-61.

23. Buettner, F., et al., *Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.* Nat Biotechnol, 2015. **33**(2): p. 155-60.
24. Stubbington, M.J., et al., *T cell fate and clonality inference from single-cell transcriptomes.* Nat Methods, 2016. **13**(4): p. 329-32.
25. Kramer, D., *Discovering new cell types one at a time.* Nature Education, 2015. **1**(1): p. 1.
26. Ascoli, G.A., et al., *Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex.* Nat Rev Neurosci, 2008. **9**(7): p. 557-68.
27. Luo, L., E.M. Callaway, and K. Svoboda, *Genetic dissection of neural circuits.* Neuron, 2008. **57**(5): p. 634-60.
28. Masland, R.H., *The neuronal organization of the retina.* Neuron, 2012. **76**(2): p. 266-80.
29. Sanes, J.R. and R.H. Masland, *The types of retinal ganglion cells: current status and implications for neuronal classification.* Annu Rev Neurosci, 2015. **38**: p. 221-46.
30. Macosko, E.Z., et al., *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.* Cell, 2015. **161**(5): p. 1202-14.
31. Furusawa, T., et al., *Gene expression profiling of mouse embryonic stem cell subpopulations.* Biol Reprod, 2006. **75**(4): p. 555-61.
32. Singh, A.M. and N. Terada, *Bypassing heterogeneity: the road to embryonic stem cell-derived cardiomyocyte specification.* Trends Cardiovasc Med, 2007. **17**(3): p. 96-101.
33. Chambers, I., et al., *Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells.* Cell, 2003. **113**(5): p. 643-55.
34. Chambers, I., et al., *Nanog safeguards pluripotency and mediates germline development.* Nature, 2007. **450**(7173): p. 1230-4.
35. Singh, A.M., et al., *A heterogeneous expression pattern for Nanog in embryonic stem cells.* Stem Cells, 2007. **25**(10): p. 2534-42.
36. Macfarlan, T.S., et al., *Embryonic stem cell potency fluctuates with endogenous retrovirus activity.* Nature, 2012. **487**(7405): p. 57-63.
37. Faddah, D.A., et al., *Single-cell analysis reveals that expression of nanog is biallelic and equally variable as that of other pluripotency factors in mouse ESCs.* Cell Stem Cell, 2013. **13**(1): p. 23-9.
38. Kalmar, T., et al., *Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells.* PLoS Biol, 2009. **7**(7): p. e1000149.
39. MacArthur, B.D., et al., *Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity.* Nat Cell Biol, 2012. **14**(11): p. 1139-47.
40. van den Berg, D.L., et al., *Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression.* Mol Cell Biol, 2008. **28**(19): p. 5986-95.
41. Toyooka, Y., et al., *Identification and characterization of subpopulations in undifferentiated ES cell culture.* Development, 2008. **135**(5): p. 909-18.
42. Eiraku, M., et al., *Self-organized formation of polarized cortical tissues from ESCs and its active manipulation by extrinsic signals.* Cell Stem Cell, 2008. **3**(5): p. 519-32.
43. Kim, J.K. and J.C. Marioni, *Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data.* Genome Biol, 2013. **14**(1): p. R7.
44. Schmiedel, J.M., et al., *Gene expression. MicroRNA control of protein expression noise.* Science, 2015. **348**(6230): p. 128-32.

45. Smith, A., *Nanog heterogeneity: tilting at windmills?* Cell Stem Cell, 2013. **13**(1): p. 6-7.
46. Brennecke, P., et al., *Accounting for technical noise in single-cell RNA-seq experiments.* Nat Methods, 2013.
47. Baker, S.C., et al., *The External RNA Controls Consortium: a progress report.* Nat Methods, 2005. **2**(10): p. 731-4.
48. Munro, S.A., et al., *Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures.* Nat Commun, 2014. **5**: p. 5125.
49. McDavid, A., G. Finak, and R. Gottardo, *The contribution of cell cycle to heterogeneity in single-cell RNA-seq data.* Nat Biotechnol, 2016. **34**(6): p. 591-3.
50. Santos, A., R. Wernersson, and L.J. Jensen, *Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1140-4.
51. Kimmerling, R.J., et al., *A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages.* Nat Commun, 2016. **7**: p. 10220.
52. Ramskold, D., et al., *Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells.* Nat Biotechnol, 2012. **30**(8): p. 777-82.
