Using cell type-specific methods to understand molecular processes in the brain

Doctoral Thesis

In partial fulfillment of the requirements for the degree
"Doctor rerum naturalium (Dr. rer. nat.)"
in the Molecular Medicine Study Program
at the Georg-August University Göttingen

submitted
by
Ashish Rajput

born in
Indore, India

Göttingen, April 2018

Members of the Thesis Committee:

Prof. Dr. Stefan Bonn (Supervisor, Reviewer)
Institute of Medical System Biology,
Center for Molecular Neurobiology Hamburg (ZMNH), University Clinic Eppendorf, Hamburg
German Center for Neurodegenerative Diseases (DZNE), Göttingen

Prof. Dr. Steven A. Johnsen (Reviewer)
Translational Cancer Research
Department of General, Visceral and Pediatric Surgery
University Medical Center Göttingen

Prof. Dr. Tiago F. Outeiro
Department of Neurodegeneration and Restorative Research,
Center for Nanoscale Microscopy and Molecular Physiology of the Brain,
University Medical Center Göttingen

Extended Committee Members/ Members of the Examination Board:

Dr. Camin Dean, Ph.D.
Trans-synaptic Signaling
European Neuroscience Institute (ENI), Göttingen

Dr. Ira Milosevic, Ph.D.
Synaptic Vesicle Dynamics
European Neuroscience Institute (ENI), Göttingen

Prof. Dr. Lutz Walter
Department of Primate Genetics
German Primate Center (DPZ), Göttingen

Date of Disputation:

AFFIDAVIT

Here I declare that my doctoral thesis entitled "**Using cell type-specific methods to understand molecular processes in the brain**" has been written independently with no other sources and aids than quoted.

Ashish Rajput

Göttingen, April 2018

I dedicate this thesis to my parents.

List of Publications

1. Halder R, Hennion M, Vidal RO, Shomroni O, Rahman RU, **Rajput A**, et al. DNA methylation changes in plasticity genes accompany the formation and maintenance of memory. Nat Neurosci. 2015;19(1):102-10.

2. Llorens F, Thüne K, Sikorska B, Schmitz M, Tahir W, Fernández-borges N,..,**Rajput A**, et al. Altered Ca 2 + homeostasis induces Calpain- Cathepsin axis activation in sporadic Creutzfeldt-Jakob disease. Acta Neuropathol Commun. 2017;1-20.

3. Hatje K, Rahman R, Vidal RO, Simm D, Hammesfahr B, Bansal V, **Rajput A**, et al. The landscape of human mutually exclusive splicing. Mol Syst Biol. 2017;1-19.

4. Paiva I, Pinho R, Pavlou MA, Hennion M, Szeg M, **Rajput A**, et al. Sodium butyrate rescues dopaminergic cells from alpha-synuclein-induced transcriptional deregulation and DNA damage. Hum Mol Genet. 2017;26(12):2231-46.

5. Wolfes AC, Ahmed S, Awasthi A, Stahlberg MA, **Rajput A**, Magruder DS, et al. A novel method for culturing stellate astrocytes reveals spatially distinct Ca 2 + signaling and vesicle recycling in astrocytic processes. J Gen Physiol. 2017;149(1):149-70.

6. Fernandes M, Wan C, Tacutu R, Barardo D, **Rajput A**, Wang J, et al. Systematic analysis of the gerontome reveals links between aging and age-related diseases. Hum Mol Genet. 2016;25(21):4804-18.

Publications in preparation

1. **Rajput A***, Wakhloo D*, Bansal V, Ibrahim A, et al. EPO induces differentiation of pre-existing non-dividing precursors in the hippocampus to enhance pyramidal neuron numbers and cognition. (* shared authorship)

2. Kaczmarczyk L, **Rajput A***, Bansal V*, Schleif M, Rahman RU, et al. Tagger mouse: a tool to dissect brain function and pathophysiology. (* shared authorship)

Table of contents

Acknowledgements

"No one can whistle a symphony. It takes an orchestra to play it."

-H.E. Luccock

"You can't connect the dots looking forward; you can only connect them looking backward."

-Steve Jobs

The time I had spent during my PhD, in the Göttingen and Hamburg, also gave me an opportunity to develop myself in many ways. If I think about this time, many people have helped me personally and professionally. Here, I would like to thank some of the people who helped me throughout this journey and made it a remarkable time of my life.

First of all, I would like to express my gratitude and a sincere thank to my supervisor **Prof. Dr. Stefan Bonn** for giving me the opportunity to come to the DZNE, Göttingen and pursue research on such exciting projects. The fun and supportive environment of the lab has always helped me to implement new ideas. Thank you for always providing me great advice and the support during the ups and downs of my PhD life. I had learned a lot from you during this time, both personally and professionally. I wish you a lot of success in the future. Thank you!

I would also like to extend my sincere thanks to my thesis committee members, **Prof. Dr. Steven A. Johnsen** and **Prof. Dr. Tiago F. Outeiro** for their constructive discussions and helpful suggestions on my thesis project. The meetings with you were also an inspiration for me and to my research. I also enjoyed the scientific talks and fun events in Molecular medicine Ph.D. program headed by Steve. I am also thankful to my **PhD examination committee** members for accepting my request and for critical assessment of my thesis.

I am thankful to the current and previous lab members of the Bonn-lab. It was always great to work with them, and I have learned a lot from my lab mates. Mainly, I would like to mention few people here. I would like to thank **Magali Hennion** and **Ramon Vidal** for teaching wet-lab and bioinformatics techniques during early days of my PhD and also for providing helpful advice on my research projects. Thanks for working with me patiently. I wish you all the best for your research career. I would like to thank **Vikas Bansal** for providing all the help in the analysis of single-cell sequencing data and the Tagger project. Thanks for collaborating with me and provide constant support during my PhD. Also thanks for your suggestions on my thesis. I wish you a lot of success in your future endeavors. I also would like to thank **Orr Shomroni** and **Raza -ur Rahman** for being a nice colleague as well as providing constant support personally and professionally whenever I needed. Also for giving reviews on my thesis. I also would like to extend my gratitude to **Michel Edwar** for collaborating with me on single cell project as well as motivating me with kind thoughts. All the best for your research projects. I thank **Joern** for his helpful advice during my thesis writing as well as for critically reviewing my thesis. I also like to thank **Sumner** for his suggestions on my thesis and helpful discussions. I wish you all the best for your PhD and for a pug ;). I would like to thank **Anna-Lena Schuetz, Susanne Burkhardt**, and **Julio Cesar Vizcaino** for always taking my requests and helping me in my experiments. I also thank **Ting Sun** for assisting me in various projects and creating a cheerful environment

5

(also for KitKats). I wish you good luck for your PhD. I would like to thank **Abhi Gautam** for being a nice friend and for a cheerful time in the lab. Best of luck in your future endeavors. I also thank **Abdul Sattar** for helpful discussions and help in the projects. I am thankful to **Vincenzo Capece, Tonatiuh Centeneo, Abdul Qadir Ibrahim** and **Maksims** for providing help in my projects and nice suggestions. I would like to thank **Anna** and **Sven** for being a nice friend and for all the nice discussions. My thanking words also go to **Desi-Re, Yu, Mohammed Marouf, Pierre Machart** and **Sabine** for creating a lovely environment in the lab.

I would also like to utilize this moment to extend my gratitude to my collaborators in the PhD projects. I am really grateful to **Evan Macosko** and **Melissa Goldman** from **Steve MaCarroll lab** at Harvard Medical School (Boston, USA) for accommodating me in their lab and trained me with the Dropseq method. It was really helpful and excellent learning experience. I would like to thank **Prof. Till Marquardt, Mudassar Khan,** and **Pitchaiah Cherukuri** for their help in single cell projects and also for helpful scientific discussions. I would like to thank **Prof. Lutz Walter** and **Olena Zaitseva** for providing support in single-cell experiments using Fluidigm system. I want to thank Gabriela, Fabian and other TAL facility members in Göttingen for helping in the samples sequencing. I also would like to thank **Dr. Bettina Schmid** and **Frauke van Bebber** for collaboration in validation studies on zebrafish model.

I am also grateful to **Prof. Andre Fischer,** his lab members, and other **DZNE staff** members. Thank you for your continuous collaboration, support in research projects and sharing lab facilities with me. I would like to thank **Rashi Halder** and **Eva Benito** for providing me helpful advice for my projects as well as nice discussions. I would like to thank **Magda, Hendrik, Gaurav** and **Christian** for sharing excellent PhD time with me and for all those really nice, thoughtful and fun science discussions and also for helpful advice. I thank **Tea, Reza, Jonas**, and **Maryam** for a nice friendship.

During the PhD time, people outside the lab also play an essential role. I would like to thank all the friends from "Indian group" as well as "Spanish group" for all the fun events, parties, and get-togethers, which filled my life in Göttingen with so many pleasant memories. I would like to thank **Magda, Gaurav, Hendrik, Bea, Odette, Aman, Amit, Prasanth, Vivek** and all other friends from Göttingen. Thank you, **Magda**, for all those memorable trips in Germany and Europe and also for your continuous support during PhD time. I wish you a great future ahead. Thank you, **Gaurav** for all the great discussions we had and for your constant support. All the best for your PhD. I also like to extend my thanks to other friends **Latha, Mangesh da** and other friends from Osnabrück. I really cherish the time spent with you guys. My thankful word also goes to old-time friends **Rishabh, Praneeth, Sudheshna, Baisakhi, Chirag, Riddhi, Babu, Rahul** and others. A Big thank you to all of you for sharing a bond of friendship with me.

Dear **Klaus** and **Siegrid**, thank you for being such a lovely and caring host and inviting me to the traditional family celebrations. The time spent during Christmas and Easter in the beautiful town Friesoythe and your house was really memorable. Thank you for all your love and care every time we met. I wish you good health and lot of happiness together.

In the end, I would like to express my heartfelt gratitude to **Amrita** for bringing so much positive energy and happiness in my life. Thank you for your continuous support and encouragement

6

during my PhD. Your friendship is very special to me. Thank you for all your wise advises, the memorable time during PhD and everything else.

Lastly and most importantly, I would like to thank my **parents** and family for their unconditional love and support throughout my life. It is because of them I could achieve such place in my life after an incredible journey so far. I would like to dedicate this thesis to my dear Papa (**Shri Indar Singh Rajput**) and Mummy (**Shrimati Ramsakhi Rajput**). I am incredibly fortunate to have such loving parents in my life. Thank you for everything you have done for me and shaping my identity. Thank you for all your encouragement throughout my life. Though thousands of miles apart but you are always near to me in my heart. I am blessed to be your son. Thank you!

I would also like to thank other people to whom somehow I could not mention here that contributed to my scientific journey and personal development. In the end, I would like to thank all the scientific community present and in the past to accumulate all the knowledge over the centuries for the betterment of humans and our environment.

-Ashish Rajput

Abstract

Biological causes of many human diseases can be understood more comprehensively by measuring molecular states of the cells. Characterization of cells based on their inherent molecular profiles as well as functional changes in their transcriptional programs, in response to environmental stimuli, could be investigated using high-throughput next generation sequencing (NGS) methodologies. However, these technologies rely on massive amounts of input material containing thousands or even millions of cells, which leads to averaging of gene profiles from all the sequenced cells. It is important to acknowledge however that tissues can contain various types of cells with different characteristics. Furthermore, even cells that are of the same cell type can exhibit very different behavior. It is possible either because of subtypes of cells or expression variation among individual cells. Cellular specialization is especially evident in tissues, which perform varieties of functions using the similar cell types. For instance, the same brain region contains millions of neuronal cells, which differ in their molecular and physiological properties and are involved in different processes. For this reason, it is essential to develop new techniques that can measure individual cells instead of cell collectives.

This thesis explores three cell type-specific techniques for obtaining molecular information and for investigation of biological mechanisms. At first, Chapter 3. describes an implementation of BiTS (Batch isolation of tissue-specific chromatin) coupled with ChIP-seq and MeDIP-seq approaches, which is used to reveal epigenetic changes associated with the formation and maintenance of memory, specifically in neuronal and non-neuronal cells. Such procedures may allow users to obtain cell type-specific genetic and epigenetic information based on a known marker. Chapter 4. introduces the Tagger system, which is the first *in-vivo* mouse system that enables cell type-specific analysis of multiple nucleic acids from the same tissue. The Tagger system is based on a single transgene insertion into the mouse genome. This system leads to the synthesis of four components (protein molecules) in specific cell types. It enables the researchers to isolate multiple nucleic acid species (such as mRNA, miRNA, 4-TU labeled RNA) as well as isolated nuclei for genetic and epigenetic studies. Finally, the Drop-seq method is implemented in Chapter 5. to characterize individual cells of the spinal cord and hippocampus at the single-cell

resolution. Using the Drop-seq method two studies are conducted, where the first study was devoted to characterizing molecular properties of individual cell types from the spinal cord based on the unbiased single-cell RNA sequencing method (Drop-seq). The second study investigated the effect of erythropoietin (EPO) on unidentified precursor cells in the brain hippocampus. These precursor cells can differentiate into neurons and oligodendrocytes and reported to give rise to ~20% increase in the neuronal cell population (Hassouna et al., 2016).

Overall, the current thesis implements cutting-edge, robust, flexible and reliable technologies to understand the molecular mechanisms at the individual cell type as well as single cell level. The knowledge acquired in this thesis could be applied to resolving the precise molecular mechanisms of several diseases and to design targeted or personalized therapeutics in the near future. Notable examples are neuro-degeneration mediated memory deprivation, nervous system diseases affecting specific cell types such as ALS, Alzheimer's disease, Parkinson's and others. These diseases could be studied in more depth, with the cell types investigated in this thesis.

10

List of figures

12

List of tables

Abbreviations

°C	Degree Celsius
%	Percent
4-TU	4-thiouracil
A	Adenine
ACC	Anterior cingulate cortex
BAM	Binary alignment/map
BiTS	Batch isolation of tissue-specific chromatin
bp	Base pair
BSA	Bovine serum albumin
C	Cytosine
cDNA	Complementary DNA
CFC	Contextual fear conditioning
ChAT	Choline acetyltransferase
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation sequencing
CNS	Central nervous system
CpG	5'-C-phosphate-G-3'
DE	Differential expression
DEE	Differentially expressed exon
DEG	Differentially expressed gene
DHPTM	Differential histone post-translational modifications
DMR	Differentially methylated region
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic acid
EB	Elution buffer
EDTA	Ethylene diamine tetraacetic acid
ENCODE	Encyclopedia of DNA Elements
FACS	Fluroscence activated cell sorter
FDR	False discovery rate
G	Guanine
g	Gram
GO	Gene ontology
HOMER	Hyper geometric optimization of motif enrichment
HPTM	Histone post-translational modication
IP	Immunoprecipitation
M	Molar
m	Milli
MedIP	Methylated DNA immunoprecipitation

MedIP-seq	Methylated DNA immunoprecipitation sequencing
min	Minutes
miRNA	Micro RNA
ml	Mili liter
mm	Mili meter
MN	Motor neuron (s)
mRNA	Messenger RNA
ms	Millisecond
NGS	Next generation sequencing
PBS	Phosphate buffered saline
PCA	Principle component analysis
PCR	Polymerase chain reaction
QC	Quality control
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
RPKM	Reads per kilobase of transcript per
RPM	Reads per million
RPM	Rotations per minute
RT	Room temperature
SAM	Sequence alignment/map
SDS	Sodium dodecyl sulfate
T	Thymine
TF	Transcription factor
tSNE	t-Distributed Stochastic Neighbor Embedding
TSS	Transcription start site
u	Micro
ul	Micro liter
vGlut1	Vesicular glutamate transporter 1
WB	Weinmann buffer
WebGestalt	Web-based gene set analysis toolkit
WT	Wild-type
x g	Gravitational acceleration (9.81)

Chapter 1. Introduction

The human body is made up of trillions of the cells. Cells are the basic building blocks of the living organisms. Different kinds of cells organize to form various tissues and organs of the body. Every cell type in the tissue plays a vital role supporting the tissue's function and at the higher perspective maintains the body functions to attain the living organism properties. A cell type could be defined as a group of cells, which perform a similar function in the body (Poulin, 2016). Although the function of many cell types in the body is still unknown (for instance, nervous system), the experimental approaches for the classification of cells could employ the properties, which are easily accessible or measurable. As we know that a cell's function is deeply connected with its molecular composition, categorization of cell type based on their gene expression level could be a pragmatic approach (Fishell & Heintz, 2013).

At the cellular level, various genetic and epigenetic factors determine the expression of the genes. To study the gene expression, high-throughput next generation sequencing (NGS) technologies were developed, which employ methodologies such as messenger-RNA sequencing (mRNA-seq) to detect the level of mRNAs in the cells. Other applications of NGS methods are to study the role of epigenetic factors in the gene expression. This could be achieved by employing techniques such as chromatin immunoprecipitation sequencing (ChIP-seq), which could determine the DNA binding sites or genomic locations of epigenetic marks and transcription factors. Similarly, methylation DNA immunoprecipitation (MeDIP-seq) detects the presence of methylated cytosine in the genomic DNA, which inhibits the transcription of a gene. These methods are widely in use and provide massive amounts of data to analyze biological modifications in healthy to diseased conditions. For a long time, NGS techniques were relying on the bulk tissue, containing thousands or millions of cells, to investigate underlying functional mechanisms. In this approach, biomolecules from individual cells are mixed during sample processing steps, and it generates an average of gene expression profiles from all the cells in the tissue. However, this is not the accurate representation of a biological system. In the biological context, different cell types interplay with each other and respond to their environment differently at the molecular level. Cell types in the tissue further differ in their functional and cell cycle stages. Therefore, averaging the

gene expression levels from all the cells present in the tissue could not provide an accurate representation of the internal state of the tissue in a given time.

Furthermore, as bulk tissue contains many different cell types, using a combination of cells in study diminishes the chances to detect the subtle changes in the gene expression or epigenetic mechanisms in the specialized cells (such as neurons) during experimental conditions (for instance, memory formation, disease progression and so on.). In the clinical settings, it could further lead to misdiagnosis of a disease, which occur in specific cell type (for instance Cancer, ALS, Alzheimer's and so on.). Similarly, age-dependent changes in specific cell types in the tissue were hard to detect using bulk tissue sequencing techniques. Therefore, this thesis aims to implement cell type-specific and single-cell NGS techniques, to delve into various research questions described in the following chapters. Nevertheless, to completely understand the significance of cell type-specific gene expression mechanisms, it is necessary to understand the underlying biological and methodological concepts behind these techniques. Hence, this chapter is dedicated to providing biological insights about the molecular mechanisms of the cell as well as core concepts, which in turn clarify the need and significance of the cell type-specific NGS workflows.

1.1 DNA structure and gene expression

One of the essential biomolecules, which carry the genetic blueprint of life, inside the nucleus of the cell is deoxyribonucleic acid (DNA). It is made up of two polynucleotide chains arranged in a double helix form, attached to each other by hydrogen bonds via their nitrogenous bases. Each nucleotide unit consists three components: a five-carbon sugar molecule (de-oxy-ribose in DNA), a phosphate molecule and one of the four nitrogenous bases, i.e., adenine (A), thymine (T), guanine (G) and cytosine (C). To hold the double-stranded DNA structure together, adenine pairs with thymine by two hydrogen bonds and guanine pairs with cytosine connected by three hydrogen bonds. Every cell of an organism carries an identical copy of the DNA, collectively known as the genome. Although, it raises a question that, how different type of cells exists in the body having the same genome. In other words, how the static genome creates such variety of changes in the cells while responding to the environment and growth of the body. These

questions could be answered by understanding the regulation of DNA functional unit (gene) expression.

A functional unit of DNA is known as a gene. The human genome consists of approximately 30,000 genes. Each gene is transcribed and leads to the formation of specific ribonucleic acid (RNA) molecule, which is further translated to a corresponding protein molecule. Proteins are the complex and extensive molecules that play many critical roles in the cell. Proteins are the actual workhorse of the cells and essential for the structure, function, and regulation of the body's tissues and organs. The process, also known as the central dogma, first suggested that gene transcribed into RNA and later translation result in a protein. However, later discoveries reveal that messenger RNA (mRNA) molecules translated into proteins, while other types of RNA molecules, ribosomal RNAs, micro RNAs and other non-coding RNAs, remain in the form of RNA and play a role in the regulation of gene expression by interacting with DNA, proteins or other RNA molecules. For instance, these non-coding RNA could form an RNA-protein complex to regulate gene expression. In general, the number of mRNA molecules, instead of calculating protein concentration, produced from a particular gene is used for the determination of expression of that gene in a cell. One of the techniques to determine the amount of mRNA in the sample is mRNA sequencing (mRNA-seq), which will be described in the later chapters.

1.1.1 Regulation of gene expression

During the gene expression, DNA transcribes into RNA molecules, some of the RNA (mRNA), translate into proteins and performs several biological functions (Gerstein et al., 2007). Transcription of a gene is a complex process and involves several factors and regulatory mechanisms inside the cell. RNA molecules are single-stranded and consist of the un-methylated form of the base thymine called uracil (U) apart from A, G and C bases. To prepare for the RNA synthesis, the DNA double helix begins with opening and unwinding of the two strands, and in effect, they are exposed to attachment of the gene regulatory factors. In eukaryotes, RNA polymerase II (Pol II) binds to the transcription start site (TSS) of protein-coding genes and start transcription process. Binding of Pol II to the TSS is mediated by several initiation factors (general transcription factors). These general transcription factors such as TFIIB or TFIID

(TATA-box binding proteins and other co-factor complexes) recruits RNA polymerase II to the transcription start site (Wood, 1996). Apart from general transcription factors, sequence-specific transcription factors (TFs) could also regulate transcription of a gene. These transcription factors bind to regulatory elements like promoters and enhancers, which contain a short DNA sequence (5 to 20 bp), which provide a specific binding site for one or more transcription factors. Promoters are present in close proximity to the TSS region and contain TATA-box binding site for Pol II, while enhancers could be present within several hundred bases to thousands of bases distant from the TSS site of the gene. TFs bind to these regulatory elements in the DNA and control the expression of the gene.

After transcription initiation, protein-coding genes are first transcribed into precursor mRNAs (pre-mRNAs) inside the nucleus, which are further processed to mature mRNAs and transported to the cytoplasm for translation. The process of mature RNA formation consists of three main modification steps: 5' capping, 3' polyadenylation, and splicing of RNA. 5' capping is a chemical alteration of mRNA molecules in eukaryotes, which stabilize the RNA as well as important for the translocation of mRNA to the cytoplasm. In 5' capping, 7-methyl-guanosine nucleotides were added to the 5' end of the mRNA. On the other hand, 3' poly-Adenylation is the process to add poly (A) tail to the mRNA by adding multiple adenosine monophosphates. The poly (A) tail is also crucial for nuclear transport, translation and protection of mRNA from degradation. In the end, splicing of pre-mRNA is the process to remove introns (non-protein-coding regions) and joining of exons (protein-coding regions), which leads to the formation of mature mRNA. The splicing process also leads to different combinations of exons and in turn produces many distinct isoforms of a gene.

1.1.2 Epigenetic regulation of gene expression

Another layer of gene expression regulation is present in the cell through epigenetic mechanisms. Epigenetics implies features "on the top" or "in addition to" the genetic basis of inheritance. Epigenetic mechanisms regulate gene expression without altering its primary nucleotide sequence. Examples of mechanisms that produce such changes are DNA methylation and histone posttranslational modifications.

In various organisms, DNA contains millions of base pairs in a cell. For instance, the human genome contains around 3 billion bps, and with each bp having the length of 340 picometers (3.4×10^{-10} meters), human DNA spans around 1.2 meters in length in its double-stranded form. On the other hand, human cells (on average) range in volume from 30 to 5000 cubic micrometers ($0.03 - 50 \times 10^{-16}$ cubic meters) (Milo, Jorgensen, Moran, Weber, & Springer, 2010). The cells are too tiny to contain such a long DNA, and it is possible only by folding of the double helical structure of the DNA to a higher structure. To do that, proteins, known as "histones" create protein complexes and wraps DNA around. In brief, the nucleosome (the basic unit of chromatin), which contains eight copies (two copies of each) of the histone proteins H2A, H2B, H3, and H4 together form a histone octamer, and about 147 base pairs of DNA wrapped around it (Kouzarides, 2007). The histone H1 binds to DNA directly near nucleosomes and allows the next higher packing of the DNA. Multiple nucleosomes condense together to form chromatin and highly condensed chromatin forms the chromosome, where the multiple chromosomes contain the entire genetic material of an organism inside the nucleus of a cell.

Histone post-translational modifications (HPTMs) regulate the gene expression by DNA compression and relaxation during various cellular functions as well as development and differentiation stages. The HPTMs could determine the chromatin structure, which can be present in either condensed heterochromatin form, blocking gene transcription or lightly packed structure called euchromatin, which allows the transcription of a gene. To study the location of various HPTMs in a cell type during various conditions, ChIP-seq could be utilized, which is further discussed in section 3.2.3. The histone proteins can be modified by many different chemical groups, which could be added and removed post-translationally; for instance, methylation, acetylation, phosphorylation, sumoylation and ubiquitination (Kouzarides, 2007). These HPTMs could alter chromatin structure and binding of specific proteins, which can, in turn, influence the gene expression. For instance, H3K4me2 and H3K4me3 are present at the euchromatin sites and indicate active gene expression (Zentner & Henikoff, 2013). Similarly, histone acetylation reduces the positive charges of histone tails and leads to open DNA-histone binding, and thus associated with the actively transcribed promoter regions (Fischer et al., 2008). However, other histone marks such as H3K9me3 and H3K27me3 indicates the presence of heterochromatin

region and silencing of genes (Zentner & Henikoff, 2013). Various enzymes alter these HPTMs. For instance, histone acetyltransferases (HATs) and histone methyltransferases (HMTs) add the acetyl and methyl groups on the histones, respectively. On the other hand, histone deacetylases (HDACs) and histone demethylases (HDMs) remove the acetyl and methyl groups, respectively (Legube & Trouche, 2003).

Another well-studied epigenetic regulation of gene expression is DNA methylation. In this process, a methyl group is added at the 5' position of the cytosine nucleotide to modify it to 5-methylcytosine (m5C). Most of the DNA methylation is present on the CpG islands sites (high frequency of Cytosine precedes a Guanine). Although, it can be present at non-CpG sites as well (Smith & Meissner, 2013). DNA methylation is generally associated with gene expression silencing (Illingworth & Bird, 2009). Interestingly, CpG islands are present in more than 50% of gene promoter sites in humans (Smith & Meissner, 2013). In general, housekeeping and development-associated genes are hypo-methylated in their promoter regions (Smith & Meissner, 2013). The DNA methylation process is carried out by DNA methyltransferases (DNMTs). Mainly DNMT1, DNMT3a and DNMT3b function to methylate DNA as well as its maintenance (Smith & Meissner, 2013). Apart from HPTMs and DNA methylation, several other epigenetic factors such as non-coding RNA, transcription factors, chromatin remodeling complexes, and others, also play an essential role in the regulation of gene expression.

1.2 Aim of the thesis

The primary aim of this thesis is to characterize cell type-specific functional responses of central nervous system (CNS) cells in health and disease, using various NGS techniques. With this in mind, different chapters of this thesis deal with particular aspects of it, as described below.

Chapter 3. DNA methylation changes in plasticity genes accompany the formation and maintenance of memory.
Here we investigated chromatin modifications during the learning and memory, using cell type-specific technique BiTS coupled ChIP-seq and MeDIP-seq. This chapter is already published (Halder et al., 2015).

Chapter 4. Cell type-specific molecular analysis using the Tagger system.
The Tagger system is developed to obtain cell type-specific mRNA, miRNA, TU-tagged RNA as well as cell nuclear materials. A manuscript is under preparation describing this chapter in detail.

Chapter 5. Investigating motor neuron subpopulations and EPO-mediated hippocampal changes using single-cell transcriptomics.
Here we have studied transcriptomics of spinal cord cell types with the single-cell resolution using the Drop-seq method. Another part is dedicated to erythropoietin (EPO) mediated molecular mechanisms that are associated with a ~20% increase in the pyramidal neurons and oligodendrocytes. This study will also be a part of a manuscript for publication.

Chapter 2. Materials

2.1 Buffers and solutions

Low sucrose buffer

> 0.32 M Sucrose
> 10 mM HEPES pH 8.0
> 3 mM $Mg(CH_3COO)_2$
> 0.1 mM EDTA
> 0.1% Triton X-100
> 1 mM DTT

Sucrose cushion

> 10mM HEPES pH 8
> 1M sucrose
> 3mM $Mg(CH_3COO)_2$
> 1mM DTT
> 6 ml cushion for 1.5 ml lysate

RIPA buffer

> 10 mM Tris-Cl, pH 8.0
> 140 mM NaCl
> 1 mM EDTA
> 1% Triton X-100
> 0.1% sodium deoxycholate
> 1% SDS
> Roche Complete protease inhibitors

IP Buffer (ChIP)

> 50 mM Tris-HCl
> 150 mM NaCl
> 1% NP-40
> 0.5% sodium deoxycholate
> 20 mM EDTA

Wash buffer (ChIP)

> 20 mM EDTA, pH 8.0
> 100 mM Tris, pH 8.0
> 1% NP-40

1% Na deoxycholate
500 mM LiCl
Dissolve in water

Lysis buffer (MeDIP)

10 mM Tris-Cl pH 7.5
10 mM NaCl
2 mM EDTA
0.5% SDS
100ug ProteinaseK

E3 Media

5 mM Nacl
0.17 mM KCl
0.33 mM Cacl$_2$
0.33 mM MgSo$_4$

TX Buffer

50 mM Tris-HCl pH 7.4
150 mM NaCl
1mM EDTA
1% Nonidet P40
0.05% SDS

Weinmann Buffer (2X)

100 mM Tris, pH 8
20 mM EDTA, pH 8
10% SDS

Reverse transcription (RT-mix) solution (Drop-seq)

Reagent	Amount (in ul)
H$_2$O	75
Maxima 5x RT buffer	40
20 % Ficoll PM-400	40
10 mM dNTPs (Clontech)	20
RNase inhibitor (Lucigen)	5
50 uM Template switch oligo (TSO)	10
Maxima H- RTase	10
Total	**200**

Lysis buffer (Drop-seq)

Reagent	Amount (in ul)
H_2O	500
20% Ficoll PM-400 (GE Healthcare)	300
20% Sarkosyl (Sigma #L7414)	10
0.5 M EDTA (Life Technologies)	40
2 M Tris pH 7.5 (Sigma)	100
1 M DTT	50
Total	**1000**

TS-SDS solution

> 10 mM Tris pH 8.0
> 1 mM EDTA
> 0.5% SDS

TE-TW solution

> 10 mM Tris pH 8.0
> 1 mM EDTA
> 0.01% Tween-20

PBST

> 0.1% Tween20 in PBS

EB buffer

> 1mM Tris pH 8

2.2 Reagents and kits

Reagent	Company #order number
0.5 M EDTA, pH 8.5	ThermoFisher Scientific
1 M Tris-HCl, pH 8.0	Sigma
2 M Tris-HCl, pH 7.5	Sigma
2X Kapa HiFi HotStart ReadyMix, 1.25 mL	Kapa Biosystems #KK2601
Advantage® UltraPure PCR Deoxynucleotide Mix ,10 mM	Clontech #639125
AMPure XP Beads	Beckman Coulter #A63880
Barcoded Bead SeqB	Chemgenes

Bolt Bis-Tris 12% gel	Novex, Life technologies
BSA, Fraction V, 7.5% solution	ThermoFisher Scientific #15260-037
DTT, 1M in H2O	Sigma
Exonuclease I, 20 U/uL	ThermoFisher Scientific #EN0581
Ficoll PM-400	GE Healthcare Life Sciences #F4375-100G
Formaldehyde	Sigma-Aldrich F1635
Maxima RT H minus Reverse Transcriptase, 200 U/uL	ThermoFisher Scientific #EP0751
Nextera DNA sample preparation kit, 24 samples	Illumina #FC-121-1030
Nextera XT DNA library preparation kit	Illumina #FC-131-1096
Nextera XT index kit v2 set A	Illumina #FC-131-2001
Nitrocellulose membrane	GE Healthcare
Nitrocellulose membrane	GE Healthcare
NxGen® RNAse Inhibitor	Lucigen #30281-1
Perfluorooctanol (PFO), 97%	Sigma #370533-5G
Qubit dsDNA HS assay kit	Life technology
Qubit dsDNA HS assay kit	Life technology
QX200 Droplet generation oil	Bio-Rad #186-4006
Sarkosyl, sodium salt solution	Sigma #L7414-50ML
SSC, 20X	Sigma #S6639
Transcriptor first strand cDNA synthesis kit	Roche

2.3 Primers

Primer	Sequence
Template Switch Oligo	AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG
SMART PCR primer	AAGCAGTGGTATCAACGCAGAGT
New-P5-SMART PCR hybrid oligo	AATGATACGGCGACCACCGAGATCTACACGC CTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*A*C
Custom Read-1 Primer	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC
MeDIP (+) Forward	TCTCCTTGCGGCATCTCTTA
MeDIP (+) Reverse	GGCGGTAAAGGGTGCTACTA
MeDIP (-) Forward	CTGGCACTGCACAAGAAGAT
MeDIP (-) Reverse	CACCATCCGGGTTCCTATAA

2.4 Instruments

Instrument	Company #order number
Agilant 2100 Bioanalyzer	Agilent technologies
Odyssey CLX imaging system	LI-COR
Bioruptor plus	Diagenode
Western blot gel running apparatus	Novex, Life technologies
BD FACS Aria III	BD Biosciences
Syringe pump, Legato 100	KD Scientific #788100
Magnetic Mixing system	VP Scientific #710D2
Mixing Disc	VP Scientific # 772DP-N42-5-2
Stand	VP Scientific #710D2-4
Needles 26G	BD Biosciences
Drop-seq microfluidic device	FlowJem
100 um, 70 um and 40 um cell strainer, Nylon, sterile	BD Biosciences
Fuchs-Rosenthal Hemocytometer	Incyto #DHC-F01

Chapter 3. DNA methylation changes in plasticity genes accompany the formation and maintenance of memory

3.1 Overview

The ability to learn and forming memory is vital for many organisms, in order to adapt and to cope with their environments. On a cellular level, these environmental factors stimulate a series of structural and functional changes in the nervous system cells, which leads to the formation and maintenance of memory. This event is also known as synaptic plasticity and could lead to change a neuronal cell response to its surroundings (Guzman-Karlsson, Meadows, Gavin, Hablitz, & David Sweatt, 2014). At the molecular level, learning and memory-related structural and functional changes are associated with learning-related genes such as Reelin, Bdnf, and Calcineurin and others (Zovkic, Guzman-Karlsson, & Sweatt, 2013). Although these genes are regulated at the transcriptional and translational levels, evidence suggests that epigenetic regulations also play a vital role in the expression of these genes (Lopez-Atalaya & Barco, 2014; Sweatt, 2013; Zovkic et al., 2013). Here we investigated the spatio-temporal changes in chromatin modification and their effects on the expression of learning and memory-related genes. The author was mainly involved in designing and performing validation studies for the results from the ChIP-Seq and MeDIP-seq experiments described in this chapter. This study is already published (Halder et al., 2015) but the results presented in this chapter are only those which are either directly obtained by myself or are necessary to explain the rationale behind my findings. All other details could be found in (Halder et al., 2015).

In this study, a learning paradigm known as contextual fear conditioning (CFC) was applied to wild-type three months old C56BL/6 (C56 black 6) male mice. Although, animals were not used for the test phase after the CFC training. In the CFC training session, animals were divided into two groups. One group was kept in a new cage for 180 Sec, which we refer to as context (C) group. Another group was kept in a new cage at the same time, however, last 2 Sec, a mild foot

shock of 0.7 mA was given, and refer them as Context-shock (CS) group. After training, the animals were returned to home cages. The third group of animals without CFC training were categorized as naive (N) controls. Furthermore, these three groups of animals were divided into sub-groups to do time point experiments. First groups of context and context-shock animals were sacrificed after 1 hour of the training. The second group of context and context-shock animals were home caged for four weeks and sacrificed afterward. An extra third group of context and context-shock mice was trained and tested after four weeks to confirm memory retrieval. All of the animal experiments were performed under the animal protection law and were approved by the District Government of Niedersachsen (Lower Saxony), Germany.

To identify cell type-specific (neuron and glia) epigenetic changes, which possibly associated with short and long-term memory formation and maintenance, BiTS (batch isolation of tissue-specific chromatin) -ChIP and -MeDIP was done after the behavioral experiments. Isolation of tissue was done as described in the method section (3.2.1) (Bonn, Zinzen, Perez-Gonzalez, et al., 2012). In this study, contextual fear conditioning (CFC) paradigm was used as a learning method because of its robustness and other known applications (Fanselow, 1990). The brain regions, hippocampal CA1 (Cornu Ammonis-1) involve in short-term memory; and ACC (anterior cingulate cortex) involve in associative memory acquisition and maintenance (Einarsson & Nader, 2012; J. Kim & Fanselow, 1992; Runyan, 2004). To further investigate memory formation molecular mechanisms during CFC, these brain regions were analyzed for the epigenetic changes. NeuN (+) neuronal and NeuN (-) non-neuronal nuclei were sorted using the fluorescence activated cell sorter (FACS) to generate cell type-specific samples. The overall experimental design is depicted in (Figure 3.1).

Cell type-specific nuclei samples were used for chromatin immunoprecipitation and sequencing (ChIP-seq) as well as methylated DNA immunoprecipitation and sequencing (MeDIP-seq) studies. The chromatin modifications, which were used in this study, are well characterized in previous studies (see Table 3.1). In brief, activity related histone marks (H3K4me3, H3K9ac, H3K27ac, and H3K79me3) are well correlated with active genes; H3K4me1 and H3K27ac are located on active enhancers; H3K27me3 and DNA methylation are linked to gene repression (Barski et al., 2007; Bonn, Zinzen, Girardot, et al., 2012; Zhou, Goren, & Bernstein, 2011).

32
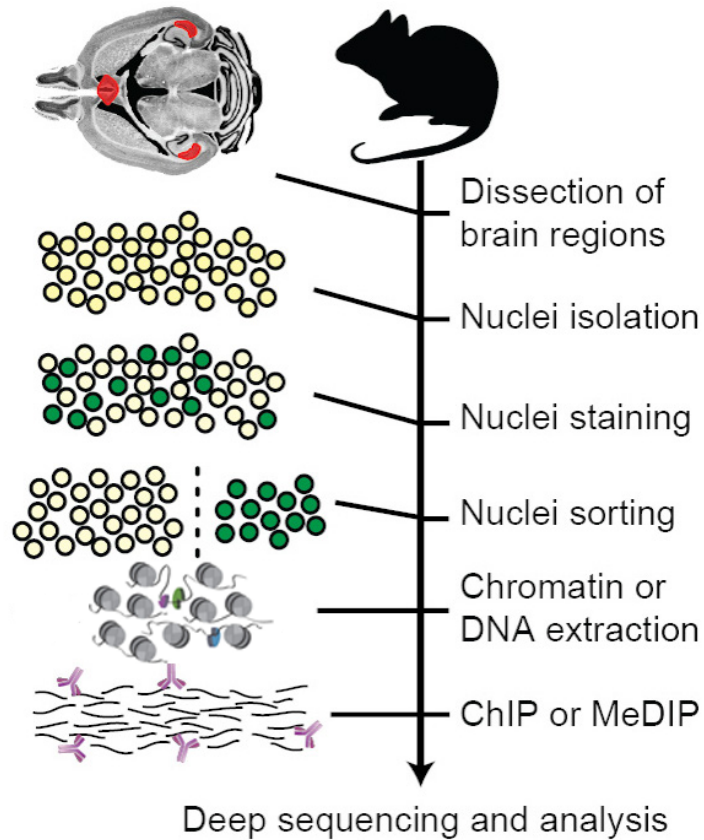
Figure 3.1: Experimental procedure.

The experimental design is depicted to explain the workflow to obtain cell type-specific (neuronal and non-neuronal) epigenetic data. This figure was initially been taken from (Halder et al., 2015) and modified to present in this thesis.

3.2 Methods

After the behavioral experiments and separating the animals into various combination of learning paradigm (N, C and CS), time-point (1 hour and 4 weeks), and cell type (NeuN+ and NeuN-), following histone modification (H3, H3K27ac, H3K27me3, H3K4me3, H3K79me3, and H3K4me1) and DNA methylation (MeDIP) experiments were done.

3.2.1 Tissue collection from mice after fear conditioning

Mice were sacrificed using the cervical dislocation method, and full brain was isolated in the chilled medium containing DPBS and EDTA-free protease inhibitor cocktail. From the brain, CA1 and ACC regions were dissected and stored in -80°C after being snap frozen in liquid nitrogen. For cell type-specific ChIP-seq and MeDIP-seq experiments, tissues from 20 mice were pooled for each replicate. In total, each treatment group was done on at least two replicates. Cell type-specific chromatin was extracted using the BiTS protocol for mouse brain tissue (Bonn, Zinzen, Girardot, et al., 2012)(Bonn, Zinzen, Perez-Gonzalez, et al., 2012). NeuN marker was used to differentiate between neuronal (NeuN +) and non-neuronal (NeuN-) cell types.

3.2.2 FACS sorting of cell type-specific nuclei

For nuclei FACS sorting, the detailed protocol could be found in the publication (Halder et al., 2015). In brief, Mice brain regions from 5 mice were taken out from -80°C freezer and homogenized using a micropestle in 500 ul of low sucrose buffer and cross-linked with 1% formaldehyde (Sigma) for 5 minutes at room temperature. After the incubation, 125 mM glycine was added and incubated for 5 minutes to quench the reaction. The nuclei were centrifuged, and the supernatant was removed. Nuclei were resuspended in 3 ml of low sucrose buffer (with protease inhibitors), and the solution was homogenized. Nuclei were passed through a sucrose cushion using centrifugation (3200g for 10 minutes) in Oak Ridge centrifuge tubes. Purified nuclei were resuspended in PBS and filtered through a 70um filter to remove clumps. The nuclei were stained using anti-NeuN mouse antibody (Millipore mab377) diluted 1:500 in PBST with 5% BSA and 3% goat serum. The nuclei were incubated for 30 minutes at 4°C; washed four times with PBST and incubated with secondary antibody anti-mouse Alexa 488 diluted (1:1000

34

in PBS) for 15 minutes. The nuclei were washed with PBST and stored in PBST with 5% BSA until the sorting. Before FACS sorting, nuclei were passed through a 26G needle (10 times) to dissociate the nuclei aggregates and filtered with 70 um pore size filter (BD Biosciences). Sorting was done on a FACSAria II (BD Biosciences), and sorted nuclei were collected into chilled conical tubes containing 1 ml of 5% BSA in PBS (Halder et al., 2015). On the basis of size and density of unstained nuclei, gate setting was determined for FACS sorting. NeuN positive and NeuN negative both fractions were collected in separate tubes. This method has yielded cell type-specific nuclei with high purity (>95%). NeuN positive fraction was containing primary excitatory neurons and interneurons while NeuN negative fraction had glial cells along with other non-neuronal cells (Halder et al., 2015).

3.2.3 Chromatin immunoprecipitation and sequencing (ChIP-seq)

ChIP protocol was optimized for low chromatin (0.5 to 1ug) input samples as described in the publication (Halder et al., 2015). ChIP-grade antibodies were used for this purpose, which was previously validated according to the Antibody Validation Database (Egelhofer et al., 2011). ChIP-seq experiments were only performed for 1-hour time point in CA1 (cellular consolidation) using detailed methods described in (Halder et al., 2015). In brief, FACS sorted nuclei were centrifuged at 3200g for 15 minutes, the supernatant was removed, and nuclei were resuspended in RIPA buffer. Nuclei were transferred into a fresh Diagenode tube and samples were sheared in a Diagenode Bioruptor plus apparatus. 4 times 5 cycles were run with 30 Sec ON/OFF setting with high power. Samples were centrifuged in between after every 5 cycles. The sheared chromatin was centrifuged, and the supernatant was aliquoted in the DNA low binding tubes (Eppendorf). Chromatin was stored in -80°C after liquid nitrogen snap freezing.

