Western Norway University of Applied Sciences

BACHELOR'S THESIS

Transcriptional Burst Kinetics in Neurons

Eva Kristine Fladhus & Anders Lindberg Sørjoten

Bachelor of Biomedical Sciences

Department of Proteomics // Department of Safety, Chemistry and Biomedical Laboratory Sciences

Gábor Juhász & Alvhild Alette Bjørkum

We declare that we wrote this thesis independently and that references for all sources used in this work are correct, in accordance with *Regulations relating to academic studies and examinations* at The Western Norway University of Applied Sciences (HVL), § 12-1.

ELTE
EÖTVÖS LORÁND
UNIVERSITY

Abstract

Neurons differ from other cells in the body in that they do not regenerate, meaning that damaged neurons need to be repaired instead of killed. Cellular function is determined by the type and amount of proteins in a cell, which in turn is determined by the type and amount of mRNA in a cell. mRNA therapy - the injection of mRNA that codes for the type and amount of proteins in a healthy cell - has been suggested as a way to fine-tune neuronal function in damaged neurons, in diseases such as schizophrenia, depression and Alzheimer's. To be able to do this, a comprehensive understanding of gene transcription is needed, as it is the function that creates the mRNA. In this study we look at the dynamics involved when a gene is transcribed to mRNA, using the Beta-Poisson model described in Vu et al. (2016) to infer kinetics on two static, single-cell RNA-sequencing (scRNA-seq) snapshot datasets, consisting of cytosolic mRNA extracted from pyramidal cells and fast-spiking cells in the pre-frontal cortex of mice. Our main focus is to investigate whether there is a statistically significant difference between the transcriptional dynamics in these two types of neurons.

Our findings suggest that at least 41 genes show a significant difference in transcriptional burst frequencies. However, we also discovered weaknesses in the model that were not accounted for by the authors. We conclude that this dataset should be re-analyzed using improved mathematical models, for example the Generalized Telegraph Model by S. Luo et al. (2023).

Acknowledgements

We would like to acknowledge the following individuals for their valuable contributions to this project:

Gábor Juhász
Alvhild Alette Bjørkum
Dániel Mittli
Vanda Tukácsz
Wilmos Tóth

We are grateful for their support and guidance throughout this project.

Eva and Anders

Eva K.

Eva Kristine Fladhus

Anders L.

Anders Lindberg Sørjoten

Contents

1. Introduction

1.1 Brain Diseases Differ from Other Diseases

The brain is the most complex organ in the human body, and understanding how it works on the cellular level is essential for treating various brain diseases. Unlike cancers, which can sometimes be treated by eradicating the malfunctioning cells, most brain diseases cannot be cured this way. Neurons lack the ability to regenerate to the same extent as other cells, and pathological changes in the neurons are largely irreversible (Steward et al., 2012). That is not to say that neurogenesis (the creation of new neurons) does not happen in the adult brain - it does, but to a much lesser degree. A more comprehensive understanding of the cellular mechanisms of healthy brain cells is needed, so that each cell can be fine-tuned to improve or optimize its cellular activity, from gene expression to neuronal firing.

1.2 Aim of Study

In this thesis, we look at two functionally different types of neurons: the pyramidal cells (PCs) and fast-spiking cells (FSCs) in the prefrontal cortex (PFC) of mice. The aim is to study their transcriptional burst kinetics to see if the fundamental differences between the cells are reflected also in the dynamics of their gene transcription. To give the reader an idea of the importance of new scientific discoveries in this field, we will begin with an overview of the human prefrontal cortex and its functions, how its dysfunction can lead to psychiatric diseases, what treatments are available today and what the future of psychiatric drug treatment may look like.

1.3 The Prefrontal Cortex

The prefrontal cortex (PFC) is the brain structure that sits behind our forehead (Figure 1). It is the most evolved brain region in humans and is not fully developed until we reach 25-30 years of age (Arnsten, 2009). The PFC controls higher-order cognitive abilities, often referred to as "executive functions", such as working memory, cognitive flexibility, emotional regulation and social interactions (Ferguson & Gao, 2018). Sometimes called the Chief Executive Officer (CEO) of the brain, the PFC receives input from other parts of the brain and decides what to use the input for ("top-down" control). Working memory is the creation of a "mental sketchpad"; the ability to keep in mind recent events or access information from long-term memory, and use it to regulate our social behaviour and emotions (Arnsten, 2009).