53. Islam, S., et al., *Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq.* Genome Res, 2011. **21**(7): p. 1160-7.
54. Hashimshony, T., et al., *CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification.* Cell Rep, 2012. **2**(3): p. 666-73.
55. Jaitin, D.A., et al., *Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types.* Science, 2014. **343**(6172): p. 776-9.
56. Soumillon, M., et al., *Characterization of directed differentiation by high-throughput single-cell RNA-Seq.* bioRxiv, 2014: p. 003236.
57. Grun, D. and A. van Oudenaarden, *Design and Analysis of Single-Cell Sequencing Experiments.* Cell, 2015. **163**(4): p. 799-810.
58. Fluidigm, *Single cell RNA-Seq workflow.* Fluidigm, 2016(1): p. 1.
59. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
60. Church, D.M., et al., *Extending reference assembly models.* Genome Biol, 2015. **16**: p. 13.
61. Robert, C. and M. Watson, *Errors in RNA-Seq quantification affect genes of relevance to human disease.* Genome Biol, 2015. **16**: p. 177.
62. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
63. Wang, K., et al., *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic Acids Res, 2010. **38**(18): p. e178.
64. Jean, G., et al., *RNA-Seq read alignments with PALMapper.* Curr Protoc Bioinformatics, 2010. **Chapter 11**: p. Unit 11 6.
65. Wu, T.D., et al., *GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality.* Methods Mol Biol, 2016. **1418**: p. 283-334.
66. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
67. Hu, J., et al., *OSA: a fast and accurate alignment tool for RNA-Seq.* Bioinformatics, 2012. **28**(14): p. 1933-4.
68. Engstrom, P.G., et al., *Systematic evaluation of spliced alignment programs for RNA-seq data.* Nat Methods, 2013. **10**(12): p. 1185-91.

69. Zhao, S. and B. Zhang, *A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification.* BMC Genomics, 2015. **16**: p. 97.
70. Aken, B.L., et al., *The Ensembl gene annotation system.* Database (Oxford), 2016. **2016**.
71. Pruitt, K.D., et al., *RefSeq: an update on mammalian reference sequences.* Nucleic Acids Res, 2014. **42**(Database issue): p. D756-63.
72. Hsu, F., et al., *The UCSC Known Genes.* Bioinformatics, 2006. **22**(9): p. 1036-46.
73. Thierry-Mieg, D. and J. Thierry-Mieg, *AceView: a comprehensive cDNA-supported gene and transcripts annotation.* Genome Biol, 2006. **7 Suppl 1**: p. S12 1-14.
74. Wilming, L.G., et al., *The vertebrate genome annotation (Vega) database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D753-60.
75. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
76. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
77. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
78. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
79. Lahens, N.F., et al., *IVT-seq reveals extreme bias in RNA sequencing.* Genome Biol, 2014. **15**(6): p. R86.
80. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biol, 2016. **17**: p. 13.
81. Quail, M.A., et al., *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* BMC Genomics, 2012. **13**: p. 341.
82. Ross, M.G., et al., *Characterizing and measuring bias in sequence data.* Genome Biol, 2013. **14**(5): p. R51.
83. Quail, M.A., H. Swerdlow, and D.J. Turner, *Improved protocols for the illumina genome analyzer sequencing system.* Curr Protoc Hum Genet, 2009. **Chapter 18**: p. Unit 18 2.
84. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.
85. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
86. Lee, J.H., et al., *Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues.* Nat Protoc, 2015. **10**(3): p. 442-58.
87. Chen, K.H., et al., *RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells.* Science, 2015. **348**(6233): p. aaa6090.
88. Lubeck, E., et al., *Single-cell in situ RNA profiling by sequential hybridization.* Nat Methods, 2014. **11**(4): p. 360-1.
89. Brennecke, P., et al., *Accounting for technical noise in single-cell RNA-seq experiments.* Nat Methods, 2013. **10**(11): p. 1093-5.
90. Sorzano, C.O.S., J. Vargas, and A.P. Montano, *A survey of dimensionality reduction techniques.* arXiv preprint arXiv:1403.2877, 2014.
91. Van Der Maaten, L., E. Postma, and J. Van den Herik, *Dimensionality reduction: a comparative.* J Mach Learn Res, 2009. **10**: p. 66-71.