DNA was isolated from an aliquot of chromatin using the SureClean method. In brief, 10 ul of chromatin was added in 20 ul of EB buffer. In the tube, 1 ul of RNase A (50 ng/ul) was added, and the solution was incubated for 30 minutes at 37°C. After the incubation, 1 ul of proteinase K (20 mg/ml) was added and incubated at 65°C for 2 hours with 800 RPM shaking. 3 ul of the co-precipitant (LPA 5 mg/ml) and 1 volume of SureClean was added. The solution was vortexed and incubated at room temperature for 10 minutes. Tubes were centrifuged at 15000g for 20 minutes

at room temperature, and the supernatant was removed. The pellet was dried, resuspended in 15 ul EB buffer and DNA content was measured using Qubit dsDNA HS assay kit. The size of the DNA fragments was also checked on an Agilent 2100 bioanalyzer using a DNA high sensitivity bioanalyzer chip. This shearing process should produce an average size of 300bp. If shearing was not good, samples were sonicated for more cycles and process was repeated.

The chromatin samples were diluted in IP buffer (10 times) including Roche Complete protease inhibitors, and chromatin was pre-cleared using BSA blocked protein A magnetic beads (Dynabeads from Invitrogen). The solution was incubated for 1 hour at 4°C. After incubation, tubes were kept on a magnetic rack and supernatant was taken. Chromatin was split into aliquots for using them in different immunoprecipitation (IP) assay. One tube with 10% Input (10% of one IP) was also aliquoted. To prepare BSA blocked protein-A magnetic beads, required volume of beads were calculated (based on samples), and taken into a 1.5 ml Vial and kept on a magnetic rack. The supernatant was removed, and beads were washed with 1 ml IP buffer. After that, beads were incubated with a solution containing 1 ml IP buffer and 0.5% BSA for 2 hours at 4°C. Beads were rewashed two times with the IP buffer and resuspended in the same volume of IP buffer as the starting volume. BSA blocked beads could be stored at 4°C for 2-3 days.

The amount of chromatin was optimized for the immunoprecipitation assay with the different ChIP-grade antibodies as described in (Table 3.2) below and incubated on a rotating wheel at 4°C for overnight.

Ab specificity	Ab ID	Input chromatin/DNA [μg]	Ab [μg] per IP	DNA recovery [ng]
H3	Abcam ab1791	0.1	0.5	6 – 15
H3K4me1	Abcam ab8895	0.5	0.5	3 – 10
H3K4me3	Abcam ab8580	1	1	1 – 5
H3K27ac	Abcam ab4729	0.5	0.5	2 – 8
H3K9ac	Millipore 07-352	0.5	1 (μL)	1.5 – 3
H3K79me3	Abcam ab2621	0.5	0.5	2 – 5

H3K27me3	Millipore CS200603	0.5	0,25 (µL)	2 − 5

Table 3.1 Antibody concentrations.

Details of the amount of chromatin and antibodies required for ChIP assays, adapted from (Halder et al., 2015).

After the overnight incubation, 15 ul of BSA blocked protein-A magnetic beads were added to each sample, and the tubes were incubated on a wheel for 2 hours at 4°C. The beads were washed two times with 200 ul cold IP buffer with 0.1% SDS and three times with 200 ul cold wash buffer. For the last two washes, samples were incubated 10 minutes on a rotator at 4°C. Beads were again washed twice with 200 ul cold IP buffer and twice with 200 ul cold TE (without any protease inhibitors). After the last washing step, the supernatant was removed, and the beads were incubated in EB buffer (1mM Tris pH 8) with the RNase A solution (0.1 ug/ul) for 30 minutes at 37°C. The RNase A treatment was also done on the Input sample that was saved in the previous step. To perform the de-crosslinking, the beads and the input samples were added with 2X Weinmann buffer (WB) without inhibitors and tubes were incubated in the presence of 1 ul proteinase K (0.5 ug/ul) and 1% SDS at 65°C for overnight shaking (800 RPM). Tubes were kept on the magnetic rack, and the supernatant was transferred to a fresh DNA low binding tube (Eppendorf). To increase the yield, the beads were washed one more time using EB and tubes were incubated for 10 minutes at 65°C with 800 RPM rotation. The supernatant was added again to the previous DNA low binding tube. The DNA was isolated using SureClean precipitation method as described earlier in the presence of LPA (linear acrylamide). The DNA pellet was washed two times using 70% ethanol. All the ethanol was removed by drying and using speedvac for 3 minutes at room temperature. The DNA was resuspended in EB (Tris 10 mM, pH 8) and the concentration was determined using Qubit dsDNA HS assay. The immunoprecipitation efficiency was also validated using qPCR with positive and negative regions.

Library preparation condition was optimized for a low amount (0.5 ng) of input materials to generate reliable and quantifiable libraries, as published in (Halder et al., 2015). The Diagenode MicroPlex kit or NEBNext Ultra DNA library preparation kit for Illumina (NEB) was used for this purpose. After template preparation and adapter ligation, the number of amplification cycles

was determined using a qPCR on a small amount of IP material to avoid over-amplification. After the library amplification PCR, libraries were purified using the SureClean method as described earlier. After ethanol wash and drying, the pellet was resuspended in 10 mM Tris pH 8. ChIP libraries were validated by qPCR assay and concentration was measured by a Qubit dsDNA HS assay kit. The library size was determined by a DNA High sensitivity bioanalyzer assay using the manufacturer's protocol, and the library was adjusted to 2 nM concentration for the sequencing on a Hiseq 2000 (Illumina) according to the manufacturer's protocol.

3.2.4 Methylated DNA immunoprecipitation and sequencing (MeDIP-seq)

The MeDIP protocol was optimized for the reliable enrichment of methylated regions with as little as 0.1 ug of DNA material as described in (Halder et al., 2015). NeuN positive and NeuN negative FACS sorted nuclei were centrifuged at 3200g for 15 minutes at $4^{o}C$. The supernatant was removed, and nuclei were resuspended in 200 ul of lysis buffer and incubated at $65^{o}C$ for overnight. Genomic DNA was isolated using the Phenol-chloroform method. Using 0.3 M sodium acetate pH 5.2 and 400 ul of 100% ethanol DNA was precipitated. Samples were centrifuged at 20,000g for 20 minutes at room temperature, and the pellet was washed with 70% ethanol. After air-drying, DNA was resuspended in 100 ul TE buffer with 20 ug/ml RNase A and incubated at $37^{o}C$ for 30 minutes and subsequently at $65^{o}C$ for 1 hour. Shearing of the genomic DNA was done in an NGS Bioruptor (Diagenode) for 10 cycles (30 sec ON, 30 sec OFF). Samples were centrifuged after 5 cycles of shearing. DNA size was analyzed using the Bioanalyzer DNA high sensitivity ChIP. DNA average size of 250-350 bp was achieved, and 700 ng of sheared DNA was used for the further MeDIP procedure.

The methylated DNA IP was done following the same procedure described later (section 4.2.1.4) in this thesis with some modification. In Brief, the DNA was end-repaired and A-tailing was done using the kit NEBNext ChIP-seq library preparation master mix (NEB-E6240 kit) according to the manufacturer's instruction. Custom synthesized paired-end sequencing adapters for Illumina (Sigma-Aldrich) were ligated using the NEB kit. 100 ng of the adapter ligated genomic DNA was used for MeDIP using an antibody anti-5-methylcytosine (5mC) as described earlier (Proudhon et al., 2012). MeDIP efficiency was checked by qPCR using the primers listed in section 2.3.

Samples were also checked for optimum PCR amplification cycles by qPCR assay following the same method as described in section xx. MeDIP libraries were prepared using the PCR amplification of MeDIP and input samples following the NEB E6240 kit protocol (9 ul DNA sample, 0.4 ul custom TruSeq PCR primer with Index sequence, 9.4 ul 2X Phusion High fidelity PCR master mix). The libraries were further purified using AMPure XP beads and resuspended in EB buffer. MeDIP libraries were quantified using Qubit dsDNA HS assay, and size analysis was done using DNA high sensitivity bioanalyzer method. Libraries were diluted to 2 nM for the sequencing in Hiseq 2000 machine.

3.2.5 CRMs validation in a zebrafish model

Biological validation has been done for the predicted *Cis*-regulatory modules (CRMs) described in (Halder et al., 2015) using the Zebrafish Enhancer Detector (ZED) vector system (Bessa et al., 2009). For this experiment, the wild-type strain of zebrafish (AB) was used. All the embryos were kept at the 28.5^{o}C in E3 media along with 10^{-5} % methylene blue and prepared according to (Kimmel, Ballard, Kimmel, Ullmann, & Schilling, 1995). These experiments were performed in Biomedical Center, Ludwig Maximilians University Munich, and accordance with animal protection standards of LMU and were approved by the government of Upper Bavaria (Regierung von Oberbayern, Munich, Germany).

From the 60,544 predicted enhancer regions, 30 sequences were chosen randomly for the validation assay (15 for CA1 and 15 for ACC). Based on the conserved regions between mouse and zebrafish 300 bp regions were selected. These enhancer sequences were synthesized and flanked by attL sites and cloned into the pMK-RQ vector from GeneArt (Halder et al., 2015). By employing LR clonase system (Gateway) enhancer sequences were cloned into the ZED vector, which had attR sites before the gata2a promoter sequence responsible for expression of GFP (Bessa et al., 2009). By checking loss of BglII site by digestion, integration of the enhancer sequence was confirmed.

The ZED vectors with the enhancer sequence for validation were injected in the zebrafish at the one cell stage with the concentration of 25 ng/ul plasmid (2-4 pl/cell). Eggs with plasmid and

controls were incubated at 28°C in E3 buffer before analysis. The zebrafish larvae (male and female) were checked for transient expression by dsRed reporter (somites) and cloned enhancer driven GFP expression after 2 and five days of fertilization (Halder et al., 2015). Using Tricaine (0.016% w/v) larvae were anesthetized and mounted on coverslips using 3% methylcellulose. Samples were analyzed using LSM710 META inverted confocal microscope (Zeiss), and fluorescent images were visualized in Photoshop 8.0 (Adobe Systems) as described in (Halder et al., 2015). Injected enhancer element from neuronal shaped and GFP expressing cells, which were present in the central nervous system were assigned as a neuronal enhancer. Further, the enhancer element which was expressing GFP in non-neuronal cells (in at least 20 injected embryos) were described as a non-neuronal enhancer.

3.2.6 Chromatin immunoblotting

To detect the global changes for the 6 HPTMs investigated in this study, immunoblotting was done on a whole CA1 region as well as using cell type-specific chromatin. For isolating CA1 regions, five mice per biological condition (Naive, Context and Context-shock in the CA1 at 1 hour) three months old C56BL/6 male mice were used. For the immunoblotting analysis of cell type-specific chromatin, the material was used from the above mentioned 20 mice pooling for each replicate (Halder et al., 2015). The CA1 and ACC regions of the mouse brain were processed for the nuclear protein enrichment. In brief, tissues were homogenized in 200 ul TX buffer and protease inhibitor using a micro-pestle. Samples were incubated on a rotating wheel for 10 minutes and centrifuged for 10 minutes at 400g. The supernatant was discarded, and the pellet was washed once with TX buffer in the presence of protease inhibitors. Samples were lysed in TX buffer with 1% SDS by 5 minutes incubation on a rotating wheel. The samples were then sheared in a Bioruptor (Diagenode) for 3 times 5 cycles (30 Sec ON/OFF). Samples were centrifuged for 10 minutes at 9300g and the supernatant, which contains enriched nuclear proteins, was collected. The Pierce BCA Protein assay kit was used to measure the protein concentration following the manufacturer's instruction in a clear bottom 96 well micro-plate. Absorbance was measured using the TECAN plate reader at 562 nm wavelength. After normalizing the samples for western blot studies, samples were denatured by incubating at 95°C for 5 minutes. 4 ug of proteins were used to run in a 12% Bolt Bis-Tris pre-casted gel, in reducing

conditions maintained by MES-SDS buffer. Samples were running for 25 minutes at 200V in western blot assembly. After that, proteins were transferred to nitrocellulose membrane (0.2 um pore size) in cold transfer assembly running for 90 minutes at 50V in cold 1X Tris-glycine transfer buffer with 20% methanol (Halder et al., 2015). The nitrocellulose membrane was incubated in the TBST buffer (TBS with 0.1% Tween 20) with 5% BSA for 1 hour at the room temperature for blocking the membrane. Protein containing membranes were incubated with primary antibodies (as described in the table below) overnight at 4°C.

Chromatin marks	Antibody	Dilutions
H4	Abcam ab31830	1/5000
H3	Abcam ab1791	1/10000
H3K4me3	Abcam ab8580	1/1000
H3K27ac	Abcam 4729	1/1000
H3K9ac	Millipore 07-352	1/1000
H4K12ac	Millipore 07-595	1/1000
H4K5ac	Millipore 07-327	1/1000
H3K27me3	Abcam ab6002	1/1000

Table 3.2: Antibody dilutions

After the overnight incubation, the membranes were washed three times in TBST buffer and incubated with the secondary antibody (IRDyeR, LI-COR with 1/10000 dilution) for 1 hour at room temperature (Halder et al., 2015). The membranes were washed in TBST buffer three times, and imaging was done using an Odyssey CLX imaging system. Images were taken with 'high quality' settings with the resolution of 84 um in 700 nm, and 800 nm channels and the image processing and signal quantification was done using Image Studio software from LI-COR (Halder et al., 2015). Using similar methods, immunoblotting studies were done using sheared chromatin from BITS applied, and FACS sorted nuclei as described in the previous section. The chromatin was diluted in RIPA buffer with 0.1% SDS and samples were denatured for 10 minutes at 99°C (Halder et al., 2015). Samples were mixed with loading buffer, and further western blotting was done similarly as described for the nuclear protein lysate using 100 ng of chromatin per well (Halder et al., 2015).

3.2.7 ChIP-, MeDIP-seq data analysis

Detailed computational analysis methods could be found in the publication (Halder et al., 2015). Following sections, describe the analysis workflows, which were relevant to this thesis.

3.2.7.1 Pre-processing and quality control of sequencing data

The data from the ChIP-seq and MeDIP-seq studies were processed through a quality control workflow. Read quality was analyzed using the *FastQC* package (v0.10.1) to assess sequencing cycles with low coverage, contaminations such as adapter sequences or repetitive sequences from PCR artifacts and so on (Andrews, 2010). Alignment quality was checked using samtools *flagstat* (v0.1.18) with default parameters (Li et al., 2009). Data quality was also visualized using an in-house developed genome browser at https://oasis.dzne.de/share/JBrowse-1.11.4/index.html (Halder et al., 2015). Samples were also checked for optimum depth of sequencing by checking for average per base coverage and the saturation correlation for all the samples by employing MEDIPS R package (Lienhard, Grimm, Morkel, Herwig, & Chavez, 2014). The biological correlation between replicates was analyzed using Pearson correlation (function MEDIPS.correlation). The samples were plotted in respect to the variance accounted for the first two principal components (PCs) to make sure that samples cluster according to their corresponding biological groups (brain region, cell type and chromatin modification). For ChIP-seq data, peak enrichment was analyzed by calculating normalized strand cross-correlation (NSC), and relative strand cross-correlation (RSC) coefficients using the in-house developed R package 'chequeR' (Halder et al., 2015). Good libraries have an NSC score >1.05 and an RSC score > 0.8. The data samples, which passed the quality control parameters were further used for downstream analysis (Halder et al., 2015).

3.2.7.2 Read alignment

The reads from ChIP-seq experiments were aligned to the mouse NCBI genome (v38) using Bowtie2 (v2.0.2) (Langmead & Salzberg, 2012). Reads were first aligned by default parameters using seed alignment with two mismatches allowed (Halder et al., 2015). For true multi-map reads, which were aligned to multiple regions with the same score, only a single alignment was returned (Halder et al., 2015). The resultant Sequence Alignment/Map (SAM) files were converted into sorted Binary Alignment/Map (BAM) files by samtools functionalities (Li et al.,

42

2009). Furthermore, reads were filtered either to high quality uniquely, and multi-mapped reads or to good quality uniquely mapped reads by using bowtie2 MAPQ scores. This step had removed all the reads, which were aligned to multiple genomic locations with the same score (Langmead & Salzberg, 2012).

The MeDIP-seq data was also aligned to the mouse genome NCBI (v38) using Bowtie2 (v2.0.2) (Langmead & Salzberg, 2012). Same as ChIP data workflow, reads were first aligned using default parameters and using seed alignment with two mismatches allowed (Halder et al., 2015). Bowtie2 first looks for end-to-end 0-mismatch alignments and end-to-end 1-mismatch alignment. Later, it performs a seed based 2-mismatch alignment (Halder et al., 2015). The resultant SAM files were converted into BAM files using the samtools (Li et al., 2009). Furthermore, reads after the alignment was filtered for high quality unique and multi-mapped reads using MAPQ score (Halder et al., 2015).

The BAM files were converted into WIG and BigWig file formats to data visualization purpose using MEDIPS R package function (MEDIPS.exportWIG) with the 50bp window size and RPM normalization (Halder et al., 2015). The BigWig files were also uploaded to the in-house developed custom browser at https://oasis.dzne.de/share/JBrowse-1.11.4/index.html (Halder et al., 2015).

3.2.8 Cell type-specific gene list

To analyze the cell type-specificity of the ChIP-seq and MeDIP-seq data, neuronal and non-neuron-specific genes were manually searched and categorized using publically available datasets (Cahoy et al., 2008; Ko et al., 2013). Cell type-specific genes were further used in the analysis of aggregate gene plots, precision and recall calculations and genome-wide prediction of cell type-specific gene lists (Halder et al., 2015).

3.2.9 Prediction of Cis-regulatory modules (CRMs)

Cell type-specific active CRMs the were identified by employing a modified random forest classifier (RFECS) along with ChIP-seq data from Naive mouse group (Rajagopal et al., 2013).

The RFECS uses a supervised machine learning technique, which needs a training dataset to identify parameters for use in the test datasets. The RFECS training dataset was built from a published primary neuronal cell culture study (T. K. Kim et al., 2010; Lienhard et al., 2014). From this study, transcription factors (CREB, CBP, and Npas4), and chromatin marks (H3K4me1, H3K4me3, H3K27ac, and H3K27me3) NGS data was acquired (Halder et al., 2015). To create a positive CRM set, genomic regions of the mouse were selected which were occupied by at least one transcription factor (Npas4, CBP or CREB) and should contain a chromatin mark peak either from H3K27ac or H3K4me1 (Halder et al., 2015). In detail, transcription factor binding sites were downloaded in bed format using the UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) and converted into mm10 coordinates. Chromatin modification peaks were obtained by using MACS2 for H3K27ac and H3K4me1 and extended by 1000 bases in each direction (Halder et al., 2015). For creating a negative set of CRM genomic locations, H3K4me3 peaks (active promoters) and known transcription start sites (TSS +/- 500bp) were combined. From these regions, transcription factor binding sites were removed, and the final negative CRMs were selected randomly from 10% of the promoter regions and remaining random regions as described in (Halder et al., 2015).

The RFACS was trained with positive and negative CRM datasets and with neuronal histone data from the naive mice ACC brain region (Halder et al., 2015). For this analysis, a window size of 2000 bases was used to generate a prediction model containing 65 trees (Rajagopal et al., 2013). To validate the predicted CRMs, the coordinates were overlapped with H3K27ac and H3K4me1 peaks. CRMs overlapped H3K4me3 and transcription start sites were labeled as false positives and later used in sensitivity or specificity analysis (Halder et al., 2015). The annotation of Cis-regulatory modules with genomic regions was performed by comparing the CRM locations bed file to genome annotation in a format using the bedtools intersection function as described in the (Halder et al., 2015) in detail (Quinlan & Hall, 2010). In addition to the CRM annotations, CRM motif enrichment was discovered by employing Homer tools (http://homer.salk.edu/homer/chipseq/). Neuronal and non-neuronal CRMs were compared by respective motif enrichment, and transcription factors for the enriched motifs were merged according to their similarities (Halder et al., 2015).

3.2.10 Prediction of novel cell type-specific neuronal and non-neuronal genes

Individual chromatin modification information from the Naive mouse neuronal and non-neuronal data (CA1 and ACC) was used to predict novel cell type-specific gene expressions. MeDIP-seq data could not be used for this purpose because of its low classification applicability. To compare neuronal and non-neuronal chromatin modifications, a matrix file containing read information corresponding to the transcription start sites or gene bodies was created. To identify statistically significant differential coverage for chromatin marks in neuronal and non-neuronal data, DEseq2 was used (Love, Huber, & Anders, 2014). The 1500 bp region upstream and downstream around the TSS was taken for H3K27ac and H3K4me3 comparison in ChIP-seq data (Halder et al., 2015). Furthermore, the full gene body was used for the chromatin modification in H3K79me3, H3K27me3 and H3K4me1 as described in (Halder et al., 2015). Following parameters were used for the filtration of the results: an FDR < 0.05, a |logFC| > 1 and a mean coverage > 50 (Halder et al., 2015). In the DEseq2 comparisons, neuronal and non-neuronal datasets were considered as treatment and control respectively. The activity related histone modifications show a positive correlation with the gene activity. For prediction of cell type-specificity a heuristic method best out-of-three classifier was used. For example, a certain gene to be annotated as a neuronal gene, at least three chromatin marks were statistically significantly enriched (or down-regulated for H3K27me3) in neuronal as compared to non-neuronal data (Halder et al., 2015).

3.2.11 Functional gene enrichment analysis

For the functional enrichment studies, the online version of WebGestalt and QIAGEN's tool Ingenuity Pathway analysis (IPA) was used as described in detail in (Halder et al., 2015).

3.3 Results

3.3.1 Cell type-specific epigenetic data

ChIP-seq and MeDIP seq data showed the high quality of the cell type-specificity of the data assessed by replicate correlation, peak enrichment rates, and other quality measures (Halder et al., 2015). Qualitative profiling of histone marks and DNA methylation of neuronal and non-neuronal genes were also showed cell type-specificity of the samples (Figure 3.2). For example, the neuronal gene Camk2a histone profiles have high levels of H3K4me3, H3K27ac, H3K9ac, H3K4me1 and H3K79me3 in neurons and as expected these histone marks were mostly not present in non-neuronal cells (Figure 3.2). Similarly, gene repression marks such as DNA methylation and H3K27me3 showed more signals in non-neuronal cells and fewer signals in neurons (Figure 3.2). Furthermore, cell type-specificity could also be confirmed on other neuron-specific genes such as Camk2b in the CA1 tissue (Figure 3.3 [a]) and the same gene in the ACC region (Figure 3.3 [b]). On the other hand, chromatin profiles were also confirmed in non-neuronal genes for instance, in Mag (Figure 3.3 [c]) and in Oaf (Figure 3.3 [d]).
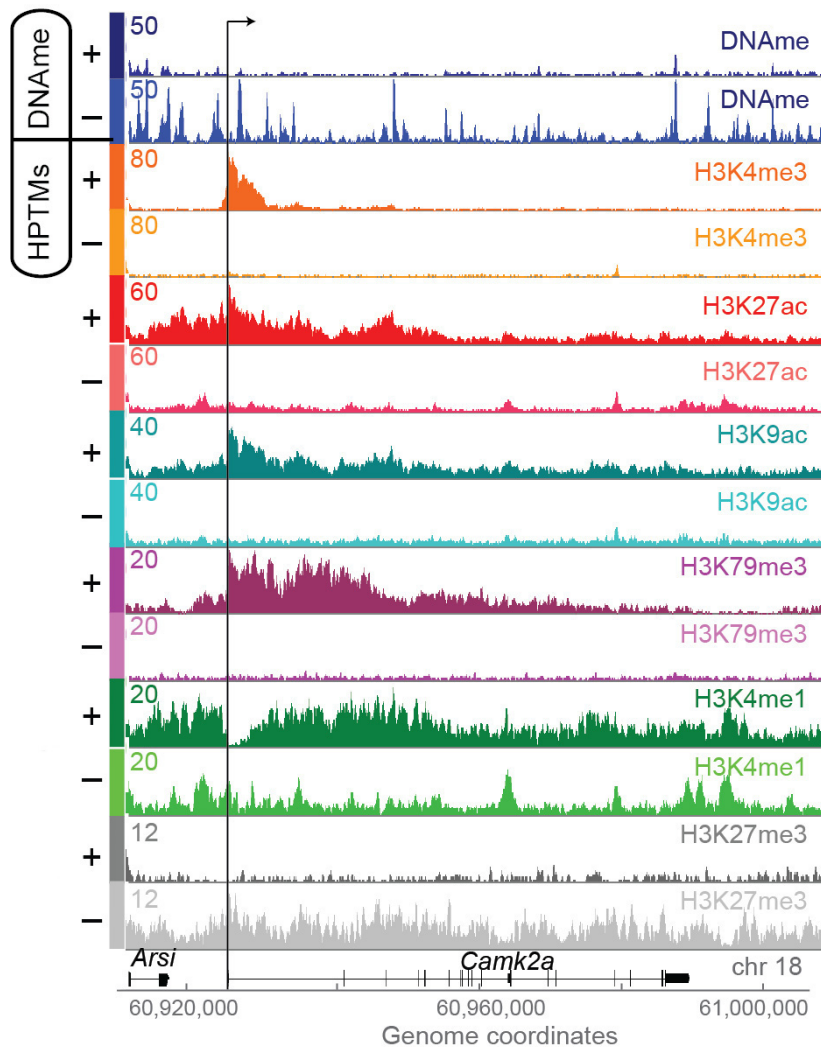
Figure 3.2: Genome browser (IGB) image for cell type-specific chromatin modification.

DNA methylation and histone posttranslational modification (HPTMs) enrichment on the neuronally expressed Camk2a gene locus is shown here. Arrows at the promoter region indicate the direction of the transcription. NeuN positive neuronal data (+) showed significant enrichment for activity related histone modifications (for example, H3K4me3, H3K27ac, H3K79me3) whereas, NeuN negative non-neuronal data (-) showed no enrichment for these histone modifications at Camk2a gene locus. This figure corresponds to Figure 1b in (Halder et al., 2015).

47
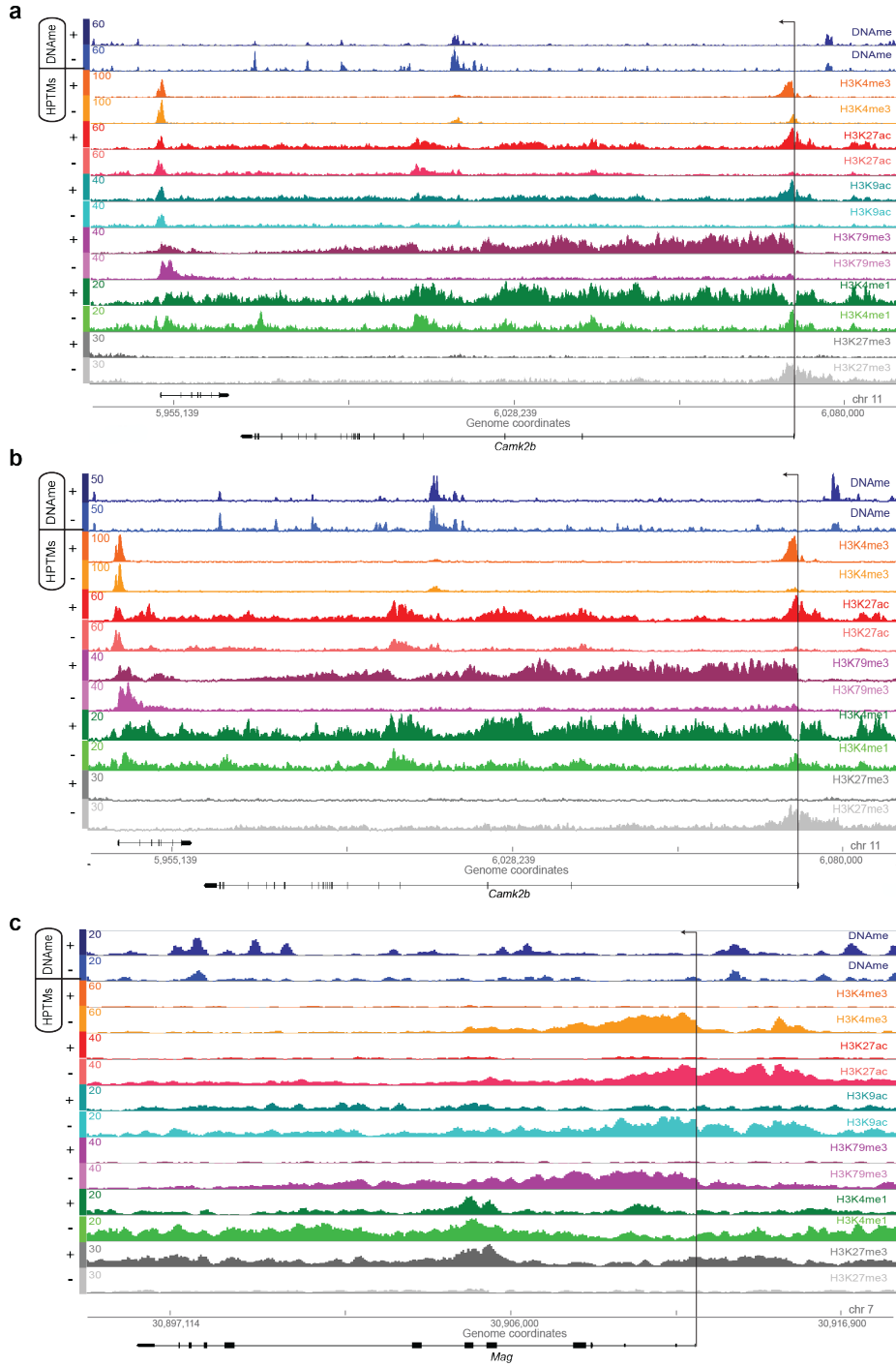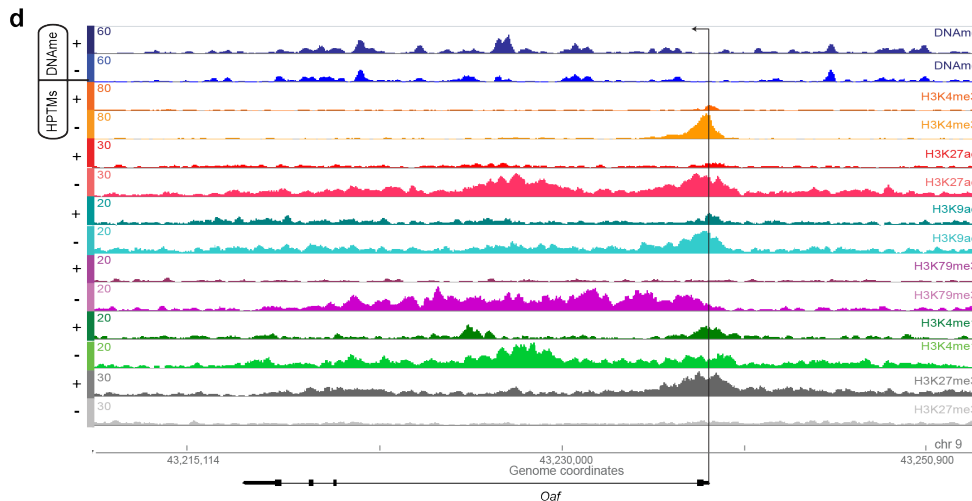
Figure 3.3: Histone modifications for cell type-specific genes.

DNA methylation changes and histone modifications for the known representative genes, which are only expressed in neurons (panel a and b), and the representative genes, which only expressed in non-neuronal (glia) cells (panel c and d) under the Naive conditions. (a) CA1 chromatin modifications for the known neuron-specific gene Camk2b are shown. (b) At the Camk2b gene locus, chromatin modifications were shown for the ACC brain region. The ACC H3K9ac mark was not included in this analysis. (c) Chromatin modifications for non-neuronal (glia) specific gene Mag is shown here. (d) Another non-neuronal specific gene Oaf chromatin profiles were depicted here. Gene expression activity marks such as H3K4me3, H3K9ac, H3K27ac and H3K79me3 and repressive marks H3K27me3 and DNA methylation showed good purity of the cell type-specific samples. This figure corresponds to Supplementary figure 5 in (Halder et al., 2015).

3.3.2 Prediction and validation of cell type-specific regulatory modules

The HPTMs were proved to be efficient to classify known cell type-specific genes (with an average precision of 88% and 69% of a recall). Therefore, HPTMs were used for the prediction of previously unknown cell type-specific genes and other *cis*-regulatory modules (CRMs) such as enhancers and other functional genomic regions (Halder et al., 2015). As described in the method section 3.2.9, using heuristic best out of three classification method, 1,647 novel neuron-specific genes (94% precision; 37% recall) and 803 novel non-neuron-specific genes (100% precision; 31% recall) were predicted (Halder et al., 2015). The predicted genes showed cell type-specific HPTMs and DNAme changes and enriched for their respective cell type related gene ontology (GO) terms.

In addition to the cell type-specific genes prediction, a random forest classifier was trained on a published set of positive and negative transcription factor binding sites (TFBS) and HPTMs in order to predict neuronal and non-neuronal CRMs (T. K. Kim et al., 2010; Malik et al., 2014). This method predicted a total of 60,544 CRMs (87.9% accuracy) in neuronal and non-neuronal cells (Halder et al., 2015). To validate biological activity of predicted CRMs using this model, 30 predicted neuronal CRMs were randomly selected, and their biological activity was tested in a zebrafish model using a reporter assay as described in the method section 3.2.5. 28 out of 30 CRMs were cloned, and 27 (96%) were detected positive for neuronal activity *in vivo* assay (Figure 3.4) (Halder et al., 2015).
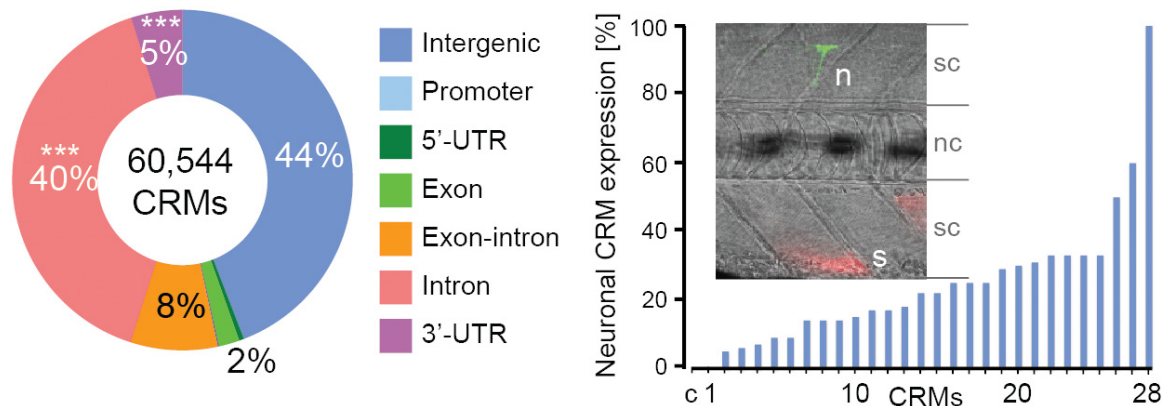
Figure 3.4: Validation of *cis*-regulatory modules.

Representation of genomic distribution of predicted CRMs in neuronal and non-neuronal data (left). 40% of all predicted enhancers were present in introns, which is known to have active enhancer sites. The right side of the figure represents the percentage of GFP positive zebrafish embryos, which showed the neuronal CRM activity. 28 predicted neuronal CRMs (Out of 30) were positive for this enhancer reporter assay (methods). The negative control (c) and CRM-14 (1) were not positive for neuronal activity. Inset, (n) is GFP positive neuronal cell image and (s) is expression of control in muscle cells (s); sc (spinal cord); nc (notochord). A permutation-based test was used for the estimation of significance (***P < 0.0001). The figure corresponds to Figure 2b and 2c in (Halder et al., 2015).

3.3.3 Learning-induced HPTMs and DNAme changes

To understand the role of epigenetic changes in memory formation and consolidation, DNA methylation and HPTM changes were measured after 1-hour and 4-weeks of the contextual fear conditioning. In this analysis, hippocampal CA1 and the ACC brain regions were investigated. Differential methylated region (DMRs) computational analysis was done as described in (Halder et al., 2015) using MEDIPS package. The differential methylation regions (DMRs) are well correlated with the spatio-temporal changes of associative memory (C-CS) as changes were mostly present in short-term memory associated brain region CA1 at 1 hour and in long-term memory related ACC region at 4 weeks in neurons (Halder et al., 2015). In total, 1137 DMGs (3216 DMRs) were changed in CA1 neurons during cellular consolidation (1h), but no DNA methylation changes could be identified during memory maintenance (4w) in CA1 (Figure 3.5[a]). On the other hand, in the ACC, no DMGs were identified during systems consolidation (1h), and 153 DMGs (365 DMRs) were detected in memory maintenance (4w) in neuronal cells (Figure 3.5[b]) (Halder et al., 2015).

Along with the DNA methylation modifications, HPTM changes during learning events were also investigated using immunoblot (IB) analysis. In contrast to the DNAme changes, no early changes in bulk HPTMs could be identified by immunoblot study of CA1 and ACC brain regions in any of the analyzed conditions (Figure 3.6). These results were also consistent with the immunoblot analysis of cell type-specific chromatin of the CA1 region (Figure 3.7) (Halder et al., 2015). To increase sensitivity and specificity to detect global changes in the HPTMs, the ChIP-seq method was used as described in the method section. Average gene intensity profiles (aggregate gene plots) showed a global increase in the activity related histone marks (such as H3K4me3 and H3K9ac), and a global decrease in the inactivity related histone mark (H3K27me3), during the cellular consolidation process (CA1 1h; N-C and N-CS) in genic regions, as shown in Fig3a and 3b in (Halder et al., 2015). On the other hand, in the intergenic regions, aggregate intensity profiles showed a global decrease in H3K4me3 and a global increase in H3K27me3 levels, compensating for the observed changes in the active areas (Halder et al., 2015).
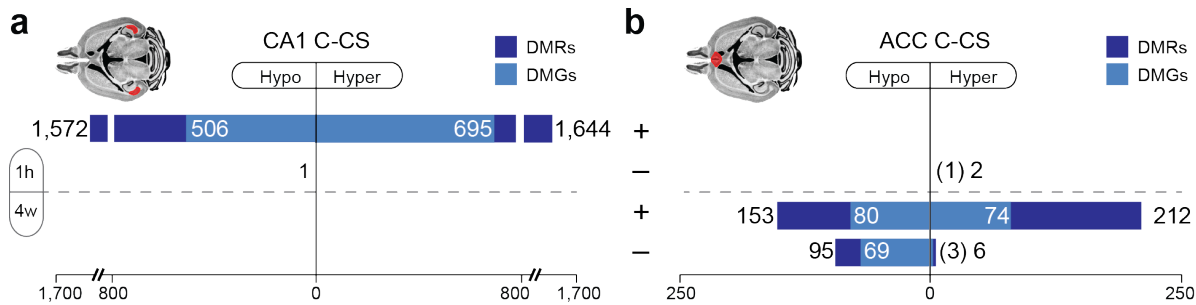
Figure 3.5: Analysis of DMRs and DMGs in CA1 and ACC brain regions.

Differentially methylated regions (DMRs) and differential methylated genes (DMGs) for the associative memory-related changes (C-CS) in neuronal (+) and non-neuronal (-) cells after contextual fear conditioning. DMRs are represented with dark blue and DMGs are with light blue bars in the figure. Total hypo- and hyper-methylated regions (black numbers) and genes (white or black in parenthesis) were represented. (**a**) Associative memory-related changes (DMRs and DMGs) in the brain hippocampal CA1 region at the 1 hour and 4 weeks time point after contextual fear conditioning. In the data analysis, DMGs were mainly present in the neurons at 1 hour after learning. (**b**) Associative memory-related changes (DMRs and DMGs) in the brain cortical ACC region at the 1-hour and 4 weeks after contextual fear conditioning. In contrast to the CA1, hypo- and hyper-methylation changes (DMGs) in the ACC region were mainly in the 4 weeks after CFC. Apart from the neuronal changes, there were few changes (69 hypo and 3 hyper-methylated) DMGs in non-neuronal cell fraction as well. This figure corresponds to Figure 5 in (Halder et al., 2015).
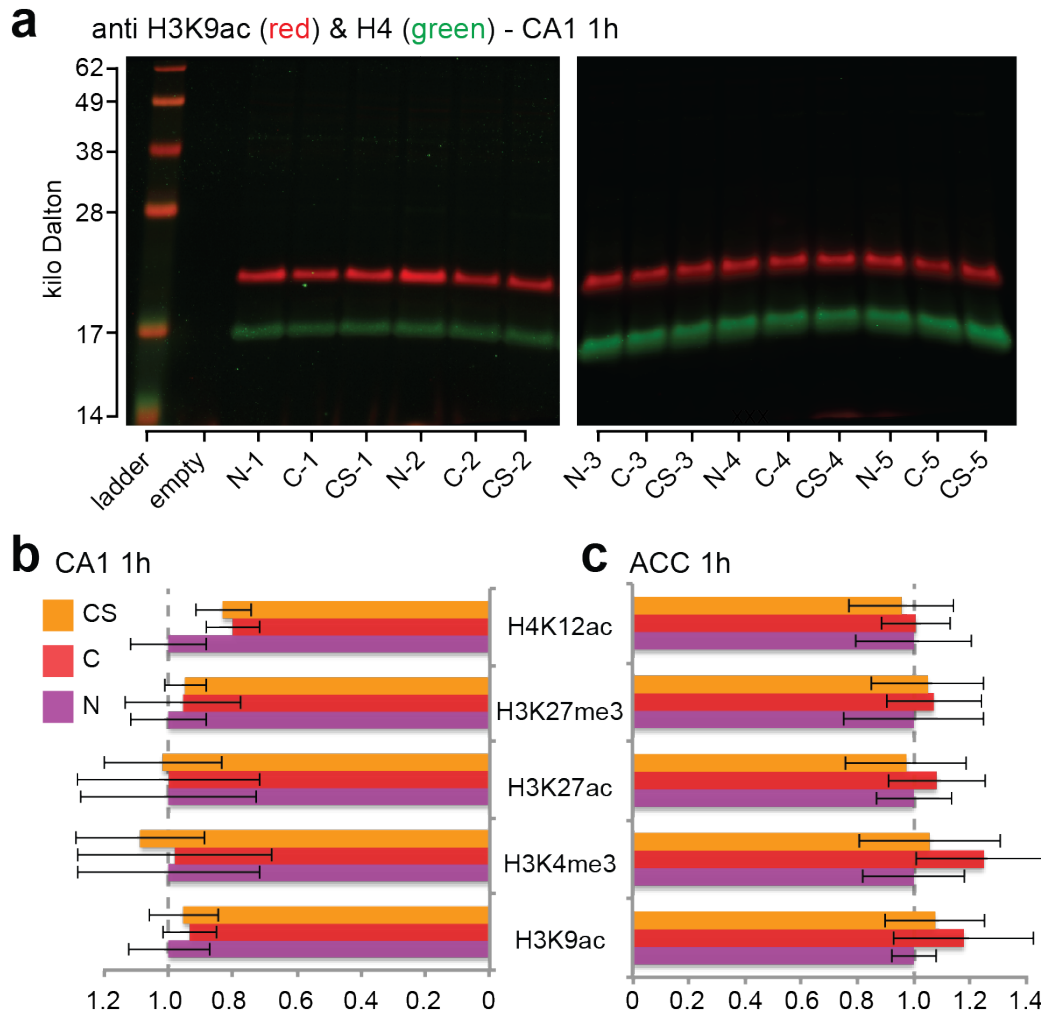
Figure 3.6: Immunoblotting analysis of HPTMs changes.

