Dissertation

submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of

Doctor of Natural Sciences

Presented by

M.Sc. Manuel Göpferich

born in Bruchsal, Germany

Oral examination: December 17th, 2020

Single cell 3'UTR analysis identifies changes in alternative polyadenylation throughout neuronal differentiation and in autism

Referees: Prof. Dr. Ana Martin-Villalba

Dr. Wolfgang Huber

Summary – English

The usage of single cell sequencing techniques has grown extensively over the last years. One commonly utilized platform to conduct these experiments is the one from 10X Genomics. The applied workflow enables the detection of gene expression in single cells. Importantly, standard analysis tools count the gene but do not consider its exact mapping position within the transcript. This can be considered as a missed opportunity since mRNAs are captured by oligo(dT) primers and sequenced from the most 3' position of the transcript (3' capturing or tagging). These 3'ends are biologically relevant as one gene can have 3' untranslated regions (3'UTRs) of alternative lengths, also referred as alternative polyadenylation (APA). From multiple polyadenylation singles within one gene, one signal is selected by the cleavage and polyadenylation machinery terminating mRNA transcription. This process yields mRNAs with shorter and longer 3'UTRs but the same protein coding sequence. In order to make use of this information I developed a bioinformatical pipeline to call 3'peaks in sequencing data with single cell resolution and analyzed alterations in APA by multinomial regression (MNR).

I applied this method on the neural stem cell (NSC) lineage of the adult mouse. This system resembles the differentiation process from a quiescent stem cell into a neuroblast. Interestingly, genes altering their poly(A) site choice with lineage progression were enriched for risk genes in neurodevelopmental disorders, amongst others, Autism spectrum disorder (ASD). Analyzing 3'UTR usage in ASD patients and a control group revealed a trend for 3'UTR lengthening in individuals diagnosed with ASD compared to controls.

Motif analysis in the murine and the human sequencing data further pointed to an enrichment of the cytoplasmic and polyadenylation element (CPE) in 3'UTRs. These motifs are recognized by CPE binding proteins. RNA immunoprecipitation showed that CPEB4 can bind this motif. Proteomics as well as ribosomal profiling data was utilized to estimate the effects of CPEB4 binding and 3'UTR shortening on mRNA translation.

Intriguingly, co-expression of CPEB4 and amyloid beta precursor-like protein 1 (APLP1), a synaptic adhesion molecule which can bind CPEBs, was observable in the mouse and the human single cell sequencing data. Finally, comparing APLP1 knockout to wildtype mice revealed CPE-dependent alterations in 3'UTR usage between both genotypes.

In summary, the results of the here developed method for 3'UTR calling shows that 3'peaks, single poly(A) signals, can be successfully detected in 10X Genomics data. In addition, the downstream analysis suggests a link between APA, neuronal differentiation, CPEB4 and the neurodevelopmental disorder ASD.

Summary – German

Die Verwendung der Methoden einzelne Zellen zu sequenzieren hat im Laufe der letzten Jahre stark zugenommen. Eine häufig verwendete Plattform um diese Experimente durchzuführen ist die von 10X Genomics. Diese erlaubt es, die Expression von Genen in einzelnen Zellen zu detektieren. Allerdings quantifiziert die standardmäßig verwendete Software nur die Gene aber berücksichtigt nicht deren exakten positionelle Zuordnung im Transkript. Dies kann als verschwendete Gelegenheit gesehen werden, da die mRNAs mit Oligo(dT)-Primern gebunden und aus Richtung der 3'Enden der Transkripte sequenziert werden (3'Capturing oder Tagging). Diese 3'Enden sind von biologischer Relevanz, weil ein Gen 3' untranslatierte Regionen (UTRs) mit unterschiedlichen Längen haben kann, auch bekannt als alternative Polyadenylierung, oder kurz APA. Von mehreren Polyadenylierungssignalen in einem Gen wird eines von der Polyadenylierungs-Maschinerie erkannt, welche die Transkription der mRNA beendet. Dieser Vorgang erzeugt mRNAs mit kürzeren und längeren 3'UTRs aber der gleichen codierenden Sequenz. Um diese Information verwenden zu können, habe ich eine bioinformatische Methode entwickelt, welche sogenannte 3'Peaks basierend auf Einzelzell-Sequenzierungsdaten identifizieren kann und Änderungen in APA mittels multinominaler Regression (MNR) analysiert.

Besagte Analysemethode habe ich auf neurale Stammzellen der adulten Maus angewendet. Dieses Tiermodell stellt den Differenzierungsprozess einer quieszenten Stammzelle zu einem Neuroblasten dar. Interessanterweise zeigten Gene, die ihre Polyadenylierung mit fort-schreitender Differenzierung ändern, eine Überrepräsentation von Risikogenen für neurologische Entwicklungsstörungen, unter anderem Autismus-Spektrum-Störung (ASS). Ein Vergleich der 3'UTRs in ASS-Patienten zeigte, dass 3'UTRs tendenziell länger sind in Autisten relativ zur Kontrollgruppe.

Motivanalysen in den Maus- und humanen Sequenzierungsdaten deuteten weiterhin auf eine Überrepräsentation des cytoplasmatischen Polyadenylierungselements (CPE) hin. Diese Motive werden von CPE-Bindeproteinen (CPEBs) erkannt. RNA Immunoprezipitierung ergab, dass CPEB4 dieses Motiv binden kann. Proteomik sowie Ribosome-Profiling wurden herangezogen, um den Einfluss der Binding durch CPEB4 und das Verkürzen der 3'UTRs auf die mRNA-Translation abzuschätzen.

Erstaunlicherweise war eine Koexpression von CPEB4 und dem CPEB-bindenden, synaptischen Adhäsionsmolekül Amyloid-Beta-Precursor-like Protein 1 (APLP1) in den Maus-

sowie den humanen Datensätzen zu beobachten. Zudem zeigten APLP1-defiziente Mäuse vom wildtyp-abweichende alternative Polyadenylierung.

Als Zusammenfassung ist zu sagen, dass die hier entwickelten Methode mit dem Ziel 3'UTRs-Signale zu bestimmen, also einzelne poly(A)-Signale, diese in 10X-Genomics-Daten erfolgreich detektieren kann. Zusätzlich deutete die Auswertung der Daten auf eine Verbindung von APA, neuronaler Differenzierung, CPEB4 und der neurologischen Entwicklungsstörung ASS hin.

Acknowledgement

In this section I would like to thank all lab members of the Ana Martin-Villalba and the Wolfgang Huber group. Without the guidance and work of my thesis advisory committee (Ana Martin-Villalba, Wolfgang Huber, Simon Anders and Michael Boutros), my project partner Nikhil Oommen George and the support I received from both groups this work would not have been possible. Special thank goes to my first supervisor Ana Martin-Villalba for giving me the opportunity to work on this project and to develop this project based on my findings and ideas. Claiming that a transmembrane protein (APLP1) can act on alternative polyadenylation is not really intuitive and defending this claim in lab meetings was, frankly speaking, challenging. In terms of organization, Susanne Kleber, Irmgard Weirich and Simone Bell (Huber group) helped me a lot. I would like to thank my colleague Lukas Kremer for writing the R package *ggpointdensity*, which allows plotting single points and displaying their density (overplotting factor) at the same time. I am certainly one of most frequent users of this tool. In general, my colleagues in the Ana Martin-Villalba lab help me a lot by taking over some projects and tasks so that I could focus on the "3'UTR project". The most dramatic setback in this PhD was however that my second supervisor Bernd Fischer died few months after I started in the Martin-Villalba lab. I am thankful that Wolfgang Huber took over and invited me to join the group meetings at the EMBL. At this point I would also like to mention my side projects about the effect of interferons on neural stem cells. I hope that we will be able to finish and submit them as soon as possible to peer-review journals. Finally, I thank everyone giving me a great time (4 years by now to be exact) at the DKFZ and at the EMBL.

Abbreviations

(v)SVZ	(ventricular) subventricular zone
APA	alternative polyadenylation
APLP1/Aplp1	amyloid beta precursor like protein 1
ASD	autism spectrum disorder
CPE	Cytoplasmic Polyadenylation Element
CPEB4	Cytoplasmic Polyadenylation Element Binding Protein 4
GESA	gene set enrichment analysis
GLM	generalized linear model
LRT	log-likelihood ratio test
MNR	multinomial regression
NSC	neural stem cell
PAS	polyadenylation signal
RBPs	RNA binding proteins
UMI	unique molecular identifier

Table of Contents

Chapter 1

Introduction

In this chapter I will introduce the neural stem cell lineage and reason why this system is a suitable model to study neurodevelopmental disorders (1.1). Next, I will give a background on the biology of alternative polyadenylation and motivate why it is highly relevant in the context of neurodevelopmental disorders and stem cells (1.2). In addition, I will explain why the two genes CPEB4 and APLP1 are of specific interest for this work (1.3 and 1.4). Also, I will introduce single-cell sequencing techniques, explain their applications, analysis (1.5) and reason why these methods are of specific interest for the analysis of alternative polyadenylation (1.6). Finally, I will summarize the scientific goals of this study (1.7).

1.1 Neurogenesis and Neural Stem Cells in the subventricular zone

The brain is considered to be the most complex organ and its tasks are executed by myriads of neurons transmitting signals through chemical synapses in highly interconnected circuits. Neuroglial cells like astrocytes fulfill supportive functions, for instance, the maintenance of water and ion homeostasis (Jäkel S. & Dimou L., 2017). In order to set this network in place the brain needs to generate millions of neurons from few progenitor cells. This process is also known as neurogenesis. It was shown that radial glia cells are the main players in embryonic neurogenesis and that some radial glia cells are kept in the adult brain (Kriegstein A. & Alvarez-Buylla A., 2011). These cells are also referred to as neural stem cells (NSCs). In adult mice, neurogenesis only takes place at certain regions of the brain (Ming G.L. & Song H., 2011). The two most important ones are the lateral walls in the ventricular-subventricular zone (vSVZ) and the sub granular zone (SGZ) in the dentate gyrus of the hippocampus (Ming G.L. & Song H., 2011). While the stem cell niche of the SGZ produces dentate granule cells, the vSVZ functions as a source of neuroblasts which migrate through the rostral migratory stream (RMS) into the olfactory bulb (OB) and eventually mature into interneurons (Ming G.L. & Song H., 2011). As a result, the animal can adopt and fine-tune odor discrimination throughout lifetime (Lledo P.M. et al., 2016). Most vSVZ NSCs reside in a state of quiescence and show high similarity to astrocytes, in their morphology as well as their gene expression program (Linnarsson S., 2015). In literature they are termed quiescent NSCs (qNSCs) or B1q-cells (Kriegstein A. & Alvarez-Buylla A., 2011). The maintenance of this population was shown to depend on signals and factors outside the stem cell niche (Bond A.M. et al., 2015). In addition, the presence of interferons which increase with aging plays a role (Kalamakis G. et al., 2019). Upon activation, qNSCs switch their gene expression program and become active NSCs (aNSCs) and eventually transient-amplifying progenitors (TAPs) that differentiate into neuroblasts (NBs) (Llorens-Bobadilla E. et al., 2015). This lineage progression is depicted in Figure 1. Utilizing single cell sequencing techniques, the studies from Llorens-Bobadilla E. et al., 2015 and Kalamakis G. et al., 2019 described the transcriptomic changes throughout the NSC lineage progression. Here, the transition from qNSCs to neuroblasts was described as a continues process, meaning that inter-mediate states are represented by the transcriptomes of single cells.

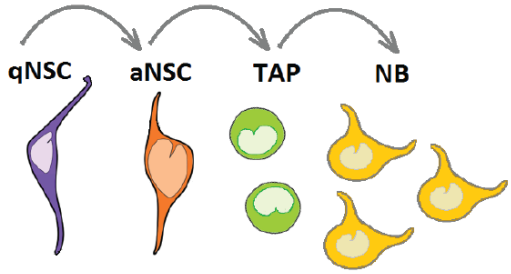


Figure 1: Schematic depiction of neurogenesis in the subventricular zone of the adult mouse, modified from Llorens-Bobadilla E. et al., 2015, neural stem cell (NSC) activation: qNSC → aNSC → TAPs; transit-amplifying progenitors (TAPs), neuroblasts (NBs) and differentiation: TAPs → NBs.

For the human brain, multiple studies reported evidence of adult neurogenesis (Eriksson P., et al. 1998; Ernst A. et al., 2014; Boldrini M. et al., 2018; reviewed by Kempermann G. et al., 2018). The study from Ernst A. et al., 2014 suggested the human orthologues of vSVZ-NSCs to be located in the striatum and linked a reduction in the number of postnatally generated neurons to the onset of Huntington's disease. This finding can be seen as a hint that the neurogenesis taking place in the vSVZ resembles a feasible model to study neurodevelopmental disorders. Apart from this aspect, vSVZ-NSCs are experimentally accessible as a mouse model. In particular, NSCs can be kept in cell cultures where they are in their active state and treated with the ligand BMP4 to shift them to a more quiescent state (Mira, H. et al., 2010; Martynoga, B. et al., 2013). In summary and most importantly, vSVZ-NSCs enable insights into the differentiation process from a glia-like cell (qNSC) into a neuron.

1.2 Why study Alternative Polyadenylation

The central dogma of molecular biology states that genes are transcribed into messenger RNAs (mRNAs) from their genomic templates and subsequently translated into proteins (Koonin E.V., et al., 2015). However, the amount of proteins produced per mRNA greatly varies as regulatory elements in untranslated regions (UTRs) can control with which rate an mRNAs is translated (Sandberg, R. et al., 2008). This is also referred as posttranscriptional regulation.

A relevant player in this regulation is the length of 3'UTRs. The process in which alternative 3'UTR lengths are created is termed alternative cleavage and polyadenylation (or short: APA). Among several polyadenylation signals (PASs) in the genomic template, one is selected (Figure 2). Around two-thirds of all mammalian genes bear multiple PASs and therefore can be subject to APA (Chen, C.-Y., et al. 2012; Miura P. et al., 2014).

Mechanistically, during the termination phase of mRNA transcription, the polyadenylation machinery is recruited. Cleavage and polyadenylation specificity factors (CPSFs) recognize the PAS, most commonly the hexamer AAUAAA, or a slight variation of this motif. Cleavage can also occur at non-canonical sites and depend less on the hexamer AAUAAA but more on auxiliary factors. Cleavage stimulation factors (CSTFs) bind to downstream elements (DSEs). The pre-mRNA is then cleaved and polyadenylated 10 to 30 nucleotides downstream of the PAS. If a proximal (upstream) PAS is selected the 3'UTR will be shorter and if a distal (downstream) one is selected it will be longer (Figure 2, Miura P. et al., 2014).

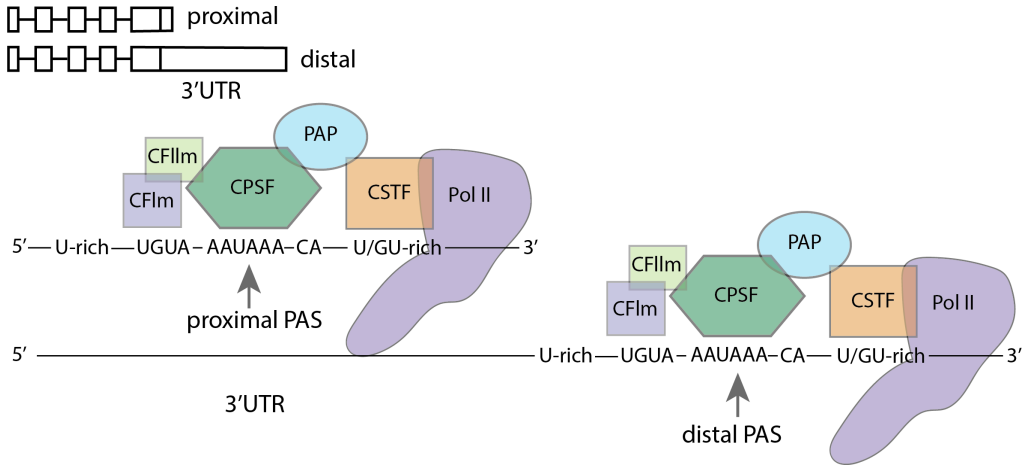


Figure 2: Schematic depiction of alternative polyadenylation, modified from Miura et al. 2014, the assembled factors (polyadenylation specificity factors, CPSFs & Cleavage stimulation factors, CSTFs) either cut and polyadenylate the nascent mRNA at a proximal polyadenylation signal (PAS) or at a distal one.

PAS selection depends on various conditions, like the expression of polyadenylation factors (Miura P. et al., 2014; Lackford B. et al., 2014), other binding proteins (Miura P. et al., 2014, Bava F.-A. et al., 2013), genetic variations (Mariella E. et al., 2019) and the sequence composition around the cleavage site. As an example, it was shown that

reduction of CPSF5 levels (also known as NUDT21 and CFIM25) affects APA in numerous genes (Alcott, C. E. et al., 2020).

Notably, APA can not only occur in terminal exons but also in introns resulting in mRNAs with truncated coding regions (Miura P. et al., 2014). This is most likely to happen in genes bearing a strong PAS in combination with a weak splice-site in an intron (Tikhonov M. et al., 2013). More commonly, mRNAs with the same coding region can have alternative 3'UTR lengths. Therefore, regulatory elements that are present in longer UTRs are absent in shorter ones (Miura P. et al., 2014). These elements can influence mRNA stability, the rate of mRNA translation and intracellular location of mRNAs (Miura P. et al., 2014). As the majority of these elements rather repress mRNA translation, shorter 3'UTRs tend to have a higher protein production (Gruber, A.R. et al., 2014). It was suggested that 3'UTR shortening in hundreds of genes impacts stem cells differentiation (Brumbaugh, J. et al., 2012; Lackford B. et al. 2014) and that even the 3'UTR length of a single gene like Pax3 can be critical for the fate decision of stem cells (De Morree A. et al., 2019). In addition, Sommerkamp P., et al. 2020 reported global 3'UTR shortening with stem cell activation and differentiation. In summary, this points to the importance of 3'UTR length regulation in stem cells.

From an evolutionary perspective, 3'UTR length was found to correlate with the complexity of organisms (Chen, C.-Y. et al. 2012). In the brain, expression of transcripts with ultra-long 3'UTRs was reported in neurons (Wang L. & Yi R., 2013; Miura P. et al., 2013). In general, 3'UTRs are longer in neurons compared to other cell-types (Guvenek A. & Tian, B., 2017; Guvenek A. & Tian B., 2018). The role of alternative 3'UTR lengths in neurons was studied in greater detail for single genes (Miura P. et al., 2014), like brain-derived neurotrophic factor (BDNF). In hippocampal neurons, the BDNF-transcript bearing the long 3'UTR is transported to dendrites, there stimulating dendritic outgrowth while the short isoform localizes to the soma (An J.J. et al., 2008). This local translation of mRNAs is relevant for the connection of neurons. A study by Jereb, S. et al., 2018 identified differential 3'UTR usage across different neuronal cell types. Also, the 3'UTRome was suggested as a hub of potential pathological variations relevant for neurodevelopmental disorders (Wanke K. et al., 2018).

1.3 The CPE motif affects translation and poly(A)-site selection

As described in the previous section, 3'UTRs can contain a variety of regulatory elements like binding sites for micro RNAs (miRNAs) or RNA binding proteins (RBPs), which can be critical for posttranscriptional regulation (Miura P. et al., 2014). A family of motifs that are of specific interest for this study are cytoplasmic polyadenylation elements (CPEs) with the core motif UUUUAU. In terms of intracellular localization, CPE-containing mRNAs were reported to be transported to dendrites in neurons (Fernandez-Moya S.M. et al., 2014, Groisman I. et al., 2006). The protein family of RNA binding proteins recognizing CPE motifs is termed CPEB, consisting of four members (Weill L. et al., 2012; Richter J.D., 2007), one of them being CPEB1. Although CPEB's have cytoplasmic functions as implied by their name, CPEB1 was shown to also have a nuclear one; it can influence poly(A)-site choice (Bava F.-A. et al., 2013). With regards to its role in mRNA translation, the presence of CPE motifs usually enhances protein outcome, meaning CPE-mRNAs are favored for translation over non-CPE-mRNAs (Weill L. et al., 2012; Richter J.D., 2007). However, CPEs can also repress translation mainly in non-neuronal cell types (Groisman I. et al., 2006). Lengthening of poly(A)-tails by CPEB4 was implied in ASD and CPEB4 binding targets are enriched for autism risk genes (Parras A. et al., 2018). In addition, in CPEB4 knockout mice autistic phenotypes were reported (Parras A. et al., 2018). Whether CPEB4 (not only CPEB1) can also influence PAS selection in the nucleus remains unclear but will be addressed in this work.

1.4 APLP1 binds CPEB and affects polyadenylation

APLP1 (amyloid beta precursor like protein 1) is a synaptic adhesion molecule (Schilling, S., et al. 2017) and belongs to the family of amyloid beta proteins, together with APP and APLP2 (Müller U.C., 2017). Amyloid beta proteins are a critical factor in Alzheimer, a neurodegenerative disease leading to dementia (Schilling, S., et al. 2017). In mice, most combinations of double knockouts for these proteins are lethal (von Koch, C. S., 1997; Schilling, S., et al. 2017). The molecular function of amyloid-beta proteins is not fully known (Müller U.C., 2017). It was suggested that APLP1 is important in maintaining dendritic splines and in synaptic signal transmission (Schilling,

S., et al. 2017). A different study showed that Aplp1 can mediate poly(A) induced translation via binding to CPEB1 with its intracellular domain in neurons (Cao Q. et al., 2005). This protein-protein interaction was identified in a yeast two-hybrid screen and confirmed utilizing co-immunoprecipitation and glutathione S-transferase (GST) pulldowns (Cao Q. et al., 2005). Combining the results from Bava F.-A. et al., 2013 about CPEB4 and Cao Q. et al., 2005, one can hypothesize that APLP1 can indirectly impact APA by enriching CPEB's at the membrane thus altering their activity on PAS selection in the nucleus.

1.5 History and analysis of single cell sequencing and other "omics" techniques

The development of high-throughput RNA sequencing techniques enabled the transcriptome-wide comparison of gene expression levels and found wide usage in medicine and biology (Hwang, B., et al. 2018). Microarrays were the first instance of this technique and quantified gene expression based on reverse-transcribed mRNAs (cDNA) as input (Hwang, B., et al. 2018). This platform is limited in the number of detected genes by predefined probes and also by the fact that expression could only be compared relative to a control sample. Next generation sequencing (NGS) techniques overcame these hard limitations (Hwang, B., et al. 2018). cDNA is clustered on flow-cells and sequenced as optical read-out when a matching base is incorporated. In NGS samples, the number of detected genes increases with sequencing depth and expression is quantified since gene counts enable the comparison to all other samples (Hwang, B., et al. 2018). Over many years NGS was restricted to bulk or cell population averages. Further technical developments in mRNA capturing and amplification facilitated sequencing the transcriptomes of single cells (Eberwine, J. et al. 1992, Hwang, B., et al. 2018). For the first time, scientists were able to study the heterogeneity in gene expression of single cells (Hwang, B., et al. 2018). This led to the discovery of unknown dynamics in gene expression and new cell-types that could not be isolated or observed by other techniques before (Hwang, B., et al. 2018, Llorens-Bobadilla E. et al., 2015). The restrictions on the other hand are mainly poor recovery of genes per single cell compared to bulk sequencing (Hwang, B., et al. 2018). Two main RNA sequencing

strategies had been established in the single cell sequencing field: full-coverage RNA approaches like the Smart-seq2 protocol (Picelli, S., et al. 2014) and droplet-based approaches (Zheng G., 2017). The latter one captures single cells in small droplets, binds mRNAs with oligo-(dT) sequences on beads and adds single cell barcodes and unique molecular identifiers (UMIs) to the reads (Figure 1.4). This workflow allows the detection of single mRNA molecules in single cells. Compared to full-coverage single-cell sequencing, the droplet-based tagging sequencing technique usually has a higher output in the number of single cells (Hwang, B., et al. 2018). One of the most commonly used platforms for droplet based single-cell RNA sequencing is the one from 10X Genomics (Zheng G., 2017).

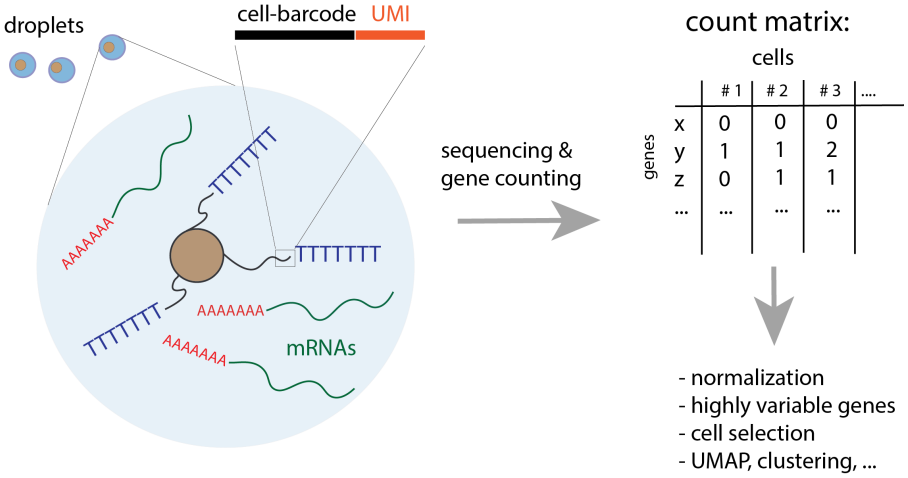


Figure 1.4: Droplet based single-cell sequencing, modified from Hwang, B., et al. 2018 for each cell: mRNAs are captured by oligo-(dT) sequences, UMI and cell barcodes are added, mRNAs are reverse transcribed and sequenced (not shown), gene counting per single cell and further down-stream analysis approaches.

A common analysis approach for single cell sequencing experiments are dimensionality reduction techniques like principle component analysis and UMAP, short for Uniform Manifold Approximation and Projection (Becht, E., et al. 2019). Both algorithms approximate the distance of single cells in the realm of gene expression and project them into a two-dimensional space. The UMAP algorithm tries find the neighbors of a given cell and then extends these local neighborhoods. For this purpose, highly variable genes are selected (Hwang, B., et al. 2018). Based on the result novel clusters of rare cell populations can be assigned and established marker genes are used to identify known cell-types (Hwang, B., et al. 2018). In the last years, a number of

dimensionality reduction algorithms centered around the idea to fit single cells onto trajectories – and therefore to model differentiation – had been introduced (Hwang, B., et al. 2018). One of these algorithms is reversed graph embedding (Trapnell C. et al., 2014). Its objective is to find structures (or trajectories) in high-dimensional data, *i.e*. high number of features and observations. From the learned embedding a latent variable can be derived referred as "pseudotime". The prefix "pseudo" is used because it is not a real time but rather a smooth transition from one state in gene expression to another (Trapnell C. et al., 2014; Hwang, B., et al. 2018). Assigning cells to discrete cell clusters or using a continues variable like "pseudotime" represent two solutions to describe cell states. Which one to use highly depends on the underlyng biology and further analysis. As single cell sequencing data has a high number of observations and features, reverse or feature-centered analyses are also feasible. For instance, signaling pathways, ligand-receptor interactions as well as other gene-to-gene interactions can be predicted based on the expression patterns in single cells (Hwang, B., et al. 2018). For the common task of computing differential expression in single cells, some scientists used modified versions of tools developed for bulk sequencing like *DESeq2* (Love M.I. et al., 2014) while others applied solutions from established single cell analysis software packages (Hwang, B., et al. 2018). Here, it has to considered that counts from classical bulk sequencing and those from single cells are quite distinct. Single cell counts represent unique molecules while bulk samples mostly have cDNA amplification biases (no unique molecular identifiers). Moreover, single cells are not independent observations, they describe the heterogeneity within an organism. Commonly, bulk RNA sequencing experiments are designed in such a way that they contain multiple replicates in order to represent the biological variability within a population (Robles, J.A., et al. 2012).

To summarize, the main objectives in single cell analysis are cell clustering and differential expression (Hwang, B., et al. 2018). But the pool of possible applications is way broader and not necessarily restricted to expression levels alone. For instance, the tool *velocyto* separates between spliced and unspliced RNAs based on the mapping to intronic regions and models the dynamics of mRNA degradation and splicing in single cells (La Manno G., et al. 2018). Another layer of gene regulation are the features of genes: coding regions, 5' and 3' untranslated regions. The same gene can be expressed in multiple isoforms having alternative compositions of exons and different 5' or 3' UTR lengths as already introduced in section 1.2. Standard single-cell analysis methods do not account for this complexity and rather treat genes as invariant functional units.