92. Pierson, E. and C. Yau, *ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis.* Genome Biol, 2015. **16**: p. 241.

93. Hicks, S.C., M. Teng, and R.A. Irizarry, *On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data.* bioRxiv, 2015: p. 025528.
94. Ilicic, T., et al., *Classification of low quality cells from single-cell RNA-seq data.* Genome Biol, 2016. **17**: p. 29.
95. Fluidigm, *Redesign of a new medium-cell IFC provides higher single-cell efficiency.* 1, 2016. **1**(1): p. 1.
96. Islam, S., et al., *Quantitative single-cell RNA-seq with unique molecular identifiers.* Nat Methods, 2014. **11**(2): p. 163-6.
97. Smith, T.S., A. Heger, and I. Sudbery, *UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy.* bioRxiv, 2016: p. 051755.
98. Zeisel, A., et al., *Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.* Science, 2015. **347**(6226): p. 1138-42.
99. Thomson, J.A., et al., *Embryonic stem cell lines derived from human blastocysts.* Science, 1998. **282**(5391): p. 1145-7.
100. Yabut, O. and H.S. Bernstein, *The promise of human embryonic stem cells in aging-associated diseases.* Aging (Albany NY), 2011. **3**(5): p. 494-508.
101. Nichols, J. and A. Smith, *Naive and primed pluripotent states.* Cell Stem Cell, 2009. **4**(6): p. 487-92.
102. Chazaud, C., et al., *Early lineage segregation between epiblast and primitive endoderm in mouse blastocysts through the Grb2-MAPK pathway.* Dev Cell, 2006. **10**(5): p. 615-24.
103. Rossant, J., *Stem cells and early lineage development.* Cell, 2008. **132**(4): p. 527-31.
104. Batlle-Morera, L., A. Smith, and J. Nichols, *Parameters influencing derivation of embryonic stem cells from murine embryos.* Genesis, 2008. **46**(12): p. 758-67.
105. Brook, F.A. and R.L. Gardner, *The origin and efficient derivation of embryonic stem cells in the mouse.* Proc Natl Acad Sci U S A, 1997. **94**(11): p. 5709-12.
106. Boroviak, T., et al., *The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification.* Nature Cell Biology, 2014. **16**(6): p. 516-528.
107. Kojima, Y., et al., *The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak.* Cell Stem Cell, 2014. **14**(1): p. 107-20.
108. Tesar, P.J., et al., *New cell lines from mouse epiblast share defining features with human embryonic stem cells.* Nature, 2007. **448**(7150): p. 196-9.
109. Brons, I.G.M., et al., *Derivation of pluripotent epiblast stem cells from mammalian embryos.* Nature, 2007. **448**(7150): p. 191-195.
110. Xu, R.H., et al., *BMP4 initiates human embryonic stem cell differentiation to trophoblast.* Nat Biotechnol, 2002. **20**(12): p. 1261-4.
111. Thomson, J.A., et al., *Embryonic stem cell lines derived from human blastocysts.* Science (New York, NY), 1998. **282**(5391): p. 1145-1147.
112. Ezashi, T., B.P.V.L. Telugu, and R.M. Roberts, *Model systems for studying trophoblast differentiation from human pluripotent stem cells.* Cell and Tissue Research, 2012. **349**(3): p. 809-824.
113. Kalinka, A.T. and P. Tomancak, *The evolution of early animal embryos: conservation or divergence?* Trends Ecol Evol, 2012. **27**(7): p. 385-93.
114. Roode, M., et al., *Human hypoblast formation is not dependent on FGF signalling.* Dev Biol, 2012. **361**(2): p. 358-63.

115. Hamilton, W.J. and J.D. Boyd, *Development of the human placenta in the first three months of gestation.* J Anat, 1960. **94**: p. 297-328.
116. Emera, D., R. Romero, and G. Wagner, *The evolution of menstruation: a new model for genetic assimilation: explaining molecular origins of maternal responses to fetal invasiveness.* Bioessays, 2012. **34**(1): p. 26-35.
117. Kunarso, G., et al., *Detailed characterization of the mouse embryonic stem cell transcriptome reveals novel genes and intergenic splicing associated with pluripotency.* BMC Genomics, 2008. **9**: p. 155.
118. Wei, C.L., et al., *Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state.* Stem Cells, 2005. **23**(2): p. 166-85.