In the CA1 and ACC naive samples (N-1 to 5), context (C-1 to 5) and context-shock (CS-1 to 5) in mice 1 hour after contextual fear conditioning. Immunoblotting studies were performed using antibodies against H4K12ac, H3K27me3, H3K9ac, H3K4me3 and H3K27ac were used on the total tissue lysate as described in the method section in this thesis. Immunoblotting analysis of H3 and H4 were used as a loading control. (a) Representative image of immunoblotting with H3K9ac (red bands) and loading control H4 (green bands) in CA1 1-hour after contextual fear conditioning. (b and c) Quantification of the immunoblot signal for the different HPTMs. Different color of the bars indicates the different biological conditions (naive in purple, context in red and context-shock in yellow). All the HPTMs were first normalized to H4 (H3 for H4K12ac) levels and later to the Naive mice (N) HPTMs level of the corresponding histone mark. The error bars represent the standard deviation. In immunoblot analysis, no significant HPTMs changes could be identified at the 1-hour after CFC in CA1 and ACC regions of the brain. This figure corresponds to supplementary figure 14 in (Halder et al., 2015).
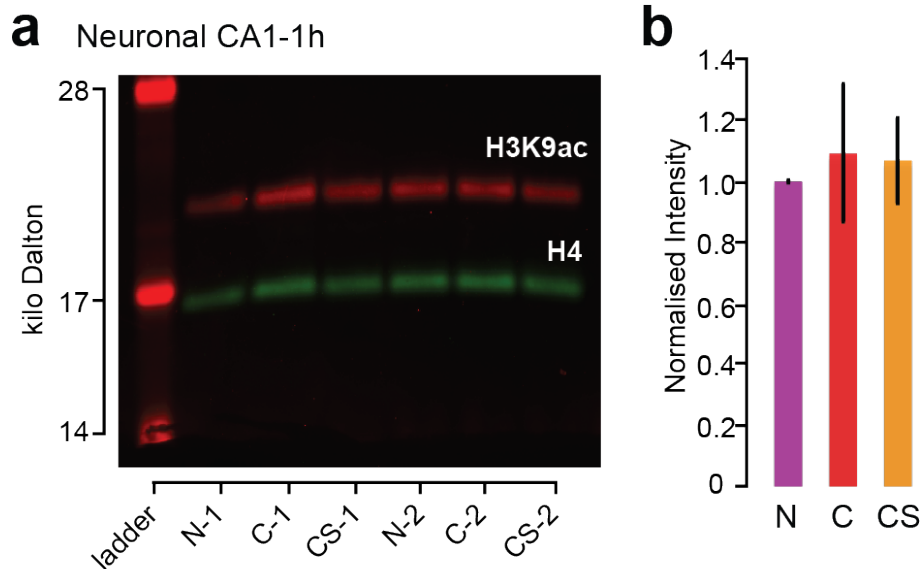
Figure 3.7: **Immunoblot analysis of neuronal chromatin from the CA1 region.**

From the CA1 region, neuronal chromatin for naive (N-1 and 2), context (C-1 and 2), context-shock (CS-1 and 2) mice 1-hour after contextual fear conditioning were used for immunoblot studies. Antibodies against the H3K9ac and H4 were used in the immunoblotting. A pool of 20 mice chromatin was used for each biological replicates. **(a)** Levels of H3K9ac (red band) and H4 (green band). **(b)** Quantification of immunoblot signals. H3K9ac levels were first normalized to their respective H4 levels and further with the level of the naive (N) mice. The error bars represent the standard deviation value (SD). In this analysis, no significant HPTM changes could be detected in H3K9ac levels. The figure corresponds to the supplementary figure 15 in (Halder et al., 2015).

3.4 Discussion

Learning and memory formation activities happen in the brain. Different brain regions are responsible for learning new information and its storage so that it can be retrieved later whenever needed. At the cellular level, neurons form a connection with the other neuronal cells, known as a synapse, to store the information. The adult human brain contains trillions of synapses. The external stimuli or learning event can fine tune as well as alter these synapses, called synaptic plasticity. At the molecular level, these structural changes are dependent on intracellular signaling networks, which regulate the expression of genes and protein synthesis. In the current chapter, we have mainly investigated chromatin modifications, which can have a role in learning and memory processes. Results showed that 5-methylcytosine DNA methylation levels correlate well with the spatio-temporal regions of the brain (Halder et al., 2015). This study highlights the possibilities of DNA methylation as a molecular correlate for the short and long-term memory.

In this chapter, after the neuronal and non-neuronal cell separation using BiTS protocol, chromatin samples were used for ChIP- and MeDIP-seq analyses. In these analyses, we looked at histone profiles for activity related marks such as H3K4me3, H3K27ac, H3K9ac, H3K4me1, and H3K79me3 as well as repressive histone mark H3K27me3. A closer investigation of these histone profiles made it possible to categorize cell type-specific genes for neuronal and non-neuronal cells. Based on the performance of data from histone post-translational modification (HPTMs; 88% precision, 69% recall) previously unknown cell type-specific genes and other *cis*-regulatory modules such as enhancers were predicted. The biological validation of the predicted enhancers in a zebrafish model system using a reporter assay confirmed the activity of the newly predicted neuronal enhancers in an *in-vivo* assay with 96% accuracy. These results demonstrate the potential of the histone marks to find cell type-specific genes and enhancer elements. Apart from the cell type-specificity, this study was also designed to understand the epigenetic changes (HPTMs and DNA methylation) associated with learning and memory formation as well as consolidation by employing an associative memory-related learning paradigm (fear conditioning). Results from these experiments suggest that neuronal cell type-specific differentially methylated regions (DMRs) correlate strongly with the spatio-temporal changes of associative memory (Halder et al., 2015). DNA methylation changes were mainly present in the short-term memory brain region CA1 at 1 hour and in the long-term memory associated brain

region ACC at 4 weeks in neuronal cells. Furthermore, average gene intensity profiles based on histone marks spread across the genome showed a global increase in the gene activation-related histone marks such as H3K4me3 and H3K9ac and a global decrease in the gene repression-related histone mark H3K27me3, during the cellular consolidation process in CA1 at 1 hour (Halder et al., 2015). In previous developmental studies, HPTMs showed a strong correlation with gene expression (Barski et al., 2007; Bonn et al., 2012; Zhou et al., 2011). However, in the current study, the analyzed HPTMs showed changes globally but very few regions specific changes. While the average gene intensity profiles for HPTMs corresponded to the gene expressions as expected (activating genes show gene activation and repressing genes show gene repression), the deregulated HPTMs between context-shock and naive mice did not correspond to the deregulated genes in the same conditions. In other words, HPTM changes seemed to be largely decoupled from differential gene expression (H3K79me3 was an exception) during learning (Halder et al., 2015). Our paper suggested some options to address this observation. (1) 1 hour is too long to capture short-pulsed HPTM changes, and (2) as this is the first cell-type-specific analysis of context shock vs. naive mice, maybe there are just not many individually deregulated HPTMs on the cell-specific level.

Similarly, there were few DMRs present in the non-neuronal cells in the ACC at 4 weeks time point. Even though it is highly unlikely that non-neuronal cells play a role in memory formation or consolidation, more experiments need to be done in order to determine whether it is actually true and if so, to understand this phenomenon. These observations suggest that DNA methylation specifically plays a role in memory formation and consolidation in the brain (CA1 and ACC) regions and correlate well with the spatio-temporal changes in memory. One shortcoming of this study is that it uses a large brain tissue. In other words, memory-forming stimuli, such as fear conditioning, stimulate changes in only a few neuronal networks in the brain regions. Merging of stimulated neurons with surrounding un-stimulated neuronal cells could dilute the actual gene expression and epigenetic signals to a great extent. Although, isolation of only activated neuronal cells is not a trivial task. One such approach could be to isolate neuronal cells based on immediate early gene (egr-1, c-fos, and Arc) expression using the FACS. These cells could provide high-resolution information about the molecular and epigenetic mechanisms underlying learning and memory.

56

Chapter 4. Cell type-specific molecular analysis using the Tagger system

4.1 Overview

Many neurodegenerative diseases affect only specific regions of the brain. This phenomenon is termed as selective vulnerability (Mattsson et al., 2016). At the cellular level, the interplay of sub-cellular factors (mRNA, microRNA, non-coding RNA, epigenetic regulators and so on) and surrounding microenvironment play a vital role in the selective vulnerability. Studying cell type-specific selective vulnerability inducing factors could be very useful to gain more knowledge about the medical condition and to find molecular targets for the treatment of the disease. Although, obtaining cell type-specific sub-cellular information is not a trivial task because of the cellular complexity and heterogeneity of the biological systems (for instance, the brain). To gain insights of the molecular mechanisms about the selective vulnerability of brain cells, in the *in-vivo* system, a transgenic mouse line called the Tagger was developed. The study was done in collaboration with the Jackson lab (DZNE, Bonn). Specifically, mouse line generation and immunoprecipitation experiments were done in the Jackson lab. The high throughput NGS library generation for sequencing was performed by myself in the Bonnlab. Furthermore, Dr. Bansal from our lab has analyzed the NGS data and helped in the interpretation of the results.

Nevertheless, this chapter deals with the sample processing (after the immunoprecipitation studies) and subsequent analysis of the NGS data. Manuscript preparation is underway describing the overall research, during the writing of this thesis.

In the Tagger system, following four molecular enrichment techniques were combined in a single transgene.

a) Co-immunoprecipitation of actively translated ribosome-associated mRNA using ribosomal protein Rpl22 tag (Sanz et al., 2009).
b) Isolation of newly transcribed RNA utilizing the 4-thiouracil (4-TU) labeling (Gay et al., 2013).

c) Purification of cell type-specific nuclei by FACS using a fluorescently labeled nuclear protein (Halder et al., 2015).

d) Co-immunoprecipitation of Argonaute2 bound miRNAs (He et al., 2012).

The Tagger system was developed as a single copy knock-in transgene, which comprises a polycistronic expression cassette encoding four different proteins. Expression of this transgene was dependent on both Flp- and Cre- recombinase system. To verify the Tagger system for cell type-specificity and purity, three cell type-specific Cre lines were developed. These lines were PV-Cre (tagger expression in parvalbumin-positive GABAergic interneurons in the nervous system; also function as inhibitory neurons), Gad2-Cre (expression in Gad2 positive GABAergic neurons) and vGlut2-Cre (expression in glutamatergic neurons; also known as excitatory neural cells).

In the experimental workflow, the Tagger mouse with a recombinase mouse of choice to produce mice driver line with particular cell type-specific Tagger transgene expression. The resultant mice line was subjected to experimental paradigms (for instance disease modeling) and optionally 4-thiouracil injections. At the desired time point, Tagger system allowed to isolate four different cell type-specific sub-cellular molecular entities from the same tissue sample. Furthermore, individually enriched molecules were processed for high throughput sequencing library preparations as described in the method section (below) in detail.
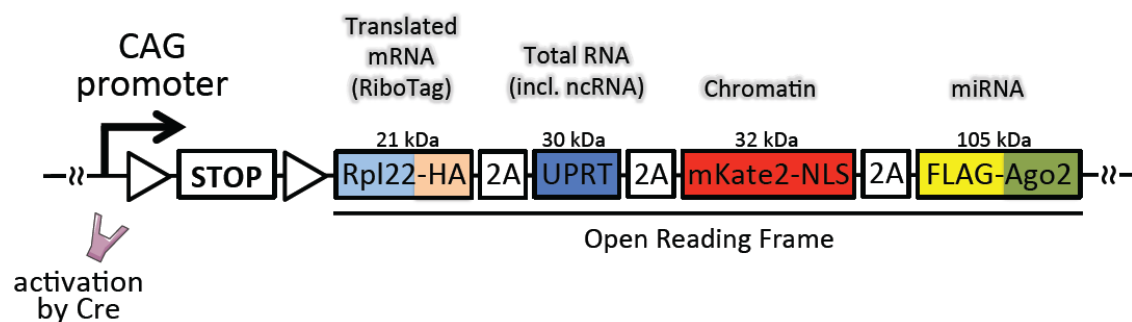
Figure 4.1 Overview of the Tagger transgene.

The Tagger transgene consists of four sub-components. HA = HA tag; 2A peptides are 18-22 amino-acid long viral oligopeptides that mediate cleavage of the polypeptide during translation in eukaryotic cells (Liu et al., 2017). NLS = Nuclear

localization signal. Cre-recombinase system regulates the Tagger transgene expression.

4.2 Methods:

4.2.1 Processing of cell type-specific nuclei samples

Nuclei were sorted in the FACS machine based on mKate2 signal expression in the nucleus. Sorted nuclei were frozen in the -80°C until the processing time. For isolation of DNA, sorted nuclei were resuspended in 30ul EB inside a 1.5 ml vial. 1ul of 50ng/ul RNase A was added to the tube and solution was incubated at 37°C for 30 minutes. After the RNase treatment, 30 ul of the 2X WB (Weinmann buffer) and 1ul of PK (proteinase K, 20mg/ml) was added to the solution and incubated at 65°C for overnight with intermittent shaking (30 sec ON 1300rpm; 15 minutes OFF). After the overnight incubation, nuclei solution was transferred into Diagenode shearing tube, and the cap was covered with parafilm. Shearing was done in Bioruptor plus with high power for 5x3 cycles (30 sec ON/OFF). After every five cycles, the tubes were removed from the Bioruptor machine and centrifuged briefly. After 15 cycles, the solution was manually triturated with p100 pipette (10 times up and down) and centrifuged for (10000g for 5 minutes). The supernatant was transferred into 1.5 ml DNA low binding vial. The remaining pellet was either freeze in -80°C or sheared again to obtain more DNA material.

In the DNA low binding tube, 3 ul of co-precipitant (LPA 5mg/ml) was added along with 1 volume of SureClean plus (Bioline, BIO-37048). The tube was briefly vortexed and incubated for 10 minutes at room temperature. Centrifugation was done at 15,000 g for 20 minutes at room temperature, and the supernatant was removed. The pellet was washed with 1 ml of 70% ethanol in water. The tube was briefly vortexed and centrifuged for 5 minutes at 10,000g. The supernatant was removed, and the pellet was dried at room temperature inside a hood (until pellet become transparent). The pellet was resuspended in 15ul EB buffer. DNA concentration was measured with Qubit 2.0 using the manufacturer's protocol. DNA samples were also checked for DNA size using High-sensitivity DNA bioanalyzer assay adjusting the concentration appropriate for High-sensitivity chip. Samples are having average size more than 300bp were subjected to another round of DNA shearing and Bioanalyzer.

4.2.1.1 Sample preparation

For the following processes, the Ultra NEB kit was used. For end repair of the DNA samples, 12 ul of the DNA (100ng) was added with 1.5 ul of End prep enzyme mix (NEB kit) and 1.5 ul of 10x buffer. The 15ul mix was incubated at 20°C for 30 minutes and 65°C for 30 minutes. Subsequently cooled down to 4°C. To ligate the adapters to the end-repaired DNA, 7.5 ul of the ligase master mix, 2 ul of the adapter (15 uM; final conc. 1.2uM) and 0.5 ul of the ligation enhancer was added. The tube was incubated at 20°C for 30 minutes. After incubation, 1.5 ul USER enzyme was added to each tube and further incubated at 37°C for 30 minutes. Adapter from Illumina are methylated and should not be used for MeDIP experiments. In current MeDIP experiments, adapters from NEB were used.

4.2.1.2 Sample cleanup

25 ul AMPure XP beads (1X) were mixed with 25 ul end repaired and adapter ligated DNA from the previous reaction and mixed by pipetting 15 times. Tubes were incubated for 5 minutes at room temperature. After incubation, tubes were briefly centrifuged and kept on the magnetic stand until all the beads were accumulated at one place and the solution becomes clear. The supernatant was removed, and beads were washed with 200 ul 80% fresh ethanol using the multichannel pipette. Washing step was repeated in total for three times. After final washing, tubes were air-dried few minutes on the magnet and afterwards 30 ul EB buffer was added on top the beads. Beads were mixed by pipetting up and down and incubated for 5 minutes at room temperature. Tubes were briefly spun down and kept on the magnet. Once, solution became clear; supernatant was transferred into a new PCR tube.

4.2.1.3 Adapter ligation checking by qPCR

Following qPCR mix was prepared to check adapter ligation quality with the DNA template.

Sample (DNA inputs diluted 1/100)	2 ul
2x Q5 NEBNext master mix	2.5 ul
Index (any)	0.1 ul
Universal primer	0.1 ul
SyBr (1/200 diluted)	0.3 ul

The qPCR was done according to Q5 NEBNext master mix cycling conditions (98°C 30sec; 40 cycles: 98°C 10sec, 65°C 75sec). After confirmation of adapter ligation, samples were used for immunoprecipitation of methylated DNA (MeDIP) assay.

4.2.1.4 Methylated DNA immunoprecipitation (MeDIP)

The MeDIP assay was done based on a highly specific anti-5-methylcytosine monoclonal antibody. It was the part of Methylated DNA IP kit provided by Zymo Research (Cat D5101) used for the following assay. For the MeDIP assay, manufacturer's protocol was followed with some modifications. In Brief, 100 ng of End repaired and adapter ligated input DNA was diluted in the DNA Denaturing buffer to a final volume of 50 ul. In a 1.5 ml DNA low binding vial 250 ul of MIP buffer, 15 ul of ZymoMag Protein A containing beads and 1 ul of Mouse Anti-5-Methylcytosine antibody (5-mc) was added, and the tube was inverted 3-4 times to mix the Protein A and antibody thoroughly. Diluted input DNA was denatured at 98°C for 5 minutes and added immediately to the Protein A and antibody containing the mixture. The final solution was incubated at 37°C for 1 hour on a Thermomixer and the tubes were inverted every 10 minutes during incubation. After that, the incubation tubes were placed in a plastic stand containing magnetic bars near to the tube. Once, the beads were clustered at one place; the supernatant was removed and discarded. On top of the beads, 500 ul of MIP buffer was added in vial and caps were closed. The tube was inverted ten times and vortexed briefly to resuspend the beads. Again tube was placed on a magnetic rack and supernatant was discarded. The washing step was repeated two more times, first with 500 ul MIP buffer and after that once with 500 ul of DNA elution buffer. After the final wash, 15 ul of DNA elution buffer was added to the tubes and beads were resuspended by pipetting up and down. The bead suspension was transferred to a new 0.2 ml PCR tubes and incubated at 75°C for 5 minutes. PCR tubes were placed on a magnetic rack for 2 minutes to aggregate the beads, and the supernatant was stored in a fresh tube. This was the recovered DNA, mostly single-stranded and suitable for further downstream DNA methylation analysis. DNA could be stored at -20°C or -70°C for long-term storage.

4.2.1.5 qPCR to determine the amplification cycles

DNA samples after MeDIP assay were used to determine amplification cycles to prepare the library for sequencing.

Sample (DNA after MeDIP)	2 ul
2x Q5 NEBNext master mix	2.5 ul
Index (any)	0.1 ul
Universal primer	0.1 ul
SyBr (1/200 diluted)	0.3 ul

The qPCR was done according to Q5 NEBNext master mix cycling conditions as mentioned before.

4.2.1.6 Final library amplification

ct1/2 was calculated based on the previous qPCR, and the value was used in library amplification PCR. Following reaction was set up for each sample.

Sample (DNA after MeDIP)	10 ul
2x Q5 NEBNext master mix	11 ul
Index primer	0.5 ul
Universal primer	0.5 ul

A number of cycles were adjusted according to the qPCR (ct1/2 – 2 cycles). Cycling condition was same as used in the qPCR. Final amplification was done in a Thermocycler.

4.2.1.7 Library cleanup

22 ul of AMPure XP beads (1X) was used for cleaning the library samples following the same protocol as used for sample cleanup (describe before). 20 ul EB was added for final dilution of samples. The final library product was stored in DNA low binding tubes and kept in -20°C.

4.2.1.8 Methylation enrichment qPCR

The qPCR assay was used to measure methylated DNA enrichment efficiency.

Reaction setup:

DNA (cleaned libraries/Input)	4.3 ul
MeDIP primers +/-	0.67 ul
SyBr Green	5 ul

Cleaned libraries and input were diluted 1:100 times to use in the enrichment qPCR assay. MeDIP primers (+ or -) were a mixture of forward and reverse primer at 3 uM concentration (See primer information 2.3). The standards were also checked in qPCR.

4.2.1.9 Preparation of libraries for sequencing

After enrichment qPCR, library size was checked using the high sensitivity DNA bioanalyzer and concentration was determined using Qubit assay using the manufacturer's protocol. Libraries were diluted to 2 nM, and sequencing was done in Hi-seq 2500 sequencer following standard procedures.

4.2.2 Processing of 4-TU labeled newly synthesized RNA

The TU-tagging method was used to investigate newly synthesized RNA molecules. 4-thiouracil (4TU) was injected into the mouse brain cortex region. 4-thiouracil converts into 4-thiouridine in the presence of an enzyme called UPRT or uracil phosphoribosyltransferase (from *Toxoplasma gondii*), which was a part of Grabber/Tagger vector cassette. 4-thiouridine could incorporate into newly transcribed RNA as an analog of uracil nucleotide. After 4-TU delivery into the tissue and required experimental time, RNA was isolated from the whole tissue. Ribosomal RNA was removed, and remaining RNA was fragmented. One fraction of RNA was used for biotinylation reaction and subsequent uMacs streptavidin-based biotinylated RNA isolation. Only TU-tagged RNA was biotinylated, and this RNA fraction was used for library preparation using ScriptSeq V2 RNA-seq library synthesis as described below. Another fraction of fragmented RNA was

taken without biotinylation reaction as total RNA and also used for library preparation using ScriptSeq V2 RNA-seq method.

4.2.2.1 Scriptseq V2 RNA-seq library preparation

rRNA-depleted RNA was first subjected to annealing of the cDNA synthesis primer. Following mixture was prepared:

Nuclease-free H2O	x ul
rRNA depleted fragmented RNA	y ul
ScriptSeq cDNA synthesis primer	2 ul
Total	**11 ul**

The tube was incubated at 65^oC for 5 minutes in a PCR machine, and after the incubation, reaction was stopped by putting the tube on the ice. To synthesizing the cDNA, following cDNA synthesis master mix was prepared.

RNA fragmentation solution	1 ul
cDNA synthesis premix	3 ul
100 mM DTT	0.5 ul
StarScript AMV reverse transcriptase	0.5 ul
Total volume per reaction	**5 ul**

The RNA fragmentation solution was added in the cDNA synthesis reaction to provide Mg^{2+} ions. cDNA synthesis master mix was appropriately mixed by pipetting up and down 15 times and added 5 ul to each reaction on ice from the previous step. The solution was mixed gently and thoroughly. It was incubated at 25^oC for 5 minutes and afterward 42^oC for 20 minutes. The reaction was cooled to 37^oC, and thermocycler was paused. In each reaction tube, 1 ul of Finishing solution was added and mixed thoroughly by pipetting. The tubes were incubated at

37°C for 10 minutes, followed by incubation at 95°C for 3 minutes. After that reaction was cooled to 25°C and thermocycler was kept on hold.

4.2.2.2 Synthesizing 3'-tagged DNA

The following mix was prepared on the ice.

Terminal tagging premix	7.5 ul
DNA polymerase	0.5 ul
Total volume/ reaction	**8 ul**

Terminal tagging master mix was viscous and thoroughly mixed by pipetting 10-15 times. From the previous step, tubes were removed from thermocycler one by one and 8 ul of Terminal tagging master mix was added to each of them. Tubes were incubated at 25oC for 15 minutes and at 95oC for 3 minutes. The reaction was cooled to 4oC in the thermocycler.

4.2.2.3 Purification of the cDNA

The di-tagged cDNA was purified using the 1.8X AMPure XP system as described previously in the method section. cDNA was eluted in 22.5 ul elution buffer (EB).

4.2.2.4 PCR amplification and addition of indexes

To generate the complementary strand of cDNA, following steps were used. These steps also incorporate an Illumina adapter sequence and an index in the library. The following mix was prepared for this PCR amplification in a PCR tube:

Di-tagged cDNA	22.5 ul
Forward PCR primer	1 ul
ScriptSeq Index PCR primer	1 ul
FailSafe PCR premix E	25 ul
FailSafe PCR enzyme (1.25 U)	0.5 ul
Total volume / reaction	**50 ul**

These steps were followed in the PCR reaction. At first, samples were denatured at 95°C for 1 minute. 15 PCR cycles was performed of 95°C for 30 seconds; 55°C for 30 seconds; 68°C for 3 minutes. After the PCR cycles, samples were incubated at 68°C for 7 minutes.

4.2.2.5 Purification of the library

1X AMPure XP system was used to purify the TU RNA-seq libraries following the same protocol described earlier. Samples were eluted in 20 ul nuclease-free water.

4.2.2.6 Quantification and quality checking of library

The libraries were quantified using the Qubit assay, following manufacturer's protocol. Furthermore, to assess the size distribution of the libraries, High sensitivity DNA bioanalyzer was used following manufacturer's protocol. Libraries were diluted to 2nM for sequencing into Hiseq-2500 system.

4.2.3 Ribosome-associated RNAs (RPL-tag) processing

The co-immunoprecipitated ribosome-associated mRNA, using a tagged ribosomal protein Rpl22, was subjected to high throughput sequencing library preparation. To prepare the ribosome-associated mRNA library, TruSeq total RNA sample preparation protocol was followed according to manufacturer's instructions. In brief, 0.25 to 0.5 ug of total RNA was diluted with nuclease-free water to a final volume of 5 ul in a PCR plate labeled as BRP. 2.5 ul of rRNA binding buffer and 2.5 ul of Ribo-Zero rRNA removal mix was added to each well. The plate was sealed and samples were incubated at 68°C for 5 minutes and cooled down to 4°C. The plate was incubated for 1 minute at room temperature. In a fresh PCR plate, 17.5 ul of rRNA removal beads were added to each well and 10 ul of the sample from the BRP plate was added. After mixing correctly by pipetting, the plate was incubated for 1 minute at room temperature and placed on the magnetic stand at RT for 1 minute. The supernatant was transferred to the corresponding well into a new plate. The new plate was again placed on the magnetic stand and supernatant was transferred to a new fresh plate. This procedure was repeated until all of the beads were removed. After that, 49.5 ul of well-mixed RNAClean XP beads were added to each well and mixed by

pipetting. The plate was incubated at room temperature for 15 minutes and placed on the magnetic stand for 5 minutes at room temperature. The supernatant was removed, and beads were washed with 200 ul of 70% ethanol by incubating 30 seconds at room temperature and then discard the supernatant. The plate was air dried for 15 minutes at RT and then 5.5 ul Elution buffer was added and mixed properly. The plate was incubated at RT for 2 minutes and then placed on the magnetic stand for 5 minutes at RT. 4.25 ul of the supernatant was transferred to a new 0.3 ml PCR plate. In the new plate, 4.25 ul of Elute, prime, fragment high mix was added in each well and mixed by pipetting. The solution was incubated at 94°C for 8 minutes in a thermal cycler. The plate was cooled to 4°C and briefly centrifuged.

4.2.3.1 Synthesizing the first strand of cDNA

To reverse transcribe the cleaved RNA fragments that were primed with random hexamers into first strand cDNA, the following process was done. In the PCR plate, 4 ul of First strand synthesis Act D and SuperScript II mix was added and mixed properly. The plate was sealed and incubated at 25°C for 10 minutes, 42°C for 15 minutes, 70°C for 15 minutes, 4°C hold in a thermal cycler.

4.2.3.2 Synthesizing the second strand of cDNA

Following process was used to remove the RNA template and to synthesize a replacement strand, incorporating dTTP in place of dUTP to generate ds cDNA. AMPure XP beads were used to separate the ds cDNA from the second strand reaction mix. At the end of this process, blunt-ended cDNA was generated. To proceed with, 2.5 ul of Resuspension buffer and 10 ul of Second strand master mix was added to the samples and mixed properly. Samples were incubated at 16°C for 1 hour and after that 45 ul of AMPure XP beads were added at the RT and mixed properly. The plate was incubated at RT for 15 minutes and after that placed on the magnetic stand for 5 minutes. The supernatant was discarded, and beads were washed with 200 ul of 80% ethanol. The plate was incubated at RT for 30 seconds, and the supernatant was discarded. Ethanol washes were repeated two more times. The plate was dried at RT for 15 minutes, and 8.75 ul of Resuspension buffer was added and mixed properly. The plate was incubated at RT for 2 minutes

and again placed on the magnetic stand for 5 minutes at RT. 7.5 ul of the supernatant was transferred to a new 0.3 ml PCR plate (ALP plate).

4.2.3.3 Adenylation of 3' end

A single 'A' nucleotide was added to the 3' ends of the blunt fragments to prevent them from ligating to one another during the adapter ligation reaction. A corresponding single 'T' nucleotide on the 3' end of the adapter provides a complementary overhang for ligating the adapter to the fragment. This strategy ensures a low rate of chimera (concatenated template) formation. For this reaction, 1.25 ul of Resuspension buffer and 6.25 ul of A-Tailing mix was added to the samples and mixed by pipetting. The plate was sealed and incubated at 37°C for 30 minutes; 70°C for 5 minutes and was held at 4°C in a thermal cycler.

4.2.3.4 Adapters ligation

Multiple indexing adapters were ligated to the ends of the ds-cDNA to enable them to hybridize onto the flow cell. To the samples, 1.25 ul of the Resuspension buffer, 1.25 ul of Ligation mix and 1.25 ul of RNA adapter index was added and mixed by pipetting. The plate was sealed and incubated at 30°C for 10 minutes. 2.5ul of Stop ligation buffer was added and mixed properly by pipetting. To the samples, 21 ul of the AMPure XP beads were added, and purification was done as described previously. At the end, 26.25 ul of the Resuspension buffer was added and mixed properly to recover the DNA molecules. Samples were incubated at RT for 2 minutes. The plate was placed on a magnetic stand for 5 minutes, and 25 ul of supernatant was transferred to a new plate (CAP plate). With 1X AMPure XP beads, samples were purified and isolated in 10 ul Resuspension buffer.

4.2.3.5 Selective enrichment of DNA fragments

To selectively enrich DNA fragments that have adapter molecules on both ends and to amplify the amount of DNA in the library following step was done. The PCR was performed with a PCR primer cocktail that anneals to the ends of the adapters. The number of PCR cycles was kept a minimum to avoid skewed representation of the library. In the adapter-ligated samples from the

68

previous step, 2.5 ul of PCR primer cocktail and 12.5 ul PCR master mix was added. Samples were mixed properly, and the plate was sealed. In a thermal cycler samples were incubated at (98°C for 30 sec; 11 cycles of [98°C for 10 sec; 60°C for 30 sec, 72°C for 30 sec]; 72°C for 5 minutes; 4°C hold). Samples were cleared using 1X AMPure XP beads and resuspended in 16.25 ul Resuspension buffer. 15 ul clear supernatant was transferred to a new 0.3 ml PCR plate (TSP1).

Samples were measured by Nanodrop, Qubit, and Agilent Bioanalyzer DNA 1000 kit. Libraries were normalized to 10nM concentration using Tris-Cl 10mM, pH 8.5 with 0.1% Tween 20 and pooled for sequencing into Hiseq 2000 sequencer.

4.2.4 Library preparation of Ago2 bound miRNA

The flag-tagged Argonaute2 (Ago2) bound miRNA, obtained from the co-immunoprecipitated reaction was subjected to high throughput sequencing library preparation. For this purpose, TruSeq small RNA library preparation kit from Illumina was used, following the manufacturer's protocol. In brief, following steps were followed.

4.2.4.1 Ligation of 3' adapter

In a 200 ul PCR tube, 1 ul of RNA 3' adapter and 5 ul of 1 ug total RNA was added. After mixing properly tube was incubated for 2 minutes at 70°C and then immediately kept on ice. In each sample tube, 2 ul of ligation buffer, 1 ul of RNase inhibitor and 1 ul of T4 RNA ligase two deletion mutant was added and mixed by pipetting. Samples were incubated at 28°C for 1 hour, and at the end, 1 ul Stop solution was added to each tube on the thermal cycler. Tubes were incubated further for 15 minutes at 28°C and placed on the ice.

4.2.4.2 Ligation of 5' adapter

In a 200 ul PCR tube, 1.1 ul per sample amount of RNA 5' adapter was incubated at 70°C for 2 minutes and then placed on the ice. In the tube, 1.1 ul per sample amount of 10 nM ATP was added and mixed properly. After that, 1.1 ul per sample amount of T4 ligase was added and

mixed by pipetting. From the mix, 3 ul was added to the 3' adapter ligated per samples. The plate was incubated at 28°C for 1 hour and placed on the ice.

4.2.4.3 Reverse transcription and amplification

In a 200ul PCR plate, 6 ul of 3' and 5' adapter ligated RNA and 1ul of RNA RT primer was added and mixed properly. The plate was incubated at 70°C for 2 minutes and placed on ice. In another tube, following mix was prepared for each sample. 2ul 5x first strand buffer, 0.5ul of 12.5nM dNTP mix, 1 ul 100 mM DTT, 1 ul RNase inhibitor and 1 ul SuperScript II reverse transcriptase; mixed properly and 5.5 ul of the mix was added to per tube of the RNA sample. Total 12.5 ul of the solution was mixed by pipetting and incubated at 50°C for 1 hour and placed on the ice. For PCR amplification, following mix was prepared in a 200 ul PCR tube. 8.5 ul ultrapure water, 25 ul of PCR mix, 2 ul of RNA PCR primer and 2 ul of RNA PCR primer index was mixed properly, and 37.5 ul of the mix was added to the 12.5 ul sample from the reverse transcription step. Plate was placed on the thermal cycler and incubated at (98°C for 30 sec; 14 cycles of (98°C for 30 sec, 60°C for 30 sec; 72°C for 15 sec) 72°C for 10 minutes; hold at 4°C).

4.2.4.4 Purification of cDNA construct

A number of samples pooled accordingly (for example, a pool of 24 samples would use 4 ul each) to run on a polyacrylamide gel electrophoresis (PAGE). In each well, 30 ul sample was loaded, and between different samples, two wells were kept empty. High-resolution ladder and custom ladder both were mixed with equal amount of DNA Loading Dye. Samples were also mixed with 20 ul of DNA Loading Dye. In the wells of the gel, 2 ul of the diluted high-resolution ladder and 2 ul of the custom ladder was loaded. 25 ul of samples were also loaded with the loading dye on the 6% PAGE gel. Electrophoresis was done for 57 minutes at 145V. The gel was stained with Midori green advance (in H_2O) for 10 minutes and observed on a UV trans-illuminator. Using a clean razor blade bands between 160 bp and 145 bp from the sample lane was cut and placed on the gel breaker tube. Tubes were centrifuged at 20,000xg for 2 minutes at RT. In the tube, 200 ul of ultrapure water was added, and DNA was eluted by shaking the tube for 2 hours at RT. Gel debris and supernatant was filtered through a 5 um filter by centrifuging it to 600xg for 10 seconds. Final libraries were concentrated by ethanol precipitation method. In the

supernatant 2 ul of Glycoblue, 30 ul of 3M NaAc and 975 ul of 100% ethanol was added. Tubes were centrifuged to 20,000xg for 20 minutes at 4°C. The supernatant was removed, and the pellet was washed with 70% ethanol and again centrifuged to 20,000xg for 2 minutes at RT. The pallet was dried on 37°C heat block for 10 minutes and resuspended in 10 ul of 10 mM Tris-HCl (pH 8.5). Libraries were validated using High sensitivity DNA bioanalyzer.

4.2.5 Sequenced data analysis

4.2.5.1 Quality control, read alignment and differential gene expression analysis

RNA-seq raw reads quality was assessed using FASTQC quality control tool (v0.10.1). The single end 50 bp sequence reads were aligned to the mouse reference genome (mm10) with Bowtie2 (v2.0.2) using RSEM (v1.2.29) with default parameters. First, the mouse reference genome was indexed using the Ensembl annotations (v72) with *rsem-prepare-reference* from RSEM software. After that, *rsem-calculate-expression* was used for read alignments and quantification of gene and isoform abundance. The output of this program provided the read counts and the TPM (transcript per million) value for all the genes and their respective isoforms separately. Differential expression of genes was analyzed by DESeq2 package using their read counts and genes with a p-value below 0.05 were determined as differentially expressed genes.

4.2.5.2 Cell type-specific gene expression (RNA-seq data) analysis

From the published literature, a cell type-specific gene list was prepared and used for benchmarking of our data. TPM values were extracted for these genes from all of the RNA-seq samples in the study. To generate the heat map, scaled TPM values (z-score) were used for the gplot R package function *heatmap.2*. For validation of cell type-specific gene expression, recently published data GSE78163 was downloaded from the Gene expression omnibus (GEO) database (Hornstein et al., 2016). In this dataset, RiboTag system was used to purify and identify actively translated genes from excitatory neurons in the brain cortex of Camk2a-Cre/RiboTag mice. The data from (Hornstein et al., 2016) contains the two biological replicates of immunoprecipitated ribosome profiling and the total mRNA from RiboTag mice. Read counts on coding sequences (CDS) were normalized by size factors generated from DESeq2. Furthermore,

enrichment scores were calculated by dividing the normalized RiboTag counts by normalized total mRNA counts.

4.2.5.3 MeDIP-seq data analysis

Sequenced reads of MeDIP libraries obtained using the Tagger system were analyzed as described in the section 3.2.7.

4.2.5.4 Small RNA (miRNA) –seq data analysis

To analyze the miRNA-seq data, in-house- developed bioinformatics tool "Oasis2" was used. It can handle the complex high throughput sequencing data and provide seamless and reliable analysis of the miRNAs in the study, for instance, differential expression of the miRNAs across multiple samples (Capece et al., 2015; Rahman et al., 2018). For the analysis, raw Fastq reads were uploaded using the front-end of Oasis (https://oasis.dzne.de/). The reference genome was selected as "mus musculus – mm10" and TruSeq 3' was used as an adapter sequence for data analysis in the Oasis web tool. Other options were set as default.

4.3 Results

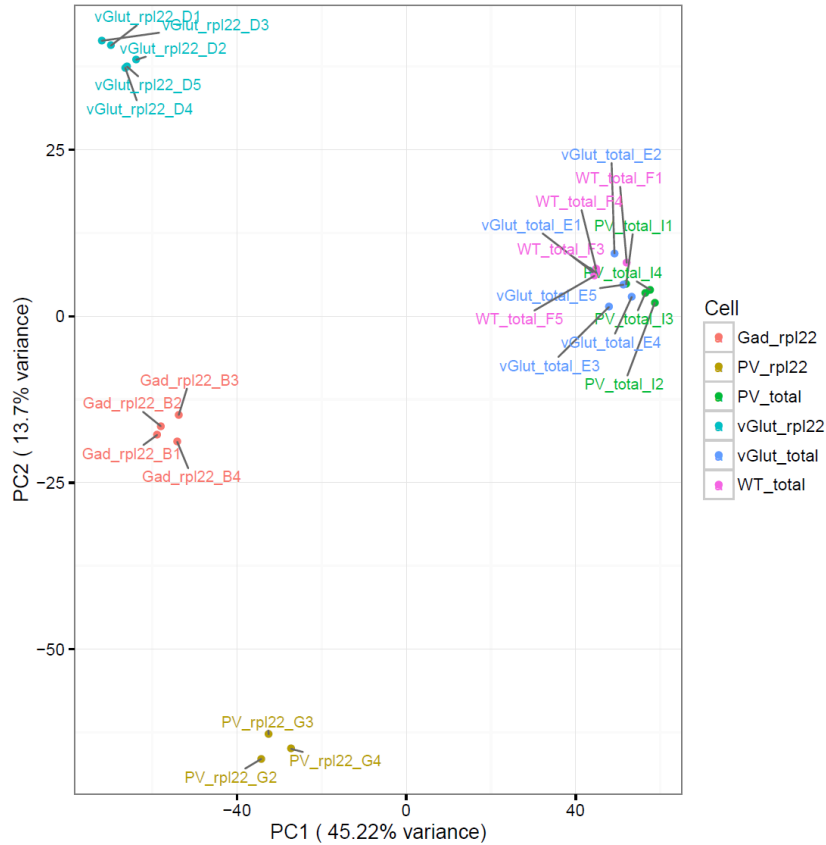## 4.3.1 Cell type-specificity of Rpl22-tag co-immunoprecipitated mRNA (polysome profiling)

The mapping quality of Rpl22 tagged (Rpl-tag) co-immunoprecipitated mRNA sequencing was analyzed as described in the section (4.2.5) and presented in the (Figure 4.2). For this analysis, PV-Cre, Gad2-Cre, and vGlut2-Cre samples were used as described in section 4.1. Apart from that 'total' samples (input for IP) were analyzed to see enrichment in the IP procedures. Wild-type (WT) mouse samples were also analyzed as a control. The mapping analysis showed that most of the reads were either uniquely or multi-mapped to the genome, which indicates good quality of the samples. Very few percentages of reads were not mapped to the genome (Figure 4.2). The samples with very low mapped reads were not included in the further downstream analysis. To benchmark the purity and verify the cell type-specificity of Rpl22 tagged co-immunoprecipitated mRNA assay for the Tagger mouse, principal component analysis (PCA) was carried out for all the three cell types Gad2, PV, and vGlut2 (Figure 4.3). The PCA analysis showed that all biological replicates of the same cell type were clustered together. Moreover, samples were distant from each other in the PCA plot for different cell type-specificity. Mixed un-immunoprecipitated samples (Input for IP; labeled as "Total") were also clustered together in a separate cluster (Figure 4.3).

In another attempt to validate the cell type-specific gene expression in the Rpl-tag co-immunoprecipitated mRNA samples, TPM values (Transcript per million) for the genes from each cell type were plotted against the previously known cell type-specific gene (Hornstein et al., 2016), as described in the method section 4.2.5.2 and presented in a heat map (Figure 4.4). The heat map showed that cell type-specific genes from Rpl-tag assay were same as their previously known genes (Hornstein et al., 2016). Furthermore, as expected genes from mixed un-immunoprecipitated samples were co-localizing with the astrocytes genes, indicating the non-specific background (Figure 4.4). These results showed the high purity and cell type-specificity of the Rpl-tag mRNA samples in the Tagger mouse.
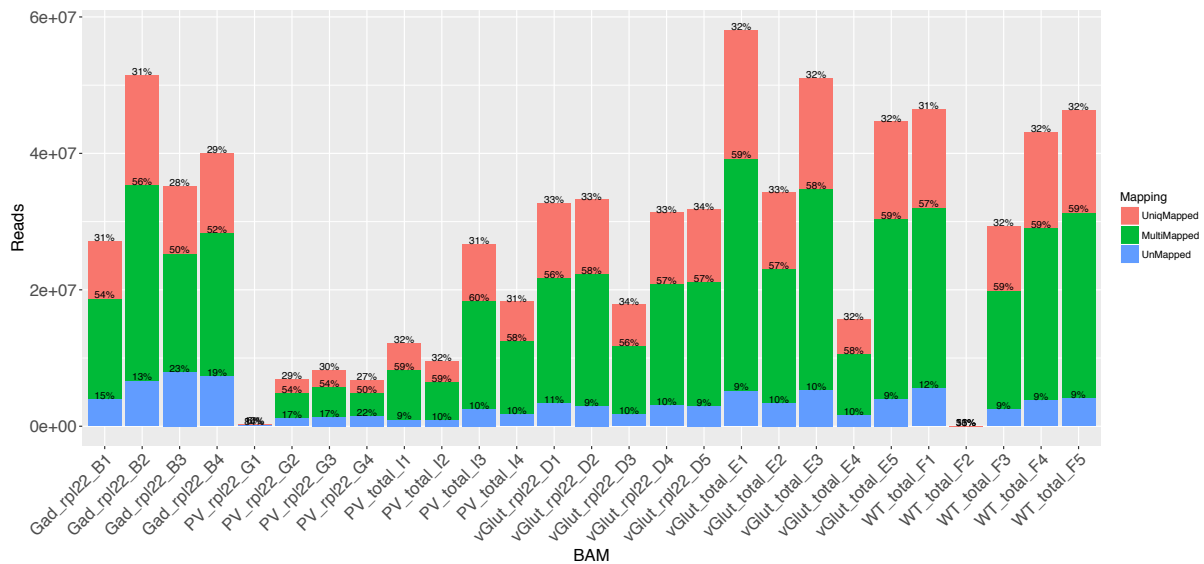
Figure 4.2: Mapping quality of the Rpl-tag co-immunoprecipitated mRNA samples.