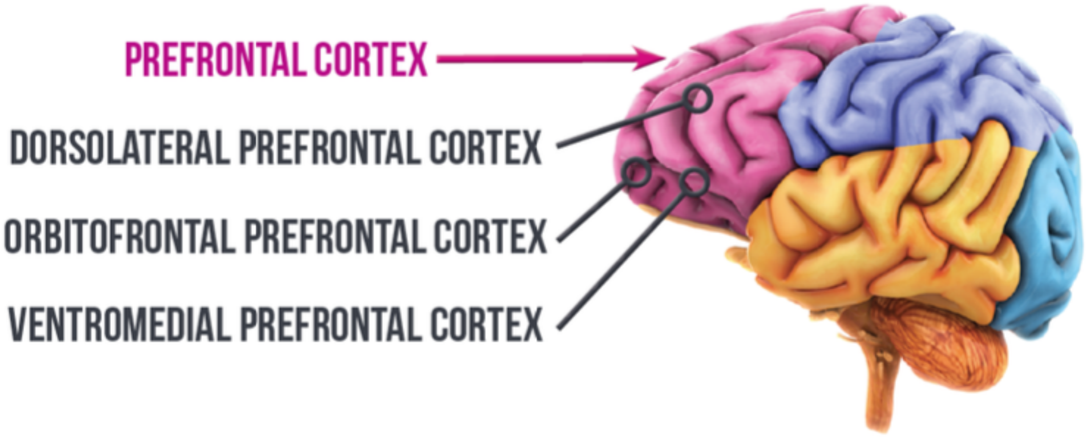
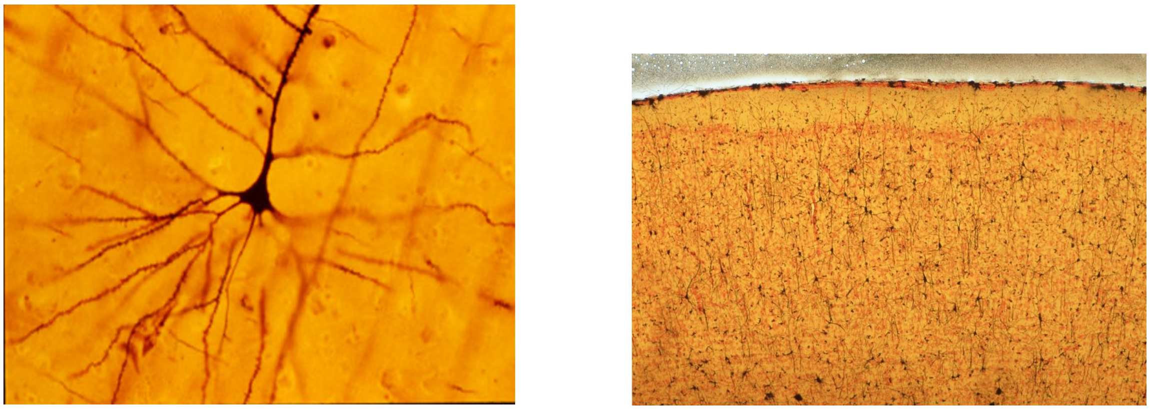

PREFRONTAL CORTEX
DORSOLATERAL PREFRONTAL CORTEX
ORBITOFRONTAL PREFRONTAL CORTEX
VENTROMEDIAL PREFRONTAL CORTEX

Figure 1: The prefrontal cortex and its subregions. This is an excerpt from an infographic by AstridCortez. License: CC BY-SA 4.0

The PFC is further divided into two regions: the lateral PFC, which consists of the dorsolateral PFC (dlPFC) and the ventrolateral PFC (vlPFC); and the ventromedial PFC (vmPFC) which is also known as the orbitofrontal PFC. The dlPFC is mainly concerned with the representation and regulation of "external state" - it regulates how we see the world. The vmPFC, on the other hand, is associated with the representation and regulation of "internal state", e.g. how we process emotions (Arnsten, 2009).

1.3.1 Neurons of the Prefrontal Cortex

Two neurons play major roles in the PFC; the pyramical cell (PC, see figure 2a and 2b), also known as a projection neuron or a principal cell; and the fast-spiking cell (FSC), a type of interneuron. The PC is the most abundant neuron in the mammalian cortex and compromises approximately two-thirds of all neurons. Its soma is cone-shaped, and it has several distinct dendrites and axons projecting long distances and targeting other brain regions and cells (Bekkers, 2011). The PC is an excitatory neuron which releases glutamate, the major excitatory neurotransmitter - or chemical messenger - in the brain. The FSC, on the other hand, makes up about 10-20% of the neurons in the mammalian cortex and its soma is about half the size as that of a PC. It has an inhibitory effect, and releases the neurotransmitter γ-amino-butyric acid (GABA) (Hu et al., 2014). The FSC's neuronal firing has a significantly higher frequency than that of the PC, which is illustrates in figure 3.