In contrast to single cell sequencing, many other "omic" methods which generally speaking try to quantify a high number of features still depend on high amounts of input materials from bulk samples. Some techniques, like DNA methylation, ATAC sequencing (Clark, S. J., et al. 2018) and proteomics (Kelly R.-T., 2020) have already been successfully applied on single cells but also have drawbacks in terms of covered features per cell and in the total number of cells (low throughput). Such experiments can be crucial for functional gene studies and be designed as follow-ups on findings from single cell sequencing studies. For instance, proteomics and ribosomal profiling were applied before to see how mRNA translation and therefore protein outcome changes with alterations in gene expression or across cell-types (Baser A. et al., 2019). In biology, this is highly critical as the functions of cells are carried out by proteins. Proteomics typically utilizes mass spectrometry methods: proteins are isolated for cells, digested into smaller peptides and detected on the spectrometer. The readouts for these peptides are then compared against databases and quantified as normalized intensities per protein (Cox J. et al., 2011; Cox J. et al., 2014). Ribosomal profiling is an RNA sequencing method and involves an isolation step to capture ribosome protected RNAs and for this reason provides a picture of which genes are actively translated and can be seen as an approximation of proteomics (Faye M.D. et al., 2014). Another useful method to study the function of a class of proteins, namely RNA binding proteins, is RNA immunoprecipitation (Wheeler, E. C., et al. 2018). Here, RNAs are pulled-down with an antibody against the protein of interest. The fraction containing the antibody will enrich for the RNA substrates of the binding protein. Next, this fraction is compared to control fractions to account for unspecific binding (Wheeler, E. C., et al. 2018).

To conclude this section, the big discrepancy in molecular biology consists of two experimental and analytic strategies. The rather classical approach aims for the comparison of features like gene expression across predefined conditions. Single cell sequencing provided a new approach focusing more on the unsupervised detection of novel cell clusters (Hwang, B., et al. 2018). Both strategies have a high synergy, for instance, in bulk sequencing one might observe differentially expressed genes. Single cell sequencing can then clarify whether this is due to the fact that either in one of the conditions a specific cell-type is overrepresented or whether the dysregulation of genes affects all cells in one condition. Ideally, additional "OMICs" like ribosomal profiling or proteomics can then address, for example, whether the changes in gene expression result in changes in protein levels.

1.6 Detection and analysis of alternative polyadenylation in high-throughput sequencing data

As described in section 1.2, alternative polyadenylation is critical for posttranscriptional regulation and therefore its detection is of high interest for biologists. In general, alternative polyadenylation can be quantified using next generation sequencing. The aim of such experiments is to compare multiple cell-types, conditions or treatments against each other. This is also called differential 3'UTR usage (Miura P. et al. 2014). Full coverage RNAseq data has been applied for such analyses (like in Szkop K. J. et al., 2017) but normally does not always allow the detection of distinct PASs (Miura P. et al. 2014). To overcome this limitation, software tools like *DaPars* (Xia Z. et al., 2014) fit functions to the 3'UTR read coverage vectors. These methods will approximate a center of gravity or a comparable measure and compare these estimates across conditions, gene-by-gene (Figure 1.5, upper panel). One reason why full-coverage RNAseq data is limited in this regard, can be explained by the cDNA fragmentation step, for instance in Smart-seq protocols (Picelli, S., et al. 2014). This protocol makes it hard to decide whether a part of a 3'UTR sequence is absent either because the cell expressed the gene with a shorter 3'UTR (proximal PAS) or because the distal 3'UTR was lost in the fragmentation step (missing at random). However, sequencing RNAs as fully covered sequences provides confidence that the whole mRNA was detected (Miura P. et al., 2014) and which exons are present.

As an alternative, 3'RNAseq techniques (Miura P. et al., 2014) were designed to read the transcript from the most 3'end (from the poly-A tail) and therefore to provide higher resolution and statistical power (Miura P. et al., 2014). In other words: the 3'ends are tagged. One recent study applying (low-input) bulk 3' sequencing is the one from Sommerkamp P., et al. 2020. Counting the most 3' positions of reads yields distinct 3'peaks reflecting the different PASs in genes (Figure 1.5, lower panel). Another representative study to name here is the one from Jereb, S. et al., 2018 in which the authors identified 3'peaks and compared them across different types of neurons. In contrast to the full-coverage approach, usually no information of the rest of the gene like its exon composition is available. The narrower the 3'peaks (meaning the smaller the peak sizes) the more likely it is to resolve PASs which are very close to each other.

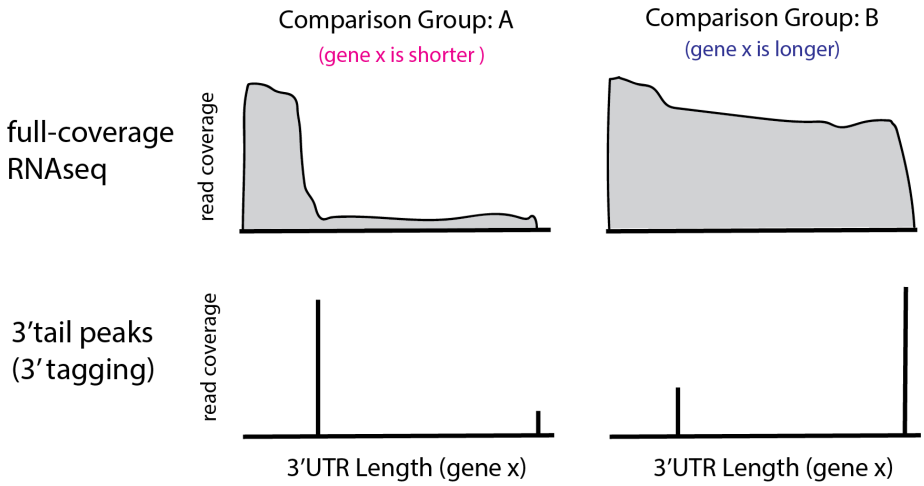


Figure 1.5: Detection of alternative 3'UTR usage, modified from Miura P. et al., 2014, the (fictitious) gene x is shorter in group A and hence longer in group B. Lower panel: the two 3'peaks mark the two PAS in the 3'UTR region of the gene, the y-axis represents the number of reads mapped per position (read coverage), upper panel: the same but for full-coverage RNAseq.

After assigning PASs by a peak calling algorithm the fractions of reads in each PAS can be directly compared across treatment conditions or cell-types. This is performed using suitable statistical methods assessing odds ratios like Fisher's Exact or Chi-Square tests. Importantly, these statistics will only be relevant if variation is estimated on biological replicates. An exception would be, for example, if the aim is to describe the heterogeneity across single cells in one sample. A statistical test to examine the confidence in odds ratios that works with replicates would be the Cochran–Mantel–Hänszel test (Agresti A., 2002). As an alternative, multinomial or ordinal models could be applied (Agresti A., 2002; see also Venables W.N. & Ripley B.D., 2002; Bohning D.,1992; Begg C.B. & Gray R., 1984). The applications of these models cover a wider range as diverse predictor variables can be fitted to the multinomial distributions allowing one to state how 3'UTR usage changes with the variable of interest (Figure 1.6). Importantly, these models will not fit alterations in total expression like the *DESeq2* tool, but compare (local) odds ratios in 3'UTR usage.

As already stated before, the benefit of bulk (3') sequencing approaches lies in the high sequencing depth which can be obtained with relatively low costs. Their downside, however, is that these approaches will reflect an average measure across cell populations and do not provide the resolution of single-cell sequencing.

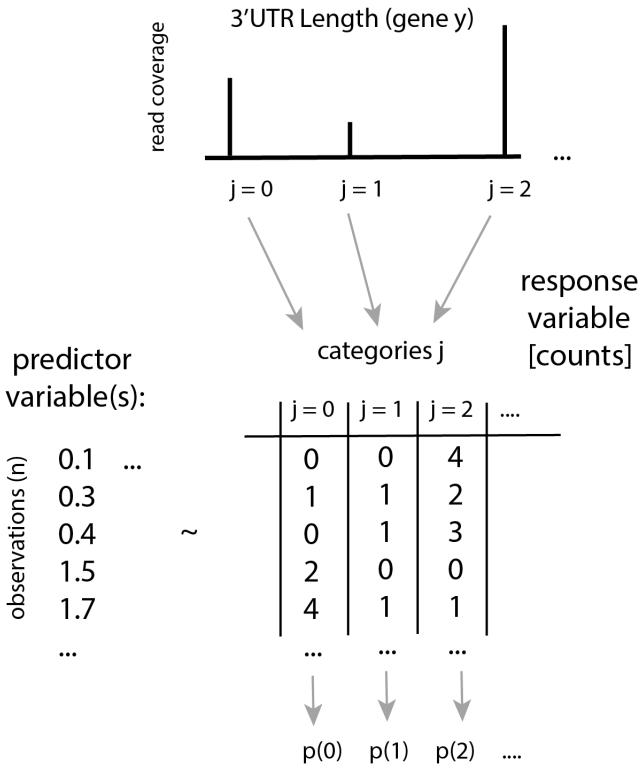


Figure 1.6: Illustration of multinomial regression models, the (fictitious) gene y has three polyadenylation signals. These are treated as categories j having probabilities p(j) to be selected, hence they are multinomially distributed. For n observations these counts (response) can be fitted against predictor variables (in this example, one continues predictor variable).

In single cells, 3'UTR usage is an understudied field. Velten L. et al., 2015 reported that single cells show greater heterogeneity in their PAS selection than expected by chance. A recently developed tool termed *Sierra* aims to call peaks in single-cell sequencing data and compares proximal to distal 3'peaks using Wilcoxon rank sum tests with single cells as statistical units (Patrick, R., et al., 2020, Discussion).

In this doctoral thesis, I will introduce a bioinformatical pipeline developed to call 3'peaks based on single-cell sequencing data from the 10X Genomics platform (Zheng G., 2017). Although mRNAs are captured by oligo(dT)-beads and sequenced from the most 3' position, the default 10X analysis pipeline *cellranger* (Zheng G., 2017) does not provide an output with the exact positional information. This can be considered a missed opportunity since the discrimination between shorter and longer 3'UTRs as well as alternative terminal exons is of high biological relevance (Miura P. et al. 2014). The benefit of using single cell sequencing data over bulk methods in the first place consists in its high resolution; different cell-types can be separated based on their

transcriptomic profiles as well as representations of lineages and trajectories reflecting differentiation processes (as introduced in the last section, Trapnell C. et al., 2014). From the two main approaches, how to analyze single cell sequencing data, namely the feature-centered approach like differential expression and the cell-centered approach like dimensionality reduction and clustering I will focus more on the first one. Doing so, I will be able to study the biology of alternative polyadenylation in the context of neuronal differentiation. In the next section I will explain he specific goals of this work.

1.7 Goals of this study

Based on the findings from Baser A. et al., 2019 that posttranscriptional regulation is crucial for the onset of NSC differentiation we intended to follow this up by focusing our studies on 3'UTR usage in NSCs. The first milestone of my PhD was to developed a bioinformatical pipeline to detect PASs in single cell RNAseq data by 3'peak calling. Next, 3'UTR alterations along the NSC lineage progression can be described applying motif detection and gene set enrichment analyses. One question was whether these genes with 3'UTR changes would show an intersection with risk genes for human diseases. To strengthen such a claim, it would be of interest to detect differential 3'UTR usage in human data by comparing a cohort of patients against a control group.

As reasoned in section 1.1 our model of the murine NSC lineage enables us to study neurogenesis and therefore neurodevelopmental disorders. Subsequently, we aimed of connecting the 3'UTR changes in NSCs and human disease (here: ASD, as we will show later) by a motif in 3'UTR sequences and its respective binding protein. Moreover, we also aimed of showing the impact of 3'UTR changes and the binding protein (here: CPEB4, as will show later) on protein outcome. I will use single cell RNA sequencing data of neural stem cells and neuroblasts collected from Aplp1 knockout mice and wildtype controls to see whether CPE-dependent PAS selection is altered when APLP1 is absent.

In summary, the goal was to link neurodevelopmental disorders, neurogenesis and an RNA binding protein to alternative polyadenylation based on different data types by developing and applying computational methods.

Chapter 2

Methods

This chapter will comprise the development of a bioinformatical workflow designed to call 3' peaks in single cell sequencing data. The goal of this method is to quantify how the usage of these 3'peaks – representing different poly(A) signals and therefore different 3'UTR lengths – changes across various conditions. I developed this method together with Dr. Simon Anders and Dr. Wolfgang Huber. Moreover, I will describe the methods for down-stream analyses (gene set and motif enrichments), differential plus co-expression as well as a comparison of 3'UTR usage to protein outcome.

2.1 Overview of datasets and general analysis strategy

The first milestone for me was to explore the 10X genomics data and to assess whether the exact 3' ends of genes can be detected. If yes, the question was how good the quality of these 3'peaks is (average peak width and positional correlation with the known PAS: AAUAAA). After having established this pipeline, I applied it on four datasets (see also Supplementary Table 1):

1 (wildtype) NSC lineage progression *(data from Kalamakis G. et al. 2019)*
2 NSCs from APLP1-/- vs. wildtype *(data from: Nikhil Oommen George)*
3 Human neurons: ASD vs. controls *(data from: Velmeshev D. et al., 2019)*
4 *in vitro* NSC lineage *(from: Nikhil Oommen George)*

The biological questions were how does APA change with NSC lineage progression (1), is APA different in APLP1-/- compared to wildtype (2) and different in human neurons from ASD patients compared to controls (3). Are the genes altering their 3'UTR usage enriched for specific gene sets and regulatory motifs? Since multiple follow-up assays (like proteomics, RNA immunoprecipitation and ribosomal profiling) were generated from *in vitro* NSCs, I also compared differential 3'UTR usage from active to quiescent NSCs as *in vivo* vs. *in vitro* (4). Other questions were whether polyadenylation factors are differentially expressed (mainly 1 and 3), what is the effect of CPBE4 binding to mRNAs (5) and how does differential 3'UTR usage as well as CPEB4 influence mRNA translation (6 and 7).

5 CPEB4 RNA immunoprecipitation in *in vitro* NSCs
(data from: Ana Domingo Muelas, Alex Bizyn & Rosa Pascual)
6 Proteomics of *in vitro* active and quiescent NSCs
(data from: Nikhil Oommen George & Daria Fijalkowska)
7 Ribosomal profiling *in vitro* NSCs
(data from: Maxim Skabkin & Damian Carvajal Ibanez)

To summarize the general strategy, datasets 1 to 4 were used to establish the 3'peak calling pipeline and to derive biological clues on differential 3'UTR usage. Datasets 5 to 7 were utilized to answer further biological questions as a follow-up.

2.2.1 Mapping of single cell 3'UTR peaks (with long reads)

As a first step, I mapped the reads from the 10X Genomics FASTQ-files to the mouse genome (mm10). Most importantly, the alignments should provide the positional information of transcript ends. I figured that running the mapper *bowtie2* (version 2.3.5.1, Langmead B. & Salzberg, S.L., 2012) with the option *--very-sensitive-local* and in paired-end mode (and not single end as in *cellranger*) would achieve this. In particular, cell barcodes and UMIs will be soft-clipped, the most 3' position of read 1 will represent the most 3' position in the mRNA and read 2 will map downstream to read 1. Figure 2.1.1 depicts this mapping strategy. In order to enable efficient processing of the alignments they were sorted by genomic region using *samtools*.

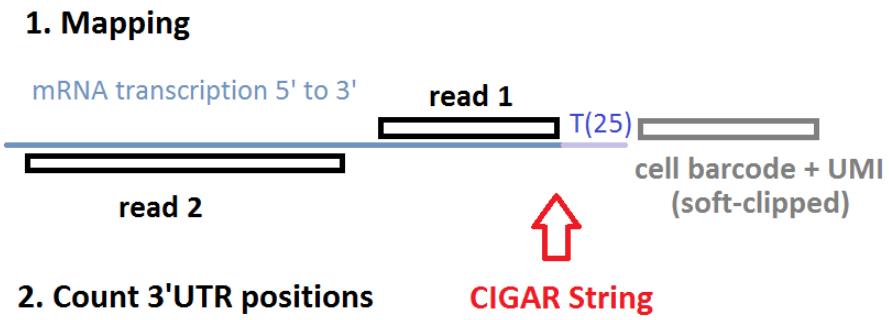


Figure 2.1.1: Mapping strategy for 3'peaks in scRNAseq data. Step 1: mapping reads in paired end mode to the human genome, step 2: count the most 3' position in the CIGAR string of read 1. The blue line represents a genomic region annotated as 3'UTR when the gene is transcribed from the forward strand (5' to 3').

2.2.2 Calling 3'UTR peaks in single cell sequencing data

The following processing steps I carried out in an R/Bioconductor environment. Firstly, I extracted the 3'UTR regions of transcripts from ENSEMBL (*GenomicFeatures* and *TxDb.Mmusculus.UCSC.mm10.knownGene*, Lawrence, M. et al., 2013). Subsequently, I resolved overlapping transcripts so that every entry in the obtained gene annotation stores unique (non-overlapping) genomic coordinates. This was done to avoid counting the same transcript multiple times.

Next, I loaded the alignments (with *GenomicAlignments*) that map to these 3'UTR regions considering only reads reported as primary alignment and mapped as proper pairs. Per region, the most 3' mapping position of read 1 was translated from CIGAR strings to local coordinates in transcripts (Figure 2.1.1). Over duplicated reads that were called by the unique molecular identifier (UMI) median mapping positions were computed. Additionally, the single cell barcode was added to each entry. The reason to process the data like this was to assign single transcript lengths to single cells. To ensure that this pipeline runs at reasonable speed I processed multiple regions at once with *BiocParallel* (Lawrence, M. et al., 2013; Huber W., 2015).

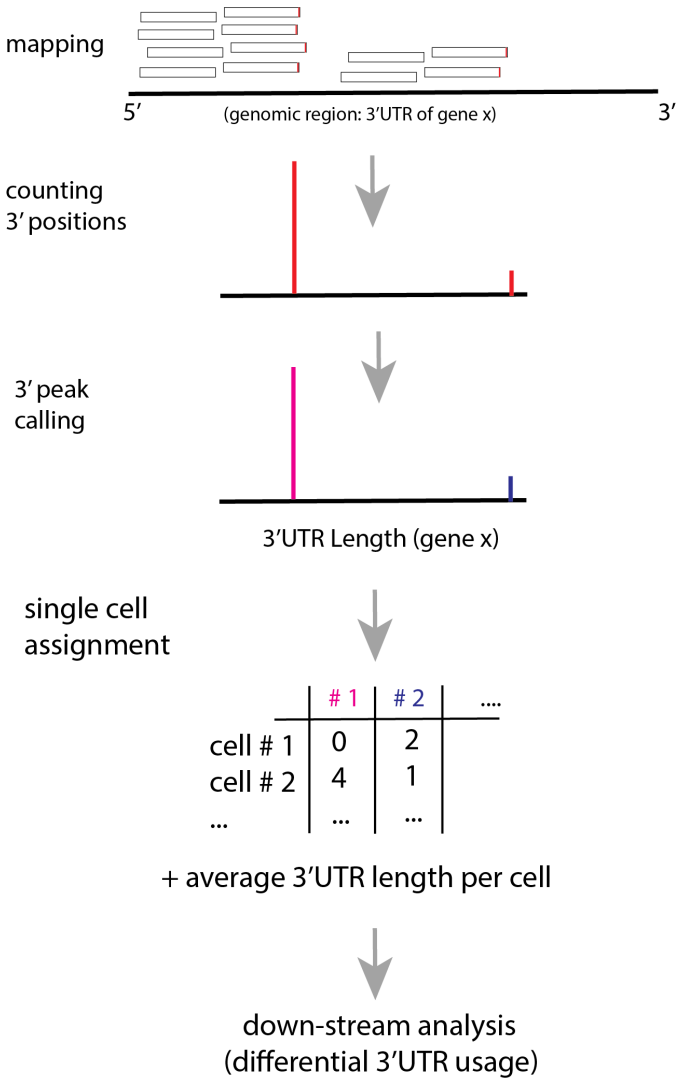


Figure 2.1.1: Workflow: calling 3'peaks in single cell sequencing data. After genomic mapping (shown in more detail in the previous Figure 2.1.1), for each

gene: most 3’ positions of reads are counted and assigned with respect to 3’peaks and single cells.

The next objective was to call 3’UTR peaks. For such tasks common peak calling algorithms like MACS2 (Zhang, Y., et al., 2008) were applied before. Due to the low mapping background (detection of sharp 3’UTR peaks) I decided to implement peak calling by clustering the non-zero positions, *i.e.* UMI counts of 3’UTR ends, (*R* functions *hclust* and *cutree*). A 3’tail peak was called if it was supported by at least 3 UMIs and had a distance to other peaks of at least 50 bp considering a mean 3’peak width of around 20 to 30 bp.

As a validation for the 3’ mapping, the distance of peak centers to the closest canonical poly(A) signal (AAUAAA) in the 3’UTR reference sequence was computed. The results (single cell 3’ mapping positions: count table and read coverage) were stored in a gene ordered list (S4 object structure). Moreover, the output inherits the single cell annotation as well as the gene annotation like the genomic coordinates (*GRanges* object, see Lawrence, M. et al., 2013) and sequence information of 3’UTRs. This structure enables easy-to-implement ‘per-gene’-downstream analyses utilizing R’s apply functionality as every entry contains the information for one gene and an index matching the single cell annotation. Another benefit is the sparsity of the object since empty entries are not contained in the object. A classical count matrix if needed can be easily reconstructed.

2.2.3 Inferring NSC lineage progression as pseudotime

As a next step in this project, a reversed graph embedding algorithm was applied to single cell RNAseq (Trapnell C. et al., 2014) data in order to obtain a latent variable that resembles NSCs lineage progression *in silico*. A commonly applied implementation for this task is available in the R package *monocle2* (Trapnell C. et al., 2014). In this work, pseudotime was computed following *monocle2*'s standard workflow. Genes used as input for the reversed graph embedding algorithm were selected by choosing those with the highest dispersion (variance) in the respective dataset. Importantly, pseudotime was computed separately for each sequencing experiment: *in-vivo* NSCs (young & old), *in-vitro* NSCs (EGF & BMP4) and APLP1 (WT & APLP1-/-). As a sanity check, marker genes for the different NSC states were explored (Kalamakis G. et al., 2019). The result of this method was already published in Kalamakis G. et al., 2019.

2.2.4 Correlation analysis of single cell 3'UTR usage with NSC lineage progression

As a first approach to see how 3'UTR lengths change with NSC lineage progression, we considered a correlation analysis. Explicitly, Pearson's correlation test and coefficient were computed inputting for every gene the variables pseudotime *t* (previous section) and average 3'UTR lengths *l* of single cells *i* denoted in (I):

$$\text{(I) } \rho = corr(t_n, l_n) \text{ with single cells being observations } i = 1, 2, \dots n$$

Extreme correlation coefficients will indicate shortening or lengthening of 3'UTRs with NSC lineage progression. As pseudotime *t* per definition increases with maturation (from qNSCs to NBs), negative coefficients correspond to shortening and positive ones to lengthening.

2.2.5.1 Inference of differential 3'UTR usage along lineage progression with MNR

As a second approach we applied multinomial regression on the 3'UTR mapping results. Here, my aim was to quantify how the fractional usage of poly(A) signals change with NSC lineage progression (quantified as pseudotime variable *t* in 2.23). For every gene detected with two or more 3'peaks, a matrix can be defined as:

(I) $U_{(t,j)} \in \mathbb{Z}^{0,+}$ given as [UMI counts]

In (I) every row represents a single cell assigned a unique pseudotime value *t* and non-zero expression of the given gene. Column entries *j* correspond to 3'peaks ordered from most proximal to most distal. Since these data is multinomially distributed we define *U* as the (polytomous) response and pseudotime as the predictor variable hypothesizing that the choice which PAS is used depends on the NSC maturation state (*t*). The most proximal 3'peak will be treated as reference category (*j = 0*) from *C* possible categories in total and probabilities π as:

(II) $\pi_i^{(0)}, \pi_i^{(1)}, \dots \pi_i^{(C-1)}$ with single cells being observations $i = 1, 2, \dots n$

In order to describe local changes in log-odd-ratios between the 3'peaks (categories *j*) a logit-function can be modelled with beta-parameters (*β)* learned by fitting the model in (III, see Agresti A., 2002; see also Venables W.N. & Ripley B.D., 2002; Bohning D.,1992; Begg C.B. & Gray R., 1984):

(III) $$\log\left(\frac{\pi_i^{(j)}}{\pi_i^{(0)}}\right) = \alpha^{(j)} + \beta_1^{(j)} U_1 + \cdots + \beta_t^{(j)} U_{ti}$$

Since it is of interest to model the trend of 3'UTR usage over *t* as an approximation of 3'UTR usage, beta-parameters should be estimated as a basis spline (Hastie T.J., 1992) matrix with three degrees of freedom (dof) as a total of *3* times *(C-1)* beta-parameters:

(IV) $\beta \rightarrow \beta' \xrightarrow{yields}$

$$\log\left(\frac{\pi_i^{(j)}}{\pi_i^{(0)}}\right) = \alpha'^{(j)} + \beta_1'^{(j)}\varphi(U,t) + \beta_2'^{(j)}\varphi(U,t) + \beta_3'^{(j)}\varphi(U,t)$$

with $\varphi()$ as B-spline kernel function and thus a matrix of beta-parameters:

$$\mathrm{B}_{dof,j} = \begin{bmatrix} \alpha'^{0} & \cdots & \beta_3'^{0} \\ \vdots & \ddots & \vdots \\ \alpha'^{(C-1)} & \cdots & \beta_3'^{(C-1)} \end{bmatrix}$$

To obtain curves depicting for each 3'peak (categories *j*) their fractional usage the B-splined pseudotime was multiplied by the transposed beta-parameter matrix from (IV). Next, the log-transformation was removed applying the exponential function to the result:

$$\text{(V) } \mathrm{T}_{n,j*} = \varphi(t,\ dof = 3) \quad with\ \mathrm{T}_{n,j_0} = 1\ (constant) \quad \text{obtain: } exp(\mathrm{T}\ x\ \mathrm{B}^T)$$

In terms of implementation, the R package *nnet* was deployed as it is the recommended software for multinomial regression models (Venables W.N. & Ripley B.D., 2002). As an alternative I also considered the *VGAM* package which gave comparable results (not shown). In *nnet*, beta-parameters are learned using feed-forward neural network with a single hidden layer and in this case UMI counts are treated as weights for each observation (Venables W.N. & Ripley B.D., 2002).

The dependence of pseudotime (t) and 3'UTR choice can be further assessed by log-likelihood-ratio tests (LRTs). In general, LRTs (Agresti A., 2002) will report high values for the test statistic if the inclusion of the critical predictor variable (here pseudotime *t*) yields model parameters for the full model that strongly deviate from the reduced model. Formally, this can be expressed in a likelihood chi-squared statistic (Agresti A., 2002) as:

$$\text{(VI) } G^2 = -2log\Lambda \quad (where\ \Lambda \text{ represents the deviation between both models})$$

To this end, the model was fitted as in IV and compared to a reduced model with a constant predictor (VII) utilizing the corresponding S3 method in R: *anova.multinom*.

$$\text{(VII)} \quad MNR(\ U_{(t,j)} \sim \varphi(t,\ dof = 3)\ ) \qquad \text{as full model}$$
$$MNR(\ U_{(t,j)} \sim 1\ ) \qquad \text{as reduced model}$$

On the one hand LR statistics are useful to evaluate whether alterations in APA usage with pseudotime are supported by the raw data, on the other hand these statistics may not represent a classical p-values since observations (single cells) are not independent of each other.

2.2.5.2 Inference of differential 3'UTR usage between genotypes with MNR

Apart from 3'UTR alterations in NSC lineage progression it is also of interest to quantify differences in 3'UTR usage between two genotypes. In this work, we will address the hypothesis that 3'UTRs are altered in APLP1-/- compared to wildtype. Given matrix $U_{(t,j)}$ (I, as the previous section) for both genotypes: wildtype and APLP1-/- plus the assignment of cells to biological replicates (next section), we may consider summing UMIs over biological replicates for each 3'peak:

(VIII) $U_{(t,j)} \rightarrow U_{(s,j)}$ $\quad s = \text{observation per genotype and replicate}$

Doing so will facilitate the computation of (classical) p-values based on independent statistical units. However, we assume that pseudotime *t* is equally represented in both genotypes which we can base on the fact that both samples were FACS sorted the same way. With multinomial regression we can notate this problem as follows:

(IX) $MNR(U_{(s,j)} \sim Genotype)$ as fully model

$MNR(U_{(s,j)} \sim 1)$ as reduced model

As a metric of differential 3'UTR usage between APLP1-/- and wildtype, the Earth Mover's distance (EMD, Levina E. & Bickel P., 2001) was defined as follows:

$$(X) \qquad \Delta_{3'UTR,=} \frac{U_{WT(j)}}{\sum_{i=0}^{C-1} U_{WT(j)}} - \frac{U_{APLP1-/-(j)}}{\sum_{i=0}^{C-1} U_{APLP1-/-(j)}}$$

$$(XI) \qquad EMD = \sum_{i=0}^{C-1} \left(CUMSUM(\Delta_{3'UTR})\right)$$

$U_{(j)} \sim all\ UMIs\ summed\ over\ single\ cells\ per\ 3'UTR\ peak\ j\ for\ one\ genotype$

As an alternative to the Earth Mover's distance, Agresti's generalized odds ratio (Agresti A., 1980) could be considered, a solution to quantify odds ratios if there are more than 2 categories ($j > 2$). Here, EMD was favored as we considered it to be more intuitive: it can be thought of the cost to shift the 3'UTR distribution observed in APLP1-/- to the one of the WT. Phrased in other words: EMD is the difference between the probability of mass function (PMF) of PAS counts between both genotypes. As defined in (*X*) indices *j* of PAS go from proximal to distal and the APLP1-/- distribution

in PAS is subtracted from WT one, hence EMD < 0, the gene is shorter in APLP1-/- vs. WT and EMD > 0, the gene is longer in APLP1-/- vs. WT. If one or more middle PASs show alterations between the genotypes the gene will get a low p-value by MNR and EMD ≈ 0.