119. Sudheer, S., et al., *FGF inhibition directs BMP4-mediated differentiation of human embryonic stem cells to syncytiotrophoblast.* Stem Cells Dev, 2012. **21**(16): p. 2987-3000.
120. Yabe, S., et al., *Comparison of syncytiotrophoblast generated from human embryonic stem cells and from term placentas.* Proc Natl Acad Sci U S A, 2016.
121. Ge, X., et al., *Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues.* Genomics, 2005. **86**(2): p. 127-41.
122. Bilban, M., et al., *Identification of novel trophoblast invasion-related genes: heme oxygenase-1 controls motility via peroxisome proliferator-activated receptor gamma.* Endocrinology, 2009. **150**(2): p. 1000-13.
123. Vassena, R., et al., *Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development.* Development, 2011. **138**(17): p. 3699-709.
124. Brunskill, E.W., et al., *Single cell dissection of early kidney development: multilineage priming.* Development, 2014. **141**(15): p. 3093-101.
125. Treutlein, B., et al., *Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq.* Nature, 2014. **509**(7500): p. 371-5.
126. Guo, G., et al., *Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst.* Dev Cell, 2010. **18**(4): p. 675-85.
127. Tarkowski, A.K., et al., *Individual blastomeres of 16- and 32-cell mouse embryos are able to develop into foetuses and mice.* Dev Biol, 2010. **348**(2): p. 190-8.
128. Yan, L., et al., *Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells.* Nat Struct Mol Biol, 2013. **20**(9): p. 1131-9.
129. Grun, D., et al., *De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data.* Cell Stem Cell, 2016. **19**(2): p. 266-77.
130. Sato, T., et al., *Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium.* Gastroenterology, 2011. **141**(5): p. 1762-72.
131. Grabinger, T., et al., *Ex vivo culture of intestinal crypt organoids as a model system for assessing cell death induction in intestinal epithelial cells and enteropathy.* Cell Death Dis, 2014. **5**: p. e1228.
132. Huch, M., et al., *Long-term culture of genome-stable bipotent stem cells from adult human liver.* Cell, 2015. **160**(1-2): p. 299-312.
133. Huch, M. and B.K. Koo, *Modeling mouse and human development using organoid cultures.* Development, 2015. **142**(18): p. 3113-25.
134. De Paepe, C., et al., *Human trophectoderm cells are not yet committed.* Hum Reprod, 2013. **28**(3): p. 740-9.
135. Blakeley, P., et al., *Defining the three cell lineages of the human blastocyst by single-cell RNA-seq.* Development, 2015. **142**(20): p. 3613.

136. Vu, V.Q. *Visualizing a million, PCA edition*. 2011 2011 March 04 [cited 2016 Access Date 2016 March 17]; 2011 March 04:[Available from: http://stats.stackexchange.com/questions/7860/visualizing-a-million-pca-edition.
137. Goetz, J.J. and J.M. Trimarchi, *Transcriptome sequencing of single cells with Smart-Seq.* Nat Biotechnol, 2012. **30**(8): p. 763-5.
138. John, N.J., M. Linke, and H.W. Denker, *Retinoic acid decreases attachment of JAR choriocarcinoma spheroids to a human endometrial cell monolayer in vitro.* Placenta, 1993. **14**(1): p. 13-24.
139. Nei, M., P. Xu, and G. Glazko, *Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms.* Proc Natl Acad Sci U S A, 2001. **98**(5): p. 2497-502.
140. Dodsworth, B.T., R. Flynn, and S.A. Cowley, *The Current State of Naive Human Pluripotency.* Stem Cells, 2015. **33**(11): p. 3181-6.
141. Takashima, Y., et al., *Resetting transcription factor control circuitry toward ground-state pluripotency in human.* Cell, 2014. **158**(6): p. 1254-69.
142. Theunissen, T.W., et al., *Systematic identification of culture conditions for induction and maintenance of naive human pluripotency.* Cell Stem Cell, 2014. **15**(4): p. 471-87.
143. Chan, Y.S., et al., *Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast.* Cell Stem Cell, 2013. **13**(6): p. 663-75.