(a) Cahass, 2006

(b) Britannica, n.d.

Figure 2: A pyramidal cell (a) and a light micrograph of a brain slice from the celebral cortex, with pyramidal cells and fast-spiking cells (b).

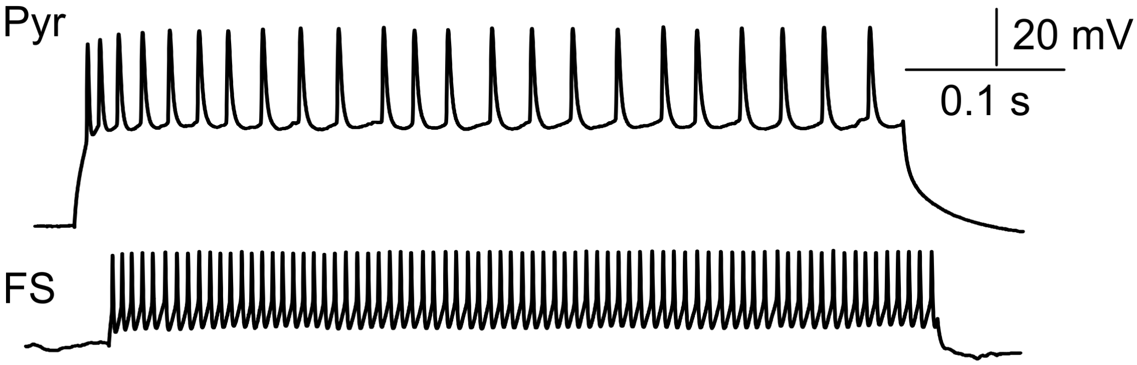


Figure 3: Neuronal burst frequency in PCs (top) and FSCs (bottom) (Ravasz, 2020a).

The co-operation of PCs and FSCs is crucial for PFC function. The PC constantly receives information from other brain regions and cells and needs to decide whether to fire an action potential or not; and the FSC helps making these decision by inhibiting unnecessary firing by the PC (Ferguson & Gao, 2018; Hu et al., 2014). The two types of neurons create different microcircuits throughout the PFC (see figure 4. The role of these microcircuits in spatial working memory was demonstrated in a study on monkeys by Funahashi et al. (1989). In this study, the monkeys were trained to look at a central spot on a TV screen while a visual cue was shown in the periphery for a short period of time (0.5 seconds). This was followed by a delay period of 1-6 seconds where the monkeys had to keep their eyes focused on the central spot. Once the delay period was over, the task was to move their eyes to the place where the cue had been shown in order to receive a reward. Through electrodes recording single neurons in the monkeys' dlPFC, the scientists could see a remarkable pattern: when the monkeys were trying to remember the position of a cue that was 45° from the reference point, certain PCs that appeared to be representing 45° kept firing even after the visual cue was gone, while PCs representing other degrees from the reference point (e.g. 90°) were suppressed by FSCs. This showed that neurons in the dlPFC can keep the representation of a stimulus "in mind" by sustained firing in the microcircuit associated with the stimulus (Arnsten, 2009; Funahashi et al., 1989).