On the graph, each bar represents a single sample. Percentages of uniquely mapped reads (salmon), multi-mapped reads (green) and unmapped reads (blue) were also plotted.

Figure 4.3: PCA analysis for Rpl-tagged co-immunoprecipitated mRNA samples.

PCA Plot for all the Rpl-tag co-immunoprecipitated mRNA samples in the study. Cell type-specific mRNA samples were clustered together indicating the good quality of the data. All mixed un-immunoprecipitated samples (labeled as "total") have also formed another cluster. Gad = Gad2-Cre; PV = PV-Cre; vGlut = vGlut2-Cre; rpl22 = Rpl22 tagged mRNA; total = input for IP. Replicates were presented as alphanumeric letters (B1, B2 and so on).
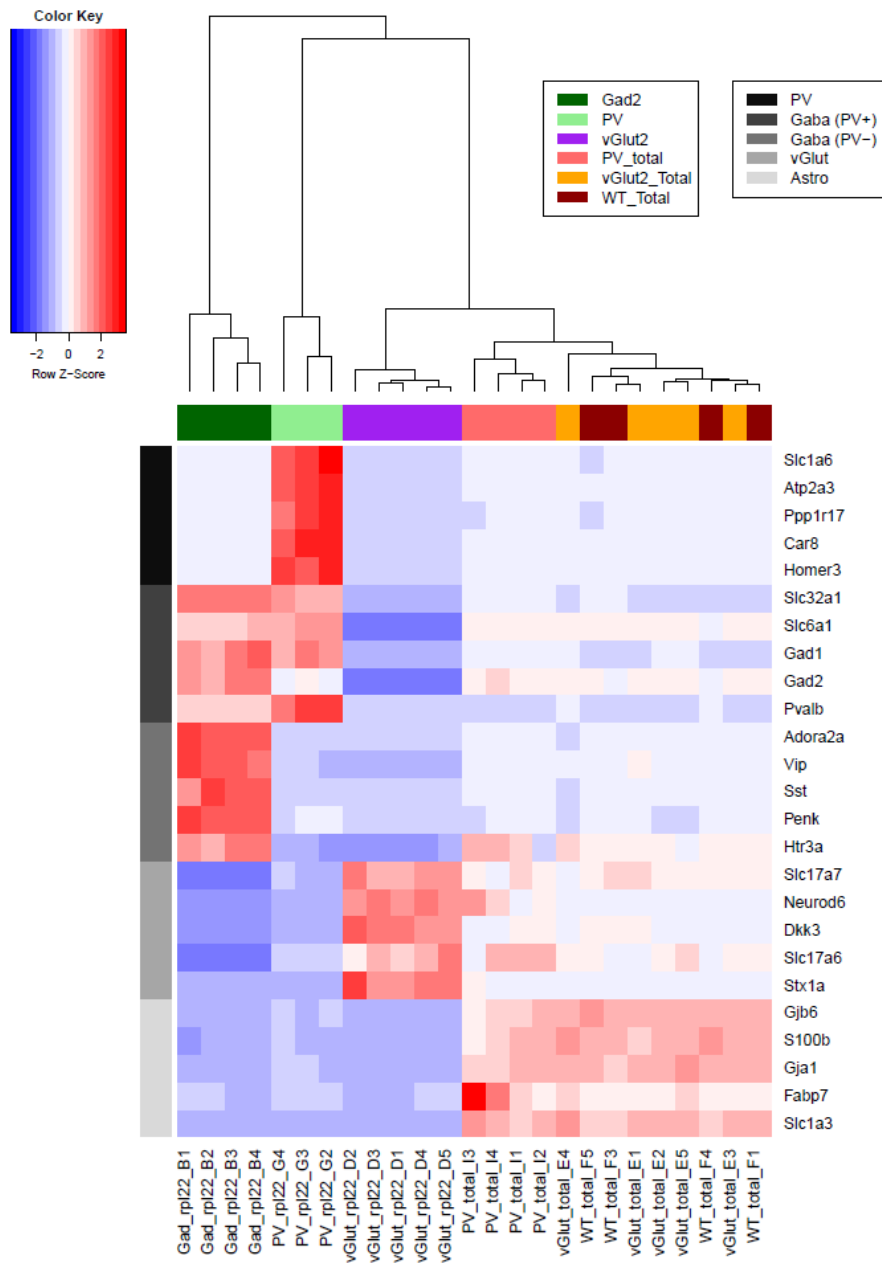
Figure 4.4: Heat map of the Rpl-tag mRNA cell type-specificity.

In this heat map, rows represent the previously known cell type-specific genes (Hornstein et al., 2016) for each cell type, shown in grey shades; and columns represent the cell type-specific Rpl-tag co-immunoprecipitated mRNA samples in this study (shown in colors). The top-left corner has a color key with the z-scores for the heat map.

4.3.2 Ago2 bound cell type-specific microRNA analysis

The FLAG-tagged-Ago2 bound miRNAs; isolated using the Tagger system, from PV-Cre and Gad-Cre samples were analyzed as described in section 4.2.5.4 using 'Oasis2' tool. The differential expression (DE) of miRNA from two cell types (PV vs. Gad) was carried out to confirm cell type-specific miRNA enrichment. A correlation study was done between the differential expressed miRNAs from in-house data to a published dataset of the similar cell type's (PV vs. Gad) miRNA DE analysis (He et al., 2012). In total 125 differentially expressed miRNAs with significant padj (<0.01) value from the published study was taken for this analysis. Correlation analysis (Pearson's correlation r = 0.68) was indicating the good overlap between the two datasets and signify the cell type-specificity of the miRNAs (Figure 4.5).
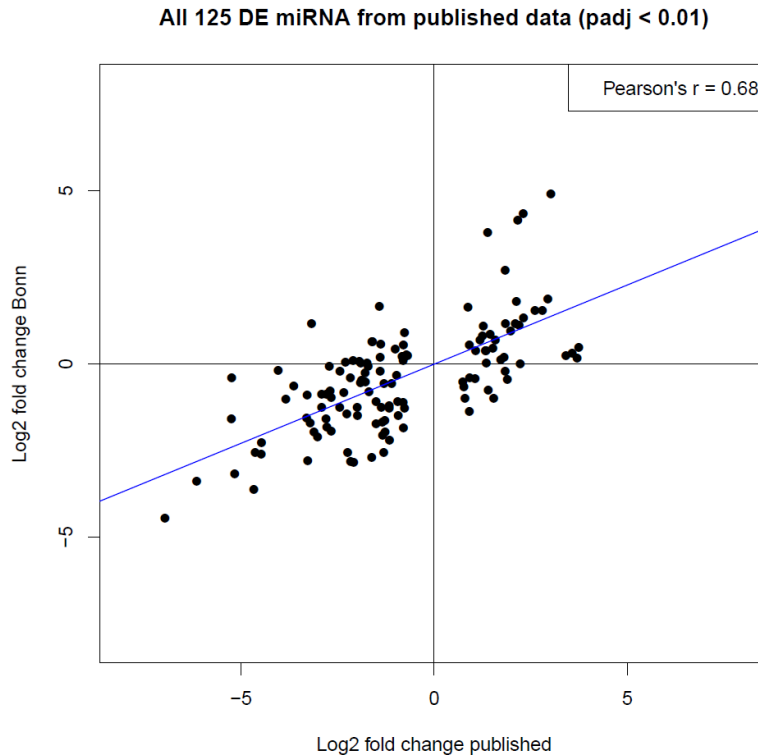
Figure 4.5: Correlation analysis for Ago2 bound differentially expressed (DE) miRNAs.

Cell type-specific DE (PV vs. Gad) data from Tagger mouse (Y-axis) and 125 differentially expressed miRNA from (He et al., 2012) (X-axis). Pearson's correlation value 'r' was depicted at the upper right of the figure.

4.3.3 Correlation between DNA methylation and gene expression

The Tagger system was used to isolate cell type-specific nuclei, and genetic material was used for MeDIP analysis (4.2.1). MeDIP-seq data were analyzed as described in the section (3.2.7). DNA methylation in gene body and upstream regions inversely correlate with gene expression, as previously presented for GABA and GLU neurons (Kozlenkov et al., 2015). In the current study, cell type-specific nuclear fraction of the cells from Tagger mouse was subjected to MeDIP analysis. The differential methylated region (DMR) analysis of Gad2 vs. vGlut2 samples was done to assess the quality of the MeDIP data. In this analysis, 1385 DMRs (padj < 0.1) were enriched that were around the gene (5kb upstream to TSS or in the gene body). Furthermore, Cell type-specific Rpl-tag co-immunoprecipitated mRNA was used to analyze differential gene expression analysis between Gad vs. Glut cells, which resulted in 609 genes (padj < 0.1). Correlation analysis between DMRs and gene expression resulted into Pearson's correlation coefficient value (r = -0.44), which was corroborating with the previously published study (Kozlenkov et al., 2015) as depicted in (Figure 4.6).
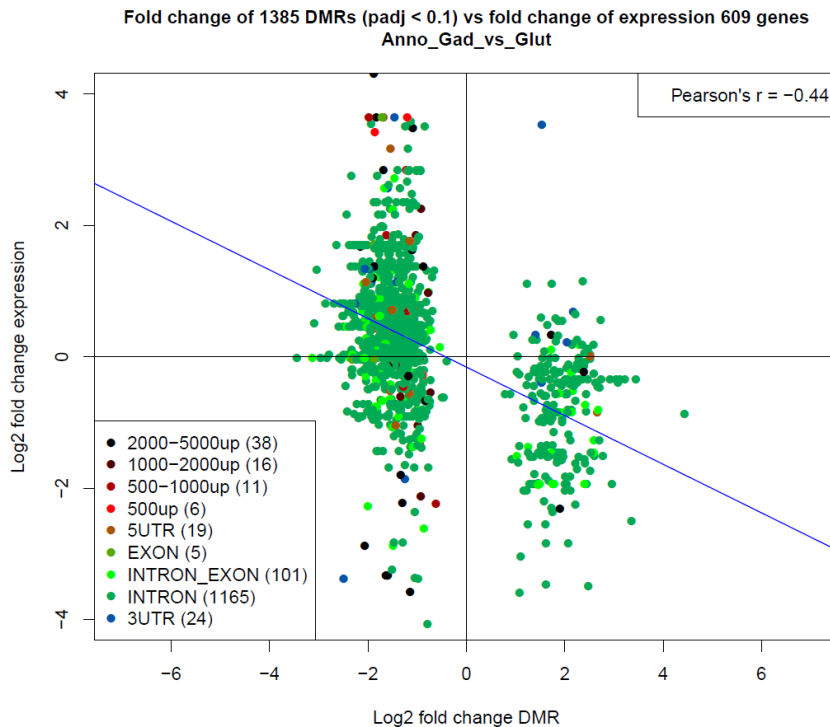
Figure 4.6: Correlation analysis between DMRs vs. gene expression.

Log2 fold change of DMRs (X-axis) and Log2 fold change of gene expression (Y-axis). Each dot in the figure represents the individual DMR, present in the vicinity of the assigned gene. The color of the dot indicates its location from the transcription start site (TSS) with respect to the annotated gene (according to the labels). The blue line represents the correlation between DMRs and gene expression in the dataset.

4.3.4 Cell type-specificity of 4-thiouracil (4-TU) labeled total-RNA

The mapping statistics of 4-TU labeled total-mRNA sequencing was analyzed as described in the section (4.2.5) and showed in the (**Figure 4.7**). Similar to Rpl-tag co-immunoprecipitated mRNA benchmarking, cell type-specificity and quality of the 4-TU labeled RNA was also confirmed (section 4.2.5.2 For this purpose, the principal component analysis was done for the samples from Gad2 and vGlut2 (Glut) expressing Tagger mice (**Figure 4.8**). In this analysis, biological replicates from the same cell type were clustered together, and different cell type samples were distant from each other in the PCA plot. Unbound fractions were clustered in the different locations in the PCA plot (**Figure 4.8**). Furthermore, cell type-specific gene expression was also checked for the known marker genes using the heat map (**Figure 4.9**). In this analysis, 4-TU labeled cell type-specific RNAs were co-localizing with the known marker genes, which was the confirmation for the high quality and cell type-specificity of the 4-TU labeled total RNA purification assay in the Tagger system (**Figure 4.9**).
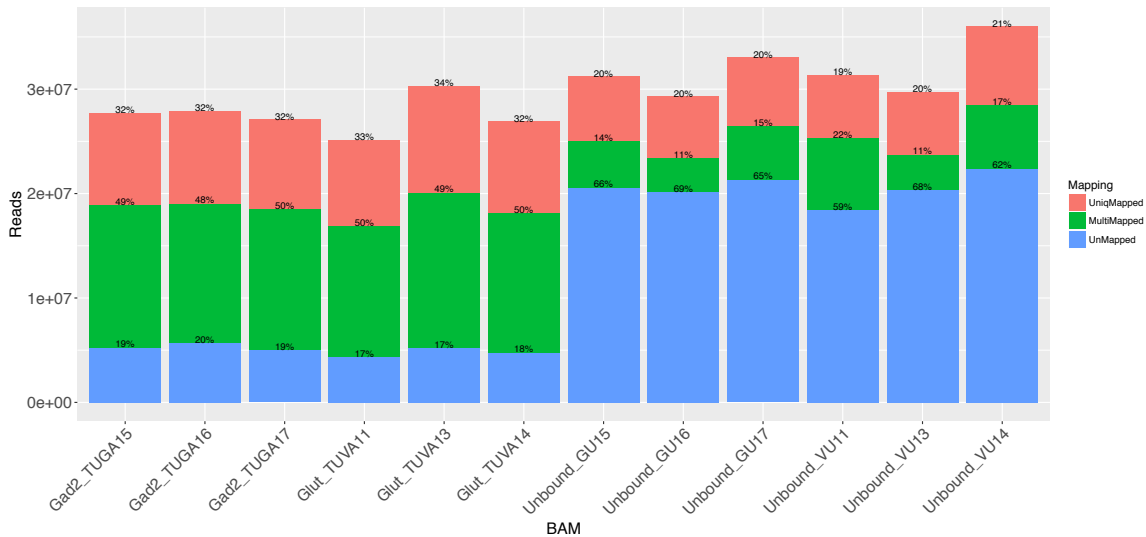
Figure 4.7: Mapping quality of the 4-TU labeled total RNA.

On the graph, each bar represents a single sample. Percentages of uniquely mapped reads (salmon), multi-mapped reads (green) and unmapped reads (blue) were also plotted.
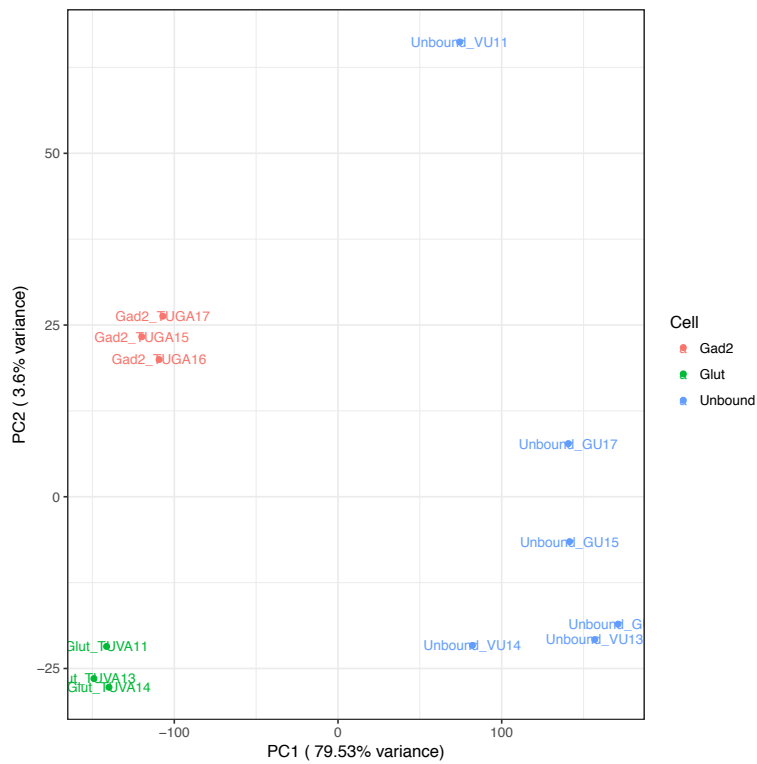
Figure 4.8: PCA analysis for 4-TU labeled RNA samples.

PCA plot for the 4-TU labeled newly synthesized RNA samples from Gad2 and vGlut2 (Glut) Tagger mouse and their unbound RNA fractions. Biological replicates of the cell type-specific 4-TU tagged RNA samples cluster together, distant from the unbound fraction. It indicates the cell type-specificity and purity of the 4-TU total RNA material.
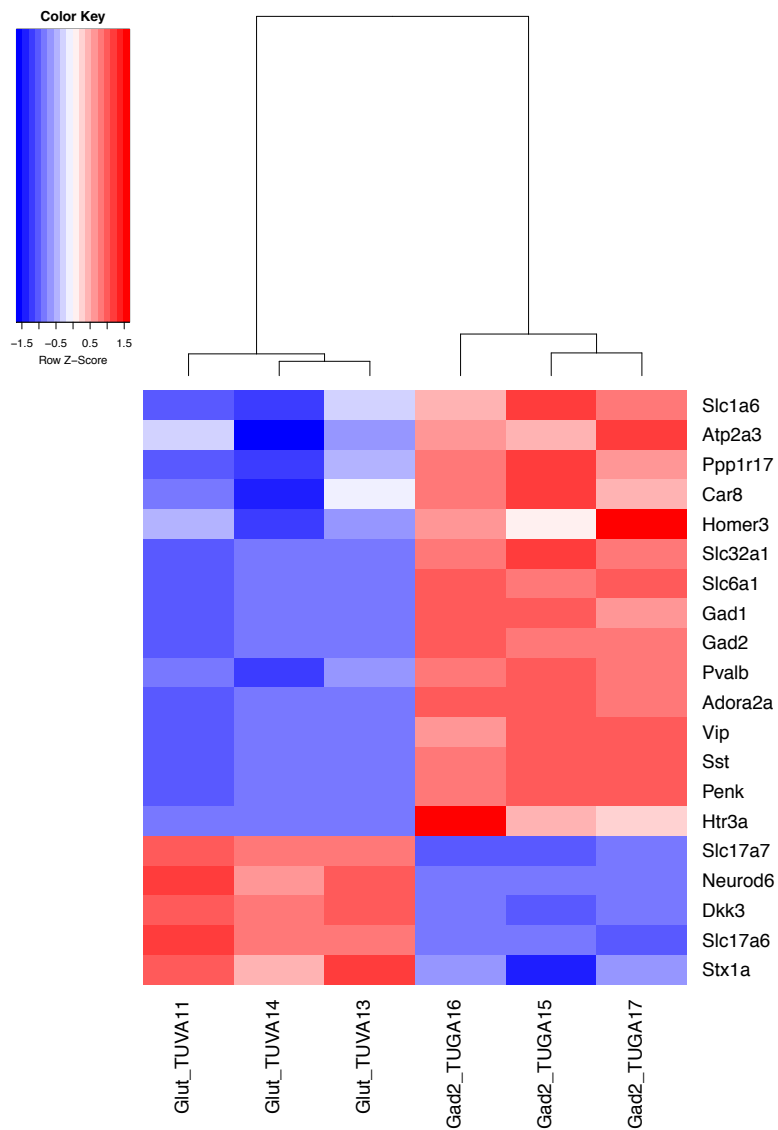
Figure 4.9: Heat map for the 4-TU labeled RNA.

In this heat map, previously known cell type-specific genes were plotted in the rows, and the columns represent the cell type-specific 4-TU labeled total RNA samples in this study. The top-left corner has a color key with the z-scores for the heat map.

4.4 Discussion

The Tagger system was developed to understand the selective vulnerability of the brain regions to diseases. *In-vivo* studies of sub-cellular factors such as mRNA, miRNA, epigenetic factors and so on could reveal the underlying mechanisms of susceptibility of the individual cell types and progression of the disease. Studying various types of nucleic acid molecules and epigenetic factors from the same biological material in a cell type-specific manner was a big challenge for the scientific community. The Tagger system is the first in-vivo mouse system to enable the study of four nucleic acid fractions by combining four different enrichment techniques in a single transgene.

One of the nucleic acid fractions, the actively translated mRNAs are isolated by co-immunoprecipitation of mRNAs associated with ribosomes (ribosomal subunit Rpl22). The advantage of this system is to capture mRNAs, which were in their active translation stage and represent the more precise functional image of the cell. One of the challenges, presented by the ribosome-associated mRNA sequencing, was associated with the contamination of rRNA in the samples. Fortunately, it was possible to efficiently remove this contamination by introducing Ribo-zero rRNA removal solution during the library preparations. The Tagger system also allows isolating cell type-specific miRNA by co-immunoprecipitation of FLAG-tagged Ago2 protein. As such, mature miRNAs bind with Ago2 protein together with RISC (RNA induced silencing complex) in the cell.

The third component of the Tagger system is the far-red fluorescent nuclear marker (mKate2-NLS), which is transported into the nucleus because of the presence of the nuclear localization signal (NLS). By harnessing fluorescence properties of the endogenous marker, cell type-specific nuclei can be FACS sorted, and nuclear material can be used for further genetic or epigenetic analysis. To validate the cell type-specificity and the purity of the samples, FACS sorted nuclei were used for the DNA isolation and subsequent MeDIP analysis as described in (section 4.3.3). To assess the cell type-specificity of the methylation signature, differentially methylated region (DMR) analysis as well as differential expression of mRNA (Rpl22 bound mRNA) was done between two available cell type samples in this experiment (Gad2 and vGlut2). Correlation

analysis between DMRs and mRNA expression resulted in negative correlation value (r = -0.44) indicating DNA methylation negatively regulate mRNA expression levels, as expected. Similar results were also observed in another study (Kozlenkov et al., 2015) which corroborates our method. For mKate2-NLS Tagger system, DNA methylation analysis served as an indirect proof of concept for cell type-specificity.

Moreover, the fourth sub-component of Tagger system is an enzyme called Uracil phosphoribosyltransferase (UPRT), which was initially isolated from *Toxoplasma gondii*. UPRT could efficiently incorporate 4–thiouracil (4TU) into newly transcribed RNA (Gay et al., 2013). Tagger mouse TU-tagging system was based on previous technological developments on Drosophila (Miller et al., 2009), cell culture system (Cleary et al., 2005) and mouse system (Gay et al., 2013). This method serves as an intersection of genetic and chemical approaches and allows covalent labeling of newly synthesized RNAs in a cell type-specific manner (Gay et al., 2013). Cell type-specificity of Tagger system was obtained by Cre-induced expression of UPRT enzyme and temporal specificity as well as experimental pulse-chase specificity obtained by injection of the uracil analog 4-thiouracil (4-TU) into the specific brain region (Gay et al., 2013). Cells expressing the UPRT efficiently incorporate 4-TU into newly synthesized RNA in the *in-vivo* conditions in the mouse system. The main advantage of this system is that it maintains the normal physiological conditions and molecular interactions inside the cells during the 4-TU RNA labeling experiment. The thio-RNA was later extracted from the total RNA using the in-vitro biotinylation and uMacs streptavidin mediated isolation of biotinylated RNA. To generate the high-throughput RNA-seq libraries, the ScriptSeq RNA-seq library synthesis kit was used because of its ability to generate libraries from the low amount of RNA, which was the case in these experiments. Another advantage of using this system is that 4-TU labeling has been shown to have a negligible effect on the *in-vivo* gene expression and molecular interplay in the cells (Cleary et al., 2005). Furthermore, ubiquitous expression of the UPRT enzyme had not shown any effect on the viability of the cells in the Drosophila (Miller et al., 2009) or in mice system (Gay et al., 2013).

To summarize, Tagger mouse system provides a simple, rapid and flexible tool to study various nucleic acid molecules and other nuclear content of the cells, from a single biological sample. It enables the researchers to understand the biological mechanisms of disease progression with better molecular insights.

Chapter 5. Investigating motor neuron subpopulations and EPO-mediated hippocampal changes using single-cell transcriptomics

5.1 Overview

Cells are the basic units of tissues, organs, and organisms. Each tissue made up of many different cell types, and each cell type can have many different biological states in a given time (E. Z. Macosko et al., 2015). In a complex biological system, each cell plays an important role, and to understand the complex tissue function, it is essential to understand the functional responses and capabilities of individual cell types (E. Z. Macosko et al., 2015). For many biological systems, cell type complexity is still not known entirely; for instance, cellular diversity of central nervous system is not fully understood and still in the investigation (Luo, Callaway, & Svoboda, 2008).

Cellular transcriptome plays a vital role in determining the cellular functions. In recent years, advances in technologies made it possible to do mRNA-seq at a single-cell level (Tang et al., 2009). Single-cell sequencing technologies were available from just hundreds of cells (Picelli et al., 2013) to thousands of cells (Jaitin et al., 2014; Shalek et al., 2014). In this study, we have implemented a microfluidic-based single-cell mRNA-seq method known as Drop-seq (Figure 5.1). This method was first developed by Macosko et al. in the McCarroll lab at Harvard medical school. In the Drop-seq method, thousands of single-cells were encapsulated in tiny oil droplets for parallel analysis of mRNA expression (E. Z. Macosko et al., 2015). The droplets compartmentalize the cells into tiny nanoliter volume reaction chamber to lyse the cells and capture all the mRNA from individual cells. To retain the molecular memory about the cell of origin, the mRNA was captured by a barcoded microparticle (bead) using the polyT primer on the surface, which binds to the polyA tail of mRNA and hybridizes it with the bead (Figure 5.1).

In brief, these steps were followed in a Drop-seq experiment. At first, a single-cell suspension was prepared by enzymatic tissue digestion method, from an intact tissue, customized according to the tissue type. After that, each cell was co-encapsulated with a uniquely barcoded

microparticle (bead) in a nanoliter volume oil droplet. Inside the droplet, lysis reagent lyses the cell and mRNA from that individual cell was captured on the surface of companion microparticle, forming the Single-cell transcriptomes attached to microparticle (STAMP). After that, reverse transcription and synthesis of cDNA processes carried out on the surface of the microparticle. In the further steps, cDNA (carrying their cell barcode identity) was removed from the surface of the microparticle, subjected to amplification, tagmentation and library preparation steps. Thousands of these STAMPs were sequenced together in Hiseq flowcell. In the end, STAMP barcodes were used to infer cells of origin and associated transcripts in the data analysis pipeline, as described in detail in the following sections.

As for technique optimization for the Drop-seq method and to determine the presence of doublet in the Drop-seq libraries with set parameters, a Human-Mouse mixed cell experiment was performed. After the successful method optimization for Drop-seq technique, it was applied to two case studies described in this thesis:

Case study I: Characterization of motor neuron subpopulations using single-cell sequencing.
Case study II: Effect of Erythropoietin (EPO) on the cells of hippocampal tissue using single-cell sequencing.

The setup of the Drop-seq platform and quality control experiments were done in the Bonnlab (DZNE, Göttingen) as previously published (E. Z. Macosko et al., 2015). The investigation of spinal cord tissue using the Drop-seq method was carried out in collaboration with Marquardt lab (European neuroscience institute, Göttingen). Dr. Edwar (Bonn-lab, DZNE, Göttingen) has helped the author in the wet-lab experiments of this study. Dr. Bansal and Abdul Qadir have also helped the author in the analysis of the NGS data in the Bonnlab.

In another study using the Drop-seq method, the molecular mechanisms were investigated, which can drive precursor cell differentiation into neurons and oligodendrocytes cells upon EPO administration. This study was conducted in collaboration with Debia Wakhloo and colleagues in the Ehrenreich lab in Clinical neuroscience department, Max Plank Institute of experimental medicine, Göttingen, Germany. Also, Dr. Bansal and Ting sun along with the author were

associated with the study specifically in the Drop-seq experiment as well as in the Drop-seq data analysis. The single-cell study was done in the Bonn-lab, German center for neurodegenerative diseases, Göttingen, Germany.
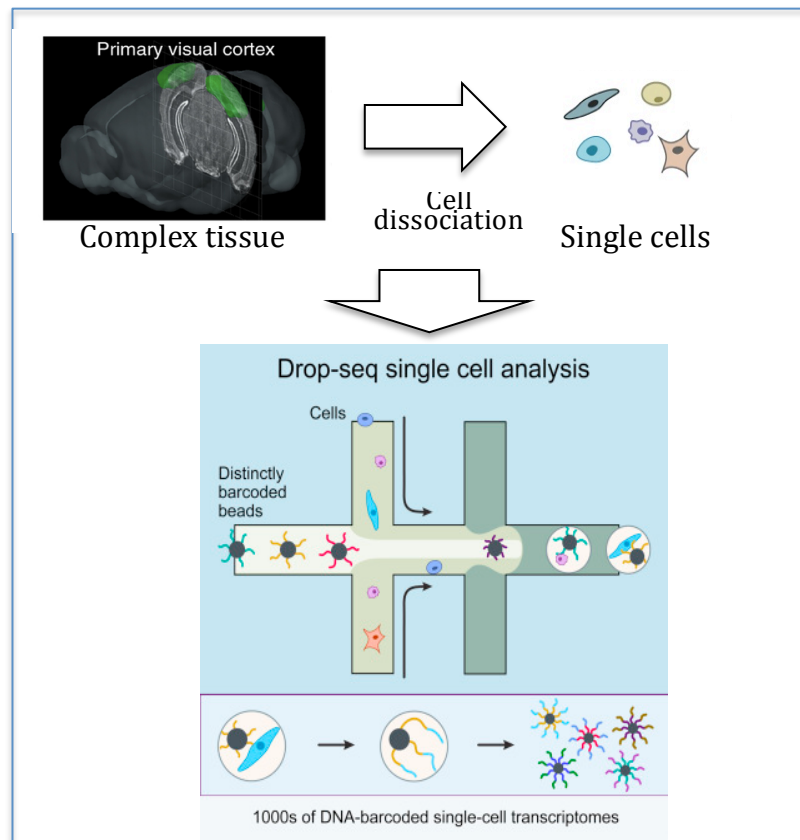
Figure 5.1: Schematic of the Drop-seq experiment.

Complex tissue was dissociated to form a single cell suspension. These single cells were captured in a tiny oil droplet together with a barcoded micro-particle (beads) during the Drop-seq experiment, as explained in this chapter. This figure was adapted from (E. Z. Macosko et al., 2015) and modified to present in this thesis.

5.2 Methods

5.2.1 Human-Mouse mixed cell experiment

Human HEK-293T cells and mouse NIH-3T3 cells were seeded in a 10 cm diameter culture plate in DMEM medium. Following day when cells are 70% confluent in a culture dish, cells were trypsinized for 5 minutes using 0.05% Trypsin and collected in a 15 ml tube. Cells were centrifuged and washed with 1 ml of PBS-BSA. Cells were again spun down at 300xg for 3 minutes, and the supernatant was removed. 1 ml of PBS was added, and cells were passed through a 40-micron filter to remove cell aggregates. 20 ul of cells were added to a C-chip (Fuchs-Rosenthal Hemocytometer) and counted using the Equation 1 (page 91). HEK-293T and NIH-3T3 cells were mixed in 1:1 ratio as a cell input for the Drop-seq experiment.

5.2.2 Motor neuron dissociation method

For motor neuron Drop-seq experiment, a transgenic mouse line having an expression of Rosa-CAG-tdTomato in motor neurons was used in process development phase. Experiment on mice was conducted in collaboration with Prof. Till Marquardt lab. Postnatal day 14 (p14) mice were anesthetized and dissected to expose the spinal cord of mice. The lumbar region of the spinal cord was isolated using the tdTomato expression as a guiding marker for motor neuron density under a fluorescence microscope. Following method was used for successful dissociation of motor neurons from the spinal cord. The spinal cord sample was washed in a vial containing 1 ml EBSS#1 solution as mentioned below (Saxena et al.,2012) and then kept into a vial containing 1 ml Papain and DNase-I solution.

EBSS#1 solution

EBSS	8.9 ml
25 mM AP-V	20 ul (0.05 mM)
100mM KA	80 ul (0.8 mM)
50% (w/v) Trehalose	1 ml
Total (for 1 mouse)	**10 ml**

To prepare Papain and DNase-I solution, 5 ml of EBSS#1 was added to papain container (Worthington kit). The papain solution was incubated for 10 min at 37°C inside the cell culture incubator after loosening the cap. The solution was equilibrated by incubator environment, having 95% O_2 and 5% CO_2. During incubation, 500 ul of EBSS was added to the vial containing DNase-I powder from Worthington kit and mixed slowly to avoid DNase-I solution degradation. 250 ul of the DNase-I solution was added in prepared papain solution to prepare final Papain DNase-I solution. Final concentration of this solution was 20 units/ml of Papain and 0.005% of the DNase I. Inside the Papain DNase-I solution, spinal cord sample was chopped into smaller pieces and the vial was incubated at 37°C alternatively for 10 minutes in a cell culture incubator with open cap and 10 minutes at Thermomixer (Eppendorf) with 500 RPM shaking. After the incubation, the sample was transferred to 15 ml falcon tube. Tissue was triturated few times with the 1 ml pipette tip. 5 ml of EBSS#2 solution was prepared as mentioned below and added to the falcon tube (Saxena et al.,2012). Tissue suspension was mixed properly without forming air bubbles.

EBSS#2 solution

EBSS	4.73 ml
25 mM AP-V	20 ul
100mM KA	48 ul
Albumin ovomucoid inhibitor	350 ul
DNase-I	250 ul
50% (w/v) Trehalose	600 ul
Total (for 1 mouse)	**6 ml**

The tissue suspension was centrifuged at 100g for 10 minutes. The supernatant was discarded, and the pellet was resuspended in 200 ul of the EBSS#2 solution. Digested tissue was mechanically dissociated with p200 pipette tip by gently pipetting up and down for around 40 times until big tissue chunks were dissociated and a cloudy suspension was prepared. 200 ul of the EBSS#2 solution was again added and pipetted for 15 more times with 1000 ul pipette tip. After this trituration step, 5 ml of Medium without serum was added (as mentioned below) and centrifuged for 10 minutes at 100g.

Medium without serum

DMEM/F12 without phenol red	18.9 ml
25 mM AP-V	21 ul
100mM KA	84 ul
50% (w/v) Trehalose	2.1 ml
Total (for 1 mouse)	**21 ml**

The supernatant was discarded, and the pellet was resuspended in 200 ul of the mixed EBSS#2 solution and pipetted up and down 20 times using p200 pipette. 5 ml of Medium without serum was added, the solution was mixed properly, and the cell suspension was washed by centrifugation at 100g for 10 minutes. The pallet was resuspended in 150 ul of YES medium (Yield enhancing separation medium) and proceeded for ClioCell treatment as mentioned in further steps.

5.2.3 Removal of impurities from the samples using ClioCell treatment

To remove cell debris and dead cells from the samples, ClioCell (Amsbio Biotechnology) treatment was performed on the spinal cord cells. ClioCell nanoparticles stock vial was vortexed for 30 seconds to make sure a homogeneous suspension. 25 ul of the resuspended ClioCell nanoparticles was transferred to a 1.5 ml vial containing 1 ml of the cell culture medium or YES medium. Tubes were placed on the magnetic rack for 5 minutes, and when the beads were aggregated, and the solution was clear, YES medium was carefully removed without disturbing the beads. The tube was removed from the magnetic platform and nanoparticles were resuspended in 100 ul of YES medium. Nanoparticle suspension was vortexed and pipetted up and down to make a homogeneous solution without forming the bubbles. Carefully 100 ul cells were added in nanoparticle suspension and triturated 4 to 5 times to make sure cells were mixed adequately with ClioCell nanoparticles. Tubes were incubated in 4°C for 25 minutes. It allowed the nanoparticles to bind the cell debris or dead cells by the dissociation and trituration processes. After the incubation, 800 ul of the YES medium or cell culture medium was added to each vial and slowly mixed them using a pipette. Tubes were placed in the magnetic rack for 5 minutes to

gather all the nanoparticles together. After that, the supernatant was removed and collected in a fresh tube, which contains healthy dissociated cells. Cells were once again washed with PBS-BSA solution and pass through a 100-micron filter. Cells were counted using the C-chip method and used for the Drop-seq experiment.

5.2.4 Preparation of Drop-seq experiment

5.2.4.1 Oil channel setup

Droplet generation oil was loaded in a 10 ml syringe using 1 ml pipette. The 26G needle was attached to the syringe and plunger was slightly pushed up to remove all the air bubbles from the syringe. The syringe was loaded into the pump and tubing was attached to the needle. The pump has a clamp for pushing the syringe plunger. The clamp was adjusted to the syringe, and the flow rate of the oil pump was set to 30,000 ul/hr. The pump was switched on until the oil dripping out of the tubing from the other side. The pump was switched off, and once oil dripping was stopped, the free end of the tube was inserted into the leftmost channel of a clean droplet generation device on the microscope stage. Cutting of tubing with a sharp angle facilitate easier insertion into the Drop-seq device. The shorter tubing was inserted into the outlet channel, which was situated at the rightmost area of the device, and the free end was kept in a waste container at the beginning.

5.2.4.2 Bead channel setup

An aliquot of the beads from stock was taken, and beads were counted using the C-chip. Using below mentioned formula (Equation 1) bead concentration was determined and adjusted to have 120,000 beads/ml.

$$\text{Concentration in beads/ul} = (\text{Beads counted/number of boxes})*5*\text{dilution factor}$$

Equation 1: To count the beads in the C-chip.

In this equation, the dilution factor was 2 because the beads were mixed with 6x DNA dye in 1:1 ratio.

Beads tended to aggregate while counting in C-chip because of a slight charge present on their surface. To counteract this problem, beads were mixed with DNA 6X dye (NEB) in 1:1 ratio (dilution factor) for efficient counting. Beads were spun down in a tabletop centrifuge, TE-TW had been removed, and beads were resuspended in lysis buffer. For a standard Drop-seq experiment requiring 1 ml of bead flow, 50ul 1M DTT (Sigma) in 950 ul lysis buffer was mixed with the beads before droplet formation. For bead solution, a 3 ml Luer-lock syringe was used, and a magnetic mixing disc was placed inside the syringe by removing the plunger from behind. Bead solution was carefully filled in the bead syringe using 1 ml pipette, excess air and bubbles were pushed out, and a 26G needle was attached to the syringe. The tubing was inserted into the syringe from one end, and it was loaded onto a syringe pump designated for the beads (stands in a vertical position above the surface of the microscopic stage). Before loading the syringe on the pump, magnetic mixing system was switched on for continuous mixing of the beads and prevents them from settling down or blocking the needle attached to the syringe. Magnetic mixing system (VP Scientific) was set at the speed of 10-15 for Europe (for the USA, 25-30 speed was recommended because of different electric voltage standards). The flow rate of the bead syringe pump was adjusted to 30,000 ul/hr and run started for a short time to push out all air from the tubing until there was liquid dripping from the other end of the tubing. The pump was switched off, and another end of the tubing was inserted into the bead inlet of the microfluidic device. The flow rate was set to 4000 ul/hr for Drop-seq experiment.

5.2.4.3 Cells loading for the Drop-seq

Cells were dissociated as described in section (5.2.2) and after post-dissociation cells were resuspended in 1 ml of PBS-BSA, spun in a microcentrifuge at 300xg for 3 minutes. The supernatant was removed, and cells were resuspended in 1 ml of PBS and filtered through 40-micron filter. Cell counting was done using a C-chip and cells were resuspended in PBS-BSA with the final concentration of 100,000 cells/ml. Cells were mixed properly and loaded in a 3 ml Luer-lock syringe with the help of 1 ml pipette. The 26G needle was attached, and air bubbles were removed by pressing the plunger while keeping the syringe in vertical position. Cell syringe was mounted into the pump in the horizontal direction, same as we did with the oil pump and tubing was attached to the needle. The flow rate was adjusted to 30,000 ul/hr and pump was short run to push out all the air and bubbles from the system. The run was halted once liquid was

dripping from the other side of the tubing. Pump flow rate was set to 4000 ul/hr, and the free end was inserted into the microfluidic device cell channel.

5.2.4.4 Flow rates:
- Oil: 15,000 ul/hr
- Cells: 4,000 ul/hr
- Beads: 4,000 ul/hr

5.2.4.5 Starting Drop-seq run:
- Start order: cells -> beads -> oil
- Stop order: beads -> cells -> oil

Drop-seq run was started by first starting cell pump, then the beads (with continuous mixing of beads) and at last the oil pump was started. The rationale behind this order was that bead solution should not flow back into the cell channel, because of the lysis reagent present in the bead channel. Lysis reagent can lyse the incoming cells before they can reach the droplet formation junction. Similarly, bead flow should be stopped first once the Drop-seq run is finished or need to be stopped in between of an experiment. It was also helpful to pull out tubing from oil, bead and cell channel simultaneously from the device after switching off the pumps in between of the two cell loading in the same device or to reuse that device in the future experiment. It takes around 10 seconds to 1-minute time for all the flows to stabilize and start production of good quality droplets. The outflow tube was positioned in a waste container until the flow was stabilized. To check droplet emulation quality, outflow tubing was tilted to one side of the Falcon so that the droplets were running down from the side. Stable droplet production could be identified by hazy and uniform line because every droplet would be the same size. Stable droplet formation could also be determined by visualizing a flickering pattern at the droplet generation junction (appears as an elongated triangle) inside the Drop-seq microfluidic device under the microscope at 4x magnification (E. Macosko & Goldman, 2015b). Apart from this, the flow downstream to this junction was appeared "blurry" because of the fast movement of droplets (E. Macosko & Goldman, 2015b).

Clear stream downstream of droplet generation junction was an indication of no droplet formation. In a higher objective than 10x, it could be seen a narrow stream of oil running in the center of outflow and the surrounding was blurry in case of the droplets formation (E. Macosko & Goldman, 2015b). Once it was sure that stable droplets were forming and reached by the end of the outflow tubing (checked by the run-down method as described previously), the droplets collection tube was switched to a fresh 50 ml falcon to collect good quality droplets for further processing. Droplet generation junction (elongated triangle) and flow downstream to this junction were continuously monitored inside the microscope to maintain the overall good quality droplets collection for the experiment. The droplets made from 1 ml of the cell and 1 ml of the bead flow was collected into each 50 ml falcon tube. More than 1 ml cell product in each falcon could have an adverse effect on downstream processing (E. Macosko & Goldman, 2015a). If droplets were not being formed after some time or a microfluidic channel was blocked, it would be necessary to stop and restart some of the flows. Stopping the bead and cell pumps while continue running the oil flow for some seconds and then restart bead and cell flow back could be helpful in droplet generation with good quality (E. Macosko & Goldman, 2015b).