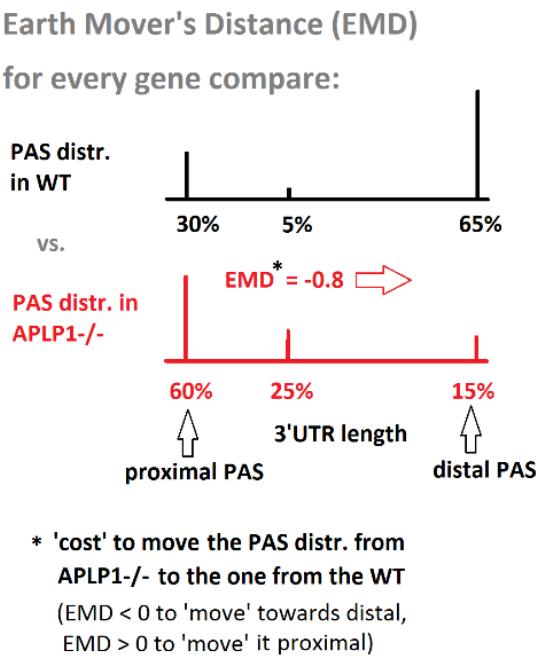


Figure 2.2: Earth Mover's Distance applied on 3'peak distributions, schematic depiction, comparison of 3'UTR usage between two genotypes.

2.2.5.3 Assignment of single cells to biological replicates (cell-hashing)

For the APLP1-/- sequencing experiment, single cells were assigned to biological replicates (mice) using hash-tag-oligos (HTOs), antibodies that bind to cell surface proteins (MHC1 and CD45) and are tagged with a small DNA sequence that can be captured by oligo-d(T) primers. Per genotype (APLP1-/- and WT control) n = 3 hash-tag-oligos were used (meaning every cell belonging to one mouse will have the same hash-tag-oligo sequence). Hash-tag-oligo sequences were detected in read 2 (*ShortRead* package) allowing one mismatch to the reference hash-tag-oligo sequences. The single cell barcodes from read 1 were compared to those from the 10X genomics output. Each cell was assigned to an individual mouse (biological replicate) in case the cell had an over-representation of one hash-tag that was higher than 1.5-fold of the median hash-tag count detected within one cell (overrepresentation of one hash-tag-oligo).

2.2.6.1 Preprocessing and 3'UTR mapping positions in single cell sequencing data of human neurons

A central hypothesis of this work was the question whether 3'UTR usage differs between ASD patients and controls. Of note, 3'UTR lengthening in ASD vs. controls was already suggested by Szkop K. J. et al., 2017 based on bulk RNAseq data. Here, I reanalyzed the single cell sequencing data from Velmeshev D. et al., 2019. In this study, samples were taken post-mortem from individuals diagnosed with ASD and control group (Figure 2.x). These samples were extracted from two locations within the brain: the prefrontal cortex (PFC) and the anterior cingulate cortex (ACC). Importantly, samples were processed with a NucSeq protocol, hence all observations will reflect changes in nuclear mRNAs.

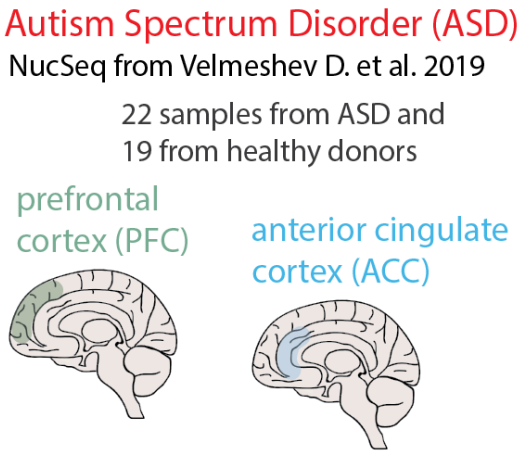


Figure 2.3: Experimental layout of the study from Velmeshev D. et al., 2019 comparing ASD to controls, schematic depiction. The presentation of the n = 41 samples is: n = 22 (ASD), n = 19 (Control), n = 23 (PFC), n = 18 (ACC), n = 32 (male), n = 9 (female), average age = 14.6 years (15 in ASD and 14.1 in Control).

To start the analysis for 3'UTR usage, I downloaded the FASTQ-files from the gene expression omnibus (accession number: PRJNA434002) and subsequently processed with the 10X genomics pipeline (*cellranger count*, version 2.2.0). Reads were mapped against the human genome hg38 (GRCh38). In contrast to the 3'UTR mapping strategy described in this chapter before, the human scRNAseq data (Velmeshev D. et al., 2019) was sequenced with a shorter read 1 (c.f. Figure 2.2) providing the information of cell barcode and UMI but not about exact 3'ends. For this reason, the most 3' position from

read 2 instead of read 1 was calculated. Gene annotations (3'UTR regions) were derived from ENSEMBL as for the mouse data (*TxDb.Hsapiens.UCSC.hg38.knownGene*). As this method will give broader 3'peaks compared to the paired-end mapping strategy (c.f. Figure 2.2) I decided not to call 3'peaks, but instead use average 3'UTR lengths for further down-stream analyses (see also Discussion of this chapter).

2.2.6.2 Differential 3'UTR usage in in single cell sequencing data comparing ASD diagnosed patients to controls

Firstly, an average or meta-gene 3'UTR length $L_{(p,ct)}$ was computed as in (I) for every combination of patients (p) and cell-types (ct), respectively. The cell-type annotation was adopted from the original paper (Velmeshev D. et al., 2019). The detected cell-types comprise a variety of neurons (N), interneurons (IN), glia-cells and potential progenitor cells (Neu-mat).

$$(I) \qquad L_{(p,ct)} = \frac{1}{m}\sum_{g=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} l_{(p,ct)}\right)$$

l = 3'UTR length p = patient ct = cell-type,
genes g = 1, 2, 3, … m single cells i = 1, 2, 3 … n

Multiple t-tests were calculated on the result comparing for every cell-type meta-gene 3'UTR lengths $L_{(p,ct)}$ between ASD and control samples. In detail, the R function *pairwise.t.test* was utilized for this purpose (two-sided, Benjamin-Hochberg correction for multiple testing and assuming that standard deviations are not equal with: *pool.sd = FALSE*). Next, per gene, linear models were fitted in II with the predictor variables *Location* (PFC or ACC), *Sex* (male or female) and *Diagnosis* (ASD or control). This was done to block for confounding effects of these variables, for instance if the gene is missing (not captured) in some samples and it would appear different between ASD and control although the difference would be due to an unequal representations of sexes or brain regions (see also Discussion of this section). The test statistics for II (ANOVA on linear model) will be high if 3'UTR length differs between ASD and control, consistently across samples.

$$(II) \qquad Length_{(3'UTR)} = \frac{1}{n}\sum_{i=1}^{n} l_{(p,ct)}$$

$LM(\ Length_{(3'UTR)} \sim Location + Sex + Diagnosis\ )$ as fully model

$LM(\ Length_{(3'UTR)} \sim Location + Sex\ )$ as reduced model

2.3.1 Gene set enrichment analysis for differential 3'UTR usage

As a biological interpretation of alterations in 3'UTR length, I computed gene set enrichment analyses based on the statistics to measure differential 3'UTR (introduced in the previous sections). Gene set enrichment analysis (GSEA) can assess whether reference gene sets are over-represented at the top or bottom of a ranked gene list (Yu G., 2012; Schriml L. M. et al., 2019). In this work I used the gene annotations from Gene Ontology (GO), Disease Ontology (DO), Disease Gene Network (DGN) and the Reactome Pathway database. For this, purpose I used the implementation *ClusterProfiler* and its extensions *DOSE* and *ReactomePA* (Yu G., 2012). Significance of enrichments is computed by permutation tests according to the GSEA method with p-value adjustment for multiple testing (*i.e.* multiple categories). A normalized enrichment score (NES) is reported since this value includes the mean of all permutations and accounts for correlations to other gene sets (Yu G., 2012). Input gene lists were ranked either by the multinomial regression statistics (multinomial regression statistics representing evidence for 3'UTR changes) as well as 3'UTR lengthening trends for the human data (linear models). For Disease Ontology (DO), Disease Gene Network (DGN), mouse gene identifiers were translated to human orthologues using the *getLDS* function from the *biomRt* package. For the translation, duplicated gene identifiers were removed, non-translatable ones not considered and all multi-match identifiers were used. This was done to estimate whether 3'UTR changes affect human disease risk genes reported in previous studies.

2.3.2 Motif enrichment analysis applied on differential 3'UTR usage

Apart from gene sets, regulatory elements (and their motifs) in 3'UTRs are of interest as they can impact posttranscriptional regulation of mRNAs. I performed motif analysis utilizing the tool homer2 (v4.9) for de-novo motif detection (Heinz S. et al., 2010). This method detects the overrepresentation of k-mers in a list of foreground sequences over background sequences and subsequently combines enriched k-mers to optimize a motif probability matrix (Heinz S. et al., 2010). For the mouse data, I used sequences flanking the 3'peaks from genes with only one 3'peak (constant 3'UTR) as background. As foreground, 3'peaks from APA that show clear variation in the UMI fraction with pseudotime were considered, irrespective whether they increase or decrease. For the human data, the top 500 lengthening (longer in ASD vs. control) to the top 500 shortening genes were compared (ranked by p-value from the linear models). For this comparison the distal regions in 3'UTRs (most 3' 250 bp in the 3'UTR) were used as input. To describe the effect of CPEB4 binding on 3'UTR length choices, a CPE motif (as regular expression: TTTTGT|TTTTGAT|TTTTAGT) was used if the 3'peak had at least one occurrence of this motif in a distance of 1 to 50 bp upstream (in its 3'UTR sequence) it was classified as CPE flanking 3'peak. The same motif was used to cluster the 3'UTRs in human genes, whether they contain this motif in their distal part (most 3' 250 bp of a 3'UTR), in a proximal/middle part or not all. A Chi-Square test (Agresti A., 2002) was used to determine whether the location of the CPE motif differs between the 3'UTRs that get shorter and those get longer in ASD vs. controls.

2.4 Analysis of differential expression and co-expression

Based on the observation that APA is different between ASD patients and controls it was of interest to see whether polyadenylation factors are differentially expressed. To this end, UMIs were summed per gene over single cells with respect to samples and cell-types in the human single cell RNA sequencing data from Velmeshev D. et al., 2019. This pseudo-bulk approach was chosen in order to avoid pseudo-replication as the intention of this analysis is to estimate the biological variation across the ASD and the control cohort (see also the Discussion section of this chapter). Per cell-type, *DESeq2* was fitted with the model: $\sim \text{Location} + \text{Sex} + \text{Diagnosis}$ to estimate p-values and log2-fold-changes between ASD and controls. Dispersion was fitted using the "local" option and *DESeq2's* default FDR correction was used. In addition, co-expression of genes was computed over single cells using Pearson's method. These correlation coefficients were estimated for each sample interpedently to see whether they are reproducible across the individuals of the cohort.

For the APLP1-/- sequencing in mice, differential expression was assessed in a similar way as for the human data, summing over cells assigned by hash-tags-oligos (as shown in the previous chapter). Accordingly, the DESeq2 model was: $\sim \text{Genotype}$.

Another question was how the gene expression patterns of cultured NSCs compare to freshly isolated NSCs. To this end, single cells from both systems were matched by (randomly) subsampling *in-vitro* NSCs. This way, every *in-vivo* cell has an *in-vitro* counterpart in pseudotime. For the gene-wise correlation tests (Pearson's moment correlation, one-sided), the expression of genes between pseudotime ordered cells (*in-vivo* vs. *in-vitro*) were computed. Importantly, pseudotime was computed independently for *in vivo* and *in vitro* to demonstrate reproducibility. In order to visualize the result, dimensionality reduction (UMAP) was performed on an integrated data set. Both datasets were inputted into the single cell batch correction algorithm mutual nearest neighbors (*MNN*, Haghverdi L. et al., 2018) for single cell batch correction. Distances of single cells in expression space were visualized with the *sleepwalk* tool (Ovchinnikova S. & Anders S., 2020) as Euclidean distances. The purpose of this was to see whether *in vitro* qNSCs are more similar to *in vivo* qNSCs than to *in vivo* aNSCs.

2.5 Analysis of CPEB4-RNA immunoprecipitation results

To assess the binding of CPEB4 to mRNAs, a CPEB4-RNA immunoprecipitation (RIP) from cultured NSCs was analyzed as follows: Reads were mapped to the mm10 genome (single reads, 50 bp) with the *bowtie* mapper and counted from BAM files applying the R/Bioconductor (Huber W. et al., 2015) workflow (function *SummarizeOverlaps* with mode 'Union'). Duplicated reads were removed. *DESeq2* was applied to compute the log2-fold-changes (LFCs) between the IP fraction and controls. Lowly expressed genes (less than 25 counts) on average were excluded. The dispersion trend as fitted in the mode 'local' and *DESeq*'s default FDR correction was applied. In this analysis CPEB4-IP, immunoglobulin (IgG) and input fractions were used (for each fraction: n = 2). The LFCs were interpreted as the affinity of CPEB4 to mRNAs: the higher the LFC for a gene, the higher its affinity/binding to CPEB4.

As a validation that CPEB4 can bind the CPE-like motif (TTTTGT, TTTTGAT, TTTTAGT) the positional information in this data was used in the following manner: per 3'UTR region of a given gene a generalized linear model (GLM) with the "poisson" family distribution was fitted to the read coverage vectors as:

$$\text{(I)} \qquad GLM\left(v_{(IP)} \sim v_{(KO)} + v_{(IgG)} \right)$$

$$v \in \mathbb{Z}^{0,+} \text{ given as read coverage vector (n = 2 for each condition)}$$

From these fits (one fit per expressed gene) the residuals were used for further down-stream analysis. The motivation why to analyze the RIP data like this can be illustrated by the following example (Figure 2.1). The residuals will be positive at a given position in the 3'UTR if the read coverage is relatively higher in the IP fraction than in the controls (here IgG and CPEB4-/- IP fractions). Next, the strategy is to observe these residuals in defined windows centered around the CPE-like motif (its occurrence in 3'UTR sequences). Averaged over the transcriptome, the residuals will be positive around the motif. As statistical control, to show that this is not an artifact coming from a bias in read coverage vectors, a random position instead of the CPE motif can be used. If the signal (average residuals) is positive for the CPE motif but not for a random motif we will interpret that CPEB4 can bind the motif.

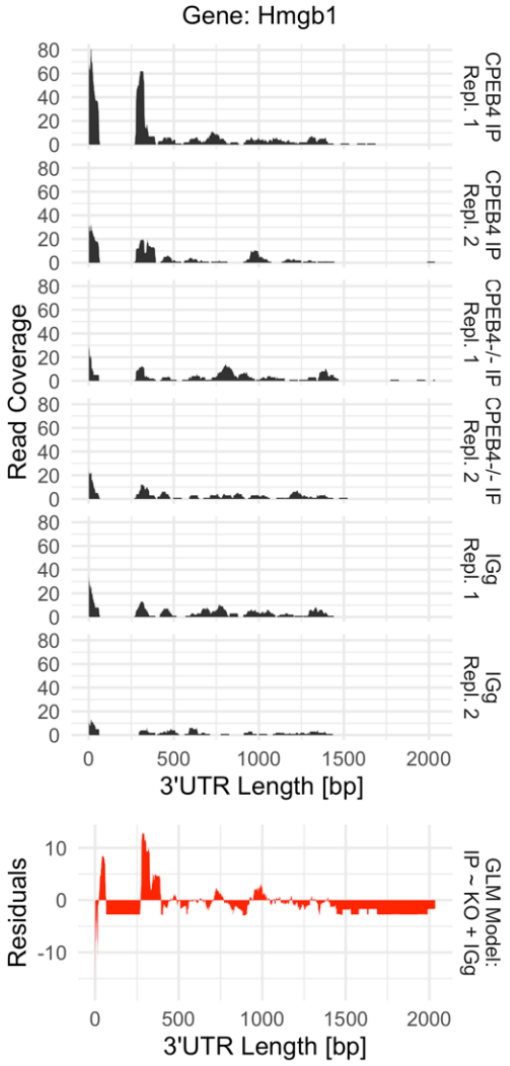


Figure 2.4: Deriving positional information of CPEB4 binding based on RNA immunoprecipitation (RIP), read coverage in 3'UTRs of example gene Hmgb1, RIP samples (in black, upper panels), fitting a GLM model yields residual read coverage (in red, lower panel) representing the signal from the RIP assay.

2.6 Analysis of proteomics (in cultured NSCs)

Proteins in cultured NSCs were isolated by Nikhil Oommen George and quantified as well as analyzed for differential abundance in NSCs treated with EGF (active) and BMP4 (quiescent) by Daria Fijalkowska using the tool Perseus45 (version 1.5.3.0, Tyanova S. et al., 2016). Subsequently, I used these proteomic results in order to estimate posttranscriptional regulation. To this end, I combined the protein mRNA expression levels into a single value referred as translation index (TI). For each

condition (EGF, BMP4), LFQ values were divided by summed UMIs and subsequently logarithmized and centered around 0. In order to show the impact of 3'UTR shortening on protein outcome, translation index values were averaged over genes getting shorter and non-shortening genes whenever the change in protein abundance was significant (FDR < 5%). These average TI values were fitted in a linear model to estimate the biological variation in the proteomics samples given as triplicates: $\sim$ Treatment + UTR_Trend + Treatment:UTR_Trend with treatment (EGF, BMP4) and UTR_Trend (gets shorter, no effect) and an ANOVA test applied to the interaction term Treatment:UTR_Trend. The biological variation in mRNA levels could not be included since the single RNA sequencing samples contained one pool of three mice for each treatment (EGF and BMP4) but no hash-tag-oligos to separate the replicates.

2.7 Ribosomal profiling

Ribosomal profiling was assayed by Dr. Maxim Skabkin and Damian Carvajal Ibanez. Briefly, ribosome protected RNA (Faye M.D. et al., 2014) as well as total RNA fractions were extracted from cultured (active) NSCs. I processed the data as follows: Reads were trimmed applying the tool *TrimGalore* version 0.4.4_dev. The adaptor sequence 'AGATCGGAAGAGC' (Illumina TruSeq, Sanger iPCR; auto-detected) as well as 3 bp from the 5'end of and 15 bp from the 3'end were removed. In addition, sequences that became shorter than 18 bp (after quality trimming) were removed using *TrimGalore*'s default settings. After trimming reads had a length of 33 bp on average. Subsequently, reads were mapped to the mm10 transcriptome build GRC38.93 from ENSEMBL using bowtie version 0.12.7 with its standard options. Reads falling into genes were counted from BAM files applying a suitable R/Bioconductor workflow (using the function *SummarizeOverlaps* with mode 'Union'). Duplicated reads were removed. For ribosomal profiling samples, reads in the coding part of genes (CDS) were counted, for total RNA samples reads in the whole gene body were counted. Next, to get an estimation of the translation efficiency (TE) per gene, log-fold-changes between ribosome protected reads and total RNA samples were computed applying *DESeq2*. The higher the log-fold-changes the higher a gene is translated. I interpreted this as a translation efficiency over the transcriptome.

Chapter 3

Results

In this chapter I will demonstrate that sharp 3'peaks can be obtained in single cell sequencing data utilizing the introduced 3'peak calling pipeline. Next, I will show how 3'UTR usage changes throughout NSC lineage progression, in APLP1-/- vs. wildtype and in ASD vs. controls. These results I will further use as input for a motif analysis and link 3'UTR changes in murine neurogenesis and human ASD to the CPE-motif. I will analyze the trends of polyadenylation signals flanked by the CPE-motif in the NSC lineage and in APLP1-/-. Analyzing an RNA immunoprecipitation suggests that CPEB4 can bind this CPE element (its consensus motif). Eventually, comparing the outcome of proteomics and ribosomal profiling to 3'UTR changes and CPEB4 binding to mRNAs, I show that genes undergoing 3'UTR shortening tend to produce more protein and that genes with a high affinity for CPEB4 tend to have a high translation efficiency. These analyses were concepted and developed together with Dr. Simon Anders and Dr. Wolfgang Huber.

3.1 Differential 3'UTR usage

In this section I will show the mapping results for 3'peaks in single cell sequencing data (human and mouse), quantify differential 3'UTR usage and apply gene set as well as motif enrichment analyses on the results.

3.1.1 Detection of 3' peaks in single cell sequencing data

Applying the developed pipeline to call 3'UTR peaks on 10X genomics data yielded sharp 3'peaks with an average width of 20 to 30 bp. Over two-third of genes were detected with multiple 3'peaks (APA genes). As illustrative example of the mapping results, Figure 3.1.1 depicts the 3'peaks detected in the gene Pea15a.

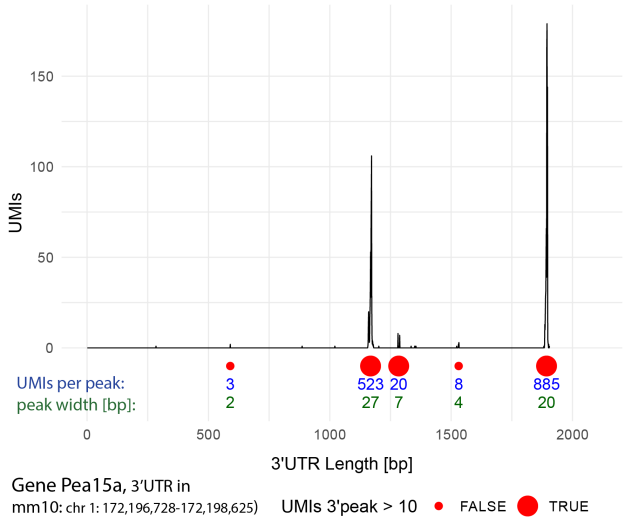


Figure 3.1.1: Mapping example gene for 3'peaks in scRNAseq data. Gene Pea15a, y-axis represents stacked UMI counts (over single cells), x-axis shows the local mapping position (3'UTR length of this gene), peak annotations added (from the implemented algorithm), 3'peaks with more than 10 UMIs considered for further downstream analysis.

On the transcriptome-wide level I observed agreement between the 3'peaks detected by the devloped pipeline and gene annotations. Figure 3.1.2a depicts the anticipated positional correlation as an average distance of around 20 bp from PAS to 3'UTR end (Miura P. et al. 2014). In addition, annotated distal 3'peaks agree with the 3'ends from Ensembl (Figure 3.1.2b).

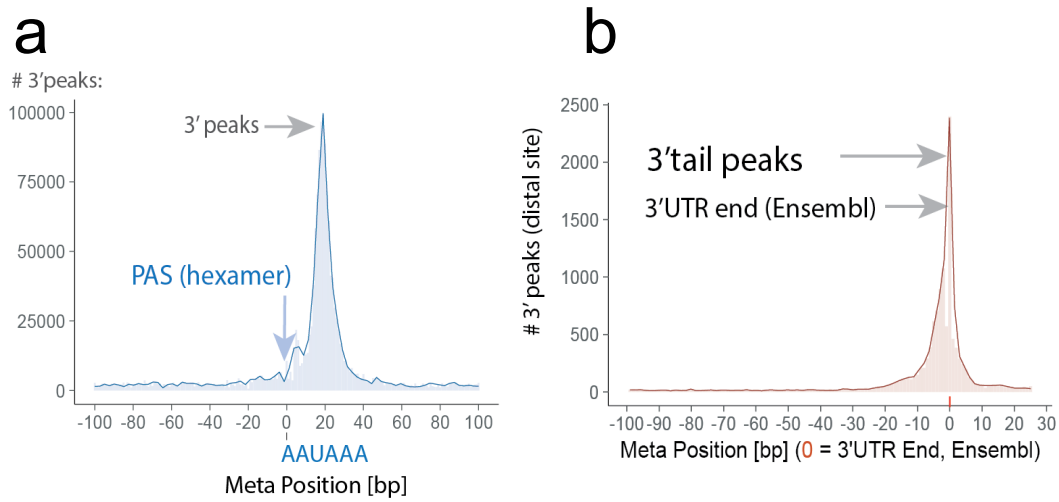
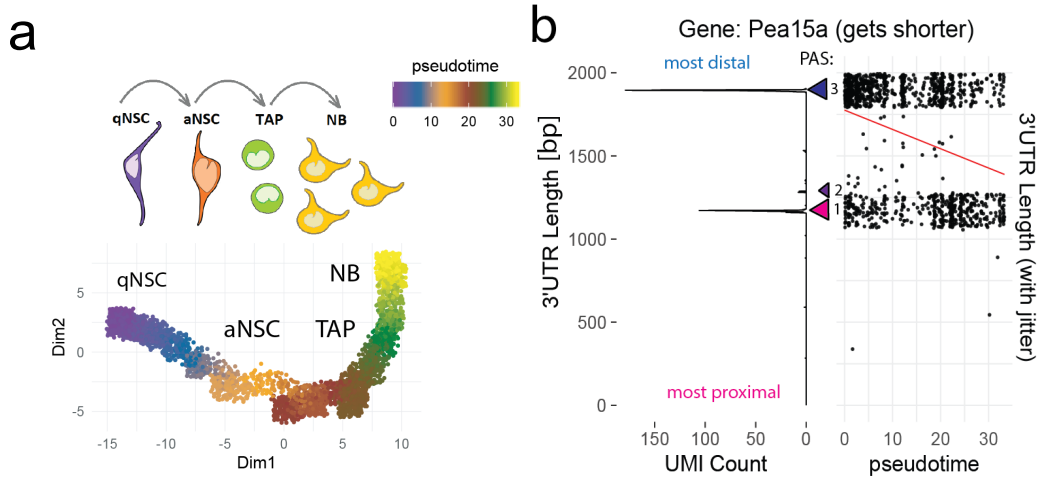


Figure 3.1.2: Annotated 3'peaks correlate with ENSEMBL gene annotations and the canonical PAS. **a**, Relative position of most distal 3' peak centers annotated with the applied peak calling algorithm to 3'UTR ends from the ENSEMBL database (at pos. 0), shown as meta-gene analysis i.e. summed for expressed genes, position < 0 upstream of 3'UTR ends, positive position > 0 downstream. **b**, Density profile of 3' mapping positions, relative to the AAUAAA hexamer (PAS), shown as meta-gene analysis *i.e.* summed for expressed genes.

3.1.2 Correlation of 3'UTR length and NSC lineage progression

To illustrate the analysis, Figure 3.2.1 shows the correlation of 3'UTR length vs. pseudotime for an example gene, Pea15a (same as in Figure 3.1.1) which gets shorter with NSC lineage progression (data from Kalamakis G. et al., 2019).



(Figure caption on the next page)

Figure 3.2.1: Example gene Pea15a getting shorter with lineage progression,

a, ordering cells in pseudotime (purple/blue: qNSCs, orange/brown: aNSCs, green: TAPs, yellow: neuroblasts) representing lineage progression, upper subpanel: schematic representation of the lineage, lower sub-panel: two-dimensional representation created with *monocle2*. **b,** gene Pea15a, left sub-panel: UMI counts and their respective 3' mapping positions. (summed over cells), right sub-panel: mean mapping pos. per cell (y axis) versus pseudotime (NSC lineage progression, x axis), linear regression line (in red).

Computing these correlations for over 8,000 (APA) genes reveals roughly 800 genes (Figure 3.2.2a) that undergo either lengthening or shortening with NSC lineage progression. These correlations were reproducible in the two samples of the NSC lineage from Kalamakis G. et al., 2019 (Figure 3.2.2b). (More gene examples can be found in Supplementary Figure 1.)