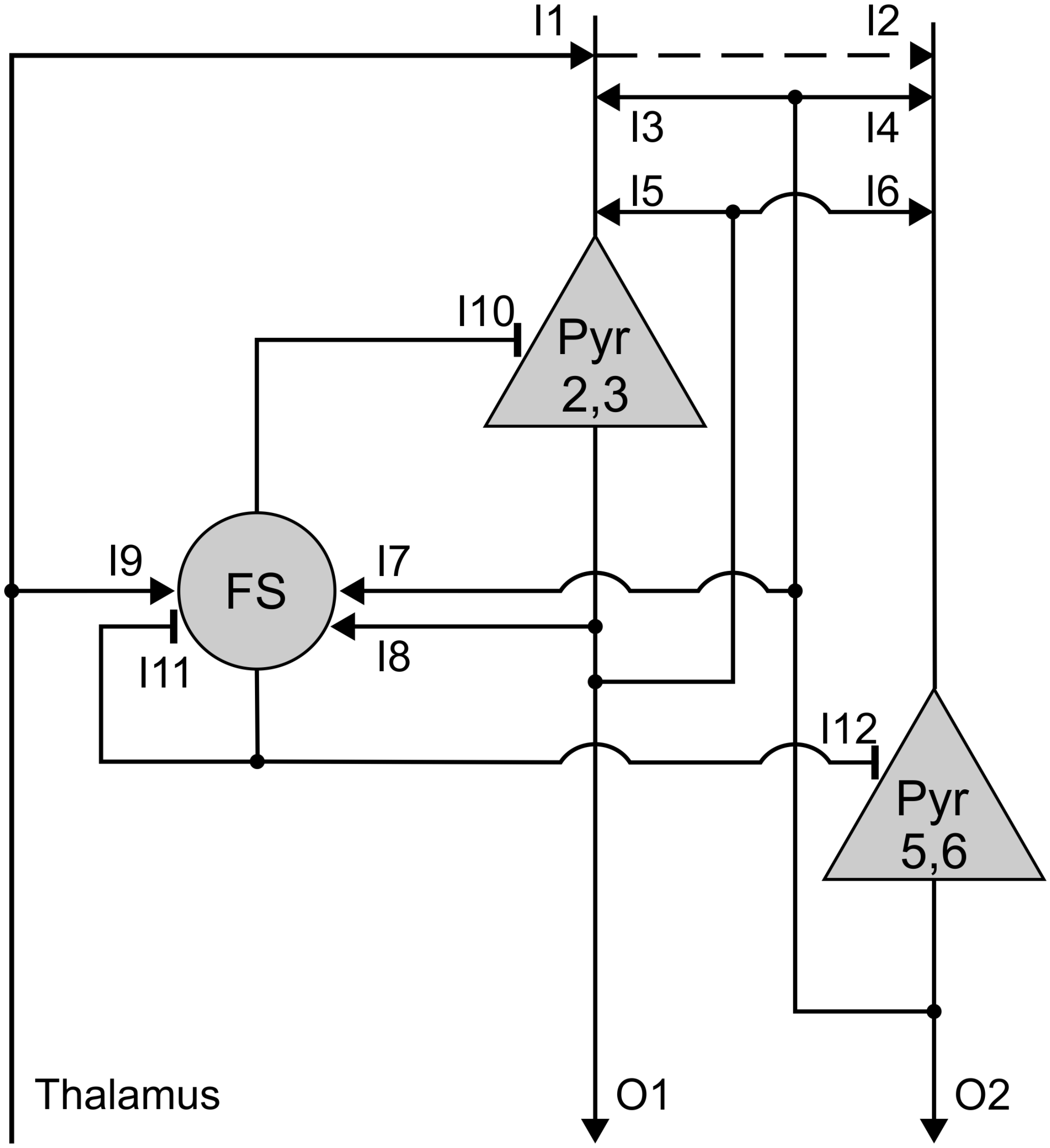


Figure 4: Microcircuits in the prefrontal cortex (PFC). Pyramidal cells (Pyr) receive information from other brain regions and cells and need to decide whether to fire an action potential or not; fast-spiking cells (FS) help making these decision by inhibiting unnecessary firing by the PCs(Ravasz, 2020b).

1.3.2 The Role of the Prefrontal Cortex in Psychiatric Disorders

The PFC is prone to psychiatric disorders, which may in part be due to its long maturation period (Kolk & Rakic, 2021). Because neurons do not have mitosis, the neurons needed for the PFC are overproduced before birth and PFC synaptic density in humans spikes around 3.5 years of age. The maturation of the PFC is in reality a gradual decline in the number and type of neurons and synapses, as well as myelination (insulation of neurons to ensure better communication; "white matter"). Neurons and neuronal networks in the PFC are chosen based on the "use it or lose it" principle - it is mainly our childhood and early adulthood experiences that shape the connections in the mature PFC, and the synapses are "pruned" according to what the PFC deems important to retain (Hebb, 1949; Kolk & Rakic, 2021). This is an example of the human brain's immense ability to adapt to its environment. However, it could also increase the risk of mental disorders because the brain stays vulnerable to change for a long period of time.

While the well-functioning PFC executes the aforementioned "top-down" control of the brain and the body, very little is needed to throw the brain's CEO off-balance. Different states of arousal affect our higher cognitive functioning in a way that is best described as an inverted U - too little arousal, such as fatigue, leads to poor functioning and so does too much arousal, such as stress. In the middle of the inverted U, we find alertness. When we experience too much stress, the PFC may resort to "bottom-up" control, meaning that it is governed by its input instead of governing its input (Arnsten, 2009). This can in some cases be useful. If the amygdala (the fear center of the brain) takes over, it can be helpful in situations where we actually need to escape from danger.