## 5.2.5 Assessment of stable droplet generation junction during a Drop-seq run

Droplet generation junction is the place in the Drop-seq microfluidic device where all three flows (cells, beads, and oil) merge to generate droplets. It appears like an elongated triangle. A good triangle was directly correlated with good quality droplet formation and a key for the Drop-seq experiment because of following reasons:

1. It indicates uniform droplet production.
2. It helps to figure out flow related problems, which could contribute to bad droplet quality.

The shape of a good triangle was different slightly from device to device, and it also depends on flow rates used in Drop-seq experiment (E. Macosko & Goldman, 2015b). If the triangle appeared little wide or small in shape, it was helpful to check that all tubing was fixed properly with the needles and there was no leaking between the junction point of the needle and tubing and device inlets of the tubing.

Here are some conditions, which can lead to the abnormal shape of a triangle:

1. **Decreased cell flow:** In this condition, the base of the triangle appears widens and outward curved (see Figure 5.2 (A)). It could occur either because of the cell solution was over, or it might indicate errors in the cell flow. For example, the cell solution might be leaking, the needle was not appropriately fixed with syringe, error in the pump, and so on.

2. **Decreased bead flow:** In this condition, the triangle appears elongated and narrow as shown in Figure 5.2 (B) (E. Macosko & Goldman, 2015b). It could occur either because of the bead solution was over, or it might be an error in bead tubing, clogging in bead needle or needle not fixed properly with bead syringe or pump errors.

3. **Pulsating triangle:** It would lead to the production of different size droplets. The reason behind this phenomenon was either magnet was touching or too close to the bead syringe or the bead pump. Sliding bead syringe could fix pulsating triangle.

4. **Tip of the triangle flicker:** It would also lead to the production of different size droplets. This phenomenon could be fixed by decreasing bead and cell flow rates. It could also emerge because of the error from outflow tubing or if it is longer than required. To stabilize flickering, tapping on the outflow tube was done.

A   B
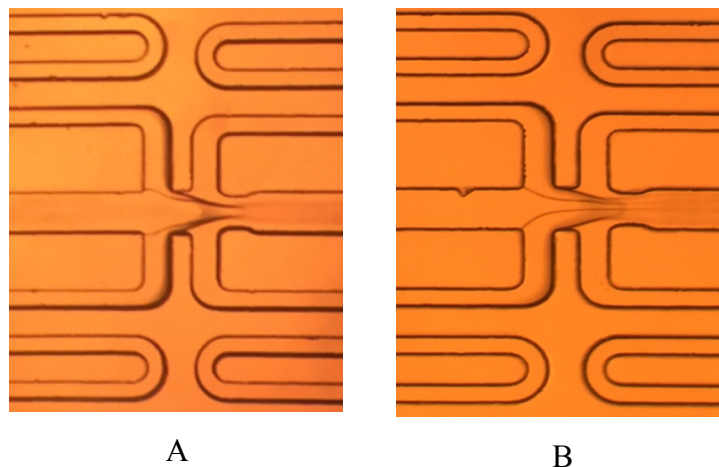
Figure 5.2: Abnormal droplet generation junctions.

(A) When cell flow decreased. (B) When bead flow decreased.

5.2.6 Quality control of droplets

To assess the quality of droplets generated in a Drop-seq experiment, 20 ul of oil containing an approximately 1/5th volume of droplet emulsion was loaded into a C-chip. It was rolled some time to equally distribute droplets on the surface. Droplets were observed using a microscope at 10x magnification. The droplets were seen as a homogeneous, circular shaped and transparent; and the beads were appearing as a smaller dark circle inside the droplets (E. Macosko & Goldman, 2015a). Cells could not be seen at this stage because of they were lysed by lysis buffer used in bead flow. A homogeneous population of droplets was considered to be high quality and droplets were counted to check for the presence of single bead or doublets of the beads in a single droplet as shown in (Figure 5.7). According to (E. Z. Macosko et al., 2015) bead doublet rate of less than 5% in an experiment is appropriate to proceed with the sample.

5.2.7 Processing of droplets

To prepare single-cell libraries for sequencing, following steps were used to recover beads, reverse transcription of RNA and generate libraries from pooled samples.

5.2.7.1 Breakage of droplets to collect the beads

To recover the beads along with captured RNA on the bead surface, the first step was to break the oil droplets and separate the oil from the beads. Because oil is heavier than the droplets, it sinks to the bottom of falcon tube. Excess oil was removed from a falcon tube using p1000 pipette. Pipette was pressed down to its first stop and pushed through the droplets layer to the bottom of the falcon tube. After that, the pipette was pressed until a second stop to remove any droplets inside the tip. Most of the oil was removed from the bottom of the falcon. In the droplets containing falcon, 30 ml of 6X SSC (Saline-sodium citrate buffer; 3M NaCl in 0.3M sodium citrate; pH 7.0) diluted in water, and 1 ml of Perflurooctanol (PFO) was added in the fume hood at the room temperature. The falcon was vertically shaken by hand 4 to 5 times for breakage of the droplets and centrifuged at 1000xg for 1 minutes. After spinning, most of the white beads were located on the interface of oil and SSC buffer. If many beads were still floating after

centrifugation, the falcon was again centrifuged for 1min at 1000g. It was also helpful to loosen the caps of the falcon to prevent the beads from floating up again.

Ice was pre-punched to keep falcon without disturbing the beads. Falcons were carefully placed into the ice after centrifugation. The supernatant was discarded from the top, leaving some liquid above the interface using a pipette. Another 30 ml fresh 6X SSC buffer was added on top of the beads to kick them up in SSC solution. After 10 seconds, once all the oil was sunk to the bottom of the falcon, supernatant above the oil surface (containing beads) were taken in a new Falcon tube. White beads could be seen in the Falcon. Beads were spun again at 1000xg for 1 minutes. All the beads were at the bottom after this centrifugation step. The supernatant was discarded, leaving 1 ml of SSC buffer. After mixing of the beads, it was transferred to a 1.5 ml DNA low binding tubes and centrifuged using a mini-centrifuge. The supernatant was removed, and beads were washed twice with 1 ml of 6X SSC buffer and one time with 300 microliters of Maxima 5X RT buffer (ThermoFisher Scientific). After removing 5X RT buffer, beads were processed for reverse transcription and further steps.

5.2.7.2 Reverse transcription of RNA

200 ul of RT mix (see 2.1) was added to the vial containing beads and incubated at room temperature for 30 minutes on a rotating wheel. The vials were further incubated at 42$^\mathrm{o}$C for 90 minutes in a lab incubator (Eppendorf) with intermediate shaking by hand after every 10 minutes. After incubation, beads were washed with 1 ml of TE-SDS and two times with 1 ml TE-TW solution. Afterward, beads were washed with 1 ml of 10 mM Tris (pH 8) prepared in water, and proceeded to exonuclease I treatment. Reverse transcription step generates cDNA on the surface of the beads using RNA as a template (hybridized by the poly A tail to the bead primer). One RT mix reaction was used for around 50K beads. Beads could be stored after RT step for some time in TE-TW solution at 4$^\mathrm{o}$C (maximum one month).

5.2.7.3 Exonuclease I treatment of beads to remove excess primers

This treatment digests excess primers on the surface of the bead that was not captured by an RNA molecule. 200 ul of the exonuclease I mix was sufficient for the processing of approximately 90,000 STAMPs. After reverse transcription steps and washing with 1 ml of 10mM Tris (pH 8),

beads were mixed with 200 ul of exonuclease mix. Reaction tubes were incubated at 37°C for 45 minutes on a thermoshaker with the rotation of tubes manually every 10 minutes intervals. After the Exonuclease-I treatment, beads were washed with 1 ml of TE-SDS and two times with 1 ml TE-TW. Beads could be stored after Exonuclease-I treatment for some time in TE-TW solution at 4°C.

5.2.8 cDNA amplification PCR (PCR-I)

cDNA amplification PCR step was employed to increase the concentration of individual full-length cDNA molecules (transcribed from RNA of single-cells) to preserve them from loss and process them for library preparation steps in downstream procedures. Beads were homogenized properly in 1 ml TE-TW solution, and 20 ul of the beads were taken out and mixed with 20 ul of 6X DNA dye in 1:1 ratio. As mentioned previously, 6x DNA dye prohibits the aggregation of beads, and it helps in evenly spreading of beads in the C-chip. Beads were counted on the C-chip using the (Equation 1). All 16 boxes were counted for the good estimation of the bead numbers. Beads were washed with 1 ml of H_2O, and approximately 5000 beads were distributed into each PCR tube. PCR tubes were spun down, the supernatant was removed, and PCR mix was added to each tube as follows:

H2O	21 ul
10 uM SMART PCR primer	4 ul
2x Kapa HiFi Hotstart Readymix	25 ul

PCR tubes were mixed well-using pipette up and down and following PCR program was used to amplify the cDNA.

98°C for 3 minutes.	1 cycle
98°C for 20s 65°C for 45s 72°C for 3 minutes.	4 cycles
98°C for 20s 67°C for 20s 72°C for 3 minutes.	9 cycles

| 72°C for 5 minutes.
 4°C forever | |

These 13 cycles were used for the cells and conditions in this thesis. A number of cycles depends on the type of cells used for the experiment and should be optimized for each cell type. Remaining beads could be stored in TE-TW at 4°C.

5.2.8.1 AMPure purification and Bioanalyzer analysis of libraries

cDNA libraries need to be purified to remove excess primers, enzymes and other components which can affect downstream library preparation procedures. The container of the AMPure beads was thawed at room temperature for 5 minutes and vortexed to mix the beads. 0.6x beads to sample ratio was used for this purification. 30 ul of AMPure XP beads were added to the amplified product (50 ul) from the previous reaction and incubated at room temperature for 15 minutes. PCR tubes were put on a magnetic stand to allow aggregation of AMPure beads. The supernatant was removed, and beads were washed with 80% ethanol three times. During each washing step beads were incubated with 80% ethanol for 30 seconds. After the last ethanol wash, tubes were centrifuged briefly, and every drop of ethanol was removed. Beads were air dried for a few minutes. 10 ul of H_2O was added to each tube, mixed properly and incubated at room temperature for 2 minutes. PCR tubes were again put on a magnetic stand to aggregate the beads. The supernatant was collected as a final product from this purification. The ampure purification process was repeated one more time using 0.6X AMPure beads. High sensitivity DNA bioanalyzer (Agilent Technologies) was used according to manufacturer's protocol to assess the quality and concentration of the cDNA libraries. 1ul of the purified product was used as an input for bioanalyzer analysis.

5.2.9 Tagmentation of cDNA using Nextera XT kit and PCR-II

Library sizes were adjusted by using Nextera XT DNA library preparation kit (Illumina). Libraries of cDNA in the previous step had an average size of 1300-2000 bp. Although, the optimum library size for NGS sequencing needs to be much less. This step employs an engineered Tagmentase enzyme (Nextera XT kit) to cut the cDNA molecules in shorter sizes

(average 500-700 bp), which is appropriate for pair-end sequencing of Drop-seq libraries. For each sample, 600 pg of purified cDNA was taken and made up to 5 ul final volume with H_2O in a PCR tube. 10 ul of Nextera TD buffer and 5 ul of Nextera Amplicon Tagmentase enzyme were added to each tube (final volume 20 ul). The solution was mixed by pipetting at least 5 times, and afterward tubes were centrifuged. Thermocycler was preheated to 55°C and tubes were incubated for 10 minutes at 55°C. After incubation, 5 ul of Neutralization buffer (Nextera XT kit) was added and mixed by pipetting 5 times. Tubes were centrifuged and incubated at room temperature for 5 minutes. Bubbles could be seen in the tube because of the viscosity of the solution. For the final library amplification, PCR-II mix was added to each tube in the following order:

Nextera PCR mix	15 ul
H2O	8 ul
10 uM New-P5-SMART PCR hybrid oligo	1 ul
10 uM Nextera Index (N70X oligo)	1 ul

PCR-II mix selectively enrich cDNA fragments from 3' end of RNA (containing cell barcode and PCR handle) so that libraries have cell identity information intact for each cDNA molecule sequenced. Following PCR program was used for PCR-II amplification

98°C for 30 sec	
98°C for 10s 55°C for 30s 72°C for 30s	14 cycles
72°C for 5 minutes. 4°C forever	

After this PCR reaction, tagmented libraries were purified with AMPure XP beads using the same protocol as described in the previous step, 2 times using 0.6x beads and 1 time using 1x AMPure XP beads. The final product was eluted with 10 ul of H_2O, and Bioanalyzer High sensitivity DNA assay was ran according to manufacturer's instructions. 1 ul of the purified cDNA libraries were used as an input. Libraries were also quantified using Qubit DNA high sensitivity reagent. Tagmented Drop-seq libraries were reasonably smooth with an average size of 500-700 bp.

100

Smaller than 400 average size libraries could have more poly A reads; larger libraries might have a lower cluster density and cluster quality for sequences.

5.2.10 Preparation of custom Read-1 primer

The stock primer was stored at 100 uM (or 100 pmoles/ul) in 10 mM Tris-HCl, pH 8.0. Immediately before loading into the sequencer, the 100 uM custom Read-1 primer was diluted to 10 uM using 10 mM Tris-HCl at pH 8. After that 30 ul of the 10 uM Custom read primer was diluted with 600 ul of the chilled HT1 buffer to make a final solution of 0.5 uM custom Read-1 primer. Primer solution was vortexed and centrifuged briefly. Primer solution was loaded into position 18 of the Miseq reagent cartridge. In the Instrument Experiment Manager sample sheet, Read-1 custom primer usage was indicated. Otherwise, the sequencer could use the Illumina Read-1 primer by default and which could lead to failure of the Drop-seq library sequencing.

5.2.11 Sequencing of Drop-seq libraries

Following specifications were used for sequencing of Drop-seq libraries in Illumina Mi-seq and Hi-seq 2500 sequencer:

Read-1: 25 bp
Read-2: 50 bp
Read-1 Index: 8bp (Multiplexing of samples)
Custom Read-1 primer

Library concentration using Qubit measurement and average size from the High sensitivity DNA Bioanalyzer was used to make a 3 nM library solution in EB buffer. Following steps were followed to prepare libraries for sequencing. To denature the Drop-seq libraries, 10 ul of 3 nM library samples were mixed with 10 ul 0.2N NaOH. It leads to 1.5 nM concentration of the samples. Solutions were vortexed briefly and centrifuged to collect contents. The sample was incubated at room temperature (25°C for 5 minutes). 20 ul denatured library was diluted with 980 ul chilled HT1 buffer to make a 30 pM library solution with 2 mM NaOH concentration. Following calculation describe this concentration:

C1V1=C2V2

1.5 nm * 20 ul = C2 * 1000 ul

C2 = 0.03nM

1nM = 1000pM

C2 = 0.03 nM * (1000 pM/1nM) = 30 pM

400 ul of 30 pM library solution was mixed with 600 ul of chilled HT1 buffer for a final concentration of 12 pM sample and <1 mM NaOH. The final solution was used for cluster generation for Mi-seq machine and also for Hi-seq 2500 sequencer.

5.2.12 Computational analysis of the Drop-seq sequences

To process Drop-seq sequence data, the Drop-seq software pipeline was implemented, provided online by the developers of the Drop-seq method (Nemesh & Macosko, 2015). This software pipeline transforms raw sequence data into a "digital gene expression" (DGE) matrix, which contains integer counts of the number of transcripts for all the genes and all of the cells. The Drop-seq data analysis includes massive de-multiplexing of the raw data, alignment of reads to a reference genome and cellular and molecular barcodes processing (Nemesh & Macosko, 2015). Sequencing libraries of Drop-seq experiments produce paired-end reads. Read-1 contains information about 8 bp cell barcode and 12 bp molecular barcode (UMI). On the other hand, Read-2 contain transcript information, which could align to the reference genome (Nemesh & Macosko, 2015).

To begin with the analysis, the raw data was processed by the Illumina's *bcl2fastq* program to generate Fastq files. These Fastq files were processed with the Picard *FastqToSam* program to generate unmapped, queryname-sorted BAM files. Afterward, following steps were executed to align the raw reads and to create a BAM file that was suitable to produce digital gene expression (DGE) analysis (Nemesh & Macosko, 2015).

    a. Cell barcodes tagging

    b. Molecular barcodes tagging

    c. 5' primer sequence trimming

d. 3' polyA sequence trimming

e. SAM to Fastq conversion

f. Alignment to the reference genome using STAR

g. STAR alignment sorting in queryname order

h. STAR aligned tagged SAM file merged with recovered cell and molecular barcodes

i. Addition of gene, exon, and other available annotation tags

j. Detection of bead synthesis errors

These programs were part of the Drop-seq software toolkit and also employ other programs like Picard tools and alignment tool STAR. After the execution of the steps mentioned above, alignment steps were completed and the pair end raw reads were converted into single end reads with the cell and molecular barcodes attached, cleaned, aligned to reference genome and prepared for digital gene expression analysis (Nemesh & Macosko, 2015).

To digitally count the gene transcripts, a list of molecular barcodes (UMIs) for each gene, within each cell, was assembled and molecular barcodes within edit distance=1 were merged. The sum of unique molecular barcodes was reported as the number of transcripts of that specific gene in a given cell. The Drop-seq program DigitalExpression was used to extract digital gene expression (DGE) data from the aligned libraries from the previous steps (Nemesh & Macosko, 2015). The input of the digital expression program was the aligned BAM file generated in the previous workflow. There were two output files presented by this program. The first was the DGE matrix and the second was a summary of the DGE matrix, which contains the number of genes and transcripts observed on a per cell level (Nemesh & Macosko, 2015). For the secondary analysis of digital gene expression matrix, Seurat package (Butler & Satija, 2017) was used following the instructions from Seurat tutorials (Satija, 2017).

5.3 Results from Drop-seq optimization

5.3.1 Optimization of Drop-seq quality using Human-Mouse cell mixing

Human-Mouse mixed cell Drop-seq experiment was performed as described in the method section (5.2.1) to ensure that parameters used in Drop-seq optimization were not producing cell doubles attached to the bead-containing a single-cell barcode. After droplet processing, reverse transcription and exonuclease-I treatment beads were counted in C-chip (Equation 1 at page 91). Approximately 5000 beads were distributed in each PCR tube, and samples were amplified using cDNA amplification (PCR-I) protocol as described in the method section (5.2.8) . Bioanalyzer profile of these samples after cDNA amplification showed that cDNA molecules are with full length (around 1200-1600 average sizes) as they reverse transcribed from RNAs (Figure 5.3)
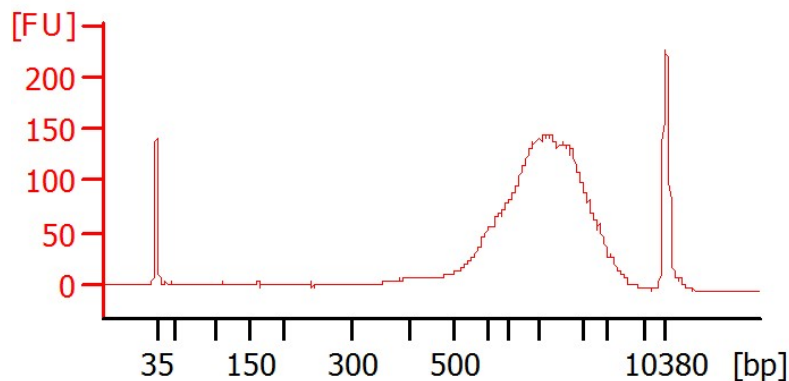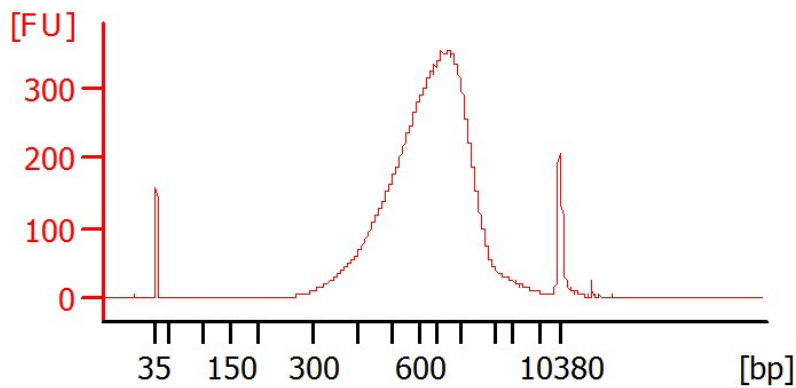

Figure 5.3: Bioanalyzer profile of Human-mice mixed cell Drop-seq samples after cDNA amplification (PCR1).

The x-axis represents the library size in base pairs, and Y-axis indicates the concentration. The first and the last peak in the graph are from low and high marker respectively.

After the PCR-I, cDNA molecules were subjected to tagmentation process and subsequent amplification. Full-length cDNA molecules create adverse effects in pair-end sequencing technique. The tagmentation process cuts the cDNA molecules in shorter sizes and adds a known flanking sequence towards both ends of resulting cDNA chunks. This sequence from *Tagmentase* enzyme was used in the second round of selective amplification, which uses PCR handle sequence from the bead primer as well as sequence added by the *tagmentase* enzyme to amplify

the cDNA libraries. In this amplification process, indexes (also contain adapters) from the Nextera XT index kit (Illumina) to attach to the Next generation sequencing platform were also added. The resulting libraries only have molecules, which have 12 bp cell barcode and 8 bp UMI intact in the cDNA. Other cDNA chunks would not amplify in this step (PCR-II). Libraries resulted after this step was relatively smooth and had an average size of 400-700bp as described in Figure 5.4.


Figure 5.4: Bioanalyzer profile of Human-mice mixed cell Drop-seq samples after PCR-II.

The x-axis represents the library size in base pairs, and Y-axis indicates the concentration. The first and the last peak in the graph are from low and high marker respectively.

5.3.1.1 STAMPs count determination

To determine the number of STAMPs present in a sample, a histogram method (Bam-Tag histogram) was used (Figure 5.5). Drop-seq program *BAMTagHistogram* was used to generate this histogram. After assigning transcript reads to individual cell barcodes, they were sorted in the descending order. The cumulative fraction of reads (of sorted cell barcodes) was plotted on the Y-axis and the X-axis, a number of cell barcodes were plotted. The knee of the plot indicates the number of the STAMPs present in a sample. In Human-mice cell mixing experiments, 180 cells were detected using this technique (Figure 5.5).
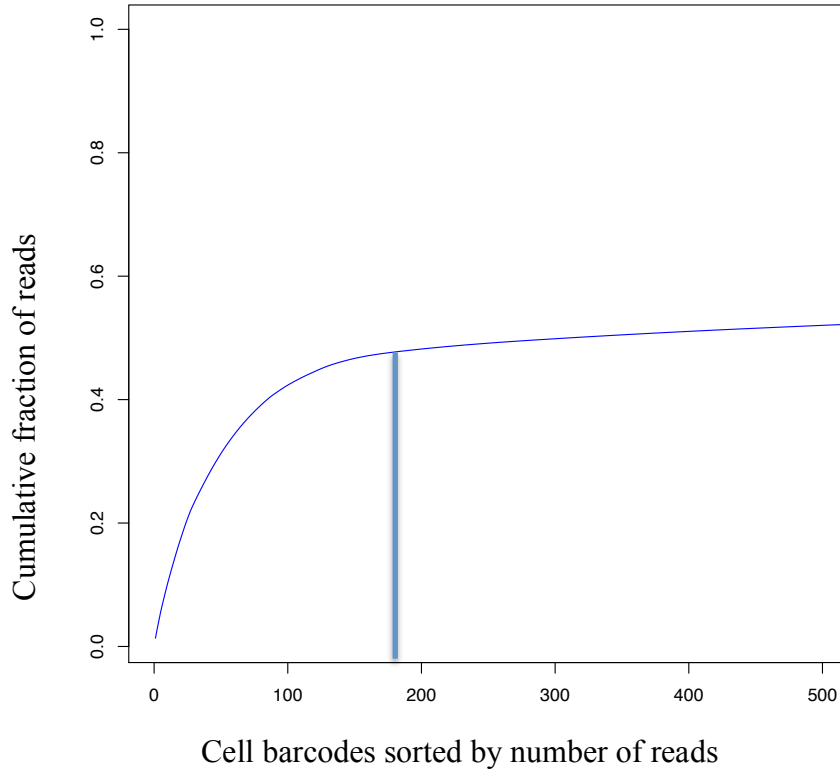
Figure 5.5: Bam-tag histogram for cell counts.

The x-axis represents the number of cell barcodes (sorted by a number of reads), and Y-axis represents the cumulative fraction of the reads.

5.3.1.2 Species mixing plot

The species mixing plot was used to determine the cross-contamination of two different species among the single-cell transcriptomes. In this method, all the reads assigned to a STAMP are mapped against a custom-made transcriptome, which has a human as well as mouse transcript information (flagged according to species). After that, if an individual cell has mapped against 80% or more human transcripts, the cell was assigned to Human cells. On the other hand, if the reads from a cell have mapped 80% or more against mouse cells, the cell was assigned as a mouse cell. In species mixing plot, different single-cells were plotted for Human transcripts on X-axis and mouse transcripts on the Y-axis. Results show that individual cells exclusively belong to one species and mixed signals which could be an indication of cell doublets or another kind of data quality complications, is not present in Human-mice cell mixing experiment (see Figure 5.6).

106
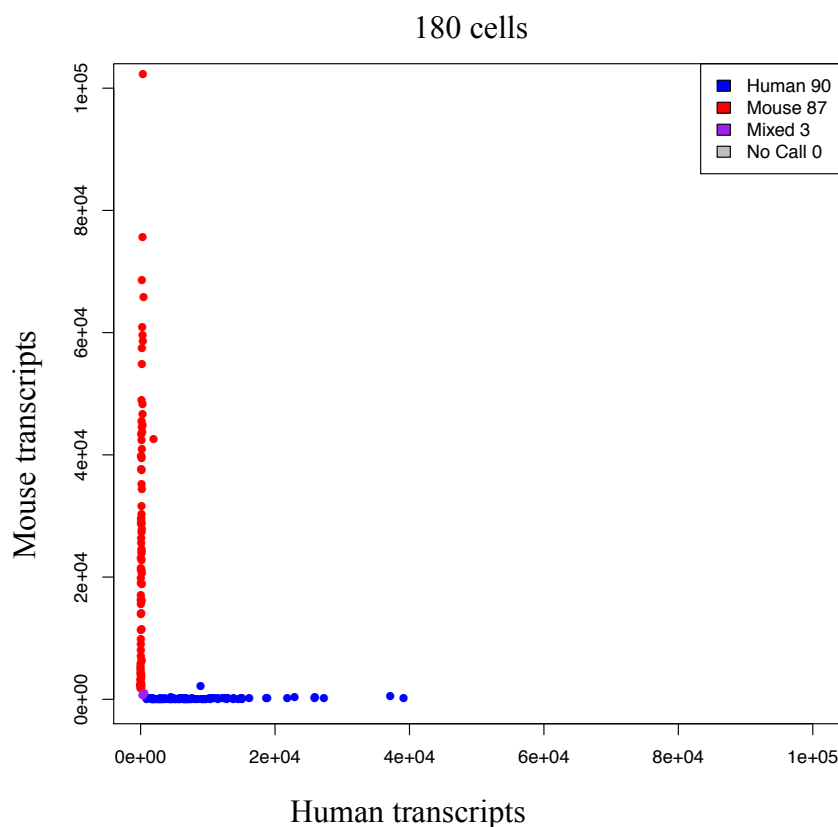
Figure 5.6: The species-mixing plot for Human-mouse cell mixing Drop-seq experiment.

X and Y-axis are representing the number of transcripts present in individual cells, mapped to human genome or mouse genome respectively. In total 180 cells were present; 90 cells were mapped to Human transcripts, and 87 cells were mapped to mouse transcripts. Three cells having a low number of transcripts could not be determined for Human or mouse, so that, depicted as mixed cells.

5.3.1.3 Droplets quality analysis for bead doublet using visual inspection

To ensure Drop-seq experiment parameters are not producing oil droplets with more than one bead, following test was done. An aliquot of 3ul droplet emulsion and 17ul oil was taken using p200 and carefully loaded onto the platform of a C-chip. It was rolled back and forth 4-5 times and visualized using a microscope. Droplets were checked for the presence of one or more beads (Figure 5.7). In this observation, droplets with only one bead could be seen, which indicates the excellent quality of the droplets with set parameters. In Figure 5.7, cells were not visible at this stage because of the lysis buffer presence inside the droplets.
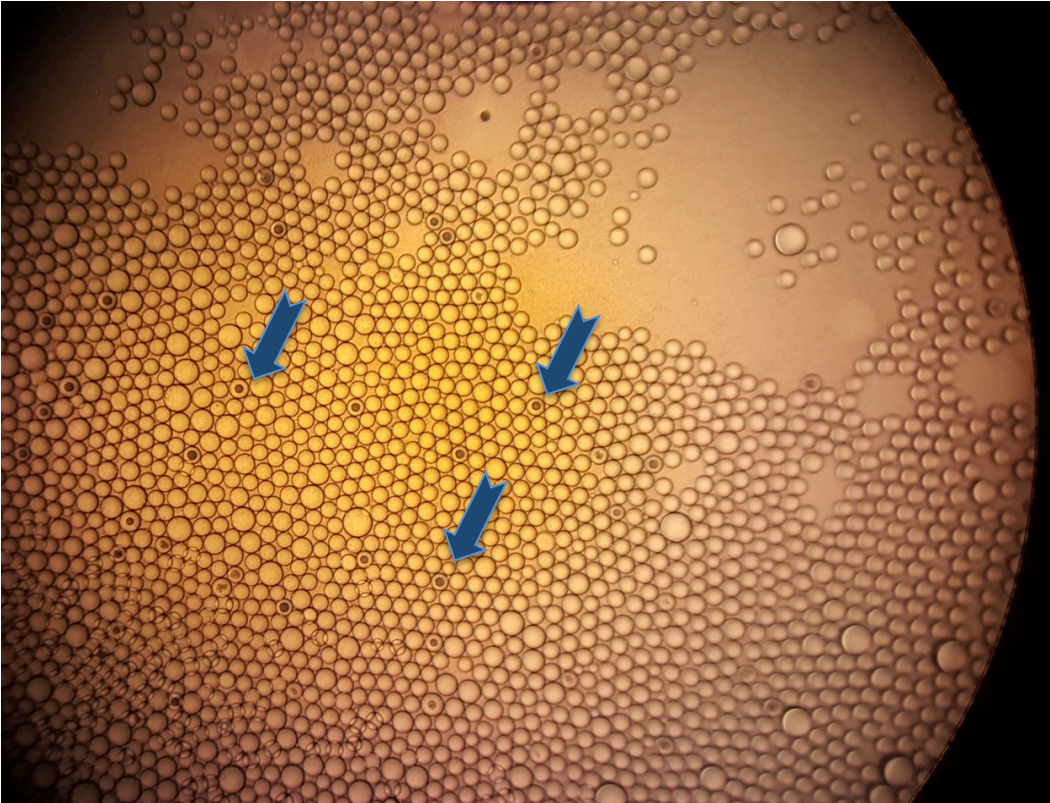
Figure 5.7: Droplets quality control for bead doublets.

Blue arrows indicate representative droplets with one bead.

5.4 Characterization of sub-population of motor neurons in spinal cord tissue using the Drop-seq method (case study-I)

After optimizing the Drop-seq parameters on the Human-mice cell mixing experiment, Drop-seq experiment on spinal cord tissue was carried out. P14 mice were used for this experiment and processed with dissection, and cell dissociation (papain, trehalose, Kynurenic acid, AP-V, and DNase) as discussed in the section (5.2.2). Followed by droplet generation, reverse transcription and exonuclease-I treatment beads were counted using the C-chip. Approximately 50000 beads were present in the sample at this stage. In PCR tubes, 5000 beads were added to each tube and samples were amplified using cDNA amplification (PCR-I) protocol as described in the method section. Remaining beads were stored for any future use at the 4°C. Figure 5.8, represents the bioanalyzer profile of motor neuron sample libraries after PCR-I.
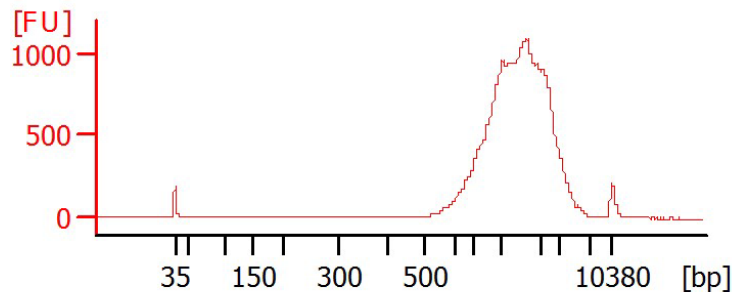
Figure 5.8: Representative bioanalyzer profile of spinal cord Drop-seq sample after cDNA amplification PCR-I.

The first and the last peak in the graph are from low and high marker respectively. The x-axis represents the library size in base pairs, and Y-axis indicates the concentration. The average size of the libraries was around 1600 bp at this stage.

After PCR-I, cDNA molecules detached from the beads and came into solution. cDNA was purified using AMPure XP beads as described in method section and after quantification using bioanalyzer, 600 pg of cDNA from each tube was used for tagmentation process and later another PCR reaction (PCR-II) together with adapters and indexes for high throughput sequencing. After PCR-II, libraries were again purified with AmpureXP beads to remove unused primers, enzyme, and other impurities. Libraries were dissolved in 10ul H_2O. Figure 5.9, represents the bioanalyzer profile of motor neuron Drop-seq libraries after PCR-II.
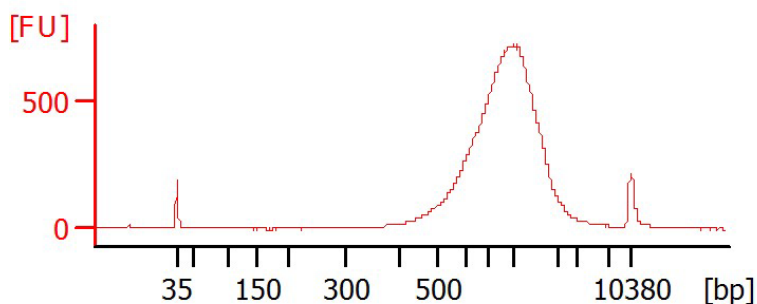
Figure 5.9: Representative bioanalyzer profile of spinal cord Drop-seq sample after PCR-II.

The first and the last peak in the graph are from low and high marker respectively. The x-axis represents the library size in base pairs, and Y-axis indicates the concentration. At this stage, libraries were relatively smooth and had an average size around 950 bp.

5.4.1 Sequenced data quality of spinal cord Drop-seq libraries

Before processing raw data generated by high throughput sequencer, data quality has been checked for per base sequencing quality and many other parameters. For this purpose, raw Fastq files were processed using the *FastQC* software. For Drop-seq libraries, pair-end sequencing generates two Fastq files:

Read-1: Cell barcode and UMI

Read-2: transcript sequence

The spinal cord Drop-seq libraries were first sequenced in Mi-seq. Table 5.1 provides information about basic statistics of Read1 of one pool of Drop-seq library. Furthermore, Figure 5.10 describes the per base sequence quality of the Read-1 for the Drop-seq library.

Measure	Value
Filename	Drop-seq_MN_sept2017_merged_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	178741389
Sequences flagged as poor quality	0
Sequence length	26
%GC	41

Table 5.1: Basic statistics for the Read-1 of a spinal cord Drop-seq library.
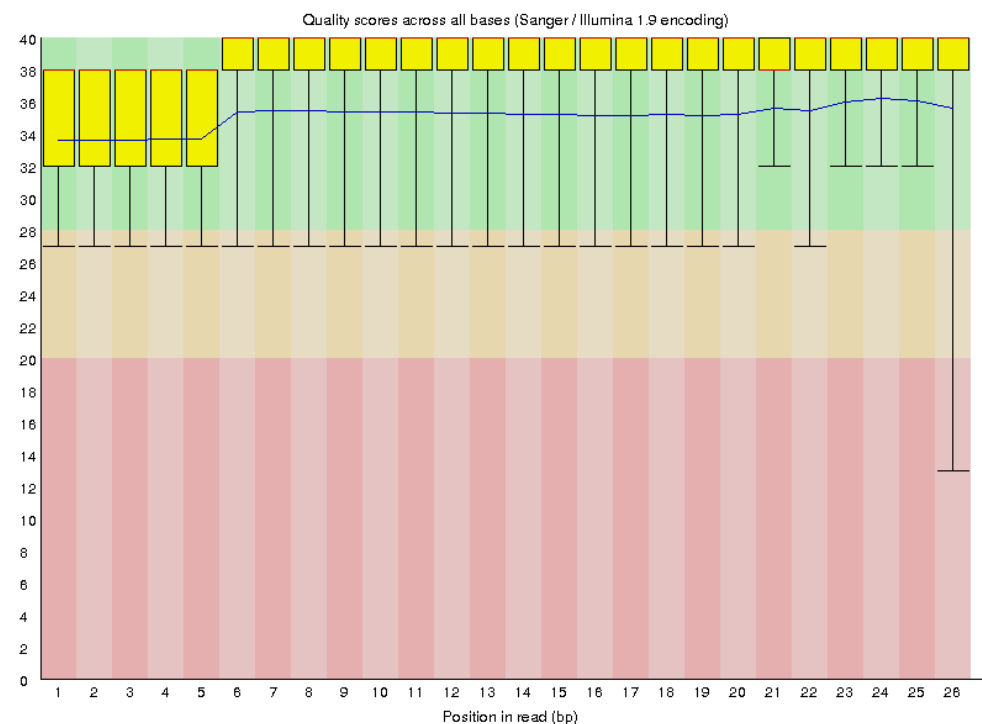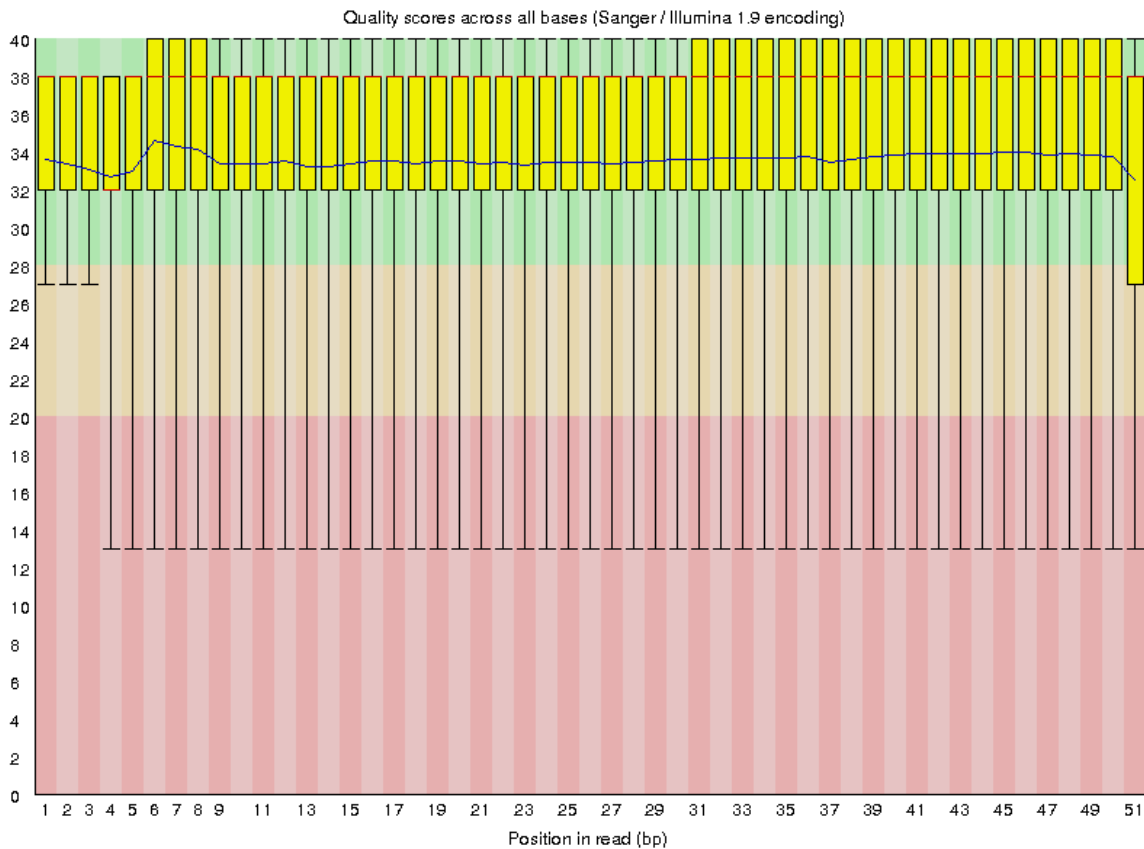

Figure 5.10: Sequencing quality (per base) for the Read-1 of a spinal cord Drop-seq library.

On X-axis, base pair positions of sequenced read are depicted. Y-axis, represent Q (Phred score) in log scale. As such, Phred score 20 and more considered as good quality.

Apart from this, several other quality scores were also presented by the FastQC program, which is not presented in this thesis. Similarly, Table 5.2 provides information about basic statistics of Read-2 of the same pool of Drop-seq library and Figure 5.11 describe the per base sequence quality of the Read-2 for the Drop-seq library.

Measure	Value
Filename	Drop-seq_MN_sept2017_merged_R2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	178741389
Sequences flagged as poor quality	0
Sequence length	26-51
%GC	43

Table 5.2: Basic statistics for the Read-2 of a spinal cord Drop-seq library.

Figure 5.11: Sequencing quality (Per base) for the Read-2 of a spinal cord Drop-seq library.

On X-axis, base pair positions of sequenced read are depicted. Y-axis, represent Q (Phred score) in log scale. As such, Phred score 20 and more considered as good quality.

5.4.2 Spinal cord Drop-seq data analysis

The primary data analysis was done following the Drop-seq analysis pipeline as described in the method section. Summary of the output results from each step is as follows (Table 5.3).

Process	Relevant Statistics
Create Unmapped BAM	
Tag Cell Barcodes	Number of Reads (Left and Right): 178,741,389
Tag Molecular Barcodes	Number of Reads (Left and Right): 178,741,389
Filter BAM	Number of Reads: 137,153,711
Trim 5' primer sequence	Number of Reads: 137,153,711
Trim 3' polyA sequence	Number of Reads: 137,153,711
Convert SAM to Fastq	
Star Alignment	Number of input reads: 137,153,711 Average input read length: 48 Uniquely mapped reads number: 111,109,025 Uniquely mapped reads %: 81.01%

Table 5.3: Statistics of reads after each step in Drop-seq pipeline.