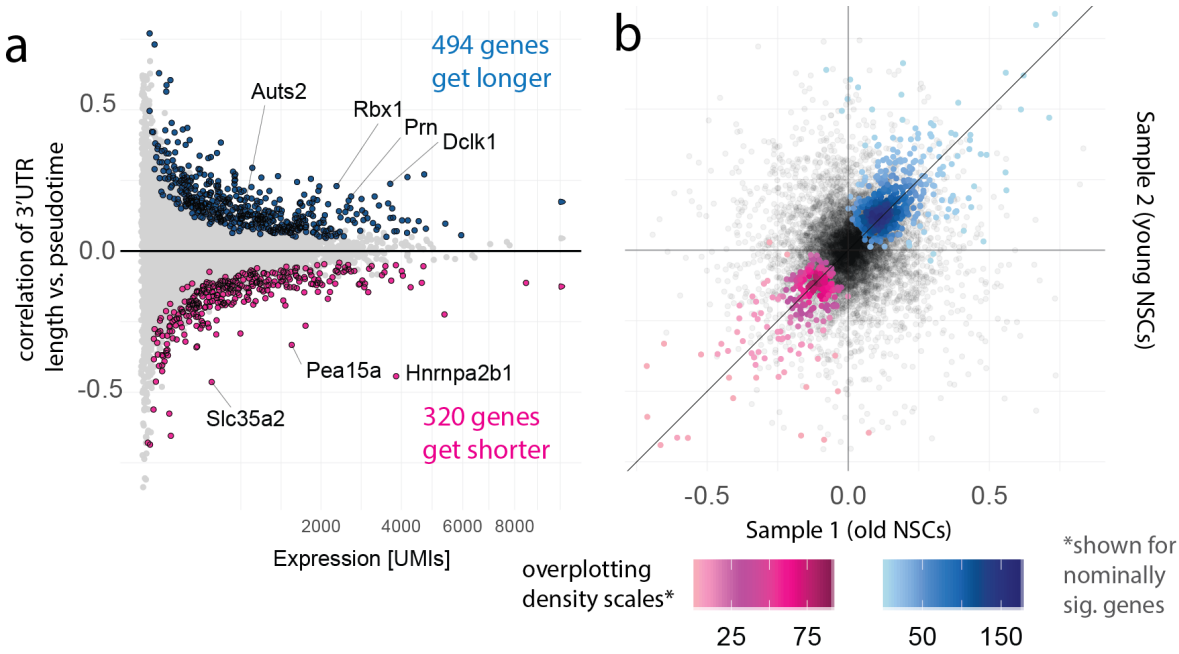


Figure 3.2.2: Several hundred of genes become longer or shorter with NSC lineage progression. For every gene with multiple PAS (APA genes): **a**, summed expression (x-axis) against correlation of mean 3'UTR length per cell vs. pseudotime (y-axis), 3'UTR shortening or lengthening marked as magenta and blue points respectively. **b**, Correlations as in section 2.23 (Figure 8), computed independently for both biological replicates as scatter plot, colored points indicate nominally sig. genes, separate color scales for shortening (magenta) and lengthening (blue).

3.1.3 Differential 3'UTR usage along the NSC lineage progression multinomial regression and gene set enrichment analysis

As an example, how the fractional usage for the PASs in the 3'UTR of Pea15a changes over pseudotime, I binned its PASs over pseudotime (Figure 3.3.1). The approximation using the multinomial spline regression approach (dof = 3) is shown in the lower panel. The relative usage of the proximal PAS increases over pseudotime (from qNSC to NBs) from roughly 20% to 70%.

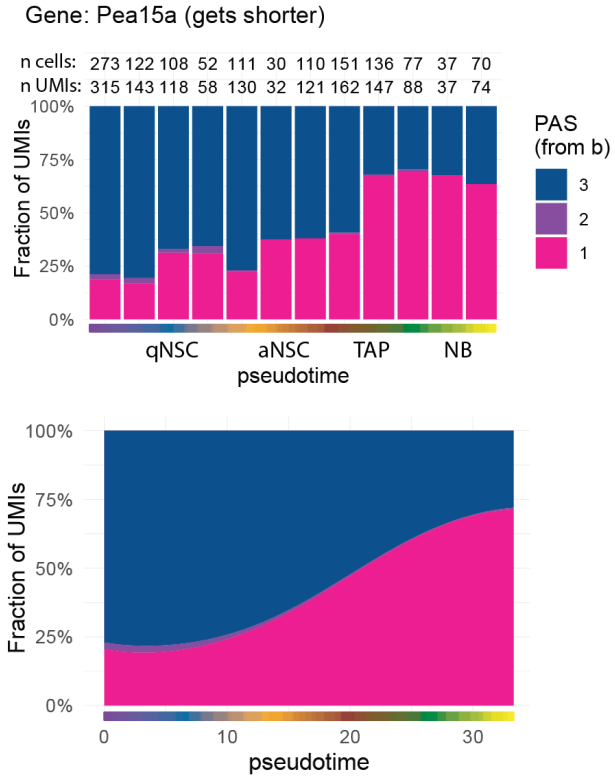


Figure 3.3.1: Multinomial splines approximate 3'UTR usage over pseudotime, example gene Pea15a, upper panel: fraction of different PASs per pseudotime window for Pea15a (blue = distal PAS, magenta = proximal PAS), lower panel: approximation of changes in 3'UTR usage over pseudotime utilizing multinomial regression splines, same 3'peaks as in Figure 3.1.1.

Testing for differential 3'UTR usage over NSC lineage progression with multinomial regression yielded LR statistics that were reproducible in both samples from Kalamakis G. et al., 2019 (Figure 3.3.2). As these statistics represent how strong a 3'UTR changes APA with lineage progression I ranked the genes accordingly and found enrichment of

various gene sets from GO (Figure 3.3.3a) and also neurodevelopmental risk genes, amongst others, autism risk genes (Figure 3.3.3b).

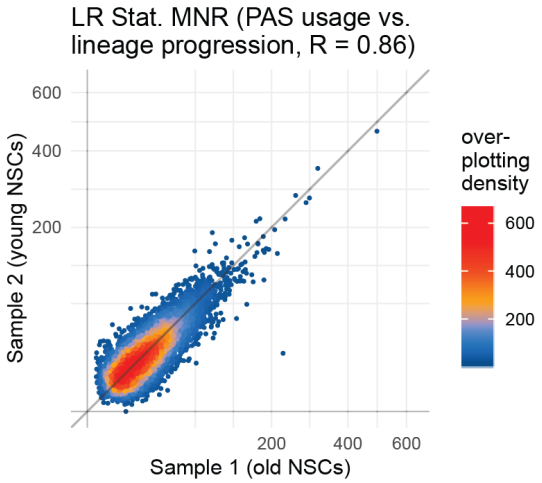


Figure 3.3.2: Regression results (MNR) agree across biological samples. Log-likelihood statistic (LR stat.) for MNR (see also section 2.24), computed independently for both samples of the NSC lineage called "sample 1" and "sample 2", the statistic will be high if 3'UTR usage changes with lineage progression.

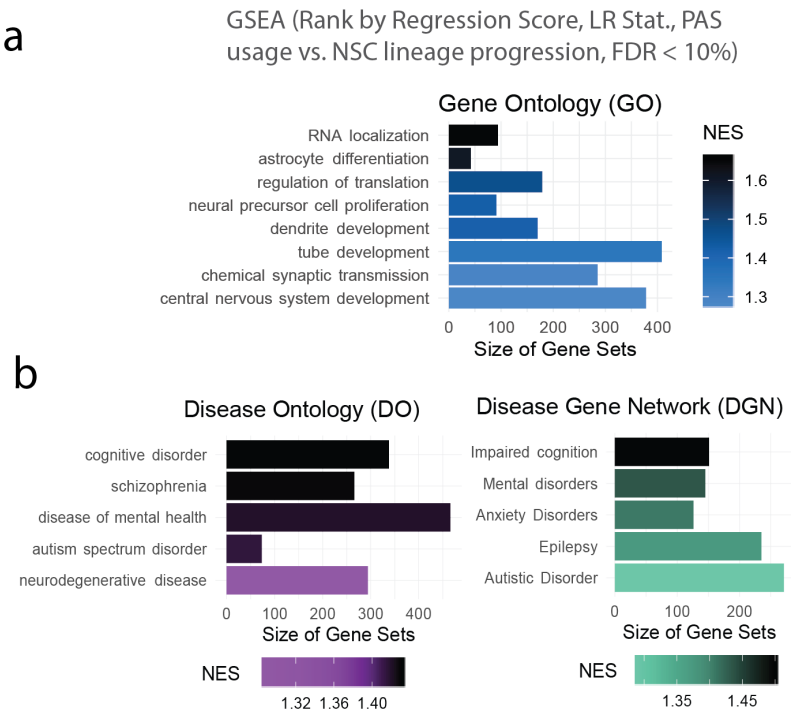


Figure 3.3.3: Gene set enrichment analysis of 3'UTR changes along the NSC lineage. Genes ranked by the strength of 3'UTR changes with NSC lineage progression (LR stat. from Figure 23 used as score), **a**, gene ontology (GO), biological process, **b**, disease ontology, a high normalized enrichment score (NES) indicates association of GO term/disease with 3'UTR changes.

3.1.4 Differential 3'UTR usage between APLP1-/- and wildtype multinomial regression and gene set enrichment analysis

As for the single cell sequencing of the NSC lineage from Kalamakis G. et al., 2019, I fitted psuedotime to the APLP1 sequencing data (APLP1-/- and wildtype) shown in Figure 3.4.1a and in addition assigned single cells to indvidual mice Figure 3.4.1b. This was done as preprocessing steps. Next, I will demonstarate that APLP1-/- and wildtype mice show differential 3'UTR usage.

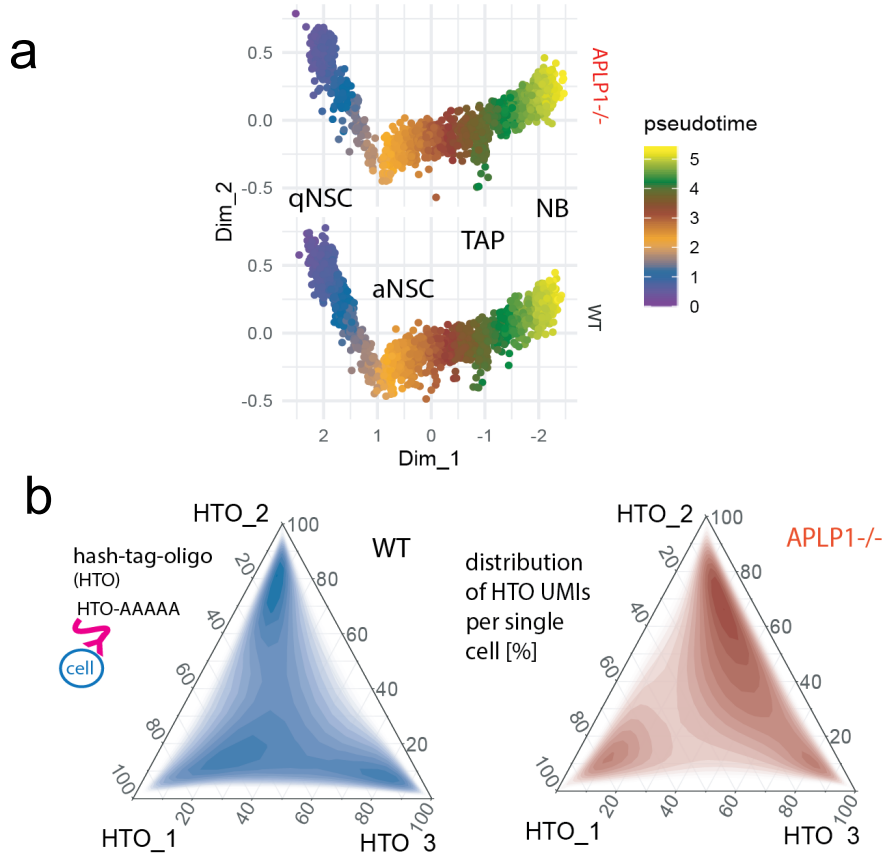


Figure 3.4.1: APLP1 single cell sequencing. **a**, comparison of pseudotime (NSC lineage progression as in Figure 21) between Aplp1-/- and WT, **b**, ternary plots for hash-tag-oligo sequencing, upper sub-panel WT sample, lower sub-panel APLP1-/- , each HTO marks one biological replicate, relative distribution of all three HTOs per single cell, as local density (dark = high, bright = low).

Applying the multinomial test statistic (Figure 3.4.2a) comparing the 3'peaks in the knockout and the wildtype called over 900 genes with differential 3'UTR usage (Figure 3.4.2b).

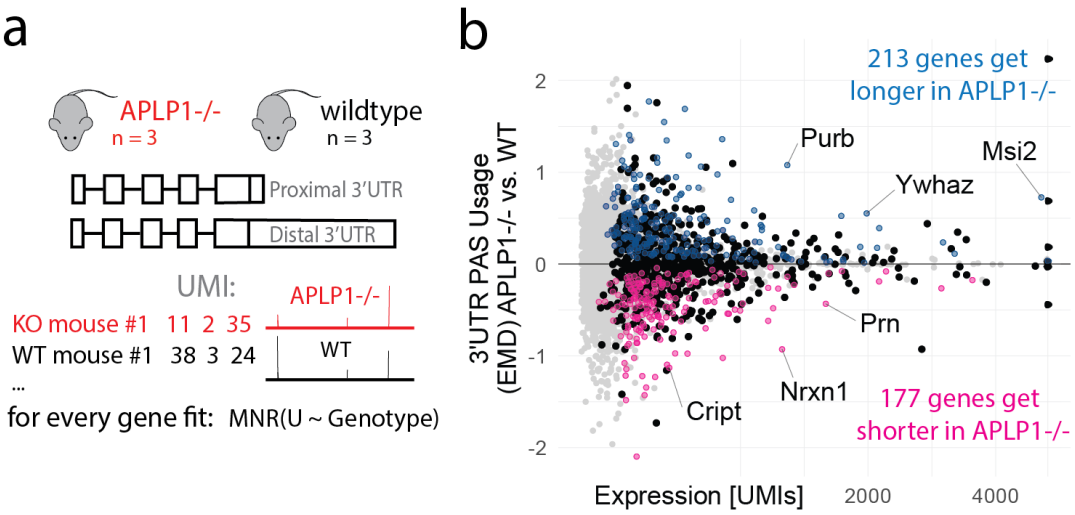


Figure 3.4.2: Differential 3'UTR usage in APLP1-/- vs. wildtype mice. a, Scheme showing the idea behind the sequencing experiment: quantifying differential 3'UTR usage APLP1-/- vs. wildtype mice, individual mice (replicates) assigned to biological replicates by hash tag oligos (see previous Figure), **b**, for every gene with multiple 3'peaks (APA genes), differential 3'UTR usage in APLP1-/- mice vs. wildtype controls (NSCs and neuroblasts), x-axis: total expression summed over single cells, y-axis: earth mover's distance showing 3'UTR trend (longer or shorter in APLP1-/-), non-grey points indicate FDR < 5% (multinomial test statistic).

The difference in 3'UTR usage between both genotypes was also visible in the raw data, the 3'mapping positions (Figure 3.4.3). More visualization of genes showing strong effects (comparison: APLP1-/- vs. WT) is provided in Supplementary Figure 3.

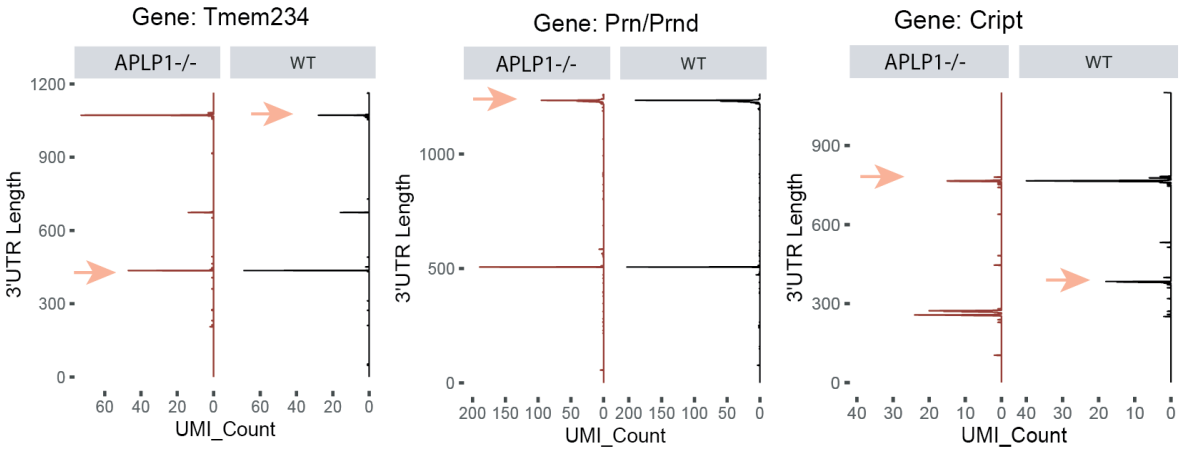


Figure 3.4.3: Differential 3'UTR usage in APLP1-/- shown on three example genes. Raw mapping positions in 3'UTRs (stacked UMIs over single cells), right sub-panels depict these for APLP1-/- (red) and left ones for WT (black) mice, respectively, arrows highlight 3'peaks differing between genotypes.

Since I averaged the 3'peaks over all cells (the NSC lineage) to use the biological replicates as statistical units (not the single cells) one could ask at which NSC state (pseudotime) the effect is strongest. Figure 3.4.4 depicts an example gene (Ywhaz, marked in Figure 3.4.2) to illustrate this. In APLP1-/-, the fraction of its distal PAS is around 25% in qNSCs and increases in the neuroblasts up to 45%, whereas in the wildtype it does not increase.

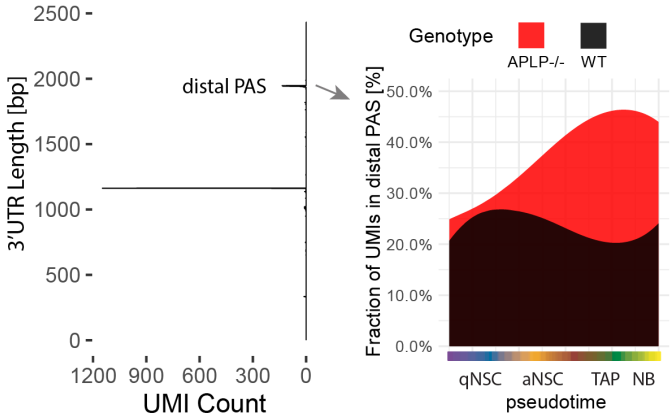


Figure 3.4.4: Multinomial spline regression comparing 3'UTR usage between genotypes. Gene: Ywhaz, fractional usage of its distal PAS over NSCs lineage progression (pseudotime) fitted for APLP1-/- (red) and WT (black).

Performing gene set enrichment analysis on the multinomial regression statistic (deviance explained by genotype) showed enrichment for genes localized to axons and dendrites and also for neurodevelopmental risk genes (Figure 3.4.5).

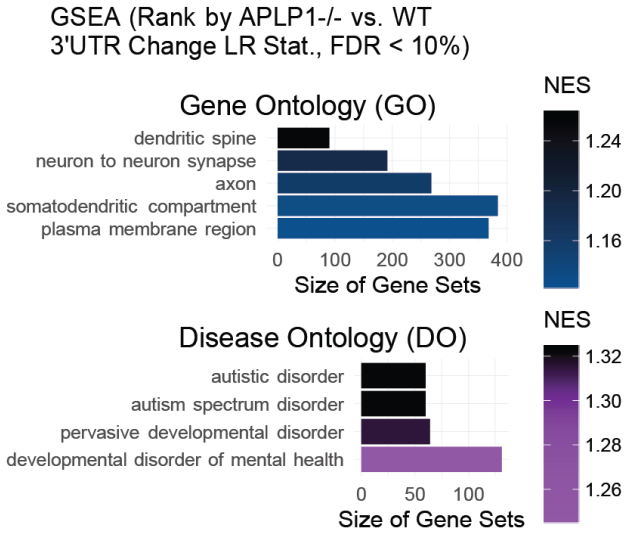


Figure 3.4.5: Gene set enrichment analysis (GSEA), genes ranked by APLP1-/- vs. wildtype 3'UTR changes, multinomial regression as LR statistic, genotype effect, normalized enrichment score (NES), upper panel: gene ontology, lower panel: disease ontology.

3.1.5 3'UTR mapping results in the human single cell sequencing data

Exploring the human single cell sequencing data, I observed even though only read 2 (not read 1) could be used to pinpoint 3'UTR mapping positions, distinct 3'UTR peaks were still visible. Figure 3.5.1 depicts the 3'UTR mapping positions obtained in this dataset for three representative genes.

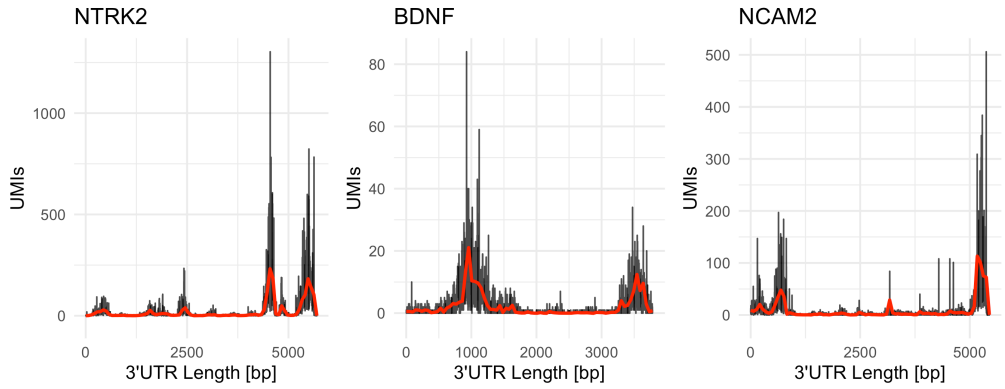


Figure 3.5.1: Example raw mapping positions in 3'UTRs in scRNAseq data from human neurons, from left to right: NTRK2, BDNF and NCAM2, raw mapping positions stacked over all samples (and all cell-types), red regression lines fitted for visualization.

Sample correlation plots of 3'UTR lengths showed very high correlations (example scatterplot shown in Figure 3.5.3) of $R > 0.95$ indicating that the mapping positions are reproducible across samples.

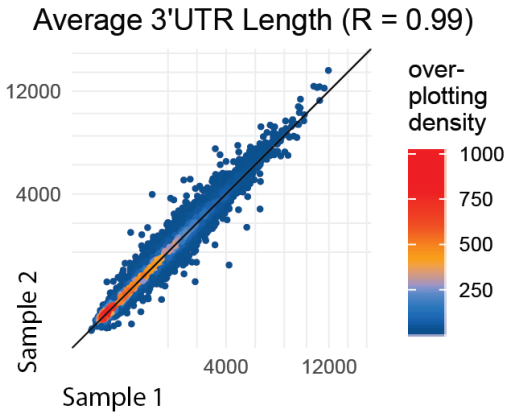


Figure 3.5.2: Comparison of 3'UTR lengths across human samples as scatter-plot, every point: average 3'UTR length of a gene (over all cell types).

3.1.6 Differential 3'UTR usage in ASD vs. Control

Based on the quality metric from the previous section (reproducible 3'mapping) I decided to proceed with the down-stream analysis comparing the 3'UTR lengths of ASD patients to controls. Averaged over the transcriptome, 3'UTRs tend be longer in individuals diagnosed with ASD vs. controls (Figure 3.5.3a). This effect was observed across all (glial and neuronal) cell-types. Since layer 2/3 excitatory neurons were reported to be of specific interest by Velmeshev D. et al., 2019 due to a high burden of differentially expressed genes I report this cell-type first. Figure 3.5.3b depicts the results of the linear model fits: more genes tend to be longer than shorter in ASD vs. controls (For the volcano plots for the other cell-types see Supplementary Figure 3).

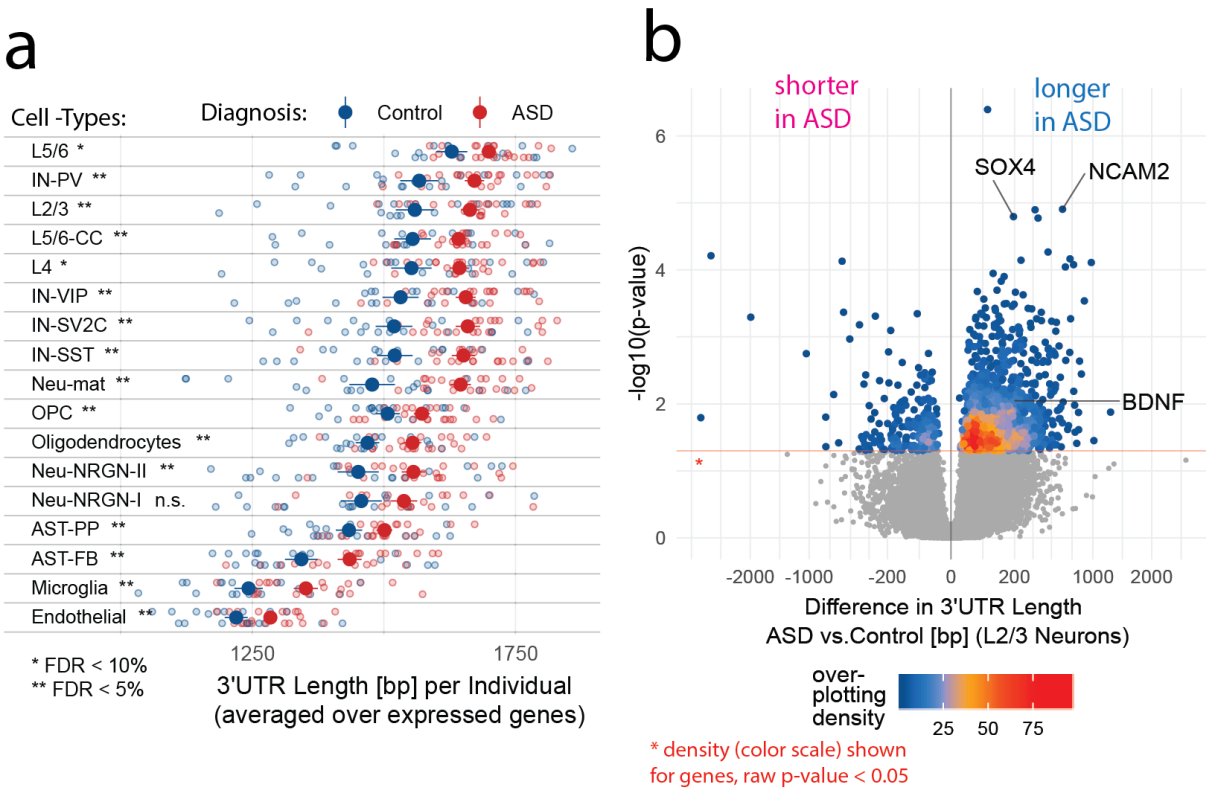


Figure 3.5.3: 3'UTRs tend to be longer in ASD vs. control. **a**, Meta-gene analysis of 3'UTR length, every point: mean 3'UTR length averaged over all expressed genes for one sample and one cell type (red for ASD, blue for healthy control, big points show group means), multiple t-tests for each cell-type with FDR control, two-sided (Benjamini-Hochberg). **b**, Volcano plot showing 3'UTR alterations in layer 2/3 excitatory neurons comparing ASD patients to controls for each gene, 3'UTR length differences (x-axis), p-value estimated from linear models, over-plotting color scale applied to genes with uncorrected p-values < 0.05 to show the overall trend.

Figure 3.5.4 depicts the linear model fits for three genes. The distributions of 3'UTR mapping positions differ between the ASD and the control group.

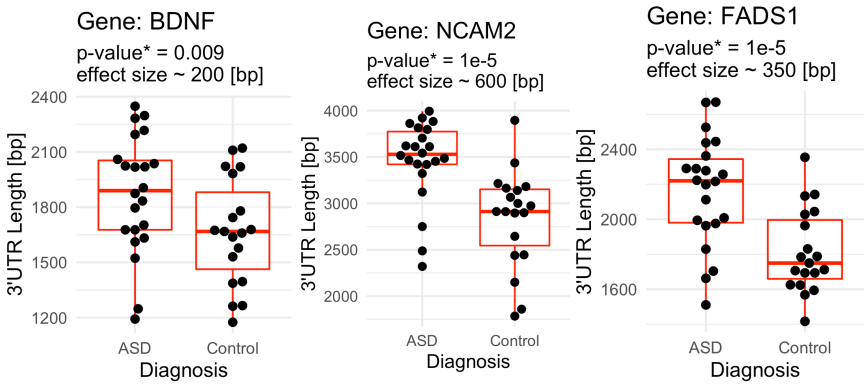


Figure 3.5.4: Illustrative examples for results of linear models fitted to 3'UTR lengths, from left to right: BDNF, NCAM2 and FADS1, * uncorrected/raw p-values, excitatory L2/3 neurons, every point represents the observation from one individual.

Applying gene set enrichment analysis on the linear model result (3'UTR lengthening trend in ASD) pointed to enrichments of genes involved in the pathways of axon guidance, translation and signal transduction (Figure 3.5.5).

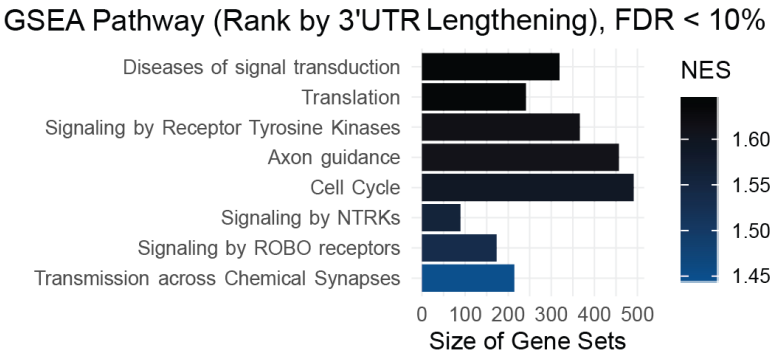
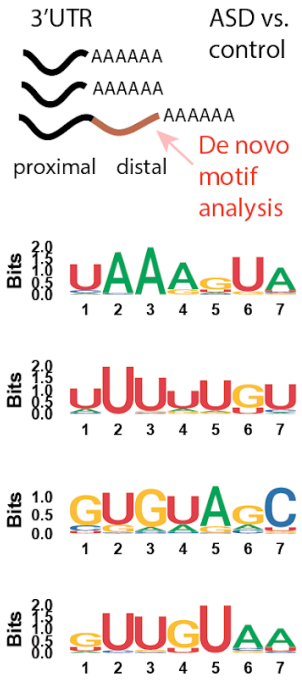


Figure 3.5.5: Gene set enrichment analysis ranking by 3'UTR lengthening in ASD vs. control, Reactome pathways, a high normalized enrichment score (NES) implies that genes of this pathway get longer in ASD vs. control.