144. Pera, M.F., *In search of naivety.* Cell Stem Cell, 2014. **15**(5): p. 543-5.
145. Warburg, O., *On respiratory impairment in cancer cells.* Science, 1956. **124**(3215): p. 269-70.
146. Hwang, P.Y., et al., *The role of extracellular matrix elasticity and composition in regulating the nucleus pulposus cell phenotype in the intervertebral disc: a narrative review.* J Biomech Eng, 2014. **136**(2): p. 021010.
147. Pollen, A.A., et al., *Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex.* Nat Biotechnol, 2014.
148. Wu, T.D. and S. Nacu, *Fast and SNP-tolerant detection of complex variants and splicing in short reads.* Bioinformatics, 2010. **26**(7): p. 873-881.
149. Flicek, P., et al., *Ensembl 2014.* Nucleic Acids Res, 2014. **42**(Database issue): p. D749-55.
150. Roberts, A., et al., *Improving RNA-Seq expression estimates by correcting for fragment bias.* Genome Biol, 2011. **12**(3): p. R22.
151. Wagner, G.P., K. Kin, and V.J. Lynch, *Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.* Theory Biosci, 2012. **131**(4): p. 281-5.
152. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
153. Kharchenko, P.V., L. Silberstein, and D.T. Scadden, *Bayesian approach to single-cell differential expression analysis.* Nat Methods, 2014. **11**(7): p. 740-2.
154. Jaskowiak, P.A. and R.J. Campello, *Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data.* Proceedings of the Brazilian Symposium on Bioinformatics Brasília, Brazil, 2011. **1**: p. 1-8.
155. Jaskowiak, P.A., R.J. Campello, and I.G. Costa, *On the selection of appropriate distances for gene expression data clustering.* BMC Bioinformatics, 2014. **15 Suppl 2**: p. S2.

156. Langfelder, P., B. Zhang, and S. Horvath, *Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R.* Bioinformatics, 2008. **24**(5): p. 719-20.
157. Vallier, L., M. Alexander, and R.A. Pedersen, *Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells.* J Cell Sci, 2005. **118**(Pt 19): p. 4495-509.
158. Grow, E.J., et al., *Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells.* Nature, 2015. **522**(7555): p. 221-5.
159. O'Rahilly, R. and F. Müller, *Developmental Stages in Human Embryos, Including a Revision of Streeter's 'Horizons' and a Survey of the Carnegie Collection.* Washington, Carnegie Institution of Washington, 1987.
160. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
161. Bjorklund, A. and S.B. Dunnett, *Dopamine neuron systems in the brain: an update.* Trends Neurosci, 2007. **30**(5): p. 194-202.
162. Wallen, A. and T. Perlmann, *Transcriptional control of dopamine neuron development.* Ann N Y Acad Sci, 2003. **991**: p. 48-60.
163. Grenhoff, J., G. Aston-Jones, and T.H. Svensson, *Nicotinic effects on the firing pattern of midbrain dopamine neurons.* Acta Physiol Scand, 1986. **128**(3): p. 351-8.
164. Hommel, J.D., et al., *Leptin receptor signaling in midbrain dopamine neurons regulates feeding.* Neuron, 2006. **51**(6): p. 801-10.
165. Schultz, W., *Dopamine signals for reward value and risk: basic and recent data.* Behav Brain Funct, 2010. **6**: p. 24.
166. Friedman, A.K., et al., *Enhancing depression mechanisms in midbrain dopamine neurons achieves homeostatic resilience.* Science, 2014. **344**(6181): p. 313-9.
167. Sulzer, D., *Multiple hit hypotheses for dopamine neuron loss in Parkinson's disease.* Trends Neurosci, 2007. **30**(5): p. 244-50.
168. Davis, K.L., et al., *Dopamine in Schizophrenia - a Review and Reconceptualization.* American Journal of Psychiatry, 1991. **148**(11): p. 1474-1486.
169. Mendez, I., et al., *Dopamine neurons implanted into people with Parkinson's disease survive without pathology for 14 years.* Nat Med, 2008. **14**(5): p. 507-9.
170. Sasaki, M., et al., *Neuromelanin magnetic resonance imaging of locus ceruleus and substantia nigra in Parkinson's disease.* Neuroreport, 2006. **17**(11): p. 1215-8.
171. Gale, E. and M. Li, *Midbrain dopaminergic neuron fate specification: Of mice and embryonic stem cells.* Mol Brain, 2008. **1**: p. 8.