5.4.2.1 STAMPs count determination

For motor neuron Drop-seq experiment, to determine the number of sequenced STAMPs, Bam-Tag Histogram was prepared (Figure 5.12). After assigning transcripts to individual STAMPs, they were sorted in descending order. The cumulative fraction of reads and the cell barcodes were plotted on the Y-axis and X-axis respectively. The knee of the plot indicates the number of cells present in the sample. In motor neuron experiment, top 1000 highest transcripts containing STAMPs were chosen for further analysis.
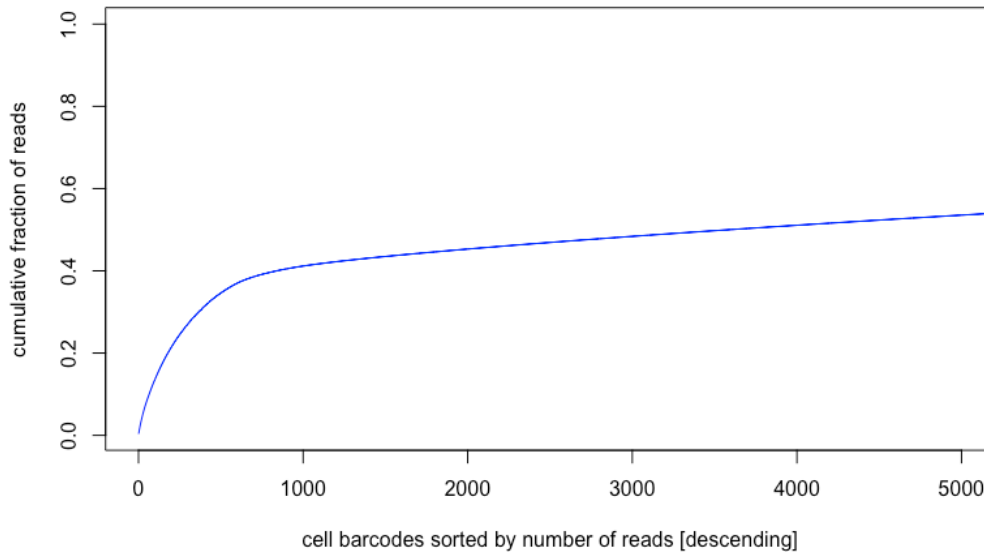
Figure 5.12: Bam-tag histogram for spinal cord Drop-seq experiment.

The X-axis represents the number of cell barcodes (sorted by the number of reads), and Y-axis represents the cumulative fraction of the reads. In this experiment, 1000 STAMPs were chosen for the downstream analysis.

5.4.2.2 Quality control and further cells selection using Seurat

The digital gene expression (DGE) file, which was obtained by running Drop-seq analysis pipeline, was further subjected to an R based toolkit Seurat (v2.0) for single-cell genomics. To process scRNA-seq data using Seurat the following steps were carried out. After creating the Seurat object using the *CreateSeuratObject* function, single-cells were filtered out based on technical and biological parameters. The number of genes and UMIs (nGene and nUMI) were calculated for each object using Seurat. The percentage of mitochondrial genes were also calculated and used as one of the QC metric criteria to filter out cells containing a high number of mitochondrial genes. Using violin plot (Figure 5.13) and gene plot (Figure 5.14) rare subset of cells which had more than 20% of the reads from mitochondrial genes, were excluded from the dataset. Similarly, the cells, which have unique gene counts more than 7000 (possible doublet of cells), or less than 1000, were filtered out before data analysis. After this quality control filtration, 837 cells were remaining in the motor neuron Drop-seq dataset.

114
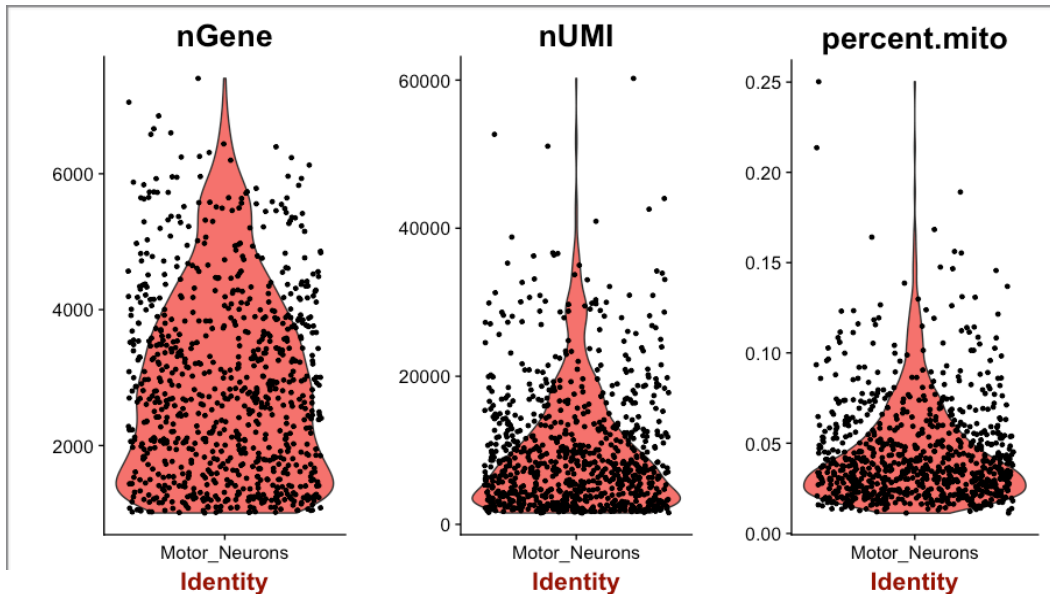
Figure 5.13: Violin plot for the gene, transcript, and percent of mitochondrial reads.

The Y-axis represents the number of genes (nGene), the number of UMIs (nUMI) and percentages of the mitochondrial reads (percent.mito) in the total reads respectively (from left to right).
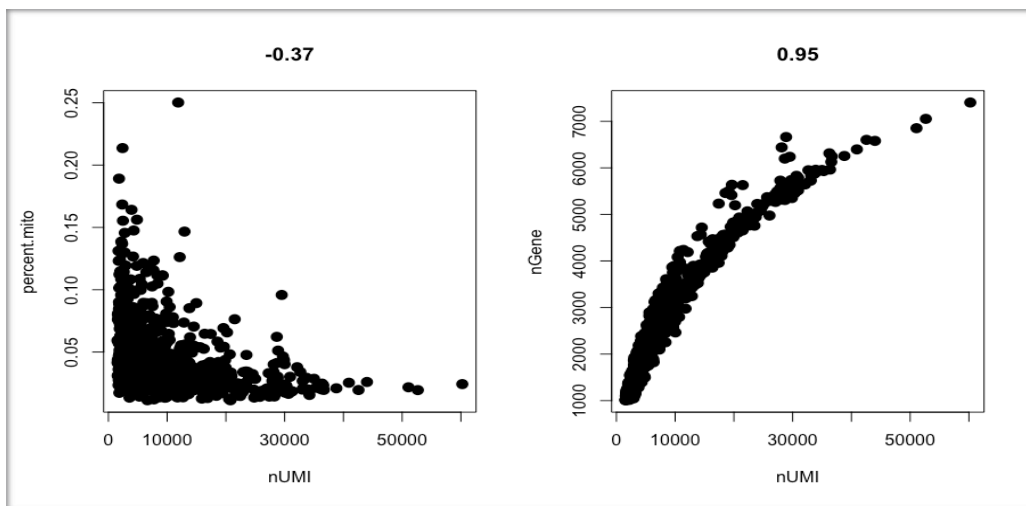
Figure 5.14: Gene plot for spinal cord Drop-seq.

The plot, at the left side, represents the distribution of mitochondrial percentage (percent.mito) and transcripts (nUMI) for the STAMPs/cells detected in the spinal cord Drop-seq experiment. The plot, at the right side, represents the number of genes (nGene) and transcripts (nUMI) for the same STAMPs. Each dot in the figure above represents a STAMP/cell.

5.4.2.3 Data normalization

After filtering out bad quality cells (mitochondrial genes > 20%; genes > 7000 and genes <1000) from the dataset, the following step was to normalize the Drop-seq data. Seurat package employs a global-scaling normalization method called *LogNormalize*. It normalizes the gene expression measurements for each cell by the total expression multiplied by a scale factor (by default 10,000) and log-transformed the result (Butler & Satija, 2017).

5.4.2.4 Variable genes detection in the Drop-seq dataset

In single-cell data analysis, variable genes are more interesting to determine cell markers or to study differential expression of genes among individual cells. Seurat package calculates highly variable genes, which could be used for downstream analysis. *FindVariableGenes* program calculates the dispersion and average expression for each gene, and place these genes into bins and then calculates a z-score for dispersion within each bin (Satija, 2017). Resulting graph from this analysis is depicted in

Figure 5.15. These results enabled the identification of variable genes while controlling for a strong relationship between variability and average gene expression.
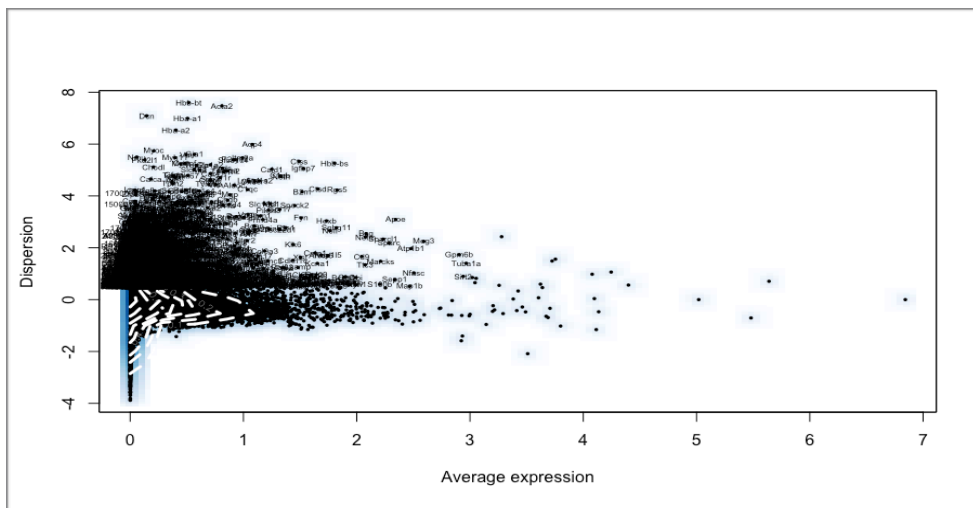
Figure 5.15: Dispersion and average expression plot for the dataset.

The X-axis represents the average expression and the Y-axis represents the dispersion of the genes. Using this graph, highly variable genes were selected for downstream analysis.

5.4.2.5 Data scaling and removing uninteresting sources of variation

Single-cell datasets contain many unwanted sources of variations, for example, technical noise, batch effects, cell cycle variation and so on. (Buettner et al., 2015) Suggested that by regress out these variations from the analysis, it could improve downstream processes such as dimensionality reduction and clustering. Seurat package implements linear models for the prediction of gene expression based on user-defined variables. The scaled z-scored residuals of linear models were used for dimensional reduction and clustering analysis.

5.4.2.6 Linear dimensional reduction

After removing technical variables and other uninteresting variations, PCA analysis on scaled data was performed. Generally, dimensional reduction on highly variable genes could increase performance. However, in UMI containing data, after removing technical variables, PCA remain the same. For motor neuron dataset, *VizPCA* function in Seurat package plotted the highly variable genes, as described in Figure 5.16 (also see PCA plot in Figure 5.17).
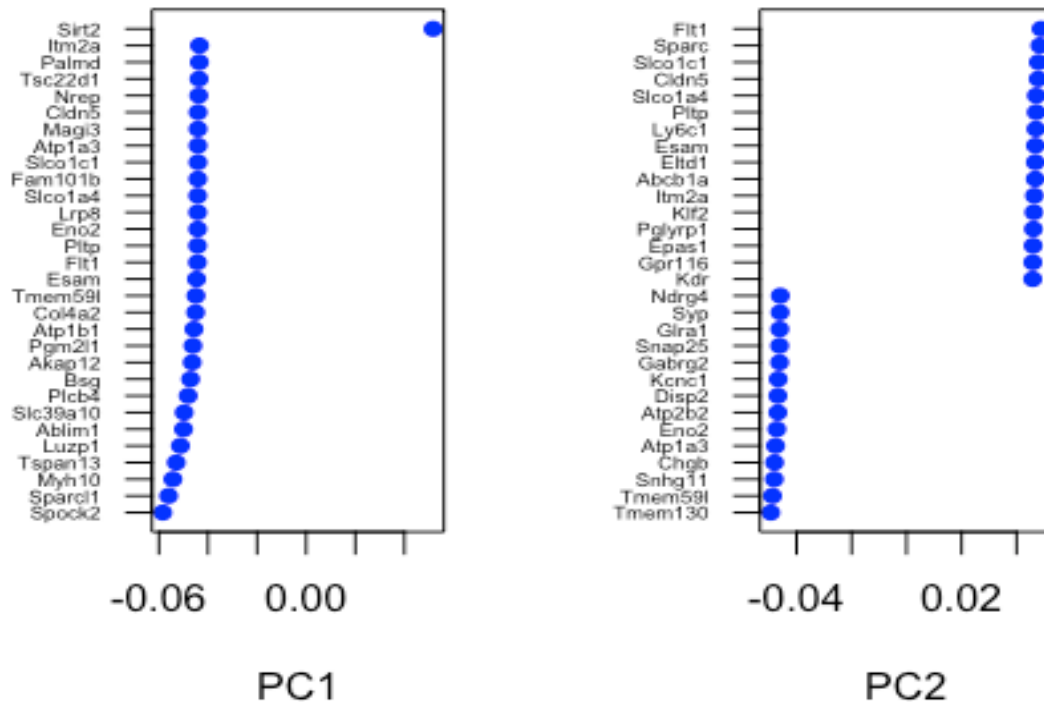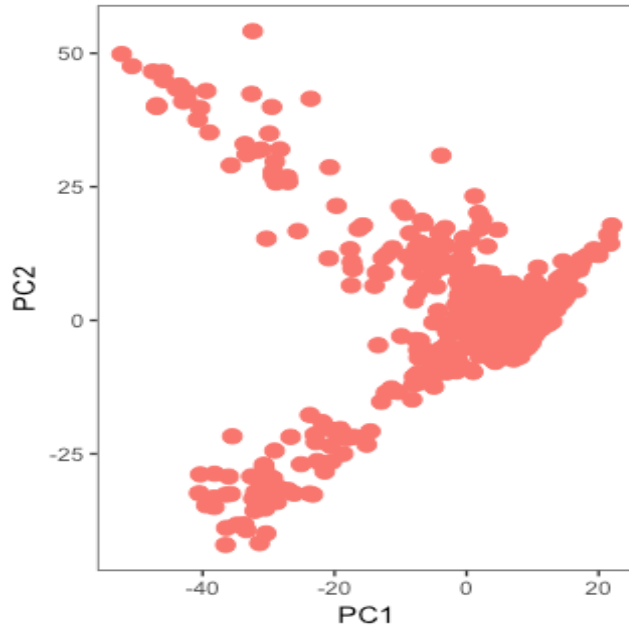

Figure 5.16: Genes enriched for PCA analysis in spinal cord Drop-seq dataset.

Figure 5.17: PCA plot for all the single-cells in spinal cord dataset.

After PCA plot analysis, motor neuron Drop-seq data was investigated for the primary sources of heterogeneity using a *PCHeatmap* function from Seurat package. It was also helpful to decide the number of PCs to add in the further downstream analysis (Butler & Satija, 2017). *PCHeatmap* ordered cells and genes according to their corresponding PCA scores (Figure 5.18). Although, this method involves a supervised analysis, it is beneficial for investigation of correlated gene sets.

Figure 5.18: PC Heat map for first 9 PCs, which were selected for downstream analysis.

5.4.2.7 Determining statistically significant principal components

To reduce technical noise in Drop-seq data, clustering of the cells was done, based on their PCA scores using the Seurat package. Each PC was representing a metagene, which combines information across a correlated gene set (Satija, 2017). Therefore, it was an important step to determine the number of PCs to include in the downstream procedures. The PCs selection was implemented by a resampling test, which was based on Jackstraw method (Chung & Storey,

2015). A subset of the data (1%) was randomly permuted, and PCA analysis was repeated several times, constructing a null distribution of gene scores each time. Significant PCs were identified which contains higher enrichment of low p-value genes (Butler & Satija, 2017). The Seurat package function '*JackStrawPlot*' provides a visualization tool for manually comparing the p-value distribution for each PC with a uniform distribution (dashed line) (**Figure 5.19**). Significant PCs were showing higher enrichment of genes with low p-value, depicted by the solid curve above the dashed line in **Figure 5.19** (Satija, 2017). Another Seurat function called *PCElbowPlot* uses more ad hoc method to determine PCs selection. It uses a plot of the standard deviations of the PCs and cutoff could be taken at the point of the apparent elbow in the graph. For the motor neuron Drop-seq data, the elbow appeared near PC16 (Figure 5.20).
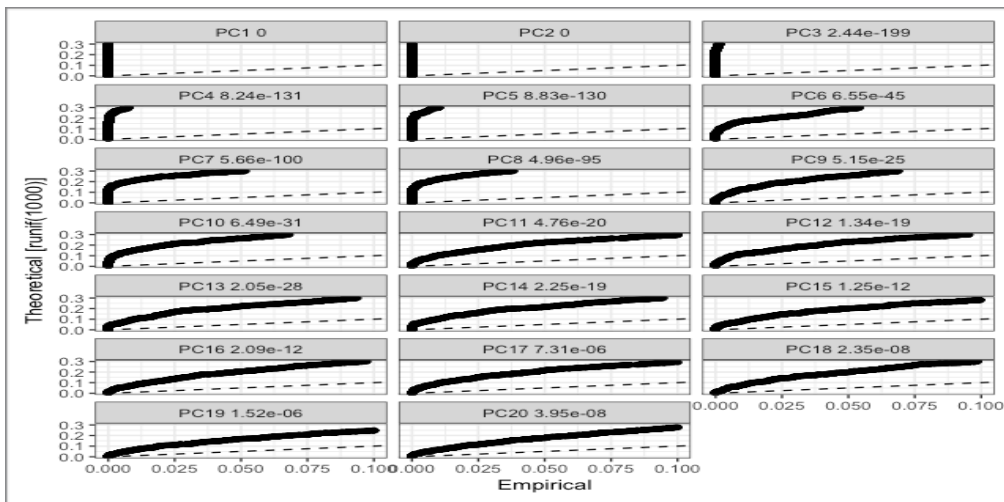
Figure 5.19: Jackstraw plot of principal components.

The dashed line represents the uniform distribution, and solid curve line represents the enrichment of genes with low p-values in significant PCs.
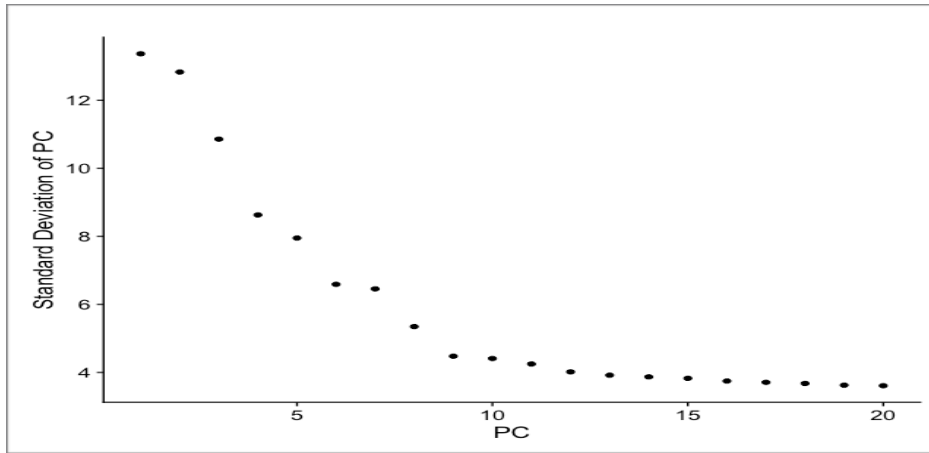
**Figure 5.20: PCElbow plot of principal components.**

The x-axis represents the number of PCs and Y-axis represents the standard deviation of principal components.

5.4.2.8 Clustering of the cells

Seurat package includes a graph-based clustering approach based on the method described in the SNN-Cliq (Xu & Su, 2015) and PhenoGraph (Levine et al., 2015) different than Drop-seq data analysis (E. Z. Macosko et al., 2015). The distance metric which drives the clustering analysis remains the same as described in (E. Z. Macosko et al., 2015). Although, partitioning the cellular distance matrix into clusters was improved because of the graph based clustering. SNN-Cliq and PhenoGraph methods embed cells in a graph structure (for instance, K-nearest neighbor or KNN graph) where graph edges between cells represent the similar gene expression patterns. Afterward, this graph was partitioned into highly interconnected 'quasi-cliques' or 'communities' (Satija, 2017). In Seurat package, a KNN graph was constructed based on the Euclidean distance in PCA space, and graph edge weights between two cells were refined based on the shared overlap in their respective local neighborhoods (also known as Jaccard distance) (Zhang, Wu, & You, 2017). To cluster the cells, Seurat package applies modularity optimization techniques to group cells together iteratively, keeping the goal of optimizing the standard modularity function (Satija, 2017). The *FindClusters* function in the Seurat package could implement this procedure, and it also contains a resolution parameter to set the granularity of the downstream clustering. The high values result in a more number of clusters (Satija, 2017).

5.4.2.9 Non-linear dimensional reduction (tSNE) plot for cell clusters

Seurat package uses tSNE (t-distributed stochastic neighbor embedding) method as an efficient visualization tool to explore single-cell datasets (Butler & Satija, 2017). Doing clustering directly on tSNE components is not appropriate. However, graph-based clusters should co-localize with the tSNE plot (with some exception). The tSNE method places the cells together in a low dimensional space, which have similar local neighborhoods in the high dimensional space (Van Der Maaten & Hinton, 2008). To generate the tSNE graph, same PCs were used, which were selected for the clustering analysis (Figure 5.20). The tSNE plot visualized the 6 clusters (cluster 0 to 5) in the motor neuron Drop-seq dataset (Figure 5.21).
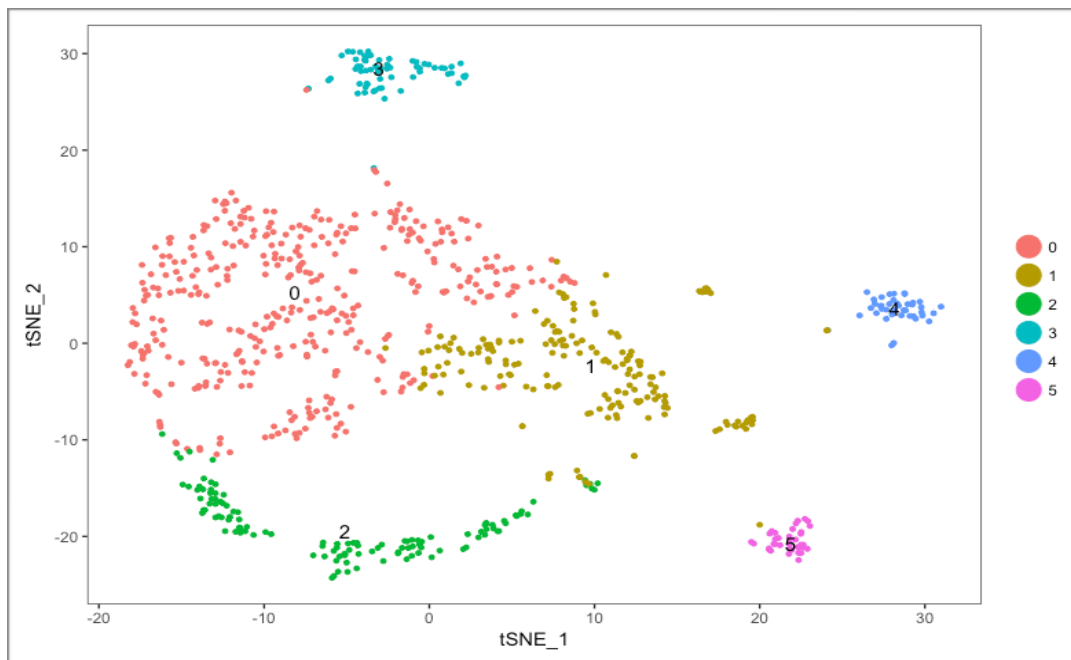
Figure 5.21: tSNE plot for spinal cord Drop-seq clusters.

Cell are colored according to their clusters in the t-SNE plot. In total, six clusters were present in the spinal cord Drop-seq dataset.

5.4.2.10 Differential expression of genes among single-cell clusters (biomarker analysis)

The identified clusters from the previous steps could be defined based on their gene markers or differential expression of genes. The Seurat package could identify positive and negative markers

of a single cluster in comparison to all other cells in the study. For the identification of marker genes in the motor neuron Drop-seq dataset, a Seurat package function *FindAllMarkers* was used (Satija, 2017). This function could also be used to test groups of clusters vs. each other or against all the single-cells. Table 5.4, summarizes the marker genes in each cluster identified for motor neuron Drop-seq dataset. Furthermore, Table 5.5 contains information about parameters and statistics used in clustering analysis and results obtained during the analysis. Based on the identified marker genes for each cluster, another Seurat function called *DoHeatmap* was used to generate an expression heat map for all the single-cells in motor neuron dataset. The top 10 marker genes for each cluster based on their respective log fold change were presented in the heat map (Figure 5.22).

Cluster	No. of Markers	No. of Positive Markers	No. Markers with p_val < 0.001
0	1111	371	371
1	587	362	116
2	1346	1052	469
3	3298	2133	2118
4	2359	1451	451
5	2778	1834	1211

Table 5.4: Statistics of marker genes in identified clusters of cells from motor neuron Drop-seq dataset.

Process	Values
Number of Cells selected	1000
Number of Genes	18.254
%age of Mitochondrial Genes used	0.2
Mitochondrial Threshold	-inf: 0.3
UMI Threshold	1000: 7000
Number of Cells After Filtering	837
No of Variable Genes Detected	3.353
No of Principal Components (PCs) Found	20
No of PCs used in Clustering	16
Clustering Resolution	0.4
No of Clusters	6

Table 5.5 Parameters selected at each step in Seurat clustering analysis.
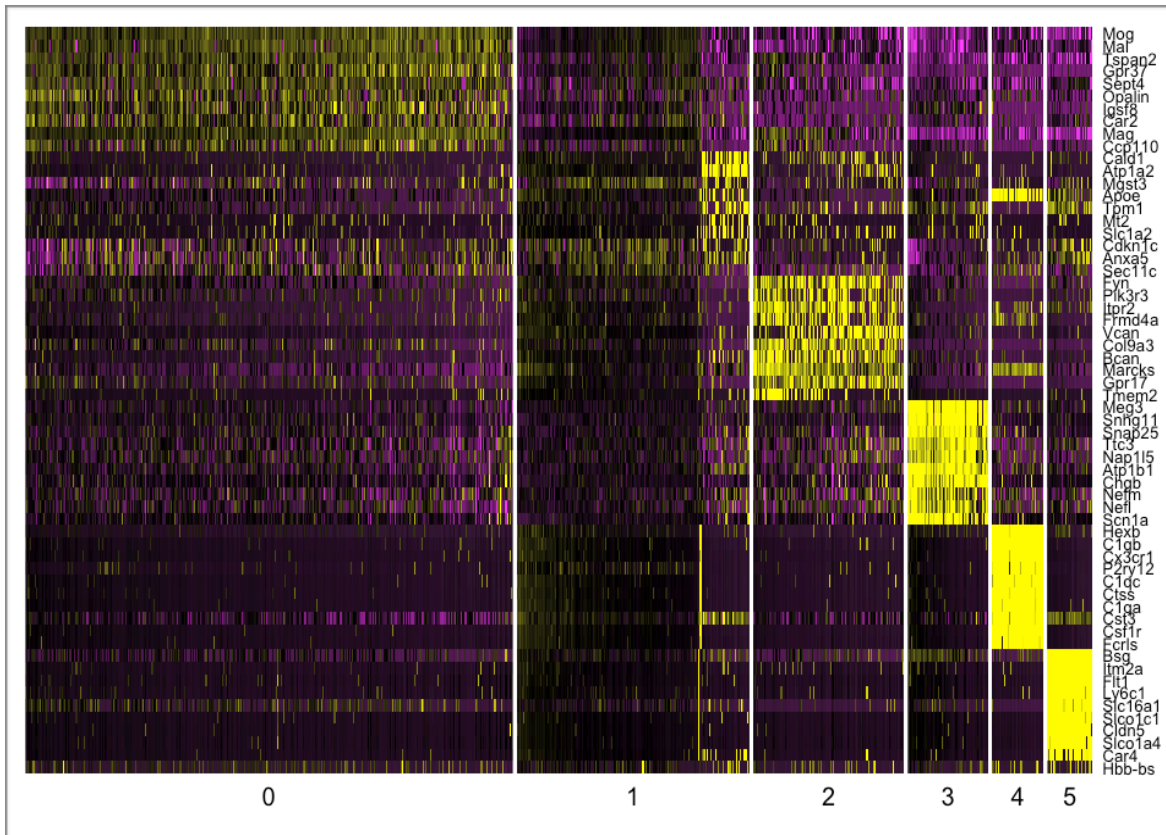
Figure 5.22: Heatmap of the top 10 biomarkers in identified clusters.

5.4.2.11 Cell type identity of the clusters

Based on previous literature (Friese et al., 2009; Shneider, Brown, Smith, Pickel, & Alvarez, 2009) , known markers of motor neuron subtypes were categorized and listed (Table 5.6). In brief, four subtypes of motor neuron markers were investigated, which are gamma motor neuron, alpha fast motor neuron, alpha slow motor neuron as well as precursor motor neuron (Table 5.6). These markers were investigated among different clusters to check their cell type identity. For instance, the violin plot shows expression probability distributions across all the clusters; the feature plot visualizes gene expression on the tSNE plot or a PCA plot. Apart from these tools, Joy-plot and Dot-plot were also employed to assess the clusters for their representative markers.

The gamma motor neuron marker Gfra1 was mainly present in cluster 0,1,2 and 3 as shown in feature plot (Figure 5.23) and violin plot (Figure 5.24) and also correlated by Dot plot and Joy

124

plot (Figure 5.25 and Figure 5.26 respectively). The alpha fast motor neuron markers Adam23, Mest, Dkk3, and Prkcb, were mainly present in cluster 3; although Adam23 was also showed some expression in the cells of cluster 0 and 1 (Figure 5.27 to Figure 5.30) Similarly, Prkcb showed some expression in cluster 1.

Furthermore, alpha slow motor neuron marker Sv2a, Crtac1, and Aldoc had high expression in the cells of the cluster 3 and mild expression in the cluster 0,1 and 2. Another alpha slow motor neuron markers Timp3 and Prkcd showed high expression in cluster 5 and cluster 4 respectively. On the other hand, other known markers for the alpha slow motor neurons such as Kcnmb4, Mmp2, Serpinf1, and Cacna1h did not appear in expression analysis in any of the clusters (Figure 5.31 to Figure 5.34).

In the end, precursor motor neuron marker Olig2 showed high expression in the cells of cluster 0,1 and 2 and did not show expression in remaining clusters in motor neuron Drop-seq dataset (Figure 5.35 to Figure 5.38). For all the six clusters of the motor neuron Drop-seq dataset, gene ontology analysis was presented in Figure 5.39.

Gamma MN	Alpha Fast MN	Alpha Slow MN	Precursor MN
Gfra1	Adam23	Sv2a	Olig2
	Mest	Kcnmb4	
	Dkk3	Crtac1	
	Prkcb	Timp3	
		Aldoc	
		Mmp2	
		Prkcd	
		Serpinf1	

Table 5.6 List of known markers for sub-types of motor neurons.

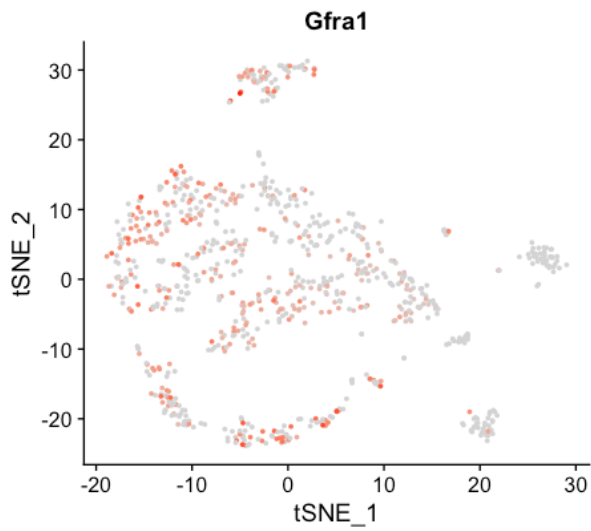**Gamma motor neuron marker**

Figure 5.23: Feature plot for Gfra1 marker across all the single-cells.
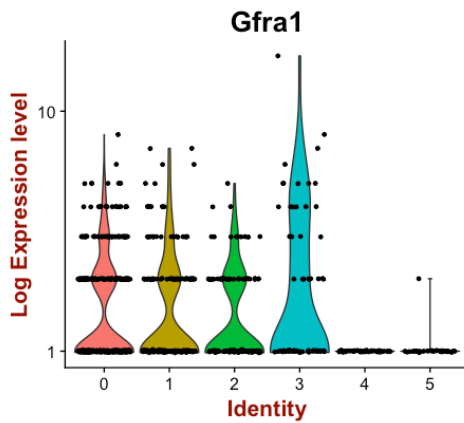

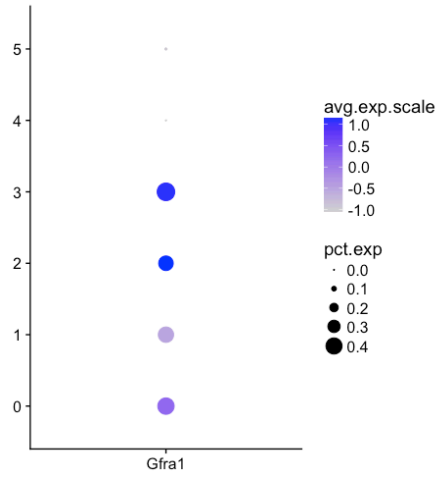**Figure 5.24: Violin plot for Gfra1 marker across all the single-cell clusters.**

126

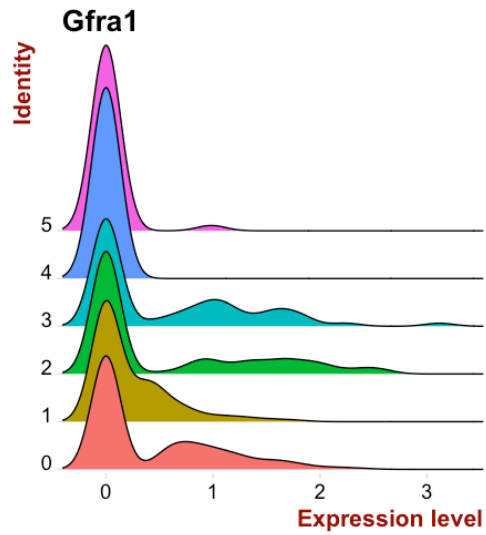**Figure 5.25: Dot plot for Gfra1 marker across all the single-cell clusters.**

Figure 5.26: Joy plot for Gfra1 marker across all the single-cell clusters.

Alpha fast motor neuron markers
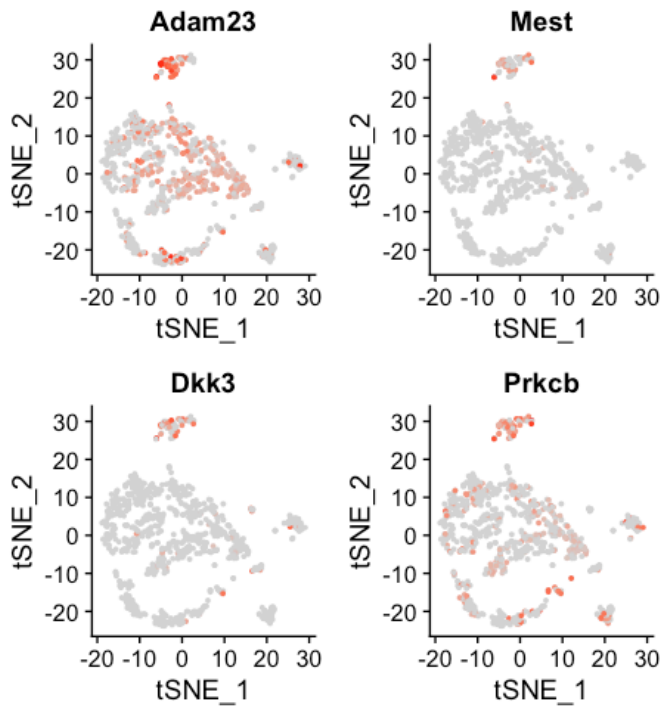
Figure 5.27: Feature plot for alpha fast motor neuron markers across all the single-cells.
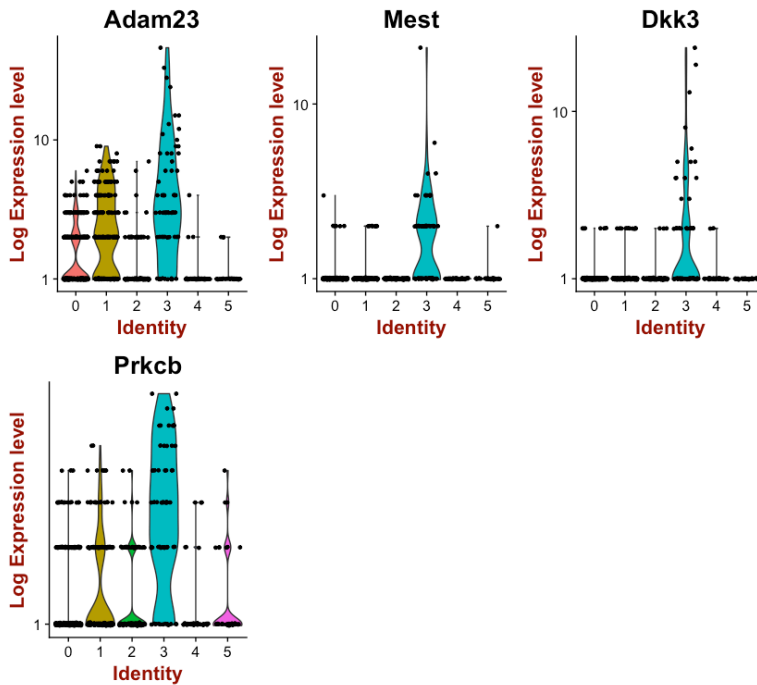
Figure 5.28: Violin plot for alpha fast motor neuron markers.

Figure 5.29: Dot plot for alpha fast motor neuron markers across all the single-cell clusters.
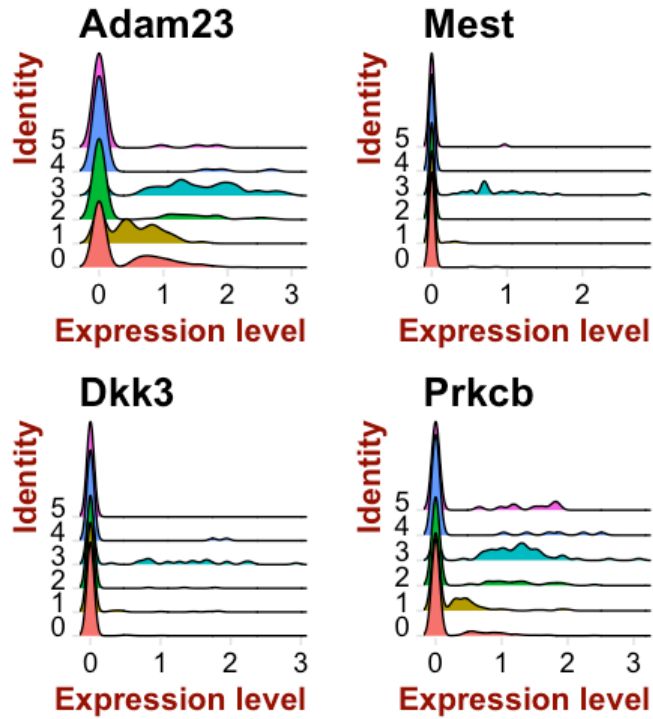
Figure 5.30: Joy plot for alpha fast motor neuron markers across all the single-cell clusters.

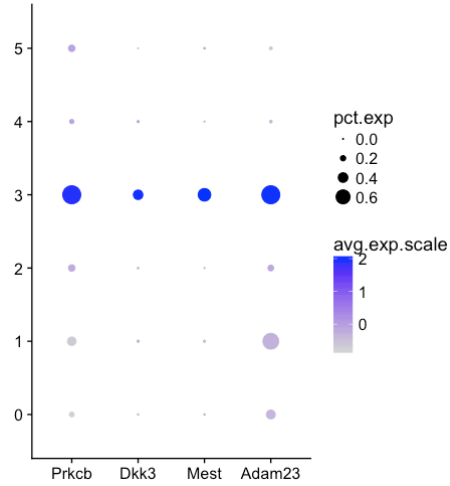## Alpha slow motor neuron markers

Figure 5.31: Feature plot for alpha slow motor neuron markers across all the single-cell clusters.
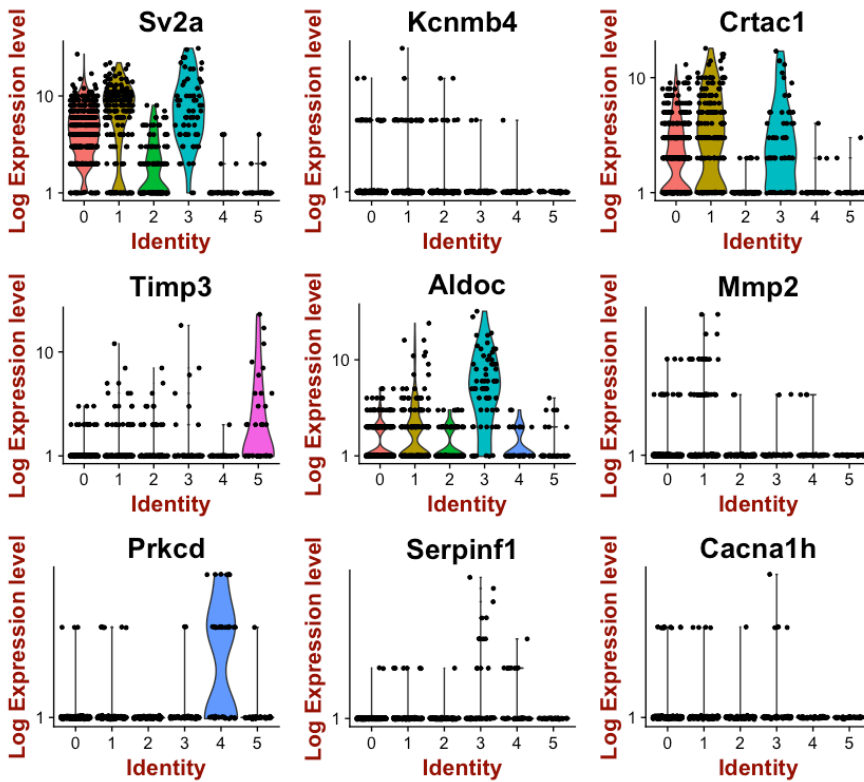
Figure 5.32: Violin plot for alpha slow motor neuron markers.