3.1.7 Motif detection in 3'UTR changing in ASD vs. Control

Comparing the distal parts of 3'UTRs that become longer in ASD diagnosed vs. controls pointed to an enrichment of motifs, amongst others, with high similarity to the CPE motif (Figure 3.6.1). CPE-like motifs were also detected in the mouse data when searching for motifs in the flanking sequences of PASs that change their usage with lineage progression (Supplementary Figure 4).



No	Motif	log_p_value	log_odds
1	**UAAAGUA**	**26**	**5,76**
2	**UUUUUGU**	**20**	**6,39**
3	**GUGUAGC**	**18**	**5,74**
4	**GUUGUAA**	**16**	**6,02**
5	CUACCGA	14	6,55
6	UGGUACG	13	6,91
7	AUACACU	13	6,44
8	UGCAUUA	13	7,28
9	AUUUUAG	12	5,35
10	ACCGUUU	11	2,16
11	GUUGACU	10	8,41
12	GAUCUCU	8	7,66
13	AUUCCAU	8	6,85
14	GCACUUG	8	8,16
15	GCAUKUU	8	7,65

Figure 3.6.1: Motif detection in 3'UTRs comparing ASD vs. controls in L2/3 excitatory neurons, de-novo motif analysis, motifs enriched in distal 3'UTRs getting longer in ASD vs. controls compared to 3'UTRs getting shorter in ASD vs. controls, list of top 15 hits (enriched motifs) as table on the right-hand side (output from the tool *homer2*), c.f. motifs enriched in 3'UTRs that get shorter in ASD vs. controls, see Supplementary Figure 4. The motif logos (position frequency metrices) for the top 4 are plotted on the left panel.

3.1.8 Usage of PASs flanked by the CPE motif

Interestingly, using the multinomial spline regression to assess whether single PAS increase or decrease their usage with the NSC lineage progression revealed a distinct pattern of PASs flanked by the CPE motif: These PAS increase in the transition from qNSCs to aNSCs and decrease from aNSCs to neuroblasts (Figure 3.6.2a). The same pattern was observed in wildtype samples, but not in APLP1-/- mice (Figure 3.6.2b).

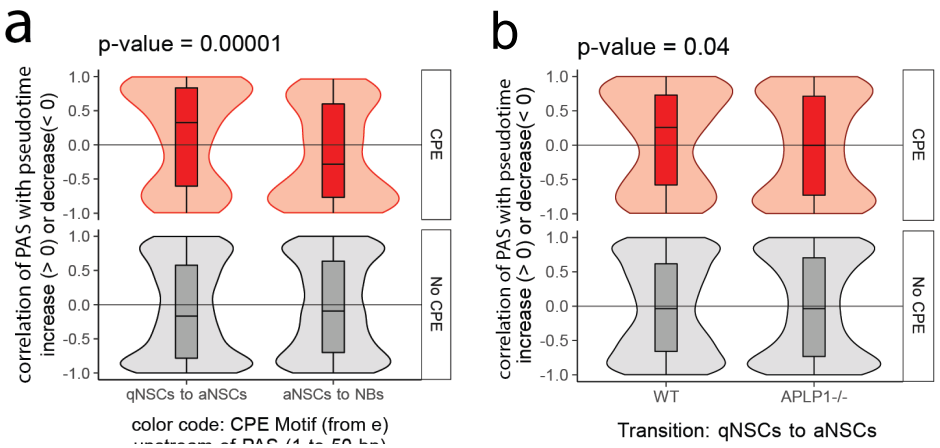


Figure 3.6.2: Selection of PASs flanked by the CPE motif, **a**, Comparison of PAS containing the CPE motif (red color) to other PAS (statistical control) depending on how the individual PAS usage changes from qNSC to aNSCs (left group) and aNSCs to NBs (right group), trend estimated by multinomial spline regression, ANOVA interaction test, two-sided. **b**, the same for the transition qNSCs to aNSCs in wildtype vs. APLP1-/- mice.

3.1.9 Differential 3'UTR usage (*in vivo* vs. *in vitro* NSCs)

Comparing the fold-changes in PASs from aNSCs to qNSCs showed that a considerable amount of these changes is reproducible in an *in vitro* system (R = 0.5, Figure 3.6.3).

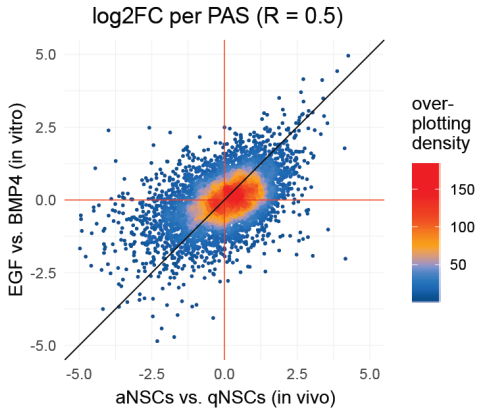


Figure 3.6.3: Comparison of *in vitro* to in vivo 3'UTR changes in NSCs, log-fold-change, every point represents one PAS (summed over cells, positive LFCs = higher PAS usage in aNSCs, negative LFCs = higher PAS usage in qNSCs), x-axis as reference (*in vivo* NSCs), y-axis for *in vitro* NSCs.

3.2 Differential expression and co-expression of polyadenylation factors in ASD and along the NSC lineage

Based on the previous findings that polyadenylation differs between ASD patients and controls, I anticipated to find differential expression of polyadenylation factors. As a summary of the DESeq2 analysis, Figure 3.7.1 indicates that across all cell-types NUDT21, CPSF1 and CSTF3 have the highest expressional changes.

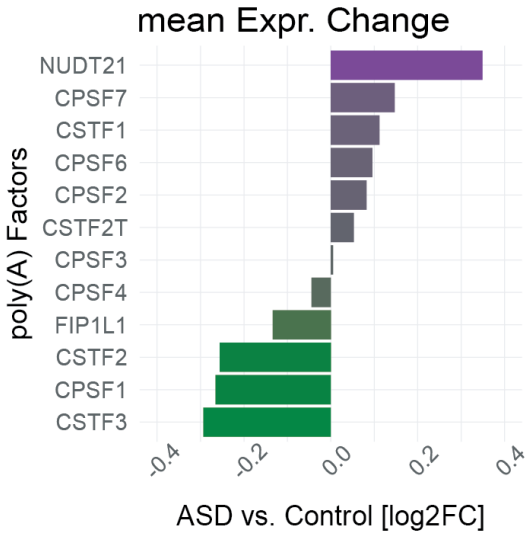


Figure 3.7.1: Differential expression of polyadenylation factors, log2-fold-changes computed between ASD and controls separately for each cell-type, the x-axis shows the average log2-fold-change (across cell-types).

I selected these three genes as potential gene candidates and also considered APLP1 and CPEB4 as genes of interest. This is reasoned with the outcome of the previous analysis that the CPE-motif was associated with 3'UTR lengthening in ASD. Figure 3.7.2 depicts the changes in expression for these genes of interest per cell-type. In L2/3 excitatory neurons CPEB4 and NUDT21 are upregulated. (A complete list of differentially expressed genes in L2/3 excitatory neurons is provided in Supplementary Table 2).

The co-expression analysis revealed that the expression patterns of CPEB4 and APLP1 in mice as well as in humans are correlated across single cells (Figure 3.7.3). These correlations (Gillis J. & Paul P., 2012) might be interpreted as a hint that APLP1 and CPEB4 operate on the same pathway (see also Discussion).

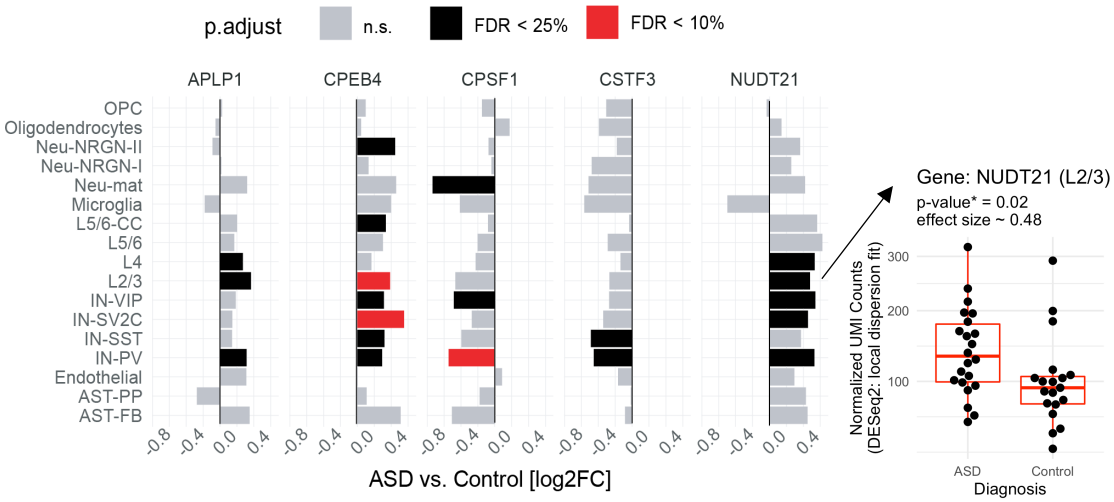


Figure 3.7.2: Differential expression of candidate genes in ASD vs. controls, fold-changes from DESeq2, y-axis, per cell type, x-axis, log2-fold-change in gene expression, color code, p-value thresholds (FDR corrected) per cell-type, raw data for the expression of NUDT21 in L2/3 excitatory neurons shown on the right panel.

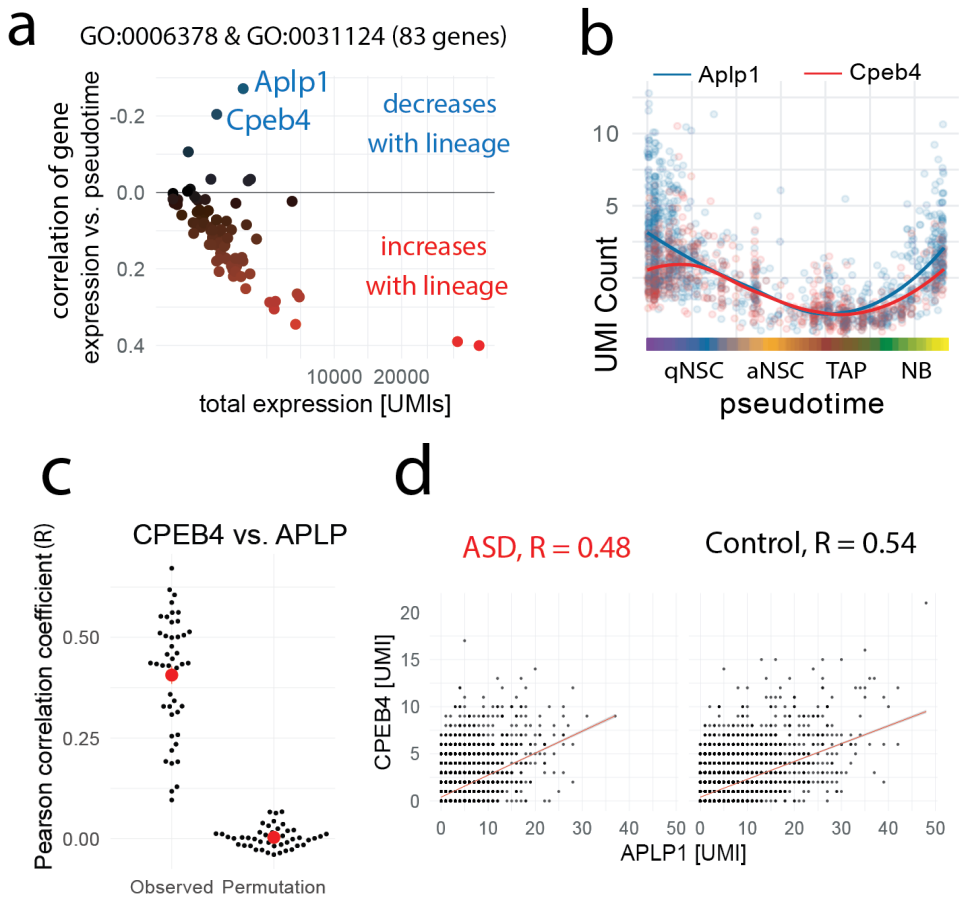


Figure 3.7.3: Correlation analysis of genes acting on 3'UTR length, in mouse and in human. **a**, Expression trend of selected genes, upstream and co-regulators of APA or poly(A)-tail length (GO categories: GO:0006378 and GO:0031124), scRNA-seq

data, total expression summed over single cells (x-axis) against correlation of expression vs. pseudotime (y-axis), **b**, co-expression of Aplp1 and Cpeb4 along the NSC lineage, each point indicates a single cell regression line fitted for both genes to show the trend, **c**, correlations of CPEB4 vs. APLP1 expression per human sample (observed vs. permutation of cell assignment as statistical control), **b**, scatterplot for correlations across all cells in ASD and control group, red lines: linear regression.

In addition, among all Cpeb genes, Cpeb4 is highest expressed in NSCs (Figure 3.7.4). As shown in Figure 3.7.3b Cpeb4's expression is highest in quiescent NSCs, decreases in active NSCs and slightly increases in neuroblasts.

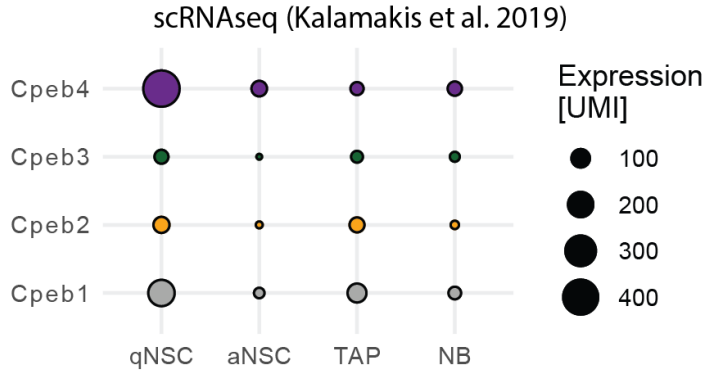


Figure 3.7.4: Expression of Cpeb genes in NSCs (scRNAseq data), summed UMI counts for Cpeb genes with respect to cell types (activation states).

Comparing the gene expression trends aNSCs vs. qNSC for *in vivo* to *in vitro cells* showed agreement, meaning genes that tend to undergo up-regulation from *in vivo* qNSCs to aNSCs also tend to be upregulated in *in vitro* aNSCs compared to qNSCs (Supplementary Figure 5). Another comparison was differential gene expressed in APLP1-/- vs. wildtype (listed in Supplementary Table 3).

3.4 CPEB4-RNA immunoprecipitation results

Next, I identified transcripts enriched for CPEB4 binding in NSCs. The analysis indicated around 900 genes with high affinity to CPEB4 (Figure 3.8.1a). Furthermore, the RNA immunoprecipitation indicated that CPEB4 binds at the assumed CPE consensus motif (Figure 3.8.1b). The genes with affinity to CPEB4 were enriched for functions as transmembrane receptors and transporters (Figure 3.8.1c).

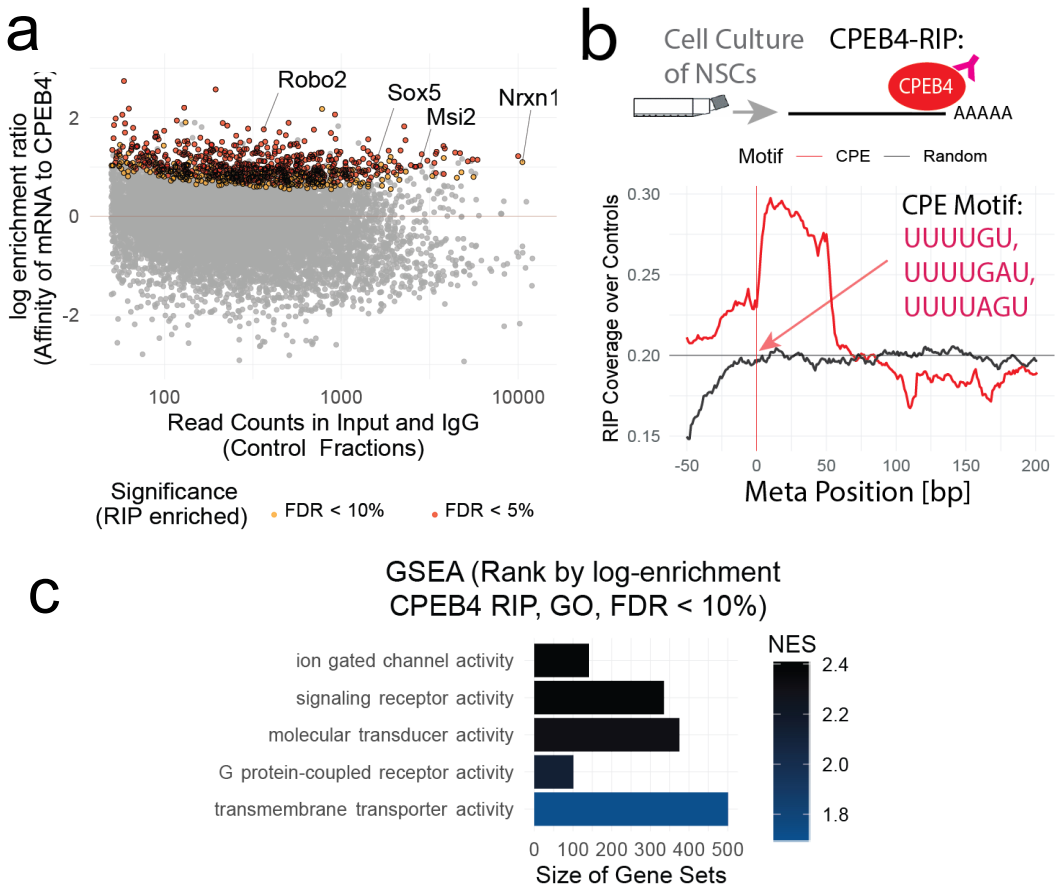


Figure 3.8.1: CPEB4 immunoprecipitation a, detection of CPEB4 substrates (CPEB4 binds mRNAs) by RNA immunoprecipitation (RIP), y-axis log-enrichment in CPEB4 IP fraction, x-axis mean mRNA expression in Input and IgG control fractions, red and yellow points mark genes called significant by DESeq2 (for CPEB4 binding), n = 2 per fraction, **b**, signal from CPEB4-RNA immune-precipitation (RIP) in cultured NSCs for the CPE motif (shown in red), mean residual coverage over RIP controls as meta-position, random position as statistical control shown as black curve, **c**, gene set enrichment analysis ranking by log-enrichment in CPEB4-IP fraction, high normalized enrichment score (NES) indicates enrichment of genes in categories for CPEB4 binding in NSCs.

Since I hypothesized that APLP1-/- can impact APA in NSCs in an CPEB4 dependent manner I also assessed CPEB4 binding to alternative 3'UTR usage in APLP1-/- vs. wildtype mice. First, I compared the APLP1-/- 3'UTR alterations to the list of CPEB1 and CPEB4 binders in neurons identified in an immunoprecipitation by Parras A. et al., 2018. Half of the genes with alterations in 3'UTRs (APLP1-/- vs. wildtype) bind CPEB4 (Figure 3.8.2, left). Importantly, CPEB1 binders were not enriched with APLP1-dependent 3'UTR changes. A corresponding enrichment in the NSC RIP assay was observed, however with a lower total number of CPEB4 binders (Figure 3.8.2, right).

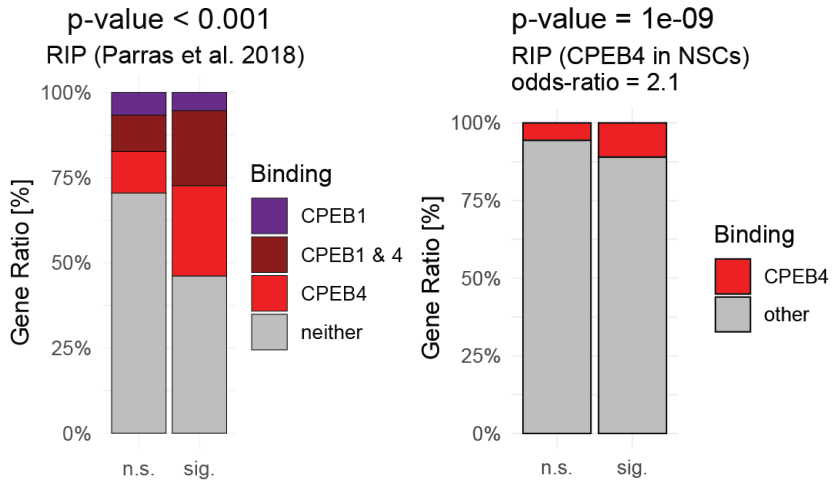


Figure 3.8.2: Genes with 3'UTR alterations in APLP1-/- vs. WT are enriched for CPEB4 binding. Left panel: Intersection of genes showing 3'UTR alterations in APLP1-/- with CPEB1 and CPEB4 binders from Parras A. et al., 2018, color code: genes bound by CPEB1, CPEB4, by both or neither of both, genes marked as significant (sig.) from Figure 37 (APLP1-/- vs. wildtype 3'UTR changes), Chi-square test; right panel: the same comparison for CPEB4 binders from Figure 3.8.1a, Fisher's exact test.

3.5 Protein detection in NSCs and comparison to 3'UTR alterations indicates higher protein outcome with 3'UTR shortening

Having described differential expression of polyadenylation factors and binding of CPEB4 I intended to integrate these findings with proteomics. We identified over 2000 proteins, observed a strong correlation to transcriptomic changes (Figure 3.9.1a) and found several hundred differentially expressed proteins in *in vitro* aNSCs vs. qNSCs (Figure 3.9.1b). As quality control of the proteomics results, sample correlation scatterplots are provided in Supplementary Figure 6. The results indicate that whenever the mRNA changes, protein levels tend to follow either the up- or down-regulation. In this context, however, I was rather interested in genes that deviate from this trend.

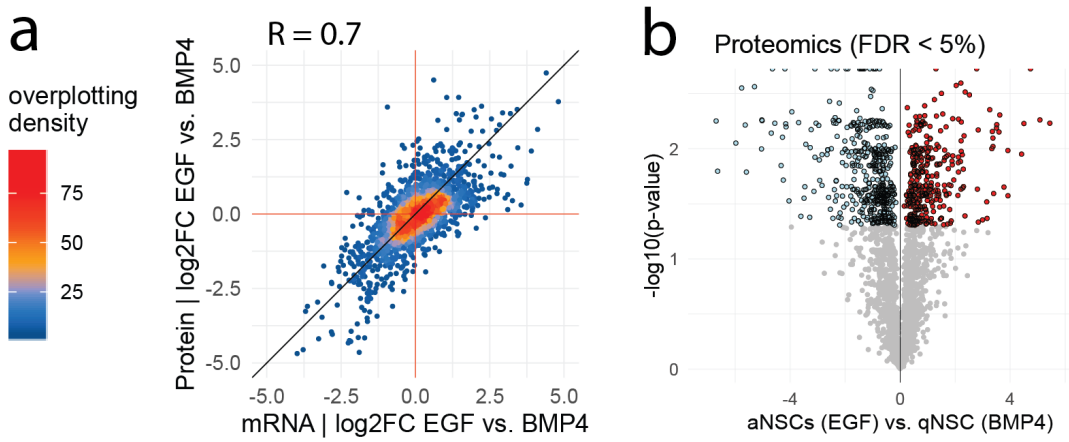


Figure 3.9.1: Protein production in aNSC vs. qNSCs. a, Log fold-change, per gene: change in mRNAs levels (x-axis) against change in protein (y-axis) between aNSCs (EGF) and qNSCs (BMP4). **b**, Volcano plot, differential protein abundance in vitro between aNSCs (EGF) and qNSCs (BMP4), proteins marked in red and blue (FDR < 5%, Perseus tool).

As described in the methods (and also depicted in Figure 3.9.2) I estimated post-transcriptional regulation by dividing per gene protein by mRNA levels, for aNSCs and qNSCs, respectively. Comparing the change in this translation index between aNSCs and qNSCs to 3'UTR length changes between both NSC subpopulations revealed that 3'UTRs getting shorter in aNSCs also tend to increase protein production in aNSCs (Figure 3.9.2). This tendency was more pronounced comparing 3'UTR lengths among significant proteins from Figure 3.9.1b. Fitting a linear model on the changes in

translation index (aNSCs vs. qNSCs) across samples indicated that this trend is consistent across samples (Supplementary Figure 7).

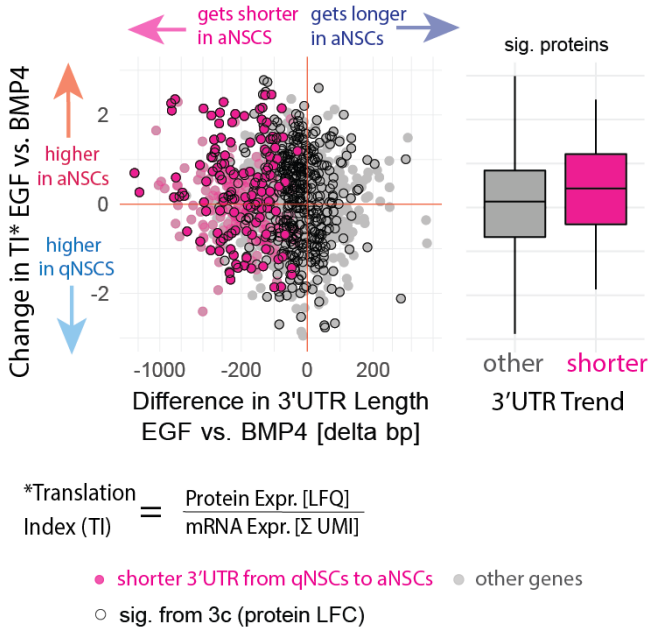


Figure 3.9.2: Genes undergoing 3'UTR shortening tend to have higher protein outcome. Scatterplot, every point is a gene, difference in 3'UTR length aNSCs (EGF) and qNSCs (BMP4) on the x-axis vs. difference in translation index (TI: LFQ values by UMI counts) on the y-axis, boxplot on the left: comparison of other genes to 3'UTR shortening genes (c.f. Supplementary Figure 7).

3.6 Binding of CPEB4 to mRNAs enhances protein production

Based on the findings of previous studies that CPEB4 impacts mRNA translation I aimed for addressing this point by combining the CPEB4 immunoprecipitation results with ribosomal profiling and the proteomics data from the previous section. As initial quality check for the ribosomal profiling data, gene counts were reproducible across samples, counts in 3'UTRs depleted for the ribosome protected but not the total RNA fraction (Supplementary Figure 8). Next, I estimated the translation efficiency, per gene dividing ribosome protected read counts by total RNA counts. This translation score correlated with the estimated affinity of CPEB4 to mRNAs from Figure 3.9.3a. This result suggests that in NSCs, binding of CPEB4 enhances mRNA translation. In order to further strengthen the claim, I compared the alterations in protein production between aNSCs and qNSCs from Figure 3.9.3b to CPEB4 binders and non-binders. Higher

protein expression for CPEB4 binders was observed in qNSCs compared to aNSCs. This agrees with previous findings since CPEB4 is higher expressed in qNSCs, hence it would increase the translation of CPEB4 binders in qNSCs.