172. Ang, S.L., *Transcriptional control of midbrain dopaminergic neuron development.* Development, 2006. **133**(18): p. 3499-506.
173. Anderegg, A., J.F. Poulin, and R. Awatramani, *Molecular heterogeneity of midbrain dopaminergic neurons--Moving toward single cell resolution.* FEBS Lett, 2015. **589**(24 Pt A): p. 3714-26.
174. Dahlstroem, A. and K. Fuxe, *Evidence for the Existence of Monoamine-Containing Neurons in the Central Nervous System. I. Demonstration of Monoamines in the Cell Bodies of Brain Stem Neurons.* Acta Physiol Scand Suppl, 1964: p. SUPPL 232:1-55.
175. Blaess, S. and S.L. Ang, *Genetic control of midbrain dopaminergic neuron development.* Wiley Interdiscip Rev Dev Biol, 2015. **4**(2): p. 113-34.

176. Chung, C.Y., et al., *Cell type-specific gene expression of midbrain dopaminergic neurons reveals molecules involved in their vulnerability and protection.* Hum Mol Genet, 2005. **14**(13): p. 1709-25.
177. Bentivoglio, M. and M. Morelli, *Chapter I The organization and circuits of mesencephalic dopaminergic neurons and the distribution of dopamine receptors in the brain.* Handbook of chemical neuroanatomy, 2005. **21**: p. 1-107.
178. Gerfen, C.R., M. Herkenham, and J. Thibault, *The neostriatal mosaic: II. Patch- and matrix-directed mesostriatal dopaminergic and non-dopaminergic systems.* J Neurosci, 1987. **7**(12): p. 3915-34.
179. Lynd-Balta, E. and S.N. Haber, *The organization of midbrain projections to the striatum in the primate: sensorimotor-related striatum versus ventral striatum.* Neuroscience, 1994. **59**(3): p. 625-40.
180. Surmeier, D.J., et al., *What causes the death of dopaminergic neurons in Parkinson's disease?* Prog Brain Res, 2010. **183**: p. 59-77.
181. Welberg, L., *Neurodegenerative disease: SNc neurons' Achilles heel.* Nat Rev Neurosci, 2011. **12**(1): p. 6.
182. Matsumoto, M. and O. Hikosaka, *Two types of dopamine neuron distinctly convey positive and negative motivational signals.* Nature, 2009. **459**(7248): p. 837-41.
183. Hegarty, S.V., A.M. Sullivan, and G.W. O'Keeffe, *Midbrain dopaminergic neurons: a review of the molecular circuitry that regulates their development.* Dev Biol, 2013. **379**(2): p. 123-38.
184. Panman, L., et al., *Sox6 and Otx2 control the specification of substantia nigra and ventral tegmental area dopamine neurons.* Cell Rep, 2014. **8**(4): p. 1018-25.
185. Madisen, L., et al., *A robust and high-throughput Cre reporting and characterization system for the whole mouse brain.* Nat Neurosci, 2010. **13**(1): p. 133-40.
186. Backman, C.M., et al., *Characterization of a mouse strain expressing Cre recombinase from the 3' untranslated region of the dopamine transporter locus.* Genesis, 2006. **44**(8): p. 383-90.
187. Nunes, I., et al., *Pitx3 is required for development of substantia nigra dopaminergic neurons.* Proc Natl Acad Sci U S A, 2003. **100**(7): p. 4245-50.
188. van den Munckhof, P., et al., *Pitx3 is required for motor activity and for survival of a subset of midbrain dopaminergic neurons.* Development, 2003. **130**(11): p. 2535-42.
189. Smidt, M.P., et al., *Early developmental failure of substantia nigra dopamine neurons in mice lacking the homeodomain gene Pitx3.* Development, 2004. **131**(5): p. 1145-55.
190. Fan, J., et al., *Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis.* Nat Methods, 2016. **13**(3): p. 241-4.
191. Zhang, J., et al., *BioMart: a data federation framework for large collaborative projects.* Database (Oxford), 2011. **2011**: p. bar038.
192. Di Giovanni, G., V. Di Matteo, and E. Esposito, *Birth, life and death of dopaminergic neurons in the substantia nigra.* J Neural Transm Suppl, 2009(73): p. 1 p preceeding table of contents.
193. Bello, E.P., et al., *Cocaine supersensitivity and enhanced motivation for reward in mice lacking dopamine D2 autoreceptors.* Nat Neurosci, 2011. **14**(8): p. 1033-8.