Figure 5.33: Dot plot for alpha slow motor neuron markers across all the cell clusters.
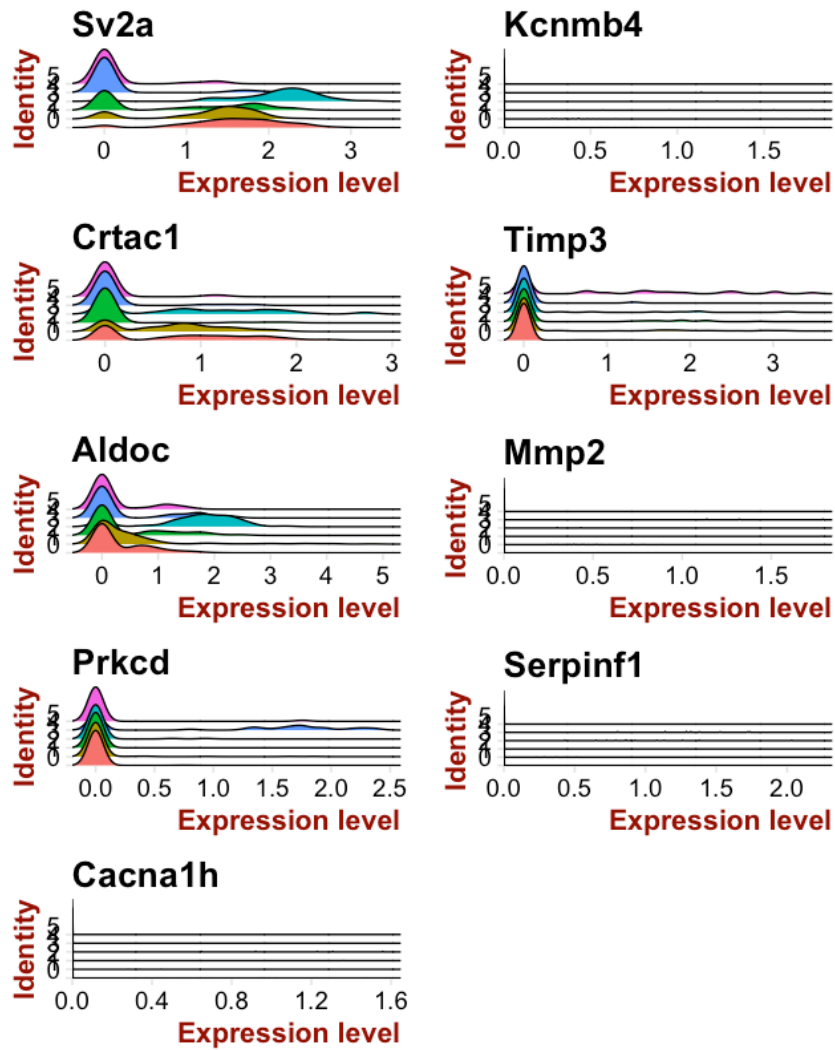
Figure 5.34: Joy plot for alpha slow motor neuron markers across all the cell clusters.

Precursor motor neuron markers
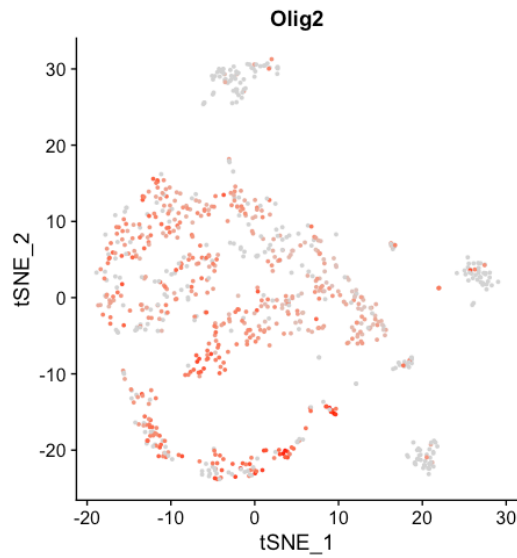
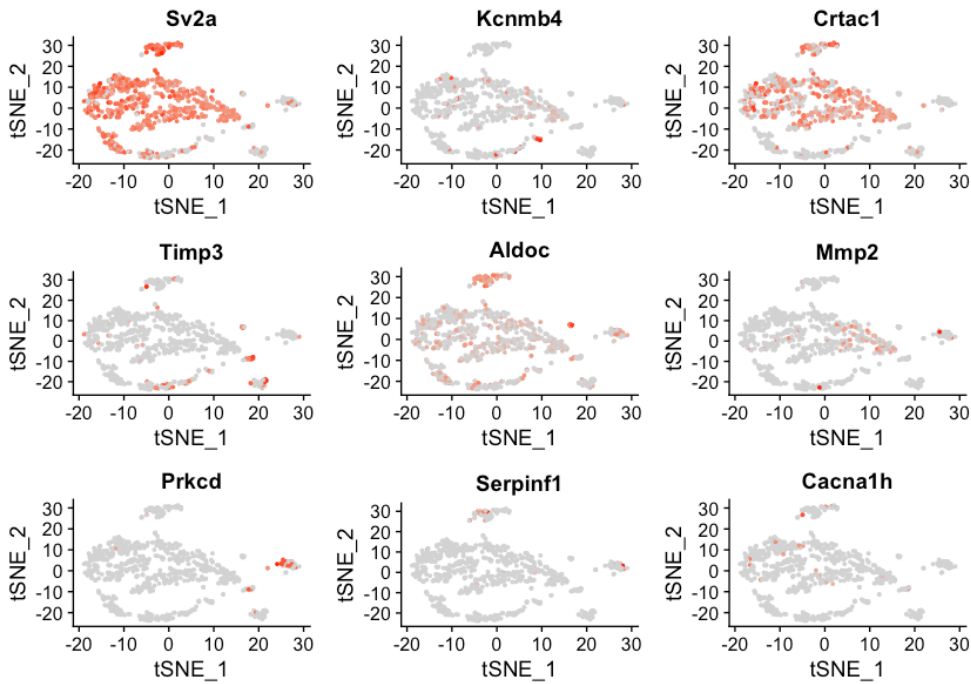**Figure 5.35: Feature plot for precursor motor neuron marker across all the single-cell clusters.**

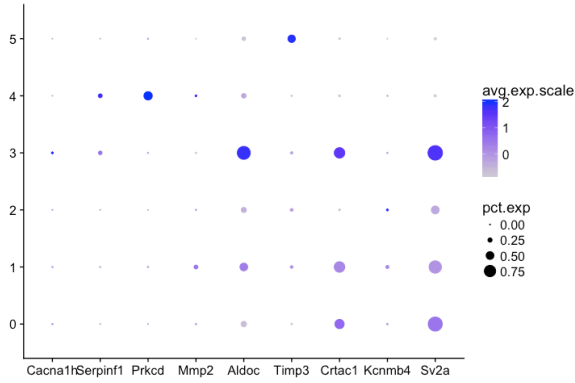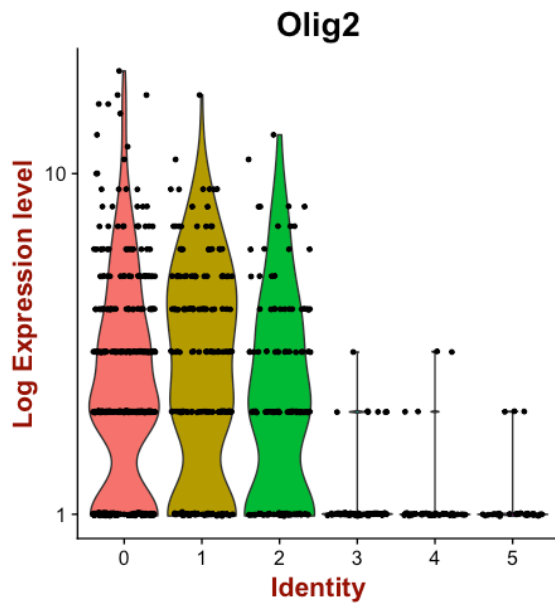**Figure 5.36: Violin plot for precursor motor neuron markers.**

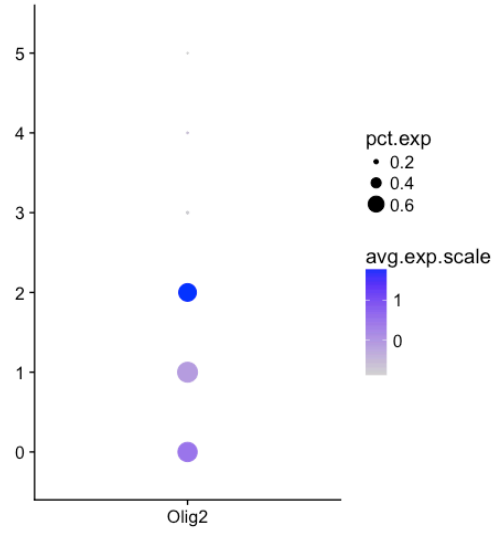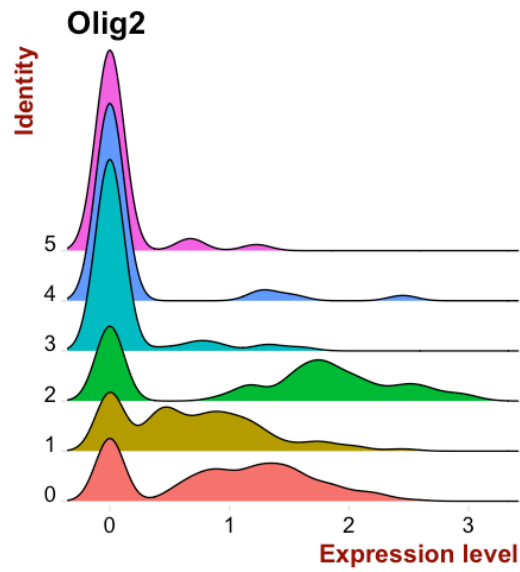**Figure 5.37: Dot plot for precursor motor neuron markers.**

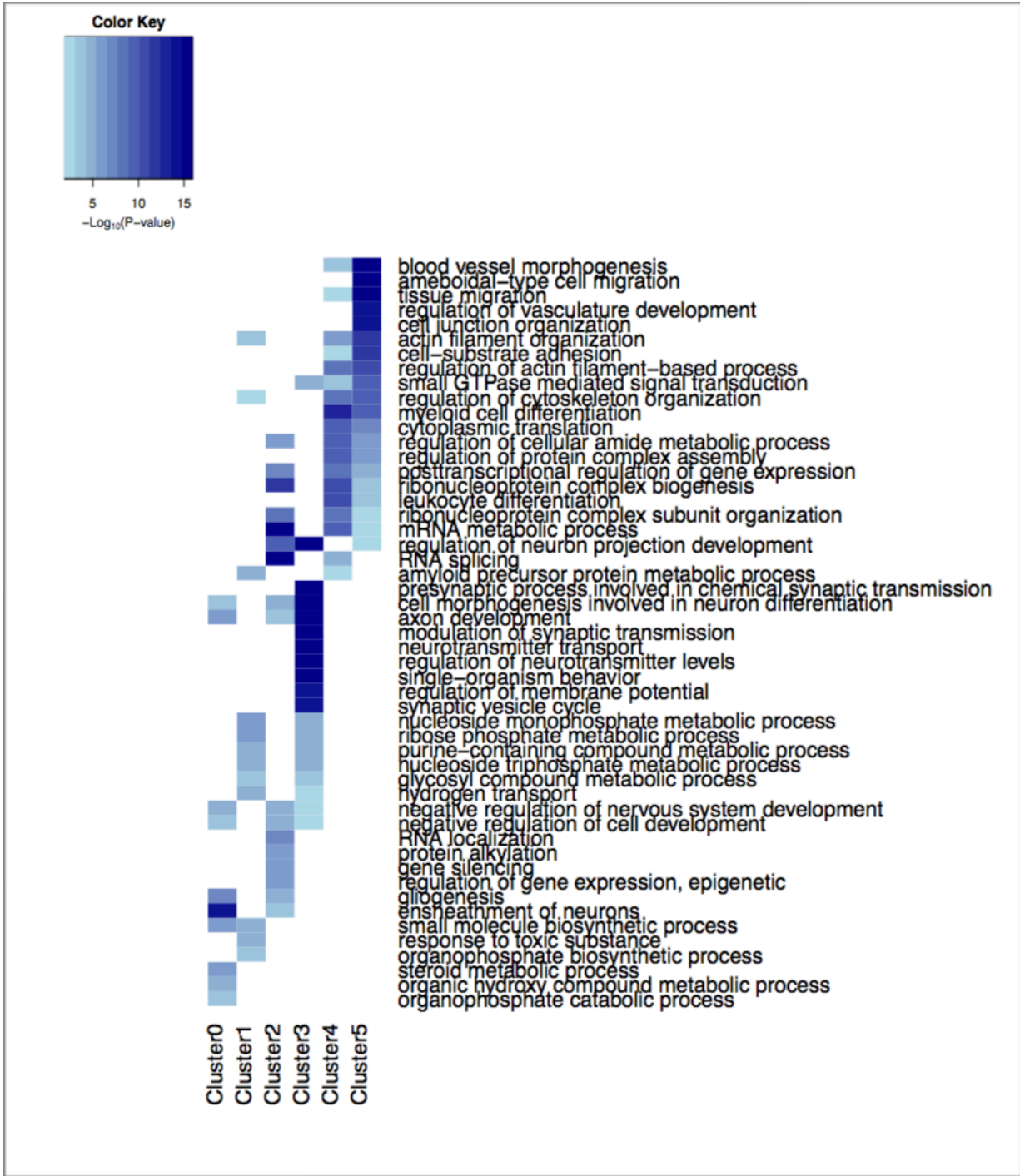**Figure 5.38: Joy plot for precursor motor neuron markers.**

Figure 5.39: Gene ontology analysis for motor neuron cellular clusters.

5.5 Effect of Erythropoietin (EPO) on the cells of hippocampal tissue using single-cell sequencing (Case study-II)

Erythropoietin (EPO; also known as hematopoietin) is a glycoprotein cytokine, which is an essential hormone for red blood cell production in the bone marrow (erythropoiesis). EPO has shown its primary function by promoting the survival of red blood cell progenitors and precursors through protecting these cells from cell death or apoptosis. EPO binds to the erythropoietin receptor (EPOR) on the surface of red blood cell progenitor cells and activates the JAK2 signaling pathways (Middleton et al., 1999). Further, it leads to activation of STAT5, PIK3, and RAS MAPK pathways, which results in differential, survival, and proliferation of the red blood progenitor cells (Hodges, Rainey, Lappin, & Maxwell, 2007). Recent studies suggest the presence of EPO receptors in the other tissues, such as heart, kidney, muscle and interestingly in central/peripheral nervous system (Genc, Koroglu, & Genc, 2004).

The expression of EPO and its receptor EPOR in the healthy postnatal brain is low but distinct in specific brain regions such as cortex and hippocampus (Digicaylioglu et al., 1995; Marti et al., 1996). During stress conditions such as brain injury, the expression of EPO is strongly upregulated (Ott, Martens, et al., 2015). Although, in healthy human and rodents, recombinant human EPO leads to improvement in cognition and increase hippocampal long-term potentiation (LTP) (Adamcio et al., 2008; Miskowiak, O'Sullivan, & Harmer, 2007). These beneficial results of EPO on cognitive functions were widely evident in clinical trials on various neurological diseases such as multiple sclerosis, schizophrenia, depression and bipolar disease (Ehrenreich et al., 2007). This consistent effect of EPO was specific to brain and independent to its hematopoiesis effects (Hassouna et al., 2016). To understand, the cellular mechanisms underlying these cognitive effects, a study was conducted in healthy young mice. This study showed that administration of EPO for three weeks is associated with a ~20% increase in pyramidal neurons and oligodendrocytes cell numbers in the hippocampus (Hassouna et al., 2016). Furthermore, neuron numbers remained elevated more than six months of age under constant cognitive challenge (Hassouna et al., 2016). Also, EPO-treated NG2-Cre-ERT2 mice showed enhanced differentiation of pre-existing oligodendrocyte precursors in the absence of elevated DNA synthesis (Hassouna et al., 2016). These results indicate towards the role of EPO as an effector

molecule, which can drive the differentiation of non-dividing precursor cells into (NG2+) oligodendroglial as well as neuronal lineages (Hassouna et al., 2016). However, the exact mechanism behind this phenomenon is not entirely clear yet. In the current study, we aimed to identify precursor cells, which differentiate into neuronal and oligodendrocytes cells followed by EPO intervention. To identify individual precursor cells with the single-cell resolution, the Drop-seq method was used.

5.5.1 Experimental design and results

The 28 days postnatal wild-type C57/BL6N mice were subjected to injection of either EPO or Placebo intraperitoneally (i.p.). The mice were injected in triplicates for each group. After 6 hours of injection, mice were sacrificed, and the hippocampal CA1 region was isolated. Brain tissues were digested following the same procedure used for spinal cord samples (see Drop-seq methods) and prepared the single-cells for the Drop-seq experiment on these samples. After sequencing the Drop-seq libraries, reads were first processed with Drop-seq pipeline to generate digital gene expression matrix and further processed using Seurat package, similar to spinal cord Drop-seq data analysis (section 5.4.2). The sequencing data from three biological replicates for each group was analyzed separately to check data quality. Once individual samples are proved to have good quality, individual replicates were merged into a single sample for secondary analysis. The parameters and the output results from essential steps in the analysis were summarized (Table 5.7). Using the Seurat package number of genes (nGene), number of UMIs (nUMI) and percentage of mitochondrial genes (percent.mito) in each cell were calculated for both of the groups (Figure 5.40). Also, Gene plot (Figure 5.41) were used to filter a rare subset of cells, which had more than 20% of the reads belong to mitochondrial genes. Furthermore, cells with unique gene count more than 8000 (possible doublet), and less than 1000 genes were filtered out from the dataset. After the quality check filtration, 973 cells (390 in group 1 and 583 in group 2) were remaining in the EPO Drop-seq dataset (Table 5.7).

Process	Values
Number of Unique Mitochondrial Genes	18
%age of Mitochondrial Genes used	0.4
Mitochondrial Threshold	-inf : 0.4
UMI Threshold	1000: 8000
Number of cells After Filtering	973 (Group 1: 390; Group 2: 583)
Groups By Experiment	GROUP 1 (M1D1,M3D1,M5D2) GROUP 2 (M2D1,M4D2,M6D2)
No of Variable Genes Detected	3008
No of Principal Components (PCs) Found	20
No of PCs used in Clustering	15
Clustering Resolution	0.2
No of Clusters	7

Table 5.7: Summary of EPO Drop-seq data analysis.
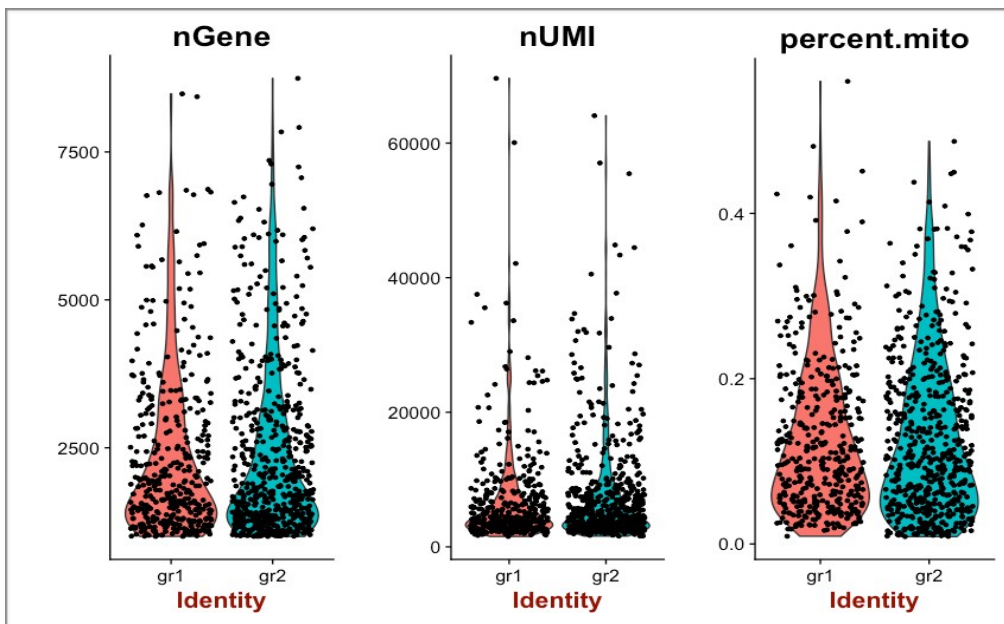
Figure 5.40: Violin plot for EPO Drop-seq data.

The Y-axis represents the number of genes (nGene), the number of UMIs (nUMI) and percentages of the mitochondrial reads (percent.mito) in the total reads respectively (from left to right). Group-1 and group-2 is placebo vs. EPO samples, respectively.
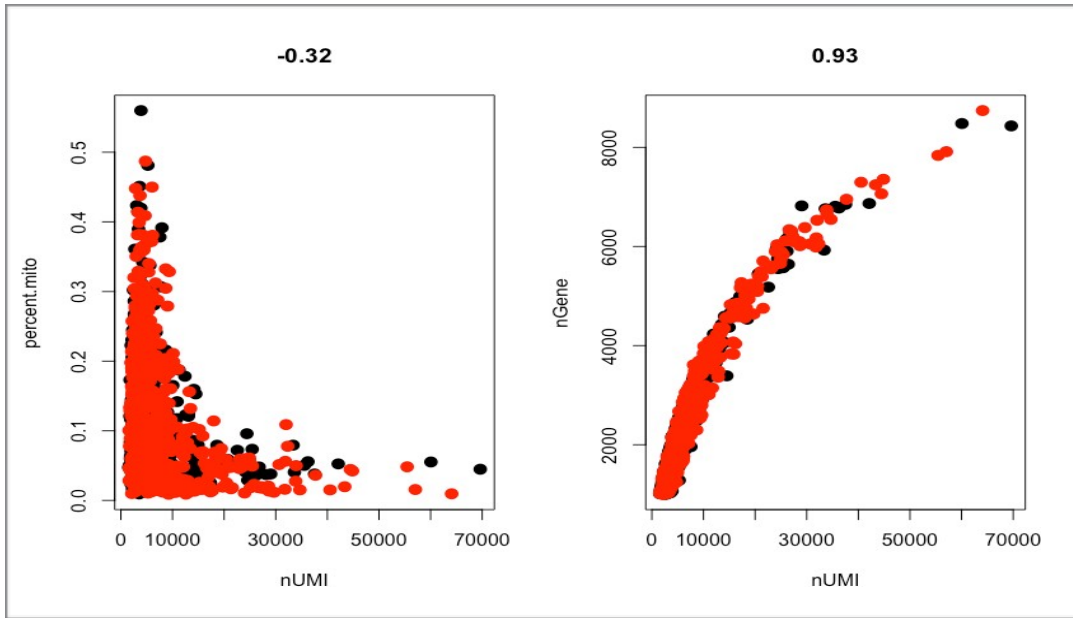
Figure 5.41: Gene plot for EPO Drop-seq data.

The plot, at the left side, represents the distribution of mitochondrial percentage (percent.mito) and transcripts (nUMI) for the STAMPs/cells detected in the spinal cord Drop-seq experiment. The plot, at the right side, represents the number of genes (nGene) and transcripts (nUMI) for the same STAMPs. Each dot in the figure above represents a STAMP/cell. Red and black colors represent two groups of the cells.

5.5.2 Data normalization and detection of highly variable genes

Using Seurat package program called *LogNormalize*, global scale normalization was performed on EPO Drop-seq dataset (see spinal cord Drop-seq analysis 5.4.2). After normalizing the data, Seurat package program *FindVariableGenes* was used to calculate the dispersion, average expression and z-score for each gene (Figure 5.42). This analysis identified the highly variable genes, which were further used in the downstream analysis of EPO Drop-seq data (Figure 5.42).
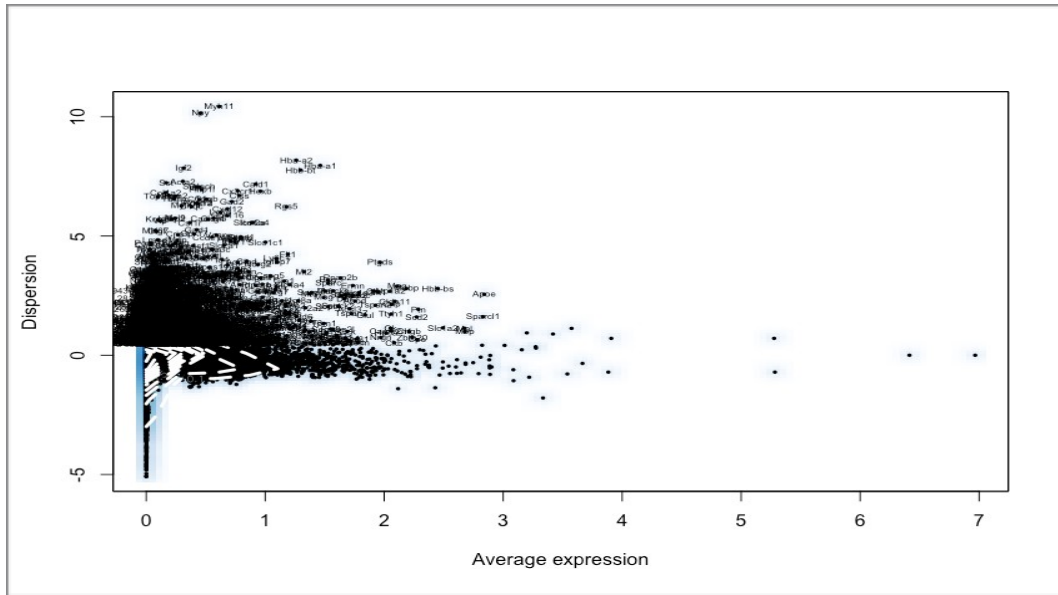
Figure 5.42: Dispersion and average expression plot for the EPO dataset.

The X-axis represents the average expression and the Y-axis represents the dispersion of the genes. Using this graph, highly variable genes were selected for downstream analysis.

5.5.3 Statistically significant principal components

After determining highly variable genes in the EPO dataset, technical and cell cycle variation were removed as described in the motor neuron Drop-seq data analysis. Linear dimensional reduction technique (PCA analysis) was performed on the scaled EPO dataset. As described earlier, significant PCs were identified, which contains higher enrichment of low p-value genes. Using the Seurat package's *JackStrawPlot* function, visualization and manual comparison for the p-value distribution of each PC (solid curve) with the uniform distribution (dashed line) was performed (Figure 5.43). Furthermore, Seurat function *PCElbowPlot* was used to plot standard deviations of the PCs, and the elbow in the graph was served as a cutoff point to determine significant PCs (Figure 5.44). For EPO Drop-seq dataset, in total 20 PCs were found and 15 PCs were selected for downstream clustering analysis.

Figure 5.43: Jackstraw plot of principal components.
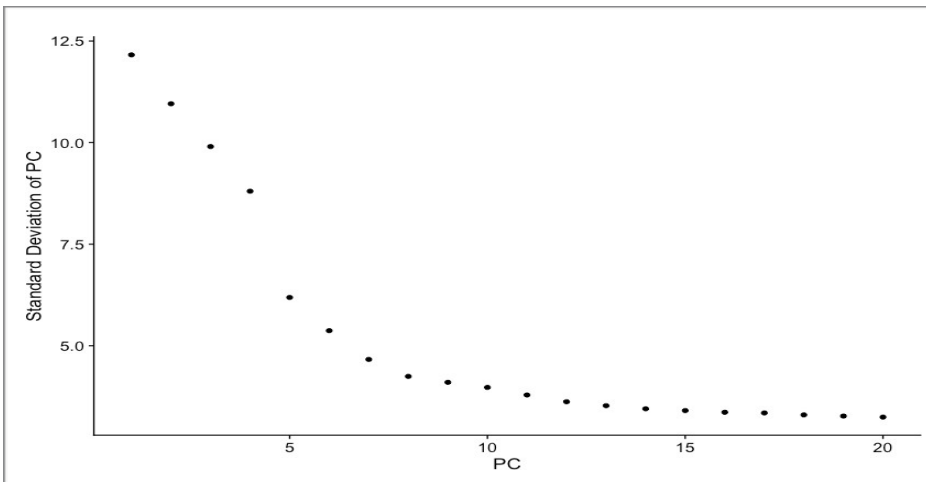
Figure 5.44: PC-elbow graph of principal components for EPO dataset.

5.5.4 Clustering of the cells

As described earlier in detail in section (5.4.2.8), based on Jaccard distance among the cells, Seurat package function *Findclusters* applied modularity optimization techniques to group cells together iteratively (also known as graph-based clustering approach). On the other hand, non-linear dimensional reduction method (tSNE) was employed as an efficient visualization tool for single-cell clusters by Seurat package. Clustering analysis revealed the seven cell clusters (cluster 0 to 6) in EPO Drop-seq dataset (Figure 5.45).
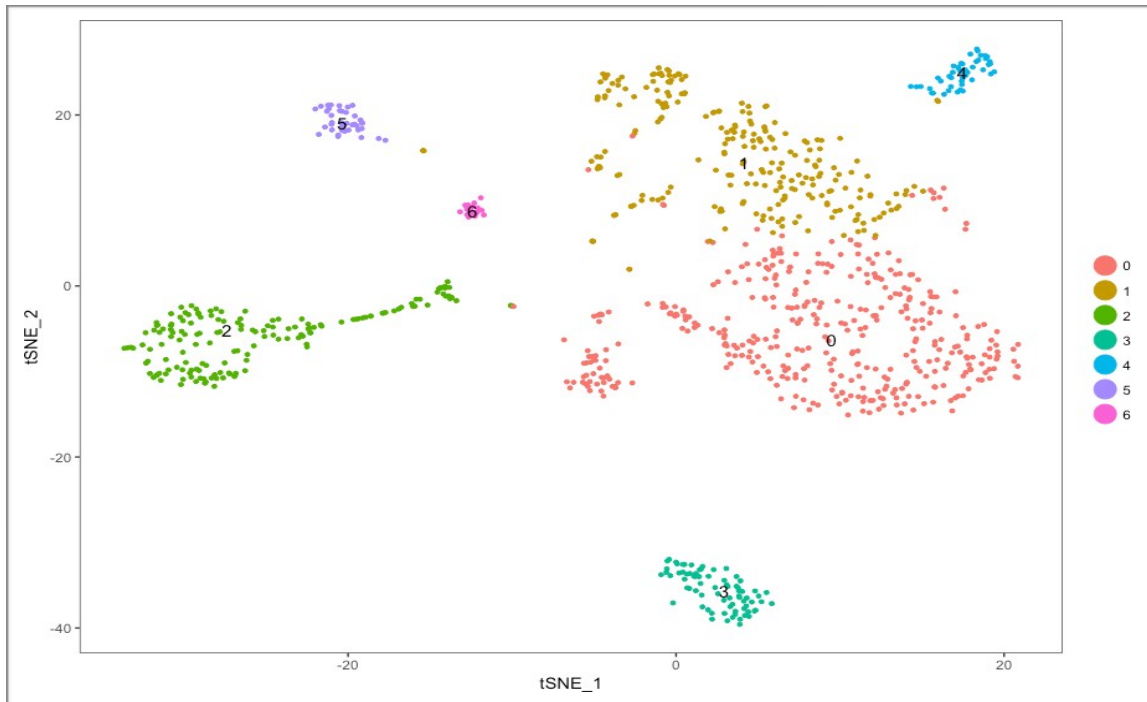
Figure 5.45: tSNE plot for the EPO Drop-seq dataset.

Cell are colored according to their clusters, in the t-SNE plot. In total, seven clusters were present in the EPO Drop-seq dataset.

5.5.5 Differential expression of genes among single-cell clusters

For the identification of differentially expressed genes among cell clusters (biomarker analysis), *FindAllMarkers* function from Seurat package was used (see case study 1). The marker genes were identified for each cluster of EPO Drop-seq dataset (Table 5.8). Furthermore, cell type-specific markers for brain tissue were taken from previously published studies (Kelsom & Lu, 2013; Wu, Pan, Zuo, Li, & Hong, 2017) and listed in Table 5.9. The expression of these markers was investigated in identified clusters in EPO Drop-seq dataset, as presented in violin plot (Figure 5.46). Based on the marker genes, each cluster was assigned a possible cell type identity, as depicted in the tSNE plot (Figure 5.47). Furthermore, based on identified marker genes for each cluster, which were determined by the *FindAllMarkers* function, a heat map was generated containing top 10 biomarkers for each cluster using the *DoHeatmap* function from Seurat

package (Butler & Satija, 2017). Gene ontology analysis was also done on the different cell clusters based on the marker genes (Figure 5.49).

Cluster	Markers	Positive Markers	p_val < 0.001	p_val < 0.00001
Glutamatergic0	1304	903	902	893
Glutamatergic1	775	444	416	306
OPCs_Oligo	2335	1113	1089	1001
Astrocytes	2219	1047	621	469
Gabaergic	1364	645	378	188
Endothelial	2846	1867	1038	751
Microglia	2650	1856	378	234

Table 5.8: Summary of biomarker genes in clusters.

Cell types	Markers
Neurons	Camk2a, Kif5c
Astrocytes	Slc7a10, Lfng
Microglia	Tmem119, C1qc
OPC (Oligodendrocyte precursor cell)	Pdgfra, Gpr17
Oligodendrocytes	Mog, Opalin
Endothelial cells	Fit1, Ly6c1
GABAergic neuron	Gad1, Gad2
Glutamergic neuron	Slc17a7, Neurod6

Table 5.9: Cell type specific markers across different cell types.

These markers were adapted from (Wu et al., 2017) and (Kelsom & Lu, 2013).
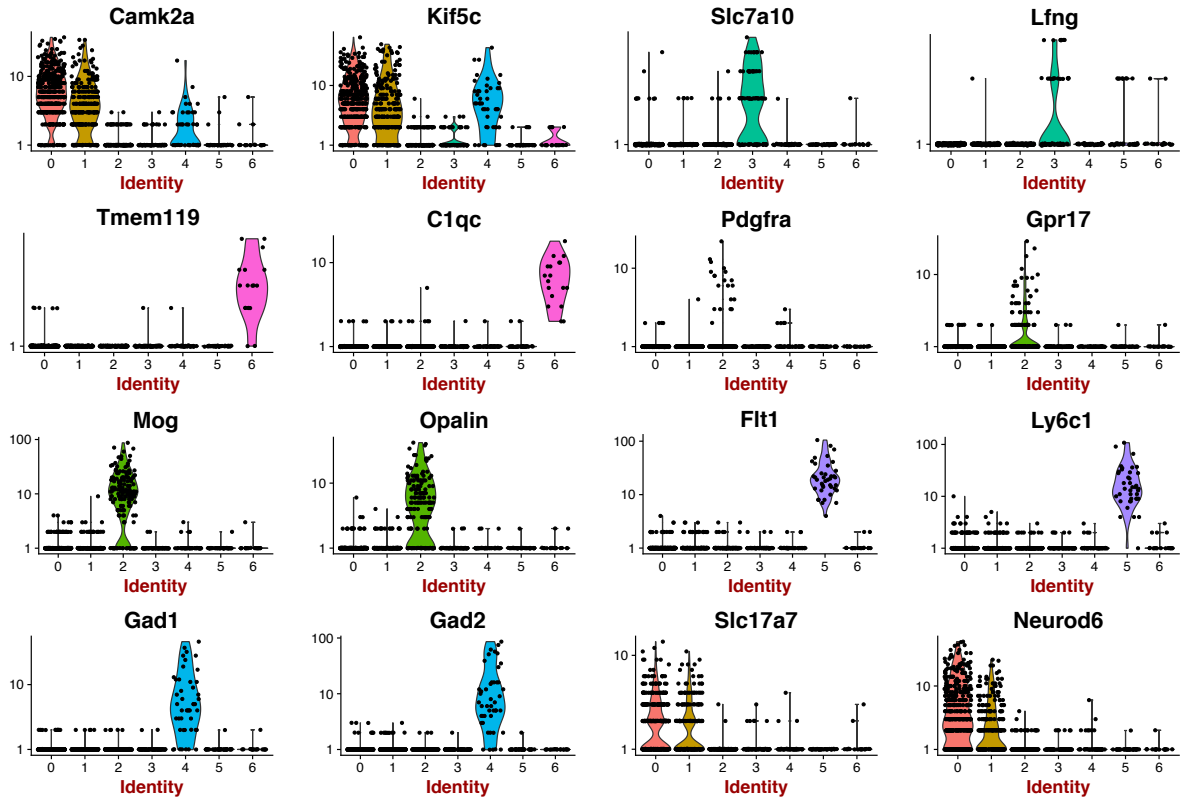
Figure 5.46: Violin plots for cell type specific markers for EPO Drop-seq dataset.

Cell type specific marker, as listed in **Table 5.9** were investigated in the identified clusters. Based on these known marker genes, identified single cell clusters were assigned a plausible class of cells.
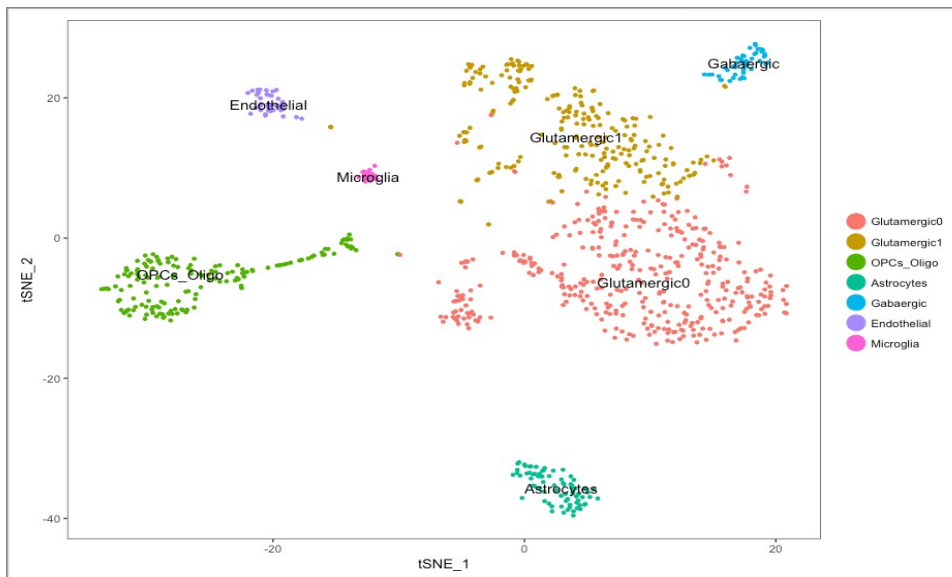
Figure 5.47: tSNE plot with cell type identity.

Based on known marker genes, these clusters were assigned a plausible class of cells.
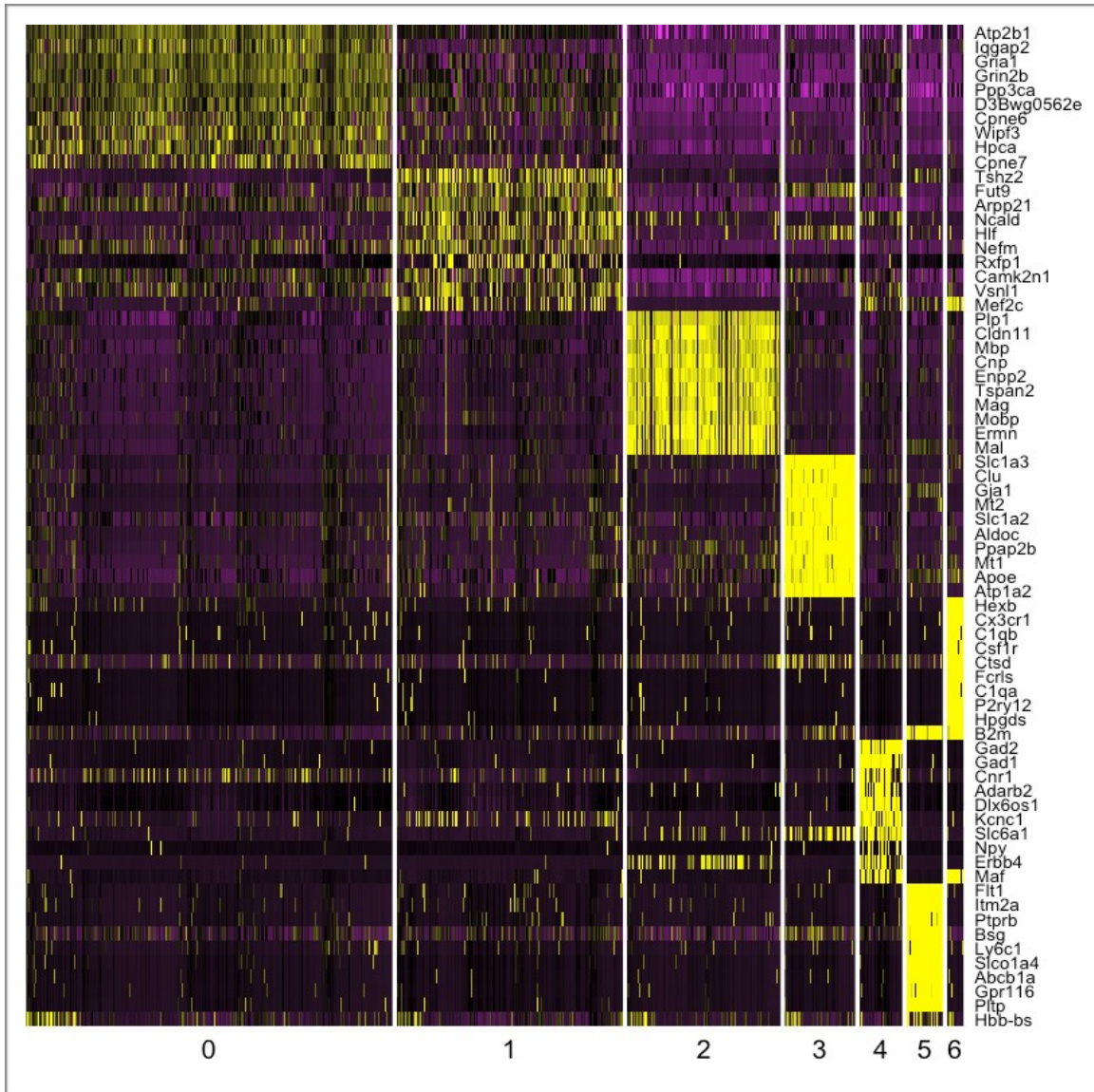
Figure 5.48: Expression heat map.

The heat map showing the top-10 biomarkers in the identified clusters in EPO Drop-seq dataset.
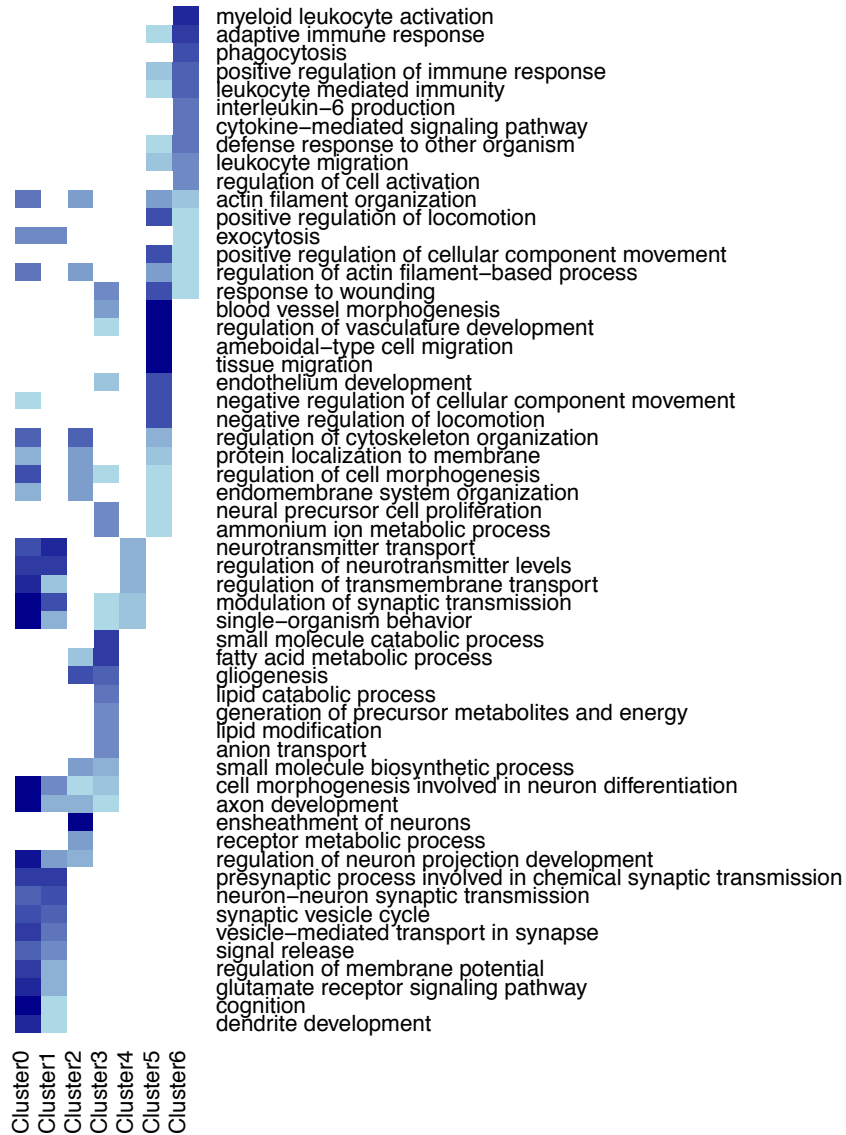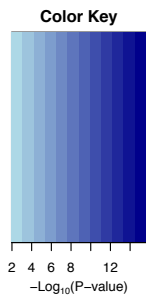
Figure 5.49: Gene ontology analysis for clusters.