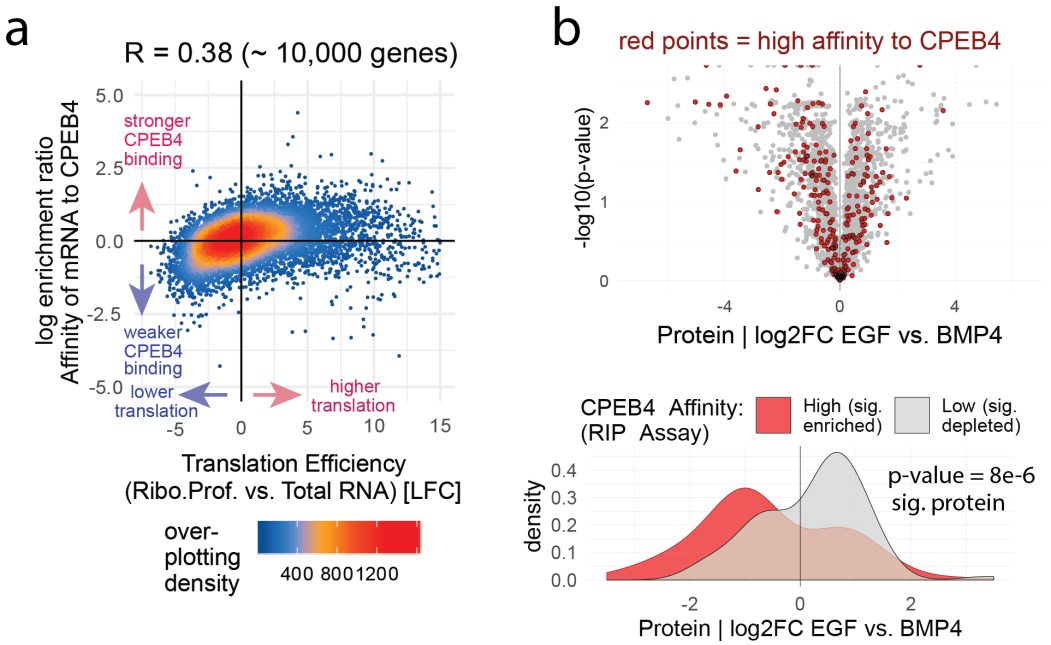


Figure 3.9.3: CPEB4 binding to mRNAs correlates with translation and differs between qNSCs and aNSCs, a, every point represents a gene, x-axis: LFC in ribosome protected vs. total RNA reads (aNSCs), y-axis: log-enrichment of in CPEB4-IP fraction (aNSCs), Spearman rank correlation, **b,** Volcano plot, LFC in protein levels, CPEB4 binders as red points (upper panel), comparison of proteomics to CPEB4-RIP, density plot, x-axis fold-change in protein expression (aNSC vs. qNSC), groups colored in red if enriched in CPEB4 binding and grey depleted for CPEB4 binding (control gene set for low-CPEB4 binding), Wilcoxson rank sum test, two-sided (lower panel).

Chapter 4

Discussion

In this chapter I will discuss the applications and limitations of the 3'peak calling pipeline in single cells by covering both, technical and biological aspects (4.1). Next, I will extend on the statistical tests for differential 3'UTR usage and place these analyses in the context of computational methods used to study neurodevelopmental disorders (4.2). In section 4.3 I will reflect on differential and co-expression in single cells. Also, I will evaluate the proteomics, ribosomal profiling and immunoprecipitation results (4.4). To conclude this thesis, I will highlight the most crucial points of this work and give an outlook on the biological results (4.5).

4.1 The 3'peak calling pipeline

Regarding the 3' mapping, the here presented pipeline can call sharp 3'peaks in single cell sequencing data that correlate with the known PAS (as demonstrated in the chapter Results). When I started my PhD, I computed differential 3'UTR usage based read coverage obtained from single cell Smart-seq2 data (from Llorens-Bobadilla E. et al., 2015). As explained in the introduction, full-coverage sequencing is less evident for 3'UTR usage than the 10X Genomics 3'tagging strategy and for this reason not shown in this thesis. However, the 3' tagging approach also has a number of limitations. For instance, potential false positive 3'peaks could arise from 3'UTR regions with either very low sequence complexity or long A-stretches in their sequences (also known as internal priming, see Miura P. et al., 2014). In general, such artifacts should be minimized by the confidence from the paired-end mapping strategy (50 bp mapped from read 1 and 100 bp mapped from read 2) and the oligo-d(T) capturing with the beads from 10X Genomics (25 bp anchoring/capturing sequence). Intronic sequences which can contain A-stretches are excluded from the analysis as only reads falling into annotated 3'UTR regions are considered. The restriction to annotated genes also excludes cases like unannotated ultra-long 3'UTRs. In such situations the assignment of the gene by 3'peaks can be deemed tricky, especially for overlapping genes. In this regard, my pipeline also checks for the mapping orientation of reads that matches to the strand (plus or minus) in which the gene is transcribed. Notably, in this work I focused on APA within the terminal exon of a gene. The alternative usage of multiple terminal exons could also be analyzed. Observed during the data exploration phase, but not reported here, are cases of alternative exon usage (roughly 10% of APA genes showed mapping to more than one terminal exon; in most cases the fraction of alternative exons was rather low compared to the main exon, not shown). However, it has to be considered that the developed 3'peak calling method also increases the dimensionality of the data as single isoforms in single cells are counted. On the one hand this workflow enables new analysis strategies but on the other hand the biological interpretability also has to be taken into account. Also, this pipeline could be easily modified to call instead of 3' peaks 5'peaks since 5' tagging became available on the 10X Genomics platform. This would enable the distinction of alternative 5' start sites (alternative 5'UTR lengths) for genes with multiple transcription start sites. As for the 3'UTR, different 5'UTR length

and regulatory elements can also impact mRNA translation efficiency. In addition, it also should be mentioned that the application of the developed pipeline is limited also in the number of datasets it can be applied on. This is due to the fact that most available datasets, like that from Velmeshev D. et al., 2019, were sequenced with a short read 1, meaning UMI and cell barcode are present, but not the part of the read mapping to the most 3'end. This is due to the intention to reduce costs and to only count the gene, not the exact 3'UTR position. For the sequencing of human neurons from Velmeshev D. et al., 2019, it could be considered to sequence these samples with longer reads in order to obtain sharp 3'peaks as for the mouse data. Most likely, doing so would improve the statistical analysis. The 3'peak calling pipeline itself could be further optimized by including C++ code for performance reasons (but runs at reasonable speed already) and published as an R package. Standardized workflows in packages also have multiple downsides as they need to cover many exceptions and provide the user with tools to ensure data quality. For non-standard analyses, this practice of applying 'black-box' workflows can also lead to wrong conclusions, one example being the estimation of p-values in one biological replicate. In general, the main focus of this work, however, were rather the biological findings. Recently, an alternative pipeline named Sierra was published calling peaks in single cell sequencing data (Patrick, R., et al., 2020). However, the peaks obtained with this tool are way broader than the peaks reported with the here presented pipeline. In their publication the authors show examples of broad peaks and also assume a peak width of 600 bp in their implemented peak calling algorithm (c.f. Patrick, R., et al., 2020). This is however not sufficient to discriminate PASs when the signals are very close. In comparison, my pipeline using the paired-end mapping approach gives way sharper 3'peaks and has therefore higher resolution. Also, their emphases on UMAPs computed on peak counts could be criticized as a strong visualization with little biological indication. The only application this would be helpful I could think of based on my experience with differential 3'UTR usage could be clustering cells by isoform usage. But even in this case, I think it would be better to directly plot the fractions of isoforms for one gene across the cell-clusters. Since differential 3'UTR usage in single cells is a new approach, this field lacks established standards. In this work I show how a down-stream analysis could look like. I will discuss this in the following section.

4.2 Down-stream analysis: differential 3'UTR usage

About the differential 3'UTR usage in NSCs, it has to be considered that the analysis can be limited, however, for lowly or for strongly differentially expressed genes (over pseudotime) as well as rather complex 3'UTR trends. The estimated correlation coefficients have a direct biological interpretation (*i.e.* the gene gets longer, shorter or does not change its length). The benefit of multinomial models over correlations and ordinal models which assume linear trends poses their sensitivity for non-linear changes. Figure 4.1 illustrates this for the example gene Sox4. Its middle PAS is more often selected in TAPs and NBs compared to qNSCs where proximal and distal PASs are predominant. A gene like this would receive a high MNR statistic but a low absolute correlation coefficient.

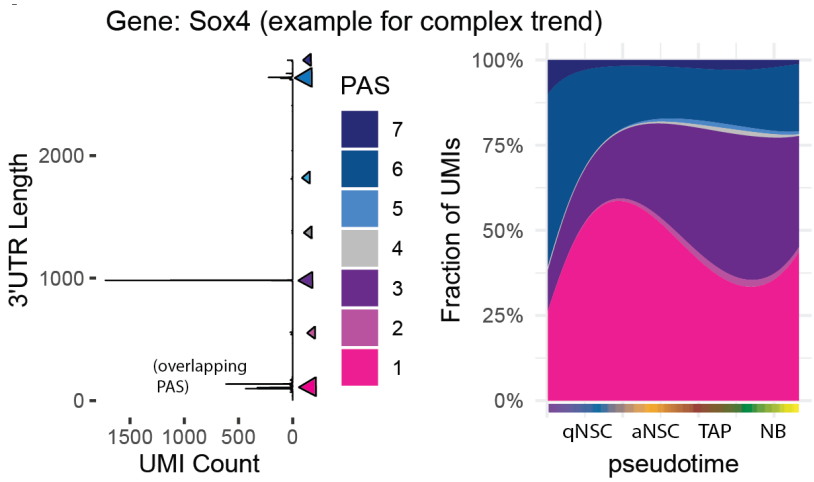


Figure 4.1: Multinomial spline regression can detect complex 3'UTR trends, Sox4 left sub-panel, raw 3' mapping pos. summed over cells, right sub-panel, approximation of changes in 3'UTR usage over pseudotime utilizing multinomial regression splines, color code as dark blue to red for most distal to most proximal PAS (color code corresponds to left sub-panel).

Interestingly, the multinomial test comparing APLP1-/- vs. wildtype could also be carried out applying the Cochran–Mantel–Hänszel test (Agresti A., 2002), the extension of the Chi-Square test for repeated observations like biological replicates. Reassuringly, computing p-values on the same genes with both methods yields highly comparable results (Figure 4.2). Here, MNR was used instead of Cochran–Mantel–Hänszel as this would be more consistent with the spline regression introduced before.

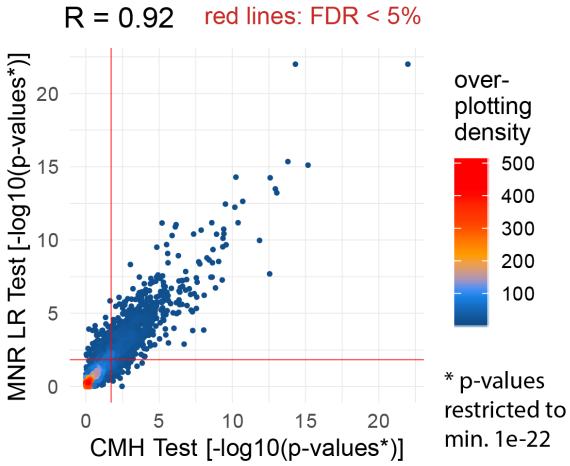


Figure 4.2: MNR and an alternative Chi-Square test give comparable results, MNR regression applied on PAS count data (3'peaks) comparing the distributions of PASs per 3'UTR between genotypes (APLP1-/- vs. WT), scatterplot p-values computed with the CMH-test vs. MNR-LRT, thresholds for Benjamini Hochberg's correction for multiple testing (FDR = 5%) added as red lines.

Moreover, I would like to mention some details regarding the statistical analysis for differential 3'UTR usage in excitatory layer 2/3 neurons. As a statistical control for the linear models, it can be shown that a permutation of the variable Diagnosis (removal of assignment) yields uniformly distributed p-values supporting the reliability of the applied test statistic (ANOVA), shown in the right panel of Figure 4.3.

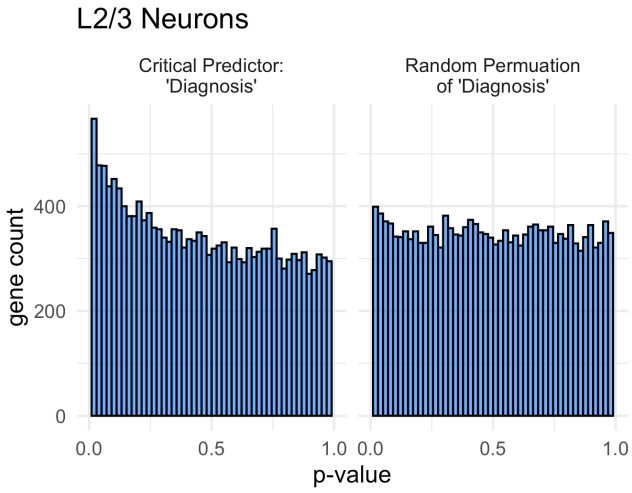


Figure 4.3: Estimation of p-values for 3'UTR changes comparing ASD patients to controls, p-value histograms, linear models fitting 3'UTR lengths, ANOVA testing with predictor variable Diagnosis (left panel) and p-values computed when *Diagnosis* was randomly permuted.

Another hypothesis was whether males and females differ in the 3'UTR alterations observed between ASD vs. control. I addressed this by adding + Sex:Diagnosis as interaction term to the model and then testing for this interaction term as critical predictor. The comparison of 3'UTR length alterations in ASD vs. control between male to females did not yield significant result (not shown) hinting that the ASD-lengthening trend is not different between sexes. As mentioned in the foreword of this chapter differential 3'UTR usage in ASD was addressed before by Szkop K. J. et al., 2017 using the *DaPars* tool (Xia Z. et al., 2014) for full-coverage bulk RNAseq. In their study the authors reported 3'UTR lengthening in ASD patients compared to controls for most of the analyzed samples. The fact that the conclusion from this work and the one from the paper of Szkop (*i.e.* 3'UTR lengthening in ASD) agrees, strengthens this biological claim as both studies are completely independent of each other. However, it remains unclear whether the reported lengthening trend is a mere consequence of ASD or whether there is a causal link. The differential expression analysis pointed to NUDT21 and CPEB4 (also with the motif analysis) as possible candidates.

Another issue poeses the comparison of human and murine 3'UTRs. In the data I see that 3'UTRs are on average roughly 500 bp longer in human neurons compared to NSCs and neuroblasts from mice. Here, the evolutionary conservation of 3'UTR sequences between both species should be considered (Miura P. et al., 2013; Guffanti, G., 2018). A direct one-to-one comparison of human and mouse 3'UTRs like changes in ASD vs. controls to APLP1-/- vs. controls appears to be difficult but rather tends to agree (significant, but weak correlation, not shown). In general, 3'UTRs in the context of ASD are understudied.

The more commonly used approaches how to address neurodevelopmental disorders with computational methods are differential expression, exome sequencing and genome-wide association studies (Wanke K. et al., 2018). These approaches are rather centered around the coding part of the genome. Wanke K. et al., 2018 suggest to also consider the non-coding parts like 3'UTRs and there to search for motifs of micro RNAs or RNA binding proteins. The here applied motif and gene set enrichment analyses demonstrate how to further interpret alterations in 3'UTRs and to generate biological hypothesis (like for the CPE motif in 3'UTRs that are longer in ASD patients). Also, it is difficult to predict what the contributions of alterations in expression

levels, 3'UTR length changes and single point mutations. For example, one commonly mutated gene in ASD is PTEN (Busch, R.M., 2019; c.f. Velmeshev D. et al., 2019).

4.3 Differential expression and correlation analysis

To evaluate differential expression, I used pseudo-bulk approaches, meaning gene expression is summed over single cells for each replicate, respectively. Applying this method, I intended to avoid pseudo-replication, an issue often seen in single cell sequencing analyses also in the paper from Velmeshev D. et al., 2019. Computing test statistics on single cells (and not biological replicates) can lead to a dramatic overestimation of confidence (overoptimistic p-values) as single cells of the same animal or individuum are not independent statistical units. Since single cell sequencing is an expensive technique usually one or few biological replicates are sequenced. This makes it *per se* difficult to decide whether a finding is significant and therefore reproducible or not. Assigning replicates by cell-hashing could be part of the solution. It might be a good idea to fit splines over cell trajectories or cell-types and to compare these spline fits across replicates and conditions to derive conclusions (Anders S., unpublished data).

A major advantage of single cell sequencing over bulk sequencing techniques is its resolution. One can observe whether a difference in expression levels between conditions comes from few or many cells. Moreover, single cell sequencing data provides statistical power for gene-to-gene correlations as these correlation coefficients are computed over thousands of single cells. Pearson's correlation is perhaps the most intuitive solution for this and easy to implement. These analyses enable the prediction of novel gene-gene interactions, which then need further validation. In this project for instance, further experiments are required to confirm the protein-protein interaction between APLP1 and CPEB4 by proximity ligation assay or co-immuno-precipitation. At the time of writing this thesis, Nikhil Oommen George and the Isabel Fariñas lab are working on these assays. Preliminary results from both labs suggest that there is an interaction between both proteins. Apart from these ongoing experiments and the mere co-expression, the pulldown of mRNAs with an antibody against CPEB4 and the position of the CPE motif could be linked to differential 3'UTR usage in APLP1-/- vs.

wildtype further supporting the hypothesis that APLP1 and CPEB4 interact. In principle, this part of the project can be seen as a follow-up study on the result of a correlation analysis. For this reason, it would be an example that analyzing co-expression in single cells can lead to the discovery of novel biological findings.

4.4 Proteomics, ribosomal profiling and immunoprecipitation

Next, I would also like to discuss the proteomics analysis. In general, quantifying total protein abundances provide an overview of the proteins produced in a cell and will change with alterations in mRNAs levels given enough time for the proteome to adopt. For the *in vitro* BMP4 (quiescent) experiment the treatment time was 3 days. An alternative to total protein levels would be to quantify newly synthesized proteins by incorporating heavy isotope labeled amino acids. Here, the high amounts of required input materials, in this case proteins isolated from NSCs, limit this method. To overcome this limitation, I estimated mRNA translation by the introduced translation index. This value will, however, reflect both, protein production and degradation. Also, the number of proteins that can be captured by mass spectrometry is limited (here roughly 2200 proteins). For this reason, the impact of 3'UTR changes in lowly expressed genes cannot be captured. One technique to estimate protein outcome and to overcome these limitations, namely the static picture of protein levels and the limited number of genes poses ribosomal profiling (Faye M.D. et al., 2014) or ribosome immunoprecipitation (Baser A. et al., 2019). These techniques allow to compute a translation efficiency by comparing the fraction of ribosome protected reads to reads in total RNA fractions (Baser A. et al., 2019). In contrast to total proteomics, these values will reflect the dynamics of translation. As an example, the translational up-regulation of mRNAs will first be visible in their ribosome protected reads and later in proteomics since the ribosomes need some time to produce enough proteins. In this project it would also be of interest to carry out ribosomal profiling in qNSCs and in NSCs from APLP1-/- mice or to do proteomics and then to correlate the results to 3'UTR changes. In general, the strongest restriction in this project poses the link between 3'UTR length changes and protein production. Studying this is only feasible for single genes as the long and short 3'UTR sequences needs to be cloned into vectors (Lackford B. et al., 2014). For this

reason, it remains unclear what the effect for a gene with 3'UTR alterations will be: either no effect, localization or a change in protein levels. By comparing the independent experiments like changes in proteomics to 3'UTR changes estimated by single cell sequencing, one can answer questions like what is the overall trend for 3'UTR shortening genes. Such data integration strategies highly depend on: 1. the intersection of genes/features in the assays, 2. low technical variability, 3. biological reproducibility of effect sizes and 4. the selection of comparison groups. To my knowledge, there are no standards available for these rather custom analyses. With the experience I gained during my time as PhD and also in the discussions with Dr. Simon Anders, I suggest: When comparing multiple high-throughput assays to first test whether the effect is reasonable strong and significant across genes and secondly, when possible, to test whether the effect – averaged over genes – is reproducible among replicates.

With regards to the CPEB4 RNA immunoprecipitation, using next generation sequencing instead of microarrays like in Parras A. et al., 2018 has the benefit that the signals from the IP fractions can be identified with nucleotide resolution. In contrast, microarrays only allow the estimation of log-enrichments per gene. On the other hand, Parras A. et al., 2018 also used an antibody against CPEB1 which poses a suitable control given that we would like to show that the alterations in APLP1-/- depend on CPEB4 and not CPEB1 in this work. An alternative idea was to fit per 3'UTR region the changes in 3'peaks between APLP1-/- vs. wildtype to the IP signal using generalized linear models. With this approach I observed agreement, meaning that CPEB4 binding and APLP1-/- dependent 3'UTR changes correlate. However, since this method would be less intuitive, we decided instead to compare the intersection of CPEB4 binding partners and genes with differential 3'UTR usage between APLP1-/- and wildtype mice. Considering the fact that RNA immunoprecipitation assays rather tend to be noisy, I think that combining the signal from multiple genes in a meta-gene analysis can provide a clearer picture than looking at individual genes.

4.5 Final summary and outlook

In this doctoral thesis I demonstrated that sharp 3'peaks can be derived from single cell sequencing data. These 3' peaks allow the distinction of different PAS, also when two PASs are as close as 50 bp. This data quality can be achieved with long reads and the paired-end mapping approach. The (multinomial) distributions of this peaks can then be fitted against a latent variable (like pseudotime representing NSC lineage progression) or compared across conditions. The future will show whether scientists will draw more attention to this isoform related approaches.

Combining the results from the 3'UTR and the down-stream analyses, explicitly the gene set enrichments and motif analysis, pointed towards CPEB4 and a possible interaction with APLP1 in regulating 3'UTR isoform choice. The absence of APLP1 impacted the selection of PASs flanked by the CPE motif. In addition, behavioral studies in APLP1-/- mice revealed that these mice have autistic-like traits compared to wildtype littermates (These studies were conducted by my project partner Nikhil Oommen George, not shown). In the context of the findings from Parras et al. 2018 that CPEB4 is an autism risk gene, our results imply that APLP1 can de-regulate CPEB4, perhaps its intracellular location, and this results in neuronal circuits that differ from wildtypes. In humans diagnosed with ASD, our data suggests that also CPEB4 plays a role. Here, these results should be seen as a starting point for further studies. For instance, one could try to rescue the behavioral phenotypes in established ASD mouse models or other mouse lines for neurodevelopmental disorders by modulation of CPEB4 or NUDT21. Such experiments could reveal a causal link between alternative 3'UTR usage and ASD.

Overall, I am convinced that looking into the 3'UTRome in the brain will provide new insights into the molecular mechanisms underlying neurodevelopmental disorders.

References

Agresti A. An introduction to categorical data analysis. John Wiley & Sons, Inc., Hoboken, New Jersey ISBN 978-0-471-22618-5 (2007).

Agresti A. Generalized Odds Ratios for Ordinal Data, International Biometric Society Stable, 36:59-67 (1980). URL: http://www.jstor.org/stable/2530495

Alcott, C. E. et al. Partial loss of CFIM25 causes learning deficits and aberrant neuronal alternative polyadenylation. Elife 9, (2020).

Alcott, C. E. et al. Partial loss of CFIM25 causes learning deficits and aberrant neuronal alternative polyadenylation. Elife 9, (2020).

An J.J., K. Gharami G.Y, Liao, N.H. Woo, A.G. Lau, F. Vanevski, E.R. Torre, K.R. Jones, Y. Feng, B. Lu, and B. Xu. Distinct role of long 3′ UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. Cell. 134:175–187 (2008).

Baser, A. et al. Onset of differentiation is post-transcriptionally controlled in adult neural stem cells. Nature 566, 100–104 (2019).

Baser, A., Skabkin, M. & Martin-Villalba, A. Neural Stem Cell Activation and the Role of Protein Synthesis. Brain Plast. (Amsterdam, Netherlands) 3, 27–41 (2017).

Bava, F.-A. et al. CPEB1 coordinates alternative 3'-UTR formation with translational regulation. Nature 495, 121–125 (2013).

Becht, E., McInnes, L., Healy, J. et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 37, 38–44 (2019). https://doi.org/10.1038/nbt.4314

Begg C.B., Gray R. (1984). Calculation of Polychotomous Logistic Regression Parameters Using Individualized Regressions.Biometrika,71, 11–18.

Blair, J.D., Hockemeyer, D., Doudna, J. A., Bateup, H. S. & Floor, S. N. Widespread Translational Remodeling during Human Neuronal Differentiation. Cell Rep. 21, 2005–2016 (2017).

Bohning D. (1992). Multinomial Logistic Regression Algorithm. Annals of the Inst. Of Statistical Math.,44, 197–200.

Boldrini M., Fulmore C.A., Tartt A.N., et al. Human Hippocampal Neurogenesis Persists throughout Aging. Cell Stem Cell. 22(4):589-599 (2018). doi:10.1016/j.stem.2018.03.015

Bond A.M., Ming G.L., Song H. Adult Mammalian Neural Stem Cells and Neurogenesis: Five Decades Later. Cell Stem Cell. 1;17(4):385-95 (2015). doi: 10.1016/j.stem.2015.09.003.

Brumbaugh, J. et al. Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling. Cell 172, 106-120.e21 (2018).

Busch, R.M., Srivastava, S., Hogue, O. et al. Neurobehavioral phenotype of autism spectrum disorder associated with germline heterozygous mutations in PTEN. Transl Psychiatry 9, 253 (2019). https://doi.org/10.1038/s41398-019-0588-1

Cao, Q., Huang, Y.-S., Kan, M.-C. & Richter, J. D. Amyloid precursor proteins anchor CPEB to membranes and promote polyadenylation-induced translation. Mol. Cell. Biol. 25, 10930–10939 (2005).

Charlesworth, A., Cox, L. L. & MacNicol, A. M. Cytoplasmic polyadenylation element (CPE)- and CPE-binding protein (CPEB)-independent mechanisms regulate early class maternal mRNA translational activation in Xenopus oocytes. J. Biol. Chem. 279, 17650–17659 (2004).

Chen, C.-Y., Chen, S.-T., Juan, H.-F. & Huang, H.-C. Lengthening of 3'UTR increases with morphological complexity in animal evolution. Bioinformatics 28, 3178–3181 (2012).

Chen, M. et al. 3' UTR lengthening as a novel mechanism in regulating cellular senescence. Genome Res. 28, 285–294 (2018).

Chen, Y.-C., Chang, Y.-W. & Huang, Y.-S. Dysregulated Translation in Neurodevelopmental Disorders: An Overview of Autism-Risk Genes Involved in Translation. Dev. Neurobiol. 79, 60–74 (2019).

Clark, S. J., Argelaguet, R., Kapourani, C. A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., & Reik, W. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature communications*, *9*(1), 781. https://doi.org/10.1038/s41467-018-03149-4

Coffey, K. R., Marx, R. G. & Neumaier, J. F. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol. 44, 859–868 (2019).

Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol. Cell. Proteomics 13, 2513–2526 (2014).

Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. 10, 1794–1805 (2011).

Eberwine, J. et al. Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. USA* 89, 3010–3014 (1992).

Elkon, R. et al. E2F mediates enhanced alternative polyadenylation in proliferation. Genome Biol. 13, R59 (2012).

Eriksson, P., Perfilieva, E., Björk-Eriksson, T. et al. Neurogenesis in the adult human hippocampus. Nat Med 4, 1313–1317 (1998). https://doi.org/10.1038/3305

Ernst, A. et al. Neurogenesis in the striatum of the adult human brain. Cell 156, 1072–1083 (2014).

Faras, H., Al Ateeqi, N., & Tidmarsh, L. Autism spectrum disorders. Annals of Saudi medicine, 30(4), 295–300 (2010). https://doi.org/10.4103/0256-4947.65261

Faye, M. D., Graber, T. E., Holcik, M. Assessment of Selective mRNA Translation in Mammalian Cells by Polysome Profiling. J. Vis. Exp. (92), e52295, doi:10.3791/52295 (2014).

Fernandez-Moya, S. M., Bauer, K. E. & Kiebler, M. A. Meet the players: local translation at the synapse. Front. Mol. Neurosci. 7, 84 (2014).

Gillis J. & Paul P. "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks. PLOS (2012), doi: https://doi.org/10.1371/journal.pcbi.1002444.

Groisman, I. et al. Control of cellular senescence by CPEB. Genes Dev. 20, 2701–2712 (2006).

Gruber, A. R. et al. Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. Nat. Commun. 5, 5465 (2014).

Guffanti, G., Bartlett, A., Klengel, T., Klengel, C., Hunter, R., Glinsky, G., & Macciardi, F. (2018). Novel Bioinformatics Approach Identifies Transcriptional Profiles of Lineage-Specific Transposable Elements at Distinct Loci in the Human Dorsolateral Prefrontal Cortex. Molecular biology and evolution, 35(10), 2435–2453. https://doi.org/10.1093/molbev/msy143

Guvenek, A. & Tian, B. Analysis of alternative cleavage and polyadenylation in mature and differentiating neurons using RNA-seq data. Quant. Biol. (Beijing, China) 6, 253–266 (2018).

Hastie T., Tibshirani R., Friedman J. (2009) The Elements of Statistical Learning: DataMining, Inference and Prediction. 2nd. edition. Springer-Verlag.

Hastie, T. J. (1992) Generalized additive models. Chapter 7 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 50, 96 (2018).

Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010).

Huang, C.-W. et al. Conditional Knockout of Breast Carcinoma Amplified Sequence 2 (BCAS2) in Mouse Forebrain Causes Dendritic Malformation via β-catenin. Sci. Rep. 6, 34927 (2016).

Huber, W., Carey, V., Gentleman, R. et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12, 115–121 (2015). https://doi.org/10.1038/nmeth.325

Hughes, C. S. et al. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. Nat. Protoc. 14, 68–85 (2019).

Hughes, C. S. et al. Ultrasensitive proteome analysis using paramagnetic bead technology. Mol. Syst. Biol. 10, 757 (2014).

Ivshina, M., Lasko, P. & Richter, J. D. Cytoplasmic polyadenylation element binding proteins in development, health, and disease. Annu. Rev. Cell Dev. Biol. 30, 393–415 (2014).