194. Anisimov, S.V., et al., *"NeuroStem Chip": a novel highly specialized tool to study neural differentiation pathways in human stem cells.* BMC Genomics, 2007. **8**: p. 46.

195. Johansson, A., *Dependence-induced changes in opioid-receptor gene expression.* 2013.
196. Abcam, *Neural markers guide.* Abcam 2016. **1**(1): p. 1.
197. Yin, M., et al., *Ventral mesencephalon-enriched genes that regulate the development of dopaminergic neurons in vivo.* J Neurosci, 2009. **29**(16): p. 5170-82.
198. Yang, Q., et al., *Ebf2 is required for development of dopamine neurons in the midbrain periaqueductal gray matter of mouse.* Dev Neurobiol, 2015. **75**(11): p. 1282-94.
199. Mendez, J.A., et al., *Developmental and target-dependent regulation of vesicular glutamate transporter expression by dopamine neurons.* J Neurosci, 2008. **28**(25): p. 6309-18.
200. Doucet-Beaupre, H., et al., *Lmx1a and Lmx1b regulate mitochondrial functions and survival of adult midbrain dopaminergic neurons.* Proc Natl Acad Sci U S A, 2016. **113**(30): p. E4387-96.
201. Bifsha, P., et al., *Rgs6 is required for adult maintenance of dopaminergic neurons in the ventral substantia nigra.* PLoS Genet, 2014. **10**(12): p. e1004863.
202. Liu, G., et al., *Aldehyde dehydrogenase 1 defines and protects a nigrostriatal dopaminergic neuron subpopulation.* J Clin Invest, 2014. **124**(7): p. 3032-46.
203. Cai, H., et al., *Aldehyde Dehydrogenase 1 making molecular inroads into the differential vulnerability of nigrostriatal dopaminergic neuron subtypes in Parkinson's disease.* Transl Neurodegener, 2014. **3**: p. 27.
204. Fabregat, A., et al., *The Reactome pathway Knowledgebase.* Nucleic Acids Res, 2016. **44**(D1): p. D481-7.
205. Sulzer, D., et al., *Dopamine neurons make glutamatergic synapses in vitro.* J Neurosci, 1998. **18**(12): p. 4588-602.
206. Zhang, S., et al., *Dopaminergic and glutamatergic microdomains in a subset of rodent mesoaccumbens axons.* Nat Neurosci, 2015. **18**(3): p. 386-92.
207. Kawano, M., et al., *Particular subpopulations of midbrain and hypothalamic dopamine neurons express vesicular glutamate transporter 2 in the rat brain.* J Comp Neurol, 2006. **498**(5): p. 581-92.
208. Dal Bo, G., et al., *Dopamine neurons in culture express VGLUT2 explaining their capacity to release glutamate at synapses in addition to dopamine.* J Neurochem, 2004. **88**(6): p. 1398-405.
209. Yamaguchi, T., et al., *Glutamatergic and dopaminergic neurons in the mouse ventral tegmental area.* Eur J Neurosci, 2015. **41**(6): p. 760-72.
210. Tritsch, N.X., J.B. Ding, and B.L. Sabatini, *Dopaminergic neurons inhibit striatal output through non-canonical release of GABA.* Nature, 2012. **490**(7419): p. 262-6.
211. Tritsch, N.X., et al., *Midbrain dopamine neurons sustain inhibitory transmission using plasma membrane uptake of GABA, not synthesis.* Elife, 2014. **3**: p. e01936.
212. Kim, J.I., et al., *Aldehyde dehydrogenase 1a1 mediates a GABA synthesis pathway in midbrain dopaminergic neurons.* Science, 2015. **350**(6256): p. 102-6.
213. Suzuki, R. and H. Shimodaira, *Pvclust: an R package for assessing the uncertainty in hierarchical clustering.* Bioinformatics, 2006. **22**(12): p. 1540-2.
214. Kleinberg, J., *An impossibility theorem for clustering.* Advances in neural information processing systems, 2003: p. 463-470.
215. Angermueller, C., et al., *Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity.* Nat Methods, 2016. **13**(3): p. 229-32.

216. Kiselev, V.Y., et al., *SC3-consensus clustering of single-cell RNA-Seq data.* bioRxiv, 2016: p. 036558.