5.5.6 Cell type analysis in EPO vs. placebo groups

To compare changes in the gene expression in the identified clusters at the single-cell level, we have calculated the percentage of the cells from each group fall into individual clusters (Table **5.10**). Also, based on the positive markers for individual clusters, cells categorized to their respective groups were also presented (Table **5.11**). Furthermore, to visualize the cells in their clusters, based on their group color code, cells were plotted based on their low-dimensional embedding produced by tSNE (Figure 5.50).

Cluster	Group 1 Cells	Group 2 Cells	Group 1 %age	Group 2 %age	Fisher Exact Test (p_value)
Glutamatergic 0	171	219	43,8 %	37,6 %	0.05307
Glutamatergic 1	77	164	19,7 %	28,1 %	0.003106
OPCs_Oligo	65	99	16,7 %	17,0 %	0.9305
Astrocytes	33	42	8,5 %	7,2 %	0.4652
Gabaergic	19	27	4,9 %	4,6 %	0.8784
Endothelial	14	25	3,6 %	4,3 %	0.6213
Microglia	11	7	2,8 %	1,2 %	0.08789
TOTAL	390	583	100,0 %	100,0 %	

Table 5.10: Summary of differential cell percentage analysis in placebo (group1) and EPO (group 2) clusters.

Cluster/Group	+ve Markers (PM)	PM with p_value < 0.001	PM with p_value < 0.00001
Glutamatergic0/1	67	40	12
Glutamatergic0/2	37	18	8
Glutamatergic1/1	111	15	3
Glutamatergic1/2	79	0	0
OPCs_Oligo/1	170	3	1
OPCs_Oligo/2	234	3	0

Astrocytes/1	482	1	0
Astrocytes/2	673	0	0
Gabaergic/1	1027	30	1
Gabaergic/2	631	4	0
Endothelial/1	963	2	0
Endothelial/2	747	3	2
Microglia/1	137	1	0
Microglia/2	477	0	0

Table 5.11: Summary of Biomarker analysis in clusters.
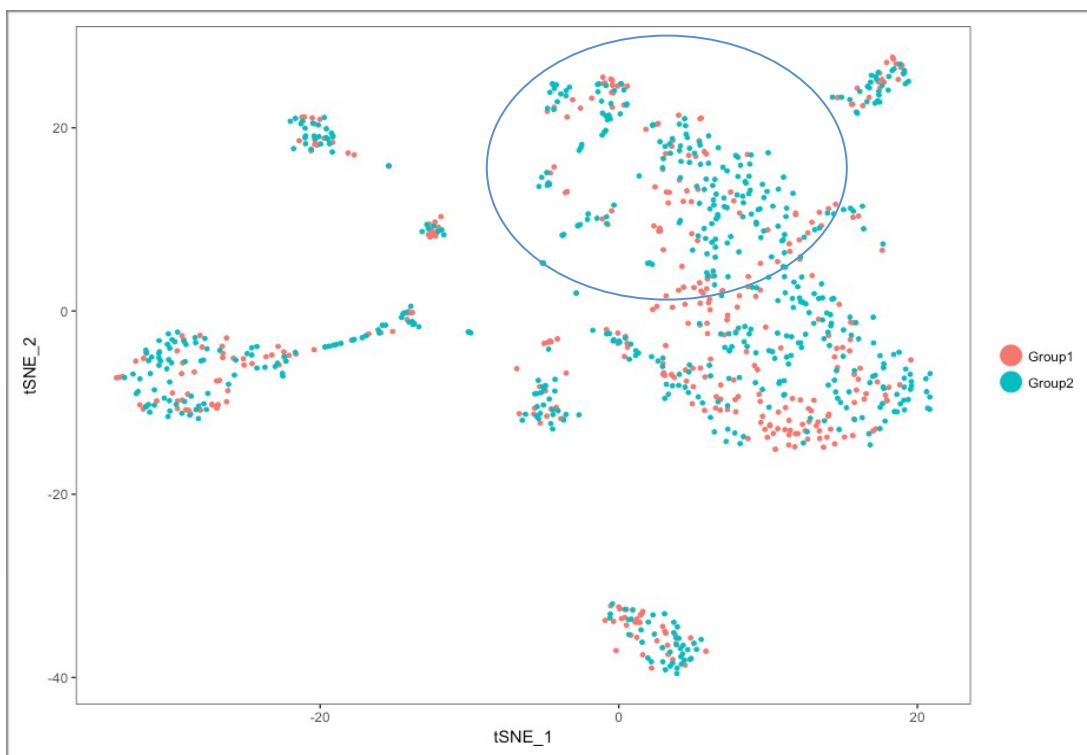
Figure 5.50: tSNE plot with the group information.

In the tSNE plot, the cells are colored according to their respective groups (group 1 and 2 for placebo vs. EPO respectively). The circle highlights the glutamatergic-1 cluster, which showed increased cells in EPO group.

5.6 Discussion

A better understanding of transcriptional variations among individual cells can be achieved by single-cell sequencing. This, in turn, leads to more molecular insights on complex tissues (for instance brain) and their functional responses. In recent years, advances in technologies made it possible to sequence mRNA from hundreds of single-cells (Picelli et al., 2013) to thousands of single-cells (Jaitin et al., 2014; Shalek et al., 2014). However, the expenses of library preparation and time involved in processing individual cells limited earlier methods. In the current chapter, we have implemented a microfluidic-based single-cell mRNA-seq method (Drop-seq), which was first developed by Macosko et al. in the McCarroll lab at Harvard medical school and used for the investigation of the retinal cells (E. Z. Macosko et al., 2015). We have used this technique to perform the single-cell sequencing of motor neurons as well as to learn the effect of a small molecule (EPO) intervention in hippocampal neurons.

At first, after setting up the physical Drop-seq setup in the lab, we have optimized the parameters by Human-mouse species mixing experiments to check the purity and presence of cell doublet rate in the libraries. Once species mixing experiments provided optimal results, Drop-seq method was further used for single-cell studies of motor neurons as well as hippocampal neuronal cells with a small molecule intervention. One of the main technical challenges of the single-cell sequencing experiments was the optimal dissociation of the cells from the tissue of interest. In complex tissues (such as the brain, spinal cord, and others.) many different cell types are interconnected with each other in close proximity, which makes it difficult to dissociate the cells and separate them from each other with high survival rate. For every tissue, dissociation methods need to be optimized for efficient processing of cells. For motor neurons, dissociation method is optimized after several experiments, as described in section 5.2.2. This method is also applied to hippocampal tissue for cellular dissociation.

Furthermore, after the Drop-seq library preparation and the pair end sequencing in Illumina high throughput sequencer (Hiseq-2500), the raw reads were analyzed using the Drop-seq pipeline for generation of digital gene expression matrix. For the second level of data analysis, the Seurat package was used, which is capable of quality filtering of the cells, normalization of the data, detection of the variable genes among single-cells, determination of statistically significant

148

principal components of the data and at the end clustering of the samples. It also plots an effective representation using the tSNE method. The Seurat package enabled us to identify and interpret sources of heterogeneity in single-cell transcriptomics data. The most important advantages of Seurat package are as follows. To begin with, it can integrate and analyze single-cell data coming from various platforms such as Drop-seq, 10x genomics technique, and others. Also, the Seurat package features three recently developed computational methods for single-cell data analysis, which are forefront techniques in this field. The first method is unsupervised clustering and discovery of cell types and states, which featured in Drop-seq publication and used in the analysis of the retinal Drop-seq dataset (E. Z. Macosko et al., 2015). It also combines dimensional reduction with graph-based clustering. The second method does a spatial reconstruction of the single-cell data and integrates single-cell RNA-seq with *in-situ* reference data to infer cellular spatial localization expression (Satija, Farrell, Gennert, Schier, & Regev, 2015). Finally, the third method is an integrated analysis of single-cell RNA-seq across conditions, technologies, and species (Butler & Satija, 2017).

At this point, we would also like to highlight a critical step in the clustering approach used by the Seurat package. In it, each principal component (PC) represents a 'metagene' (a linear combination of hundreds or thousands of transcripts), and it is robust towards drop-out events in any individual gene (Satija, 2017). Nevertheless, estimating the true dimensionality of a dataset is a non-trivial task in machine learning. In order to select a set of principal components (PCs) for the downstream clustering analysis, the Seurat package approaches with three methods to estimate this selection parameter. The first approach is the jackstraw method, which is essentially a statistical resampling procedure. It is used to construct a null distribution for PC scores, and afterward, it associates each PC with a p-value to enable the significance assessment in a more formal statistical framework (Satija, 2017). The next approach is using the Elbow plot in the Seurat package (Figure 5.20). Plotting the eigenvalues in decreasing order (also known as scree plot) is a valuable heuristic. In this method, the 'elbow' of the plot represents a transition from variable and more informative PCs to comparatively less variable PCs. The Elbow point is correlated well with the significant PCs identified by the jackstraw method. Nevertheless, the Elbow plot is much faster to obtain. The third approach suggested by the Seurat package employs supervised analysis by the user. The *PCHeatmap* function can be used to visualize heat-maps of

PCs across both genes and cells and can be helpful to exclude the PCs, which is primarily driven by mitochondria, ribosome or cell cycle genes (Satija, 2017).

The molecular insights with single-cell resolution, which we obtained thanks to the Drop-seq method is very valuable and an addition to the knowledge in this scientific domain. Analysis showed that the motor neuron Drop-seq dataset (case study I) contains 6 clusters of different cell types in total, as detected by the Seurat package analysis (Figure 5.21). These clusters are mainly investigated for four subtypes of motor neuron marker genes, which were curated from the previous literature. Individual marker genes could be present in different ratios in more than one cellular sub-type. This presents an obstacle to determining absolute cell type identity of recognized clusters, also evident in single-cell clusters from spinal cord tissue. Nevertheless, the search of gamma motor neuron marker Gfra1 among all the clusters reveals that it is highly present in cluster 3 (Figure 5.24) while also showing some presence in clusters 0,1 and 2. The next subtype, characterized by alpha fast motor neuron markers is divided among clusters 0, 1 and 3 (Figure 5.28). These results are indicating the presence of cell type sub-populations within the alpha fast motor neurons, which were not fully characterized in the previous studies. However, more validation studies need to be done before any conclusive remarks. Furthermore, another cell type containing alpha slow motor neuron markers is also bifurcated. Among these markers, three genes (Sv2a, Crtac1, Aldoc) are present in cluster 3, and other two genes (PrKcd and Timp3) have high expression in clusters 4 and 5 respectively. As discussed above, these observations are indicating that alpha slow motor neurons could also have more sub-types, which were not identified earlier. At last, precursor motor neuron marker Olig2 was expressed in the cells of clusters 0,1 and 2 (Figure 5.36). Overall, single-cell sequencing using Drop-seq method provided an unprecedented amount of data for the cell types present in spinal cord samples. This data could be used for many further applications; for example, in disease studies, comparative analysis of gene expression (specifically for individual cell types) could be possible.

As discussed in Case study-II (section 5.5), administration of recombinant erythropoietin (EPO) in mice for three weeks was associated with a ~20% increase in cell numbers of pyramidal neurons and oligodendrocytes in the hippocampus (Hassouna et al., 2016). However, the published study did not address, which specific precursor cells give rise to this increment in neuronal cell numbers upon EPO treatment. That is why we have decided to address this question

150

to investigate potential precursor cells at the single-cell level resolution, which can differentiate into neurons and oligodendrocytes followed by EPO intervention. We tried to determine the molecular mechanisms behind this phenomenon using the Drop-seq method. The two groups (EPO and placebo) of dissociated cells were subjected to Drop-seq preparation and analysis. The results showed that there are 7 clusters detected in hippocampal neuronal single-cell samples (Figure 5.45). Based on previously known markers, individual clusters are annotated and labeled with a plausible cell type. In this hippocampal neuronal dataset, the largest cluster of the cells is annotated as glutamatergic cells; however, because of the high variability within the cell cluster, it is manually divided into two sub-clusters (Glutamatergic 0 and 1) based on the tSNE plot (Figure 5.47). In the differential analysis, the Glutamatergic-1 cluster showed a statistically significant increase in the neuronal cell percentage after EPO treatment (Table 5.10). However, the Glutamatergic-0 cluster showed a decrease in the neuronal cell population, overall nullifying any increase in the glutamatergic cell cluster. Nonetheless, further investigations are needed to understand these observations. Furthermore, one branch of oligodendrocytes seems to be acquiring neuronal cell characteristics upon EPO treatment as shown in tSNE plot (Figure 5.50), which is quite unusual to comprehend and would require further research to understand this phenomenon.

Chapter 6. Summary and conclusion

Molecular information from individual cell types could help to better understand disease mechanisms as well as several other functions of tissues. However, the biological complexity and physical properties (such as proximity with other cell types and cell junctions like gap junction, tight junction, and desmosomes) of the different cell types in a tissue makes it experimentally challenging to obtain such high-resolution information on a molecular scale. In the current thesis, we have implemented the techniques, which have enabled us to explore cell type-specific genetic and epigenetic information, and at the end provided a deeper understanding of various molecular mechanisms. In Chapter 3. the BiTS technique enabled us to isolate neuronal and non-neuronal cell type nuclei based on the antibody labeling of the neuronal marker NeuN on the surface of the nuclear membrane. After FACS sorting and isolation of chromatin, several epigenetics studies were done on the sample sets. In Chapter 4. the Tagger mouse system was developed to gain information on multiple entities of nucleic acids in a specific cell type. Furthermore, in Chapter 5. the single-cell sequencing method Drop-seq enabled us to gain molecular insights at a resolution of single-cells from the spinal cord and hippocampal tissue respectively.

6.1 Outlook

The cell type-specific data for neuronal and non-neuronal cells generated in BiTS-ChIP and BiTS-MeDIP experiments followed by high throughput sequencing analysis were uploaded and maintained in an online web platform at https://oasis.dzne.de/share/JBrowse-1.11.4/index.html. In general, the data contains information on two different cell types, two different brain tissues, and three-time points along with different epigenetic marks as described in Chapter 3. Exploring this data with novel aspects and research questions can benefit the neuroscience research community to a greater extent. Results from this study also showed newly predicted cell type-specific genes for neuronal and non-neuronal cells, which could be useful to gain new molecular insights and to reveal hidden layers of disease mechanisms (such as for neurodegenerative diseases).

Moreover, the Tagger system can be employed to understand cell type-specific molecular mechanisms in *in-vivo* systems. This system can provide information from various nucleic acid sources as well as epigenetic regulators (histone marks) in the cell, which play a role in regulating the gene expression. By combining the Tagger system with disease mouse models or in treatment intervention studies, it could revolutionize the understanding of the disease mechanisms as well as drug discovery process. The TU-tagging component of the Tagger system could also address many time point related questions using the pulse-chase experiments.

In the end, BiTS coupled epigenetic studies as well as the Tagger system have advantages in specific experimental setups and are based on the pre-defined markers of the cell types. On the other hand, the single-cell RNA sequencing (scRNA-seq) techniques could be applied to any previously undefined cell types or tissue samples and provides far more in-depth information about the biological samples in the study. The Drop-seq technique already proved to be efficient in cell type characterization as well as in finding novel and rare cell types (E. Z. Macosko et al., 2015). Apart from that, this technique could be useful in linking specific cell populations with the genes involved in the disease mechanisms previously identified in genome-wide genetic studies. Furthermore, combining Drop-seq with gene expression modification techniques, for instance, mutations, small molecule intervention, pathogen or induction of cells, could provide multi-dimensional cellular response information for experimental conditions. Finally, because scRNA-seq experiments need fewer samples and provide information at single-cell resolution, these techniques could revolutionize the personalized medicine field.

6.2 Limitations and future directions

One of the limitations of the BiTS-ChIP and –MeDIP experiments are the dependencies on the known markers for specific cell populations. Expression of these marker genes could be altered by cellular stress or other environmental factors and therefore can alter the experimental observations. This method is also hard to implement with multiple markers at the same time, which is essential when one wants to obtain the sub-populations of the cells. Because of the nuclear isolation step, another limitation of this method was the loss of cytosolic information. During the experimental protocol, cell nuclei are subjected to formaldehyde cross-linking which

could change the structure and specificity of the epitopes present on the nucleus surface. It can further lead to variations in downstream procedures such as FACS or immunoprecipitation. Antibodies should be thoroughly checked for cross-reactivity before using in the experiments.

Regarding the Tagger system, its bottleneck is to develop the mouse line with the help of pre-defined markers for the desired cell types. Developing the mouse line is both costly as well as time-consuming. Furthermore, different cellular sub-types could express the different proportions of marker genes, although, the Tagger system cannot differentiate between these gene expression variations. The Tagger system also depends on pre-defined markers for the particular cell type studies, which limits its use for the previously undefined novel cell types.

Moreover, one of the limitations of the Drop-seq method is that it still needs many more cells as input compared to the output results. The fundamental principle of mixing of the cells and beads, in the Drop-seq method, is based on the diluting cells to Poisson-limiting concentration; therefore a vast majority of beads and cells (90%-99%) are not exposed to each other (E. Z. Macosko et al., 2015). This limits the Drop-seq applicability for cells that are very difficult to obtain. Another limitation of Drop-seq experiments is the information loss of the spatial location of the single-cells in the tissue. As the Drop-seq method captures only mRNAs from the single-cells, the genetic and epigenetic information is also not obtained from the single-cells. It is also not trivial to extract absolute gene expression levels and to remove noise signals from the single-cell samples, which makes it difficult to detect rare cell types.

Some of the limitations from the BiTS and the Tagger systems could be addressed by first analyzing the tissue with the Drop-seq method to gain information about the markers of all the cell types present in the tissue in an unbiased manner. Later, these markers can be used in the BiTS coupled ChIP-seq or MeDIP-seq as well as in the Tagger system to obtain much deeper genetic and epigenetic information about a specific cell type in the tissue. The Tagger system could also be used for the cell type-specific ChIP-seq experiments for histone marks or transcription factor binding site analysis. To overcome the inherent limitations of the Drop-seq method some technique developments were happening in the single-cell-sequencing field recently. To do single-cell sequencing with low input and rare samples, the Seq-well technique

was developed (Gierahn et al., 2017). In this technique, barcoded mRNA capture beads and single-cells were sealed in an array of nanoliter volume wells using a semi-permeable membrane, which facilitates the effective cell lysis and capture of transcripts (Gierahn et al., 2017). Furthermore, for the preserved tissues or the cells, which were very hard to dissociate, the single nucleus RNA-seq (DroNC-seq) method was developed (Habib et al., 2017). As an addition to DroNC-seq, the Div-seq technique was developed, which can capture rare dynamic processes (for instance adult neurogenesis) by labeling of proliferating cells by EdU (5-ethynyl-2'-deoxyuridine) to profile single dividing cells (Habib et al., 2016). In future, single-cell techniques like scChIP-seq, scBS-seq (single-cell bisulfite sequencing for DNA methylation) and other genetic and epigenetic methods need to be developed.

References:

Adamcio, B., Sargin, D., Stradomska, A., Medrihan, L., Gertler, C., Theis, F., … Ehrenreich, H. (2008). Erythropoietin enhances hippocampal long-term potentiation and memory. *BMC Biology*, *6*, 37. https://doi.org/10.1186/1741-7007-6-37

Andrews, S. (2010). FastQC A quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., … Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Bessa, J., Tena, J. J., De La Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A., … Gómez-Skarmeta, J. L. (2009). Zebrafish Enhancer Detection (ZED) vector: A new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Developmental Dynamics*, *238*(9), 2409–2417. https://doi.org/10.1002/dvdy.22051

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., … Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, *44*(2), 148–156. https://doi.org/10.1038/ng.1064

Bonn, S., Zinzen, R. P., Perez-Gonzalez, A., Riddell, A., Gavin, A. C., & Furlong, E. E. M. (2012). Cell type-specific chromatin immunoprecipitation from multicellular complex samples using bits-chip. *Nature Protocols*, *7*(5), 978–994. https://doi.org/10.1038/nprot.2012.049

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., … Stegle, O. (2015). A n a ly s i s Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells, *33*(2). https://doi.org/10.1038/nbt.3102

Butler, A., & Satija, R. (2017). Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2017/07/18/164889.abstract

Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., … Barres, B. A. (2008). A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *Journal of Neuroscience*, *28*(1), 264–278. https://doi.org/10.1523/JNEUROSCI.4178-07.2008

Capece, V., Garcia Vizcaino, J. C., Vidal, R., Rahman, R.-U., Pena Centeno, T., Shomroni, O., … Bonn, S. (2015). Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics*, *31*(13), 2205–2207. https://doi.org/10.1093/bioinformatics/btv113

Chung, N. C., & Storey, J. D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics (Oxford, England)*, *31*(4), 545–554. https://doi.org/10.1093/bioinformatics/btu674

Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R., & Boothroyd, J. C. (2005). Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nature Biotechnology*, *23*(2), 232–237.

https://doi.org/10.1038/nbt1061

Digicaylioglu, M., Bichet, S., Marti, H. H., Wenger, R. H., Rivas, L. A., Bauer, C., & Gassmann, M. (1995). Localization of specific erythropoietin binding sites in defined areas of the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(9), 3717–3720.

Egelhofer, T. A., Minoda, A., Klugman, S., Lee, K., Kolasinska-Zwierz, P., Alekseyenko, A. A., … Lieb, J. D. (2011). An assessment of histone-modification antibody quality. *Nature Structural and Molecular Biology*, *18*(1), 91–94. https://doi.org/10.1038/nsmb.1972

Ehrenreich, H., Fischer, B., Norra, C., Schellenberger, F., Stender, N., Stiefel, M., … Bartels, C. (2007). Exploring recombinant human erythropoietin in chronic progressive multiple sclerosis. *Brain : A Journal of Neurology*, *130*(Pt 10), 2577–2588. https://doi.org/10.1093/brain/awm203

Einarsson, E. O., & Nader, K. (2012). Involvement of the anterior cingulate cortex in formation, consolidation, and reconsolidation of recent and remote contextual fear memory. *Learning & Memory*, *19*(10), 449–452. https://doi.org/10.1101/lm.027227.112

Fanselow, M. S. (1990). Factors governing one-trial contextual conditioning. *Animal Learning & Behavior*, *18*(3), 264–270. https://doi.org/10.3758/BF03205285

Fischer, J. J., Toedling, J., Krueger, T., Schueler, M., Huber, W., & Sperling, S. (2008). Combinatorial effects of four histone modifications in transcription and differentiation. *Genomics*, *91*(1), 41–51. https://doi.org/10.1016/j.ygeno.2007.08.010

Fishell, G., & Heintz, N. (2013). Perspective The Neuron Identity Problem : Form Meets Function. *Neuron*, *80*(3), 602–612. https://doi.org/10.1016/j.neuron.2013.10.035

Friese, A., Kaltschmidt, J. A., Ladle, D. R., Sigrist, M., Jessell, T. M., & Arber, S. (2009). Gamma and alpha motor neurons distinguished by expression of transcription factor Err3. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(32), 13588–13593. https://doi.org/10.1073/pnas.0906809106

Gay, L., Miller, M. R., Ventura, P. B., Devasthali, V., Vue, Z., Thompson, H. L., … Doe, C. Q. (2013). Mouse TU tagging: A chemical/genetic intersectional method for purifying cell type-specific nascent RNA. *Genes and Development*, *27*(1), 98–115. https://doi.org/10.1101/gad.205278.112

Genc, S., Koroglu, T. F., & Genc, K. (2004). Erythropoietin and the nervous system. *Brain Research*, *1000*(1–2), 19–31. https://doi.org/10.1016/j.brainres.2003.12.037

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., … Snyder, M. (2007). What is a gene , post-ENCODE ? History and updated definition, *1905*, 669–681. https://doi.org/10.1101/gr.6339607.Freely

Gierahn, T. M., Wadsworth, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., … Shalek, A. K. (2017). Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods*, *14*(4), 395–398. https://doi.org/10.1038/nmeth.4179

Guzman-Karlsson, M. C., Meadows, J. P., Gavin, C. F., Hablitz, J. J., & David Sweatt, J. (2014). Transcriptional and epigenetic regulation of Hebbian and non-Hebbian plasticity. *Neuropharmacology*, *80*, 3–17. https://doi.org/10.1016/j.neuropharm.2014.01.001

Habib, N., Basu, A., Avraham-Davidi, I., Burks, T., Choudhury, S. R., Aguet, F., … Regev, A. (2017). Deciphering cell types in human archived brain tissues by massively- parallel single nucleus RNA-seq. *bioRxiv*, 1–39. https://doi.org/http://dx.doi.org/10.1101/115196

Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J. J., … Regev, A. (2016). Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn

neurons. *Science.* Retrieved from http://science.sciencemag.org/content/early/2016/07/27/science.aad7038.abstract

Halder, R., Hennion, M., Vidal, R. O., Shomroni, O., Rahman, R. U., Rajput, A., … Bonn, S. (2015). DNA methylation changes in plasticity genes accompany the formation and maintenance of memory. *Nature Neuroscience*, *19*(1), 102–110. https://doi.org/10.1038/nn.4194

Hassouna, I., Ott, C., Wüstefeld, L., Offen, N., Neher, R. A., Mitkovski, M., … Vreja, I. C. (2016). Revisiting adult neurogenesis and the role of erythropoietin for neuronal and oligodendroglial differentiation in the hippocampus. *Molecular Psychiatry*, *21*(12), 1752–1767. https://doi.org/10.1038/mp.2015.212

He, M., Liu, Y., Wang, X., Zhang, M. Q., Hannon, G. J., & Huang, Z. J. (2012). Cell-Type-Based Analysis of MicroRNA Profiles in the Mouse Brain. *Neuron*, *73*(1), 35–48. https://doi.org/10.1016/j.neuron.2011.11.010

Hodges, V. M., Rainey, S., Lappin, T. R., & Maxwell, A. P. (2007). Pathophysiology of anemia and erythrocytosis. *Critical Reviews in Oncology/Hematology*, *64*(2), 139–158. https://doi.org/10.1016/j.critrevonc.2007.06.006

Hornstein, N., Torres, D., Sharma, S. Das, Tang, G., Canoll, P., & Sims, P. A. (2016). Ligation-free ribosome profiling of cell type-specific translation in the brain. *Genome Biology*, 1–15. https://doi.org/10.1186/s13059-016-1005-1

Illingworth, R. S., & Bird, A. P. (2009). CpG islands--'a rough guide'. *FEBS Letters*, *583*(11), 1713–1720. https://doi.org/10.1016/j.febslet.2009.04.012

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., … Amit, I. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, *343*(6172), 776–779. https://doi.org/10.1126/SCIENCE.1247651

Kelsom, C., & Lu, W. (2013). Development and specification of GABAergic cortical interneurons. *Cell & Bioscience*, *3*(1), 19. https://doi.org/10.1186/2045-3701-3-19

Kim, J., & Fanselow, M. (1992). Modality-specific retrograde amnesia of fear. *Science*, *256*(5057), 675–677. https://doi.org/10.1126/science.1585183

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., … Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. https://doi.org/10.1038/nature09033

Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., & Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Developmental Dynamics : An Official Public*, *203*(3), 253–310. https://doi.org/10.1002/aja.1002030302

Ko, Y., Ament, S. A., Eddy, J. A., Caballero, J., Earls, J. C., Hood, L., & Price, N. D. (2013). Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), 3095–100. https://doi.org/10.1073/pnas.1222897110

Kouzarides, T. (2007). Review Chromatin Modifications and Their Function, 693–705. https://doi.org/10.1016/j.cell.2007.02.005

Kozlenkov, A., Wang, M., Roussos, P., Rudchenko, S., Barbu, M., Bibikova, M., … Dracheva, S. (2015). Substantial DNA methylation differences between two major neuronal subtypes in human brain. *Nucleic Acids Research*, *44*(6), 2593–2612. https://doi.org/10.1093/nar/gkv1304

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Legube, G., & Trouche, D. (2003). Regulating histone acetyltransferases and deacetylases. *EMBO Reports*, *4*(10), 944–947. https://doi.org/10.1038/sj.embor.embor941

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E. A. D., Tadmor, M. D., … Nolan, G. P. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, *162*(1), 184–197. https://doi.org/10.1016/j.cell.2015.05.047

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lienhard, M., Grimm, C., Morkel, M., Herwig, R., & Chavez, L. (2014). MEDIPS: Genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics*, *30*(2), 284–286. https://doi.org/10.1093/bioinformatics/btt650

Liu, Z., Chen, O., Wall, J. B. J., Zheng, M., Zhou, Y., Wang, L., … Liu, J. (2017). Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. *Scientific Reports*, *7*(1), 2193. https://doi.org/10.1038/s41598-017-02460-2

Lopez-Atalaya, J. P., & Barco, A. (2014). Can changes in histone acetylation contribute to memory formation? *Trends in Genetics*, *30*(12), 529–539. https://doi.org/10.1016/j.tig.2014.09.003

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Luo, L., Callaway, E. M., & Svoboda, K. (2008). Genetic Dissection of Neural Circuits. *Neuron*, *57*(5), 634–660. https://doi.org/10.1016/j.neuron.2008.01.002

Macosko, E., & Goldman, M. (2015a). *Drop-Seq Laboratory Protocol*. https://doi.org/http://mccarrolllab.com/dropseq/

Macosko, E., & Goldman, M. (2015b). Drop-seq Tutorial & Troubleshooting. Retrieved from http://mccarrolllab.com/drop-seq-troubleshootingtutorial-with-pictures-videos/

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Malik, A. N., Vierbuchen, T., Hemberg, M., Rubin, A. A., Ling, E., Couch, C. H., … Greenberg, M. E. (2014). Genome-wide identification and characterization of functional neuronal activity-dependent enhancers. *Nature Neuroscience*, *17*(10), 1330–1339. https://doi.org/10.1038/nn.3808

Marti, H. H., Wenger, R. H., Rivas, L. A., Straumann, U., Digicaylioglu, M., Henn, V., … Gassmann, M. (1996). Erythropoietin gene expression in human, monkey and murine brain. *The European Journal of Neuroscience*, *8*(4), 666–676.

Mattsson, N., Schott, J. M., Hardy, J., Turner, M. R., & Zetterberg, H. (2016). Selective vulnerability in neurodegeneration: Insights from clinical variants of Alzheimer's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, *87*(9), 1000–1004. https://doi.org/10.1136/jnnp-2015-311321

Middleton, S. A., Barbone, F. P., Johnson, D. L., Thurmond, R. L., You, Y., McMahon, F. J., … Jolliffe, L. K. (1999). Shared and unique determinants of the erythropoietin (EPO) receptor are important for binding EPO and EPO mimetic peptide. *Journal of Biological Chemistry*, *274*(20), 14163–14169. https://doi.org/10.1074/jbc.274.20.14163

Miller, M. R., Robinson, K. J., Cleary, M. D., & Doe, C. Q. (2009). TU-tagging: Cell type-

specific RNA isolation from intact complex tissues. *Nature Methods*, *6*(6), 439–441. https://doi.org/10.1038/nmeth.1329

Milo, R., Jorgensen, P., Moran, U., Weber, G., & Springer, M. (2010). BioNumbers — the database of key numbers in molecular and cell biology, *38*(October 2009), 750–753. https://doi.org/10.1093/nar/gkp889

Miskowiak, K., O'Sullivan, U., & Harmer, C. J. (2007). Erythropoietin enhances hippocampal response during memory retrieval in humans. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(11), 2788–2792. https://doi.org/10.1523/JNEUROSCI.5013-06.2007

Nemesh, J., & Macosko, E. (2015). Dropseq Core Computational Protocol. Retrieved from http://mccarrolllab.com/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf

Ott, C., Martens, H., Hassouna, I., Oliveira, B., & Erck, C. (2015). Widespread Expression of Erythropoietin Receptor in Brain and Its Induction by Injury. *Molecular Medicine*, 803–815. https://doi.org/10.2119/molmed.2015.00192

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*(11), 1096–1100. https://doi.org/10.1038/nmeth.2639

Poulin, J. F., Tasic, B., Hjerling-Leffler, J., Trimarchi, J. M., & Awatramani, R. (2016). Disentangling neural cell diversity using single-cell transcriptomics. *Nature Neuroscience*, *19*(9), 1131–1141. https://doi.org/10.1038/nn.4366

Proudhon, C., Duffié, R., Ajjan, S., Cowley, M., Iranzo, J., Carbajosa, G., … Bourc'his, D. (2012). Protection against De Novo Methylation Is Instrumental in Maintaining Parent-of-Origin Methylation Inherited from the Gametes. *Molecular Cell*, *47*(6), 909–920. https://doi.org/10.1016/j.molcel.2012.07.010

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rahman, R.-U., Capece, V., Gautam, A., Sattar, A., Pena Centeno, T., Fehlmann, T., … Bonn, S. (2018). Oasis 2: online analysis and querying of small RNA-seq data. *Bioinformatics*, 1–10.

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., … Ren, B. (2013). RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Computational Biology*, *9*(3). https://doi.org/10.1371/journal.pcbi.1002968

Runyan, J. D. (2004). A Role for Prefrontal Cortex in Memory Storage for Trace Fear Conditioning. *Journal of Neuroscience*, *24*(6), 1288–1295. https://doi.org/10.1523/JNEUROSCI.4880-03.2004

Sanz, E., Yang, L., Su, T., Morris, D. R., Mcknight, G. S., & Amieux, P. S. (2009). Cell-type-specific isolation of ribosome-associated mRNA from complex tissues, *106*(33), 13939–13944.

Satija, R. (2017). Seurat - Guided Clustering Tutorial. Retrieved from http://satijalab.org/seurat/pbmc3k_tutorial.html

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). A n a ly s i s Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5). https://doi.org/10.1038/nbt.3192

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., … Regev, A. (2014).

Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, *510*(7505), 363–369. https://doi.org/10.1038/nature13437

Shneider, N. A., Brown, M. N., Smith, C. A., Pickel, J., & Alvarez, F. J. (2009). Gamma motor neurons express distinct genetic markers at birth and require muscle spindle-derived GDNF for postnatal survival. *Neural Development*, *4*, 42. https://doi.org/10.1186/1749-8104-4-42

Smith, Z. D., & Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews. Genetics*, *14*(3), 204–220. https://doi.org/10.1038/nrg3354

Sweatt, J. D. (2013). The emerging field of neuroepigenetics. *Neuron*, *80*(3), 624–632. https://doi.org/10.1016/j.neuron.2013.10.023

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., … Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Van Der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605. https://doi.org/10.1007/s10479-011-0841-3

Wood, R. (1996). The general transcription factors of RNA polymerase II, (7), 2657–2683.

Wu, Y. E., Pan, L., Zuo, Y., Li, X., & Hong, W. (2017). Detecting Activated Cell Populations Using Single-Cell RNA-Seq. *Neuron*, *96*(2), 313–329.e6. https://doi.org/10.1016/j.neuron.2017.09.026

Xu, C., & Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, *31*(12), 1974–1980. https://doi.org/10.1093/bioinformatics/btv088

Zentner, G. E., & Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. *Nature Structural & Molecular Biology*, *20*(3), 259–266. https://doi.org/10.1038/nsmb.2470

Zhang, S., Wu, X., & You, Z. (2017). Jaccard distance based weighted sparse representation for coarse-to-fine plant species recognition. *PloS One*, *12*(6), e0178317. https://doi.org/10.1371/journal.pone.0178317

Zhou, V. W., Goren, A., & Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, *12*(1), 7–18. https://doi.org/10.1038/nrg2905

Zovkic, I. B., Guzman-Karlsson, M. C., & Sweatt, J. D. (2013). Epigenetic regulation of memory formation and maintenance. *Learning & Memory*, *20*(2), 61–74. https://doi.org/10.1101/lm.026575.112

Ashish Rajput

Kegelhofstrasse, 36, 20251 Hamburg, Germany
Mobile: +49 (0) 176-72691524
Email: ashish.medbt@gmail.com

Profile

- International experience of 6 years in research & development and project management.
- Expertise in the single cell and next generation sequencing tools, genetic and epigenetic methodologies, tissue culture and cell imaging techniques.
- Extensive experience in molecular biology and bioinformatics analysis of sequence data.

Professional experience and education

11/13-Present	**University Clinic Eppendorf (UKE), Hamburg, Germany**
	University of Göttingen, German center for neurodegenerative diseases (DZNE), Göttingen, Germany
	PhD thesis with Prof. Stefan Bonn (5 scientific papers; 2 more in preparation)

- *Established single cell sequencing technique (Drop-seq) in the lab and characterized sub-population of motor neurons and hippocampal neurons using this method in mouse models.*
- *DNA methylation and transcriptomics studies on a collaboration project to develop the Tagger mouse system: a tool to dissect brain function and pathophysiology in in-vivo conditions.*
- *Epigenetics studies to understand Learning and memory molecular mechanisms (Resulted into Nature Neuroscience publication)*
- *Small RNA sequencing study on >250 human patient and healthy individuals to investigate early stage biomarkers for Parkinson's disease (DENOPA)*

03/13-10/13	**University of Liverpool, UK**
	Virtual internship, in the group of Prof. João Pedro de Magalhães, Integrative Genomics of Ageing Group (1 scientific paper)

- *Systematic analysis of the gerontome to reveal links between aging and age-related diseases*

10/11-10/13	**Indian Institute of Science, Bangalore, India**
	Research associate with Prof. Subba Rao

- *Microarray gene expression data analysis to find a crosstalk in MiT/TFE family transcription factors in regulating the lysosome and melanosome biogenesis.*

07/08-06/10   **Pravara Institute of Medical sciences, India**
                     Master's of Science in Medical Biotechnology (1 Scientific paper)

- *Biofilm characterization and quorum quenching studies in pathogenic strains of S. aureus and P. aeruginosa*

Research skills

Next generation sequencing: Library preparation and optimization for the Next generation sequencing (NGS) studies (ChIP-seq, MeDIP-seq, RNA-seq, small RNA and miRNA-seq, script-seq). ∗ DNA methylation studies using MeDIP and whole genome bi-sulfite sequencing. ∗ Single cell sequencing using SMARTseq2, Fluidigm C1, Drop-seq and 10X genomics.

Cell Culture: Cell line maintenance, Transfection optimization, Retrovirus and lentivirus -based stable cell line, Gene overexpression and knockdown (RNAi) assay.

Animal handling: Mouse and rat handling, Embryonic cortical neurons isolation and culture.

Microscopy: Fluorescence microscopy, Confocal microscopy, Immunofluorescence (IF) studies, Image analysis (ImageJ)

Molecular biology: Plasmid and genomic DNA isolation, gene cloning from bacteria, site-directed mutagenesis, PCR, qPCR, FACS, RNA isolation from human blood and brain samples.

Biochemistry: Western blotting, 2D-Gel electrophoresis, Recombinant protein purification, ELISA.
Computational: NGS data analysis using R packages. ∗ Quality analysis, Mapping to the genome, Differential expression analysis, Pathway analysis ∗ Single cell transcriptomics data analysis. ∗ Basic knowledge of machine learning and artificial intelligence methods

Programming language: C, C++, R, Java, shell
Writing tools: MS office suite, Adobe Photoshop, Illustrator, Graph Pad Prism

Communication and management skills

Leadership: Headed 4 scientific projects individually and along with research collaborators. ∗ Treasurer and head of University of Göttingen Indian association (2015). ∗ Organized PhD program Molecular Medicine yearly scientific meeting and retreat (2017).

Teaching: Guided fellow co-workers, technical assistants and master level students in scientific goals. ∗ University of Göttingen master level course on Microarray data analysis using R.

Professional training: Single cell sequencing using the Drop-seq method hands on training (Steve MaCarroll lab at Harvard Medical School, Boston, USA). ∗ Grant writing course (University of Göttingen, Germany).

Language: English (Professional proficiency), German (A1 level), Hindi (Mother tongue), Sanskrit (Medium level).

164

Fellowships and awards

- Research grants from German Center for Neurodegenerative Diseases (DZNE) to pursue doctorate studies (2014-2017).
- Travel grants from DZNE to travel to Harvard Medical School, Boston, USA (2015).
- Conference and travel grant from University Klinikum Göttingen, Germany (2017).
- Research fellowship from the Wellcome Trust (UK) and Department of Biotechnology (India) sponsored projects (Dr. Subba Rao lab, IISc, Bangalore; 2011-2013)
- Scored 94 percentiles in Graduate Aptitude Test in Engineering (GATE), India, 2009.
- GRE (307/340) ∗ TOEFL (97/120)

Scientific publication

1. Halder R, Hennion M, Vidal RO, Shomroni O, Rahman RU, Rajput A, et al. DNA methylation changes in plasticity genes accompany the formation and maintenance of memory. **Nat Neurosci**. 2015;19(1):102-10.
2. Llorens F, Thüne K, Sikorska B, Schmitz M, Tahir W, Fernández-borges N,..,Rajput A, et al. Altered Ca 2 + homeostasis induces Calpain- Cathepsin axis activation in sporadic Creutzfeldt-Jakob disease. **Acta Neuropathol Commun**. 2017;1-20.
3. Hatje K, Rahman R, Vidal RO, Simm D, Hammesfahr B, Bansal V, Rajput A, et al. The landscape of human mutually exclusive splicing. **Mol Syst Biol**. 2017;1-19.
4. Paiva I, Pinho R, Pavlou MA, Hennion M, Szeg M, Rajput A, et al. Sodium butyrate rescues dopaminergic cells from alpha-synuclein-induced transcriptional deregulation and DNA damage. **Hum Mol Genet**. 2017;26(12):2231-46.
5. Wolfes AC, Ahmed S, Awasthi A, Stahlberg MA, Rajput A, Magruder DS, et al. A novel method for culturing stellate astrocytes reveals spatially distinct Ca 2 + signaling and vesicle recycling in astrocytic processes. **J Gen Physiol**. 2017;149(1):149-70.
6. Fernandes M, Wan C, Tacutu R, Barardo D, Rajput A, Wang J, et al. Systematic analysis of the gerontome reveals links between aging and age-related diseases. **Hum Mol Genet**. 2016;25(21):4804-18.

Publications in preparation:

1. Rajput A*****, Wakhloo D*, Bansal V, Ibrahim A, et al. EPO induces differentiation of pre-existing non-dividing precursors in the hippocampus to enhance pyramidal neuron numbers and cognition. (* shared authorship)
2. Kaczmarczyk L, Rajput A*****, Bansal V*, Schleif M, Rahman RU, et al. Tagger mouse: a tool to dissect brain function and pathophysiology. (* shared authorship)