Jäkel S., Dimou L. Glial Cells and Their Function in the Adult Brain: A Journey through the History of Their Ablation. Front Cell Neurosci. 13;11:24 (2017).

Jereb, S., Hwang, H. W., Van Otterloo, E., Govek, E. E., Fak, J. J., Yuan, Y., Hatten, M. E., & Darnell, R. B. (2018). Differential 3' Processing of Specific Transcripts Expands Regulatory and Protein Diversity Across Neuronal Cell Types. eLife, 7, e34042. https://doi.org/10.7554/eLife.34042

Kalamakis, G. et al. Quiescence Modulates Stem Cell Maintenance and Regenerative Capacity in the Aging Brain. Cell 176, 1407-1419.e14 (2019).

Kazdoba, T. M. et al. Translational Mouse Models of Autism: Advancing Toward Pharmacological Therapeutics. Curr. Top. Behav. Neurosci. 28, 1–52 (2016).

Kempermann G., Gage F.H., Aigner L., Song H., Curtis M.A., Thuret S., Kuhn H.G., Jessberger S., Frankland P.W., Cameron H.A., Gould E., Hen R., Abrous D.N., Toni N., Schinder A.F., Zhao X., Lucassen P.J., Frisén J. Human Adult Neurogenesis: Evidence and Remaining Questions. Cell Stem Cell. 23(1):25-30 (2018).

Kriegstein A. & Alvarez-Buylla A. The glial nature of embryonic and adult neural stem cells. Annu Rev Neurosci. 32:149-84 (2009). doi: 10.1146/annurev.neuro.051508.135600.

La Manno, G., Soldatov, R., Zeisel, A. et al. RNA velocity of single cells. Nature 560, 494–498 (2018). https://doi.org/10.1038/s41586-018-0414-6

Lackford B., Yao C., Charles G.M., et al. Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. EMBO J. 33(8):878-889 (20014).

Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

Lawrence, M. et al. Software for computing and annotating genomic ranges. PLoS Comput. Biol. 9, e1003118 (2013).

Levina E. & Bickel P. The EarthMover's Distance is the Mallows Distance: Some Insights from Statistics. Proceedings of ICCV 2001. Vancouver, Canada: 251–256.

Linnarsson S. Sequencing Single Cells Reveals Sequential Stem Cell States. Cell Stem Cell. 17(3):251-252 (2015). doi:10.1016/j.stem.2015.08.016.

Lledo, P. M., & Valley, M. Adult Olfactory Bulb Neurogenesis. Cold Spring Harbor perspectives in biology, 8(8), a018945 (2016). https://doi.org/10.1101/cshperspect.a018945

Llorens-Bobadilla E., Zhao S., Baser A., Saiz-Castro G., Zwadlo K., Martin-Villalba A. Single-Cell Transcriptomics Reveals a Population of Dormant Neural Stem Cells that Become Activated upon Brain Injury. Cell Stem Cell. 17(3):329-340 (2015). doi:10.1016/j.stem.2015.07.002

Love M.I., Huber W. and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biology, 15:550 (2014).

Martynoga, B. et al. Epigenomic enhancer annotation reveals a key role for NFIX in neural stem cell quiescence. Genes Dev. 27, 1769–1786 (2013).

Mariella, E., Marotta, F., Grassi, E., Gilotto, S., & Provero, P. The Length of the Expressed 3' UTR Is an Intermediate Molecular Phenotype Linking Genetic Variants to Complex Diseases. Frontiers in genetics, 10, 714 (2019). https://doi.org/10.3389/fgene.2019.00714

Mayr, C. & Bartel, D. P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138, 673–684 (2009).

Ming, G.L. & Song, H. Adult neurogenesis in the Mammalian brain: significant answers and significant questions. Neuron, 70 (2011), pp. 687-702

Miura, P., Sanfilippo, P., Shenker, S. & Lai, E. C. Alternative polyadenylation in the nervous system: to what lengths will 3' UTR extensions take us? Bioessays 36, 766–777 (2014).

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O. & Lai, E. C. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res. 23, 812–825 (2013).

Moy, S. S. et al. Sociability and preference for social novelty in five inbred strains: an approach to assess autistic-like behavior in mice. Genes. Brain. Behav. 3, 287–302 (2004).

Müller U.C., Deller T., Korte M. Not just amyloid: physiological functions of the amyloid precursor protein family. Nat Rev Neurosci. 2017 May;18(5):281-298. doi: 10.1038/nrn.2017.29. Epub 2017 Mar 31. PMID: 28360418.

Ovchinnikova, S. & Anders, S. Exploring dimension-reduced embeddings with Sleepwalk. Genome Res. 30, 749–756 (2020).

Parras, A. et al. Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing. Nature 560, 441–446 (2018).

Patrick, R., Humphreys, D.T., Janbandhu, V. et al. Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. Genome Biol 21, 167 (2020).

Picelli, S., Faridani, O., Björklund, Å. et al. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc 9, 171–181 (2014). https://doi.org/10.1038/nprot.2014.006

Piñero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 48, D845–D855 (2020).

Piqué, M., López, J. M., Foissac, S., Guigó, R. & Méndez, R. A combinatorial code for CPE-mediated translational control. Cell 132, 434–448 (2008).

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Richter, J. D. CPEB: a life in translation. Trends Biochem. Sci. 32, 279–285 (2007).
Ripley, B. D. Pattern Recognition and Neural Networks. Cambridge (1996).

Robles, J.A., Qureshi, S.E., Stephen, S.J. et al. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics 13, 484 (2012). https://doi.org/10.1186/1471-2164-13-484

Kelly R., T., Single-Cell Proteomics: Progress and Prospects Molecular & Cellular Proteomics August 26, 2020, mcp.R120.002234; DOI: 10.1074/mcp.R120.002234

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science 320, 1643–1647 (2008).

Scattoni, M. L., Gandhy, S. U., Ricceri, L. & Crawley, J. N. Unusual repertoire of vocalizations in the BTBR T+tf/J mouse model of autism. PLoS One 3, e3067 (2008).

Schilling, S., Mehr, A., Ludewig, S., Stephan, J., Zimmermann, M., August, A., Strecker, P., Korte, M., Koo, E. H., Müller, U. C., Kins, S., & Eggert, S. (2017). APLP1 Is a Synaptic Cell Adhesion Molecule, Supporting Maintenance of Dendritic Spines and Basal Synaptic Transmission. The Journal of neuroscience : the official journal of the Society for Neuroscience, 37(21), 5345–5365. https://doi.org/10.1523/JNEUROSCI.1875-16.2017

Schriml, L. M. et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. 47, D955–D962 (2019).

Shin J., Berg D.A., Zhu Y., et al. Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. Cell Stem Cell. 17(3):360-372 (2015). doi:10.1016/j.stem.2015.07.013

Silverman, J. L., Yang, M., Lord, C. & Crawley, J. N. Behavioural phenotyping assays for mouse models of autism. Nat. Rev. Neurosci. 11, 490–502 (2010).

Sommerkamp P, Altamura S, Renders S, Narr A, Ladel L, Zeisberger P, Eiben PL, Fawaz M, Rieger MA, Cabezas-Wallscheid N, Trumpp A. Differential Alternative Polyadenylation Landscapes Mediate Hematopoietic Stem Cell Activation and Regulate Glutamine Metabolism. Cell Stem Cell. 2020 May 7;26(5):722-738.e7. doi: 10.1016/j.stem.2020.03.003. Epub 2020 Mar 30. PMID: 32229311.

Sungur, A. Ö. et al. Aberrant cognitive phenotypes and altered hippocampal BDNF expression related to epigenetic modifications in mice lacking the post-synaptic scaffolding protein SHANK1: Implications for autism spectrum disorder. Hippocampus 27, 906–919 (2017).

Sungur, A. Ö., Schwarting, R. K. W. & Wöhr, M. Early communication deficits in the Shank1 knockout mouse model for autism spectrum disorder: Developmental aspects and effects of social context. Autism Res. 9, 696–709 (2016).

Szkop, K. J. et al. Dysregulation of Alternative Poly-adenylation as a Potential Player in Autism Spectrum Disorder. Front. Mol. Neurosci. 10, 279 (2017).

Tikhonov, M., Georgiev, P., & Maksimenko, O. Competition within Introns: Splicing Wins over Polyadenylation via a General Mechanism. Acta naturae, 5(4), 52–61 (2013).

Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386 (2014).

Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods 13, 731–740 (2016).

Velmeshev, D. et al. Single-cell genomics identifies cell type-specific molecular changes in autism. Science 364, 685–689 (2019).

Velten, L., Anders, S., Pekowska, A., Järvelin, A. I., Huber, W., Pelechano, V., & Steinmetz, L. M. Single-cell polyadenylation site mapping reveals 3' isoform choice variability. Molecular systems biology, 11(6), 812 (2015). https://doi.org/10.15252/msb.20156198

Venables, W. N. and Ripley, B. D. Modern Applied Statistics with S. Fourth edition. Springer (2002).

von Koch, C. S., Zheng, H., Chen, H., Trumbauer, M., Thinakaran, G., van der Ploeg, L. H., Price, D. L., & Sisodia, S. S. (1997). Generation of APLP2 KO mice and early postnatal lethality in APLP2/APP double KO mice. Neurobiology of aging, 18(6), 661–669. https://doi.org/10.1016/s0197-4580(97)00151-6

Wanke, K. A., Devanna, P. & Vernes, S. C. Understanding Neurodevelopmental Disorders: The Promise of Regulatory Variation in the 3'UTRome. Biol. Psychiatry 83, 548–557 (2018).

Wheeler, E. C., Van Nostrand, E. L., & Yeo, G. W. (2018). Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. Wiley interdisciplinary reviews. RNA, 9(1), e1436. https://doi.org/10.1002/wrna.1436

Weill, L., Belloc, E., Bava, F.-A. & Méndez, R. Translational control by changes in poly(A) tail length: recycling mRNAs. Nat. Struct. Mol. Biol. 19, 577–585 (2012).

Xia, Z., Donehower, L., Cooper, T. et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat Commun 5, 5274 (2014), doi: https://doi.org/10.1038/ncomms6274.

Yee T. J. The VGAM Package for Categorical Data Analysis. Journal of Statistical Software, 32(10) (2010).

Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16, 284–287 (2012).

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S. (2008) Model-based Analysis of ChIP-Seq (MACS), Genome Biology, 2008;9(9):R137.

Zheng, G., Terry, J., Belgrader, P. et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 8, 14049 (2017). https://doi.org/10.1038/ncomms14049

Publication related to this paper:

Single cell 3'UTR analysis identifies changes in alternative polyadenylation throughout neuronal differentiation and in autism

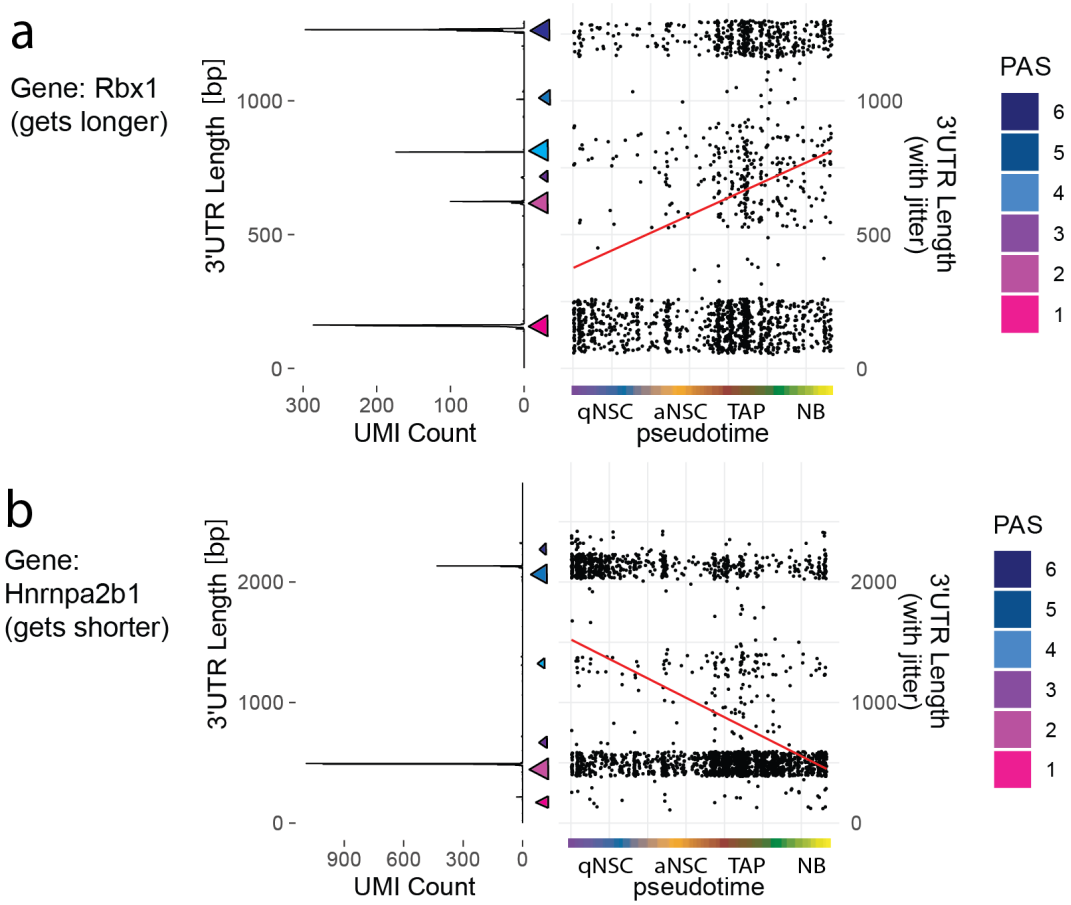
Manuel Göpferich, Nikhil Oommen George, Ana Domingo Muelas, Alex Bizyn, Rosa Pascual, Daria Fijalkowska, Georgios Kalamakis, Ulrike Müller, Jeroen Krijgsveld, Raul Mendez, Isabel Fariñas, Wolfgang Huber, Simon Anders, Ana Martin-Villalba

bioRxiv 2020.08.12.247627; doi: https://doi.org/10.1101/2020.08.12.247627

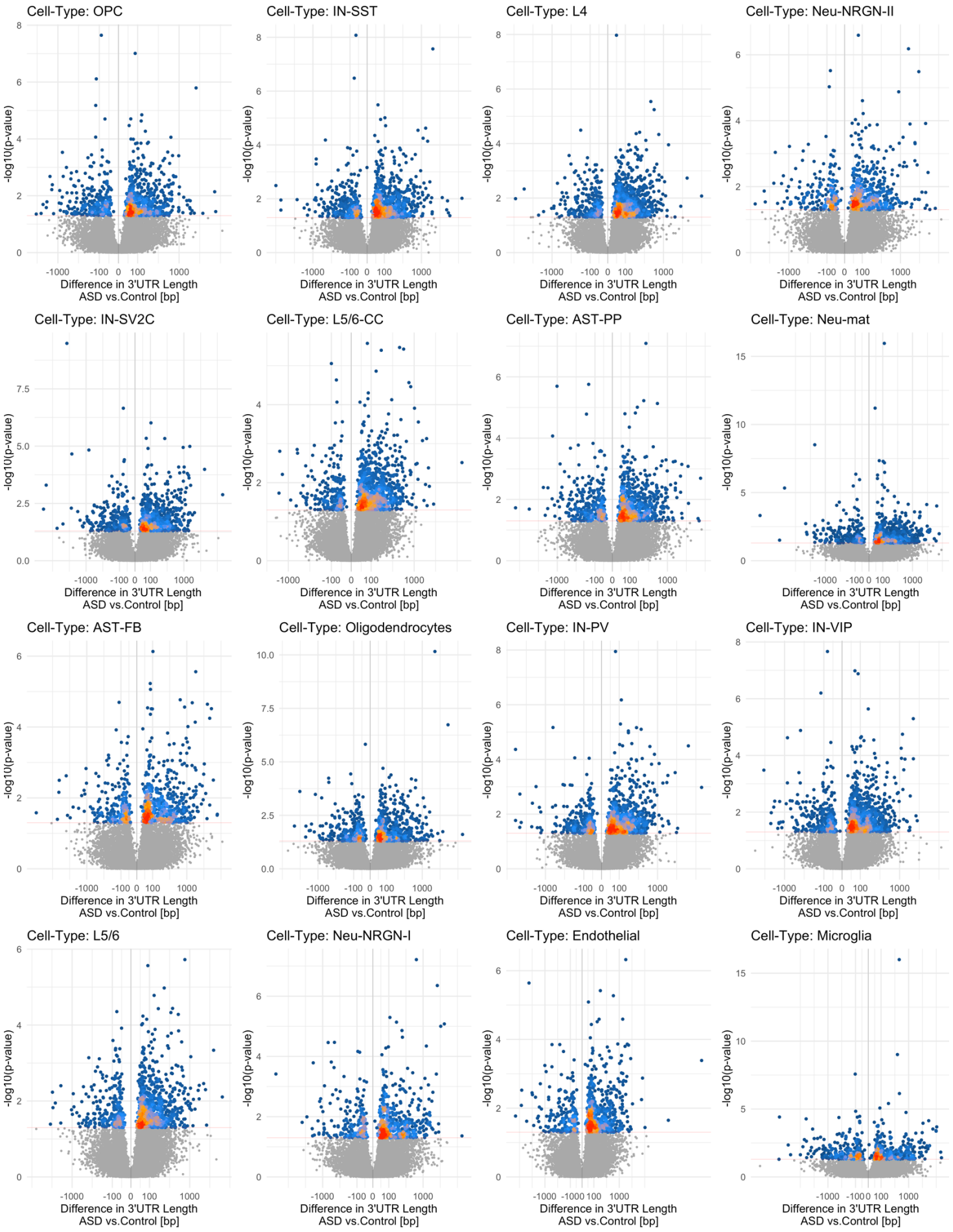
Many of the results of this preprint are presented in this doctoral thesis. My contribution to this paper was developing and executing virtually all bioinformatical analyses (calling peaks 3'UTR, enrichment analysis, etc.). Single cell RNAseq data was generated by Georgios Kalamakis and Nikhil George (see Supplementary Table 1 for details) who also conducted proteomics, behavioral studies in mice and further wet-lab validation experiments. Proteomics data was preprocessed by Daria Fijalkowska. The RNA immunoprecipitation assay was done by Ana Domingo Muelas, Alex Bizyn and Rosa Pascual. Protocols and methods applied for these wet-lab experiments can be retrieved from the preprint paper and are not contained in this thesis as they were not my own work. Ribosomal profiling data was generated in the Ana Martin-Villalba lab by Maxim Skabkin and Damian Carvajal Ibanez (unpublished data). Many of the data analyses presented here were discussed and (partly) developed together with Simon Anders (member of my thesis advisory committee), Ana Martin-Villalba (1st supervisor) and Wolfgang Huber (2nd supervisor). Furthermore, Ulrike Müller, Jeroen Krijgsveld, Raul Mendez and Isabel Fariñas contributed to the project mainly by sharing their advice regarding the biological clues. A detailed list of contributors and their affiliations is presented in the before mentioned preprint paper. Single cell RNAseq is available with the GEO accession number: GSE115626 (Kalamakis G. et al., 2019). The newly generated single cell RNAseq datasets and the immunoprecipitation assay can be accessed under: GSE156767, as private submission only for the peer-review process. Proteomics data is deposited under the identifier PXD020971 on the *ProteomeXchange Consortium* sever.

Supplements

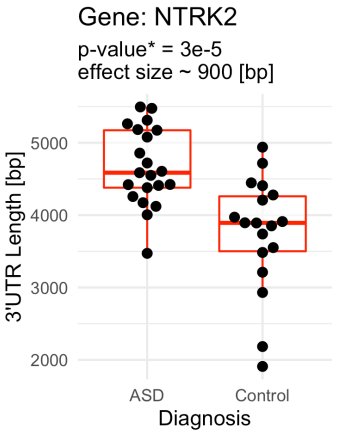
Supplementary Figures



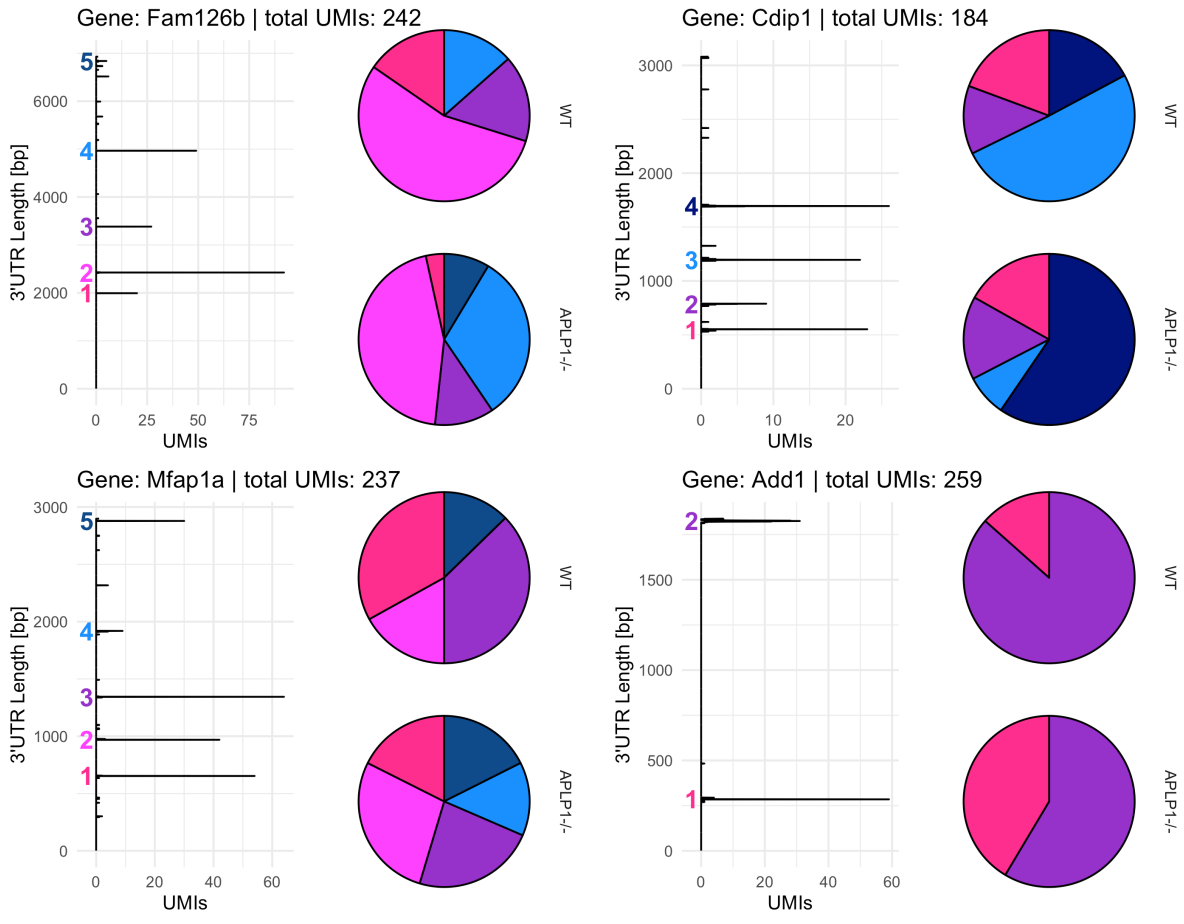
Supplementary Figure 1: a-b, example genes for 3'UTR changes in NSCs, raw 3' mapping positions summed over cells, mid sub-panels: mean mapping position per cell, linear regression lines fitted (in red), color code as dark blue to red for most distal to most proximal PAS (color code corresponds to left sub-panels), **a**, Rbx1 (ring-box 1, also known as ROC1) example for 3'UTR lengthening, **b**, Hnrnpa2b1 (heterogeneous nuclear ribonucleoprotein A2/B1) example for 3'UTR shortening.


Cell-Type: OPC
Cell-Type: IN-SST
Cell-Type: L4
Cell-Type: Neu-NRGN-II
Cell-Type: IN-SV2C
Cell-Type: L5/6-CC
Cell-Type: AST-PP
Cell-Type: Neu-mat
Cell-Type: AST-FB
Cell-Type: Oligodendrocytes
Cell-Type: IN-PV
Cell-Type: IN-VIP
Cell-Type: L5/6
Cell-Type: Neu-NRGN-I
Cell-Type: Endothelial
Cell-Type: Microglia
-log10(p-value)
Difference in 3'UTR Length
ASD vs.Control [bp]

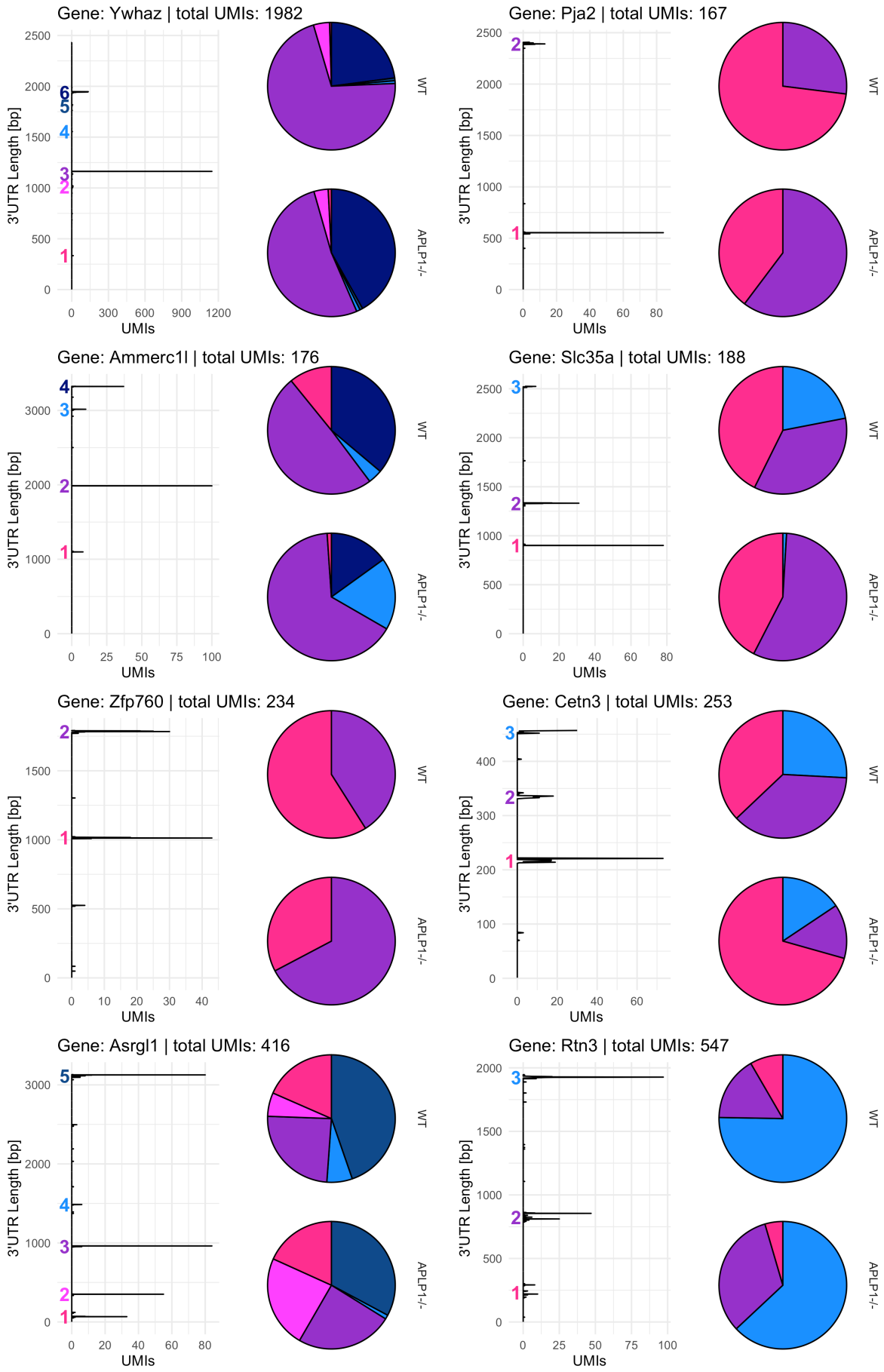
Supplementary Figure 2: 3'UTR lengthening in ASD vs. control group; each entry (volcano plot) represents one cell-type: x-axis depicts difference in 3'UTR length [bp] between ASD and controls; y-axis shows -log10 of p-values (estimated from linear models), over-plotting color scale applied to genes with uncorrected p-values < 0.05 to show the overall trend, extremely low p-values were restricted to the value of 1e-16.



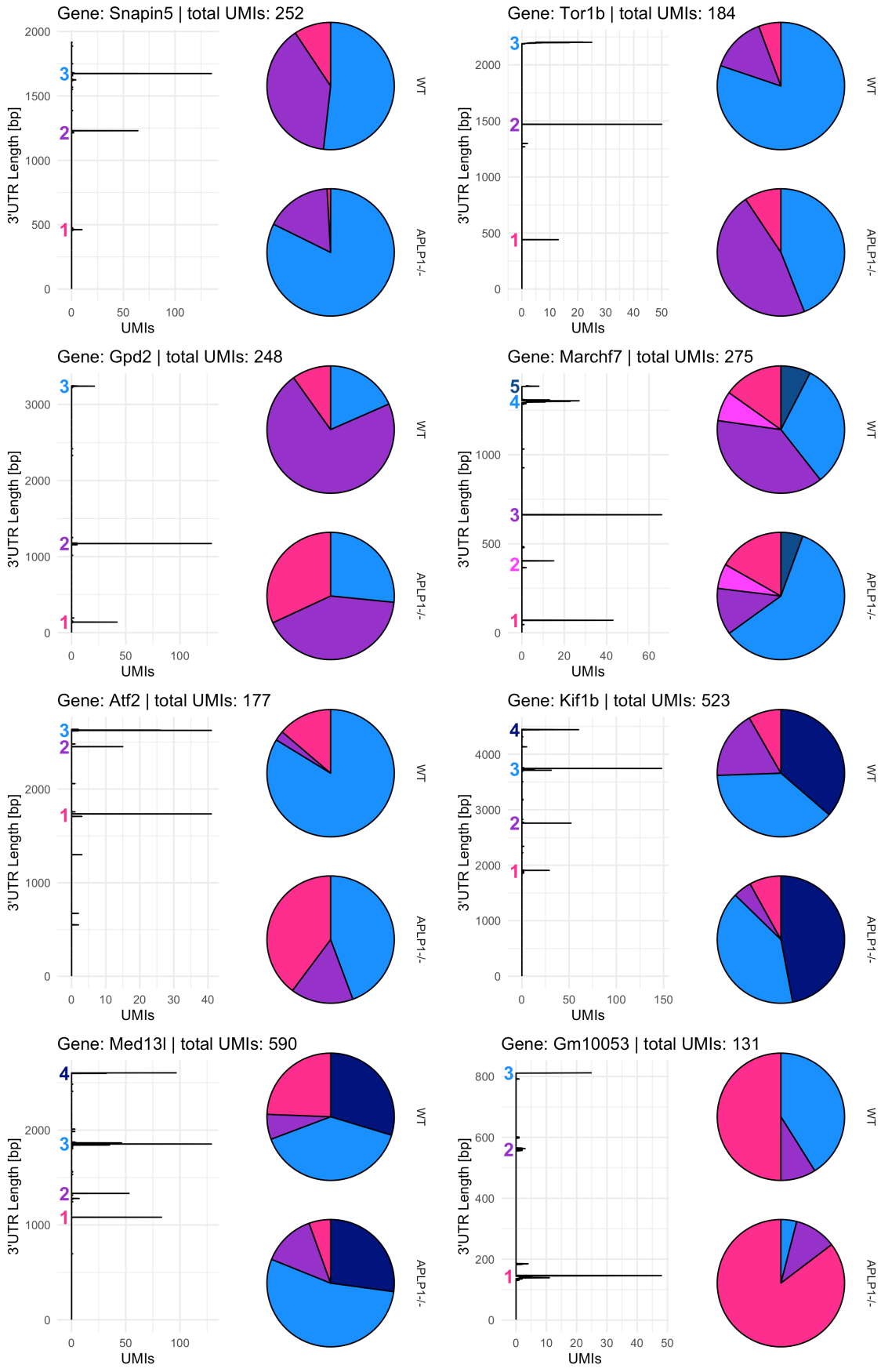
Supplementary Figure 3, example gene NTRK2 for 3'UTR lengthening in ASD vs. control in astrocytes (AST-FB), linear model fit on 3'UTR length.



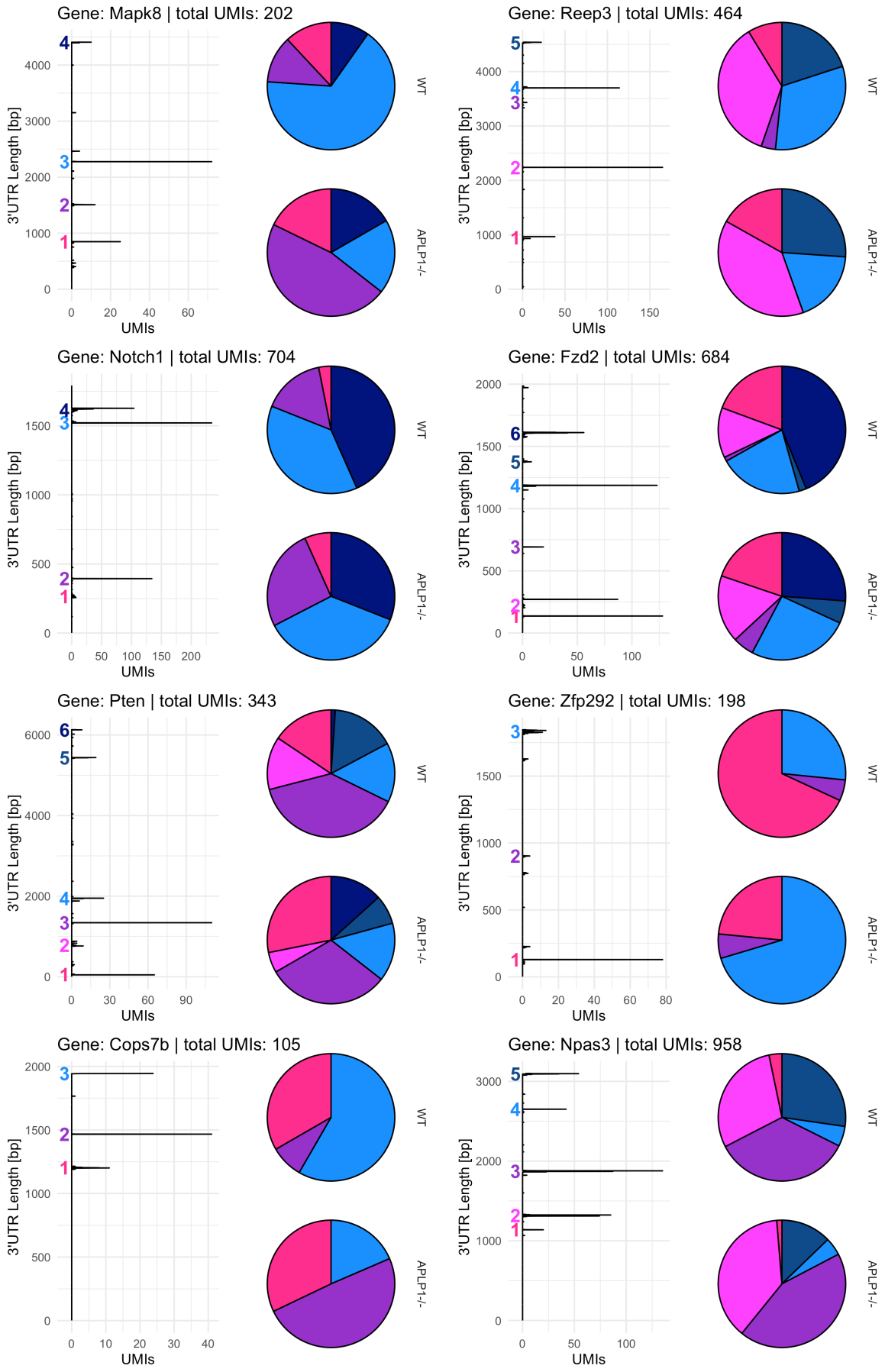
(Figure continuous on the next page)



(Figure continuous on the next page)

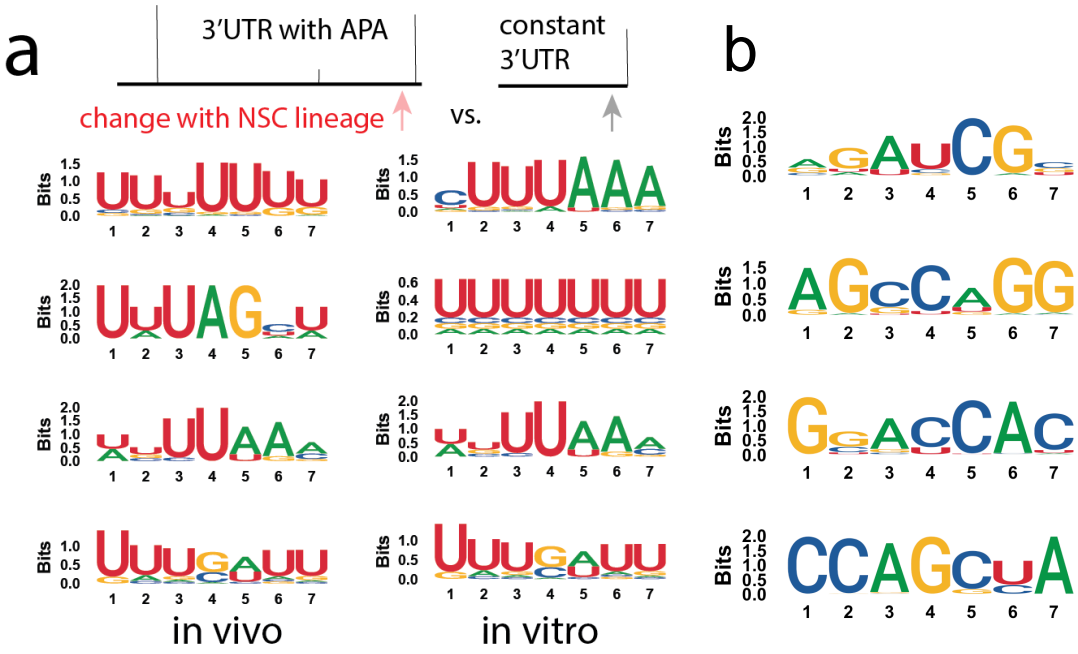


(Figure continuous on the next page)

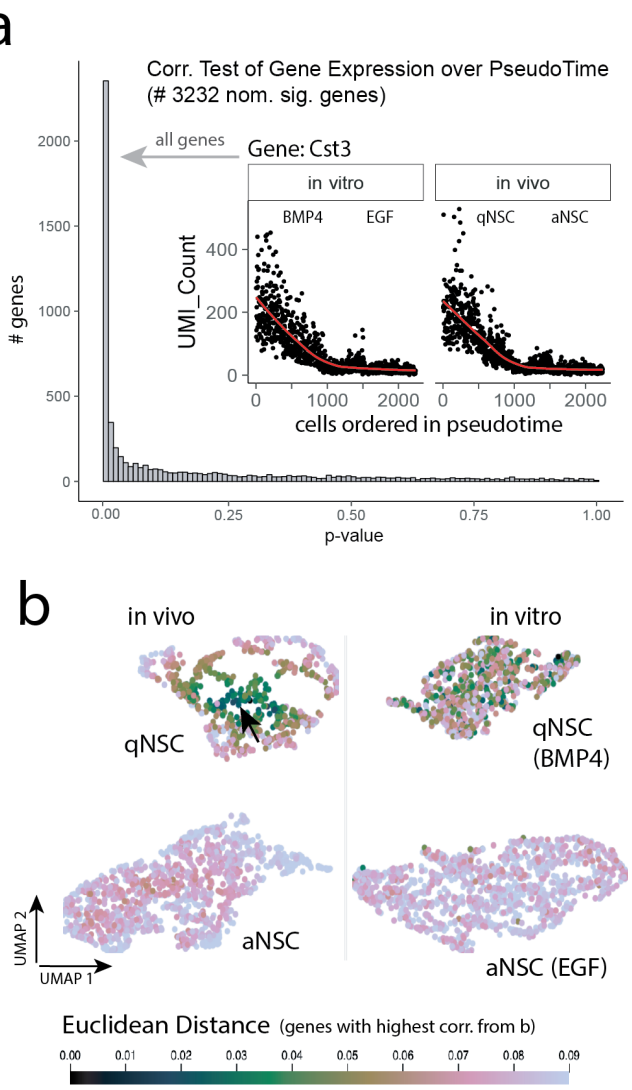


(Figure caption on the next page)

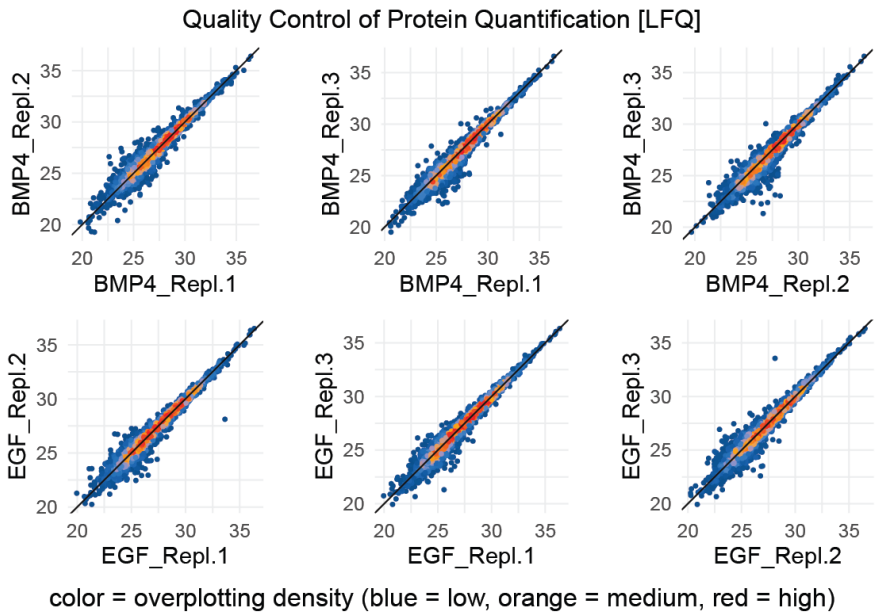
Supplementary Figure 3: Differential 3'UTR usage in APLP1-/- vs. WT, examples for significant hits from the MNR model, every entry: one gene with its 3'peak distribution (left panels), total UMIs for this gene, annotated PAS as numbers from most proximal to most distal, right sub-panels: fraction of UMIs in each PAS (by color code from magenta to dark blue), plotted as pie charts for APLP1-/- and WT, respectively.



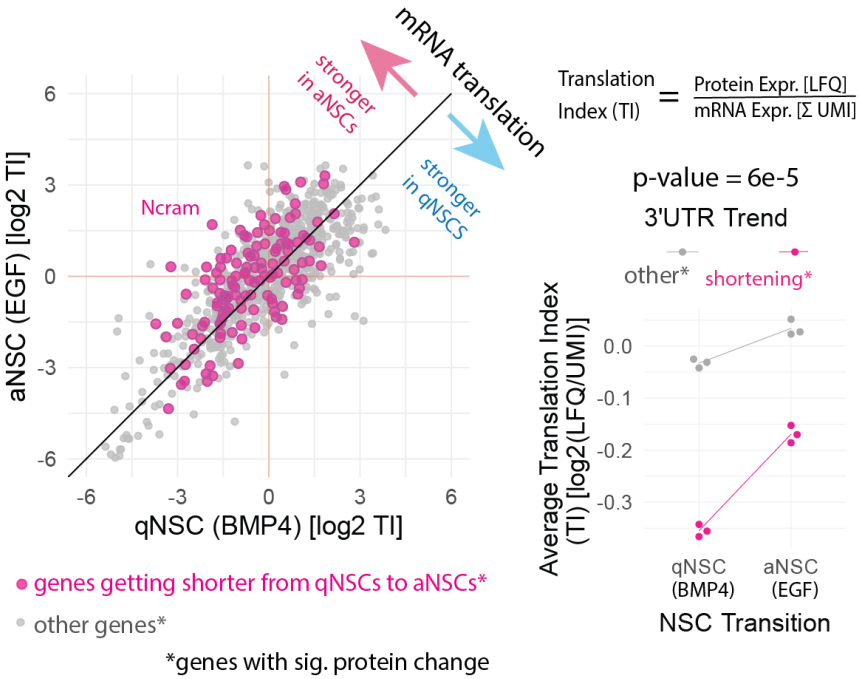
Supplementary Figure 4: a, Motif detection in 3'UTRs (murine NSC lineage), De-novo motif analysis utilizing the homer2 tool (Methods) in 3'UTRs that change along the NSC lineage, left result for in vivo and right for in vivo NSCs. **b**, Motif detection in 3'UTRs comparing ASD vs. controls in L2/3 excitatory neurons, de-novo motif analysis, motifs enriched in distal 3'UTRs getting shorter in ASD vs. controls compared to 3'UTRs getting longer in ASD vs. controls



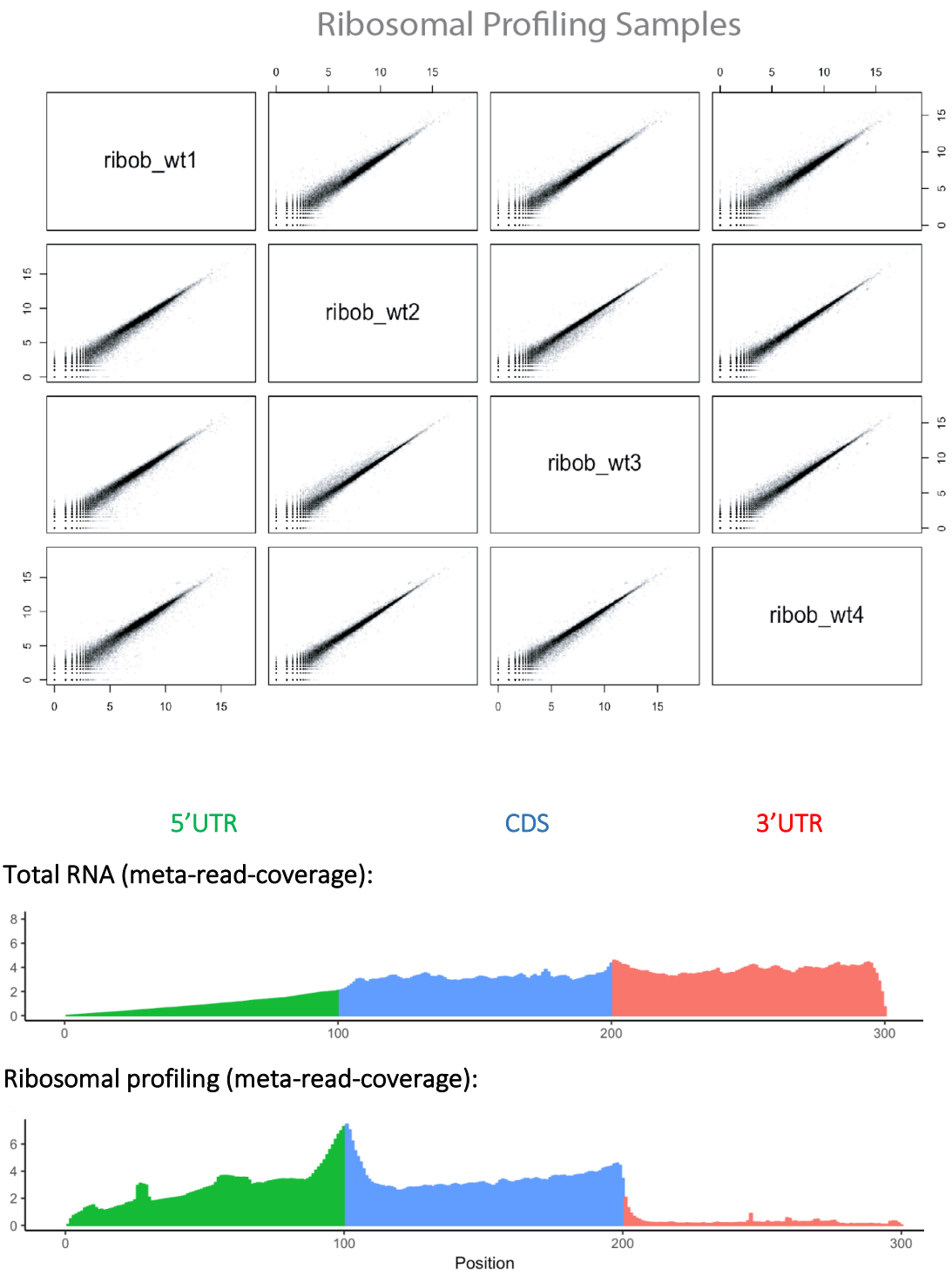
Supplementary Figure 5: Comparison of expression trends (in in vitro vs. in vivo NSCs) Statistical test whether genes follow the same expression trend along the NSC lineage in in vitro NSCs as they do in vivo NSCs, outer panel, p-value histogram, inner panel, example gene Cst3 (Cystatin C), every point is a single cell, red regression line to show the trend, nominal p-value. c, 2-D projection of in vivo and in vitro NSCs on the same UMAP, plot split for visualization, batch correction by mutual nearest neighbors (MNN), color code: Euclidian distance of in vivo qNSCs (see cursor position) to all other cells (black/ blue if cells are similar) by the sleepwalk tool.



Supplementary Figure 6: Proteomics results, scatterplot matrix of biological replicates, protein quantification in cultured NSCs by mass spectrometry, upper row BMP4 (qNSCs), lower row EGF (aNSCs), quantified as LFQ values.



Supplementary Figure 7: Comparison of translation index to shortening of 3'UTRs, left sub-panel, per gene translation index: non-degraded protein [label free quantification (LFQ)] per non-degraded mRNA [UMIs], for aNSCs and qNSCs, shown for with significant protein changes comparing aNSC to qNSCs, shortening genes in magenta, right sub-panel, average translation index per gene cluster (UTR shortening vs. others), every point represents one replicate, ANOVA interaction test.



Supplementary Figure 8: Ribosomal profiling of in vitro active NSCs (EGF treated), n = 4 replicates, sample correlation scatter matrix, every point: ribosome protected reads mapping to the CDS (coding region) of a particular gene [counts shown as logarithmic scale] (upper panel), representative (same) sample with total RNA sequencing and ribosome protected reads (lower panels), meta-transcriptome coverage (y-axis), per binned position (x-axis).

6.2 Supplementary Tables

Supplementary Table 1: Overview of scRNAseq datasets, species: Mus musculus, data generated in the Ana Martin-Villalba lab, sequenced at the DKFZ, 3'peak calling applied; mapping rate to the mm10 genome (*bowtie2*); 'mice' is the number of mice pooled in one sequencing run (sample); HTO = hash-tag-oligos were used to assign single cells to individual mice (see Methods), yes (TRUE) or not (FALSE).

Sample_ID	Experimenter	Genotype	NSC isolation	Mapping Rate	mice	cells	HTOs used
old_NSC	Kalamakis G	WT	freshly, FACS	95.55%	8	1716	FALSE
young_NSC	Kalamakis G	WT	freshly, FACS	95.45%	4	2035	FALSE
EGF_NSC	George N	WT	cultured	95.34%	3	1410	FALSE
BMP4_NSC	George N	WT	cultured	95.65%	3	887	FALSE
APLP1_KO	George N	APLP1-/-	freshly, FACS	96.04%	3	1879	TRUE
APLP1_WT	George N	WT	freshly, FACS	96.17%	3	2209	TRUE

Supplementary Table 2: Differential expression in L2/3 neurons (ASD vs. control, FDR < 5%, pseudo-bulks, DESeq2), genes sorted by increasing LFC (from most downregulated in ASD to most upregulated)

Gene	LFC	p.adjust	UMI
CXorf40B	-1,04	0,02	645
AL158154.2	-1,04	0,034	516
AC008403.3	-0,9	0,036	554
SLC25A45	-0,88	0,02	1019
NANS	-0,86	0,021	999
SAT2	-0,84	0,02	15548
SNHG12	-0,81	0,02	978
PTRHD1	-0,8	0,046	1692
FSIP1	-0,8	0,031	817
VILL	-0,79	0,029	931
TRAPPC6A	-0,78	0,028	777
MRPL57	-0,74	0,047	4233
RRP9	-0,71	0,02	692
C16orf91	-0,69	0,046	1278
MRPL9	-0,67	0,008	1587
KMT5C	-0,67	0,034	1701

UBXN1	-0,64	0,046	3244
FASTKD3	-0,6	0,044	574
NUDT22	-0,6	0,02	1613
LRRC23	-0,6	0,044	978
BICDL1	-0,6	0,022	4828
SDHB	-0,58	0,036	2236
NELFE	-0,58	0,044	2308
HDDC3	-0,55	0,037	1543
CORO7	-0,54	0,02	1364
CIRBP	-0,52	0,021	20761
MRPS15	-0,51	0,044	2353
CIAO1	-0,51	0,02	3430
MBD4	-0,51	0,02	3293
TSC2	-0,51	0,02	7285
TCF25	-0,51	0,021	18242
NOSIP	-0,51	0,046	2235
MRPS25	-0,5	0,021	4618
SDR39U1	-0,5	0,011	2874
COQ6	-0,47	0,028	1459
RNF8	-0,46	0,046	3439
SRA1	-0,45	0,046	1776
DHPS	-0,44	0,046	4106
PSMG1	-0,43	0,046	1276
SMARCA4	-0,36	0,046	5754
NDUFA10	-0,29	0,036	10734
YY1	0,32	0,046	7459
PRRC2B	0,36	0,02	10656
USP10	0,39	0,046	3234
THRB	0,4	0,046	12087
ZBTB4	0,4	0,01	5503
DNAJC16	0,43	0,046	4208
SSBP2	0,43	0,021	7153
PRICKLE1	0,43	0,044	4389
CIPC	0,43	0,046	2739
AJAP1	0,45	0,049	5944
HPCAL4	0,45	0,034	8629
KLHL2	0,45	0,049	6409
HERC3	0,47	0,048	2215
GTF3C4	0,47	0,046	1823
NR2C2	0,51	0,046	3060
ZBTB44	0,51	0,036	6980
PGAM5	0,51	0,046	1673

GLRA3	0,52	0,036	8300
RYBP	0,53	0,046	3730
ABLIM3	0,53	0,02	1320
NOL4	0,53	0,046	6308
MOB1B	0,54	0,049	4389
PPP3CA	0,56	0,036	56807
SLITRK4	0,56	0,044	14964
RAPH1	0,57	0,046	7159
TMED7	0,58	0,031	2399
ZBTB6	0,59	0,034	1026
ZNF770	0,59	0,02	5512
RAB8B	0,59	0,036	1620
TMEM151A	0,6	0,036	5376
ZFY	0,61	0,046	789
NFIA	0,65	0,044	2864
ADGRF5	0,65	0,046	1978
AL365361.1	0,67	0,044	2084
ZMIZ1	0,69	0,046	4183
SPATA2	0,69	0,044	1920
ATP1B2	0,7	0,044	844
GPR12	0,72	0,048	937
DACH1	0,89	0,012	1071
ZBTB34	1	0,011	1054
WNT7B	1,21	0,02	674
MT-ND3	1,21	0,046	84738

Supplementary Table 3: Genes differentially expressed in NSCs, Aplp1-/- vs. WT (DESeq2).

Gene	ENSEMBL ID	LFC	p.adjust	UMI
Pmpcb	ENSMUSG00000029017	-2,23	0,05	122
Lypd6	ENSMUSG00000050447	-2,13	0,05	112
Ostc	ENSMUSG00000041084	-1,95	0,03	142
Tmem237	ENSMUSG00000038079	-1,94	0,09	129
Sdc2	ENSMUSG00000022261	-1,93	0,04	175
Cstf2	ENSMUSG00000031256	-1,89	0,08	103
Phactr3	ENSMUSG00000027525	-1,89	0,10	122
Rplp1	ENSMUSG00000007892	-1,88	0,03	145
Mpc1	ENSMUSG00000023861	-1,87	0,10	121
Selenos	ENSMUSG00000075701	-1,83	0,01	251
Acadm	ENSMUSG00000062908	-1,82	0,09	130
1110065P20Rik	ENSMUSG00000078570	-1,80	0,07	165
Pyroxd1	ENSMUSG00000041671	-1,75	0,08	137
Ramp1	ENSMUSG00000034353	-1,69	0,04	282
Slc39a12	ENSMUSG00000036949	-1,54	0,05	214
Atp8a1	ENSMUSG00000037685	-1,52	0,08	214
Ogdh	ENSMUSG00000020456	-1,39	0,08	294
Grm3	ENSMUSG00000003974	-1,34	> 0,01	777
Tril	ENSMUSG00000043496	-1,31	> 0.01	666
Dbt	ENSMUSG00000000340	-1,27	0,08	266
Dynlrb1	ENSMUSG00000047459	-0,99	0,05	406
Eno1b	ENSMUSG00000059040	-0,81	0,07	1112
Gnao1	ENSMUSG00000031748	-0,74	0,07	1537
Usp22	ENSMUSG00000042506	0,76	0,07	852
Psip1	ENSMUSG00000028484	1,28	0,01	321
Zfp451	ENSMUSG00000042197	1,34	0,09	287
Rlim	ENSMUSG00000056537	1,42	0,07	174
Pnn	ENSMUSG00000020994	1,66	0,07	159
Tpx2	ENSMUSG00000027469	1,71	0,01	209
Ier2	ENSMUSG00000053560	1,87	0,07	127
Ccnb2	ENSMUSG00000032218	1,89	0,01	159
Scrib	ENSMUSG00000022568	1,94	0,05	161
Ubtf	ENSMUSG00000020923	2,05	0,01	125