Even after the PFC is considered fully developed, it still has the capacity to change; this is referred to as synaptic plasticity and happens not just in the PFC but in the brain as a whole. In other words, the PFC is still capable of adapting to new input and create new synapses, thus making new memories or learning new skills. Much like in the developing PFC, these connections are then strengthened or weakened based on how often they are used. This is the reason why we sometimes forget things we learned or we're able to perform certain tasks almost automatically

(Hebb, 1949). Pathological changes due to synaptic plasticity in adulthood may also occur (Crabtree & Gogos, 2014).

Psychiatric disorders with PFC involvement are multifactorial, heterogenous and numerous, including but not limited to schizophrenia, bipolar disorder, major depressive disorder (MDD), substance abuse, obsessive-compulsive disorder, post-traumatic stress disorder, autism spectrum disorder, attention deficit hyperactivity disorder (ADHD), Parkinson's, Huntington's, traumatic brain injury, fronto-temporal dementia and Alzheimer's diseases. For simplicity's sake, we will introduce only three of them: Schizophrenia, MDD and Alzheimer's disease. All of these diseases have genetic risk factors, i.e. our genes can make us more or less prone to developing the disease; however, environmental factors such as exposure to alcohol, drug abuse, smoking and poor diet during prenatal development or in early life are all known to elevate the risk of developing psychiatric diseases (Crabtree & Gogos, 2014; Kolk & Rakic, 2021).

Schizophrenia, a debilitating condition causing symptoms such as hallucinations, delusions, thought disorder, cognitive impairment and flat affect, is associated with fewer GABA-synthesizing enzymes (GAD61) in the PFC which is believed to diminsh the function of FSCs in the PFC. Studies show that the dlPFC is underactive as schizophrenic patients perform working memory tasks and neuropathological studies show that neurons in dlPFC layer 3 lose their dendritic spines (Arnsten, 2009). As mentioned, the dlPFC plays a role in understanding the world around us (external state), and so we can imagine how dysfunction in this part of the brain can lead to hallucinations and delusions. The most commonly used drugs to treat schizophrenia are first- or second-generation antipsychotics, that block some of the dopamine receptors in the brain. Both medication types are associated with severe side-effects such as agranulocytosis (severe lack of granulocytes), weight gain, hyperlipidemia and diabetes mellitus (Patel et al., 2014).

MDD is characterized by symptoms of persistent sadness, hopelessness, loss of interest in daily activities, anxiety and mental paralysis. Some research has shown that depression leads to less serotonin release in the brain, and patients with MDD show lower activity in an area of the vmPFC called "Broadmann Area 25", which is rich in serotonin transporters (Arnsten, 2009). Serotonin is a neurotransmitter that is involved in the regulation of wakefulness, sleep, alertness and mood. The most commonly used drugs to treat MDD are selective serotonin reuptake inhibitors (SSRIs), which cause serotonin to stay longer in the synapse without being reabsorbed. Common side-effects of SSRIs include nausea, dry mouth, headaches, insomnia, fatigue and anxiety.

Alzheimer's disease is the most common cause of dementia. It spreads from other parts of the brain and affects the PFC as a whole, causing symptoms like impaired recall, disorganization, language deficits and personality changes; for example increased aggression and/or hypersexuality (Arnsten, 2009; Knopman et al., 2021). The two main causes of Alzheimer's are believed to be beta-amyloid plaques and neurofibrillary tangles caused by malfunctioning tau proteins. Both proteins inhibit the signaling system in the neuron, and block transport of nutrition to the cell, ultimately causing the death of the cell. The beta-amyloid theory is currently under some controversy, which we do not have time to discuss in this thesis (Piller, 2022).

In summary, the most used therapeutical target for these disorders are single-molecule receptors (L. Ravasz et al., 2020). Blocking a receptor to prevent the release or reuptake of a neurotransmitter can lead to the creation of new receptors through synaptic plasticity. This could be a reason why drugs such as antipsychotics that block the release of dopamine and SSRIs that block the reuptake of serotonin from the synapse seem to work for a while and then lose their effect.

1.4 Gene Expression

1.4.1 Same DNA, Different Cells

Neuronal functions like the microcircuits of PCs and FSCs in the PFC are molecularly organized via the cellular proteome. We were taught from textbooks that "DNA makes RNA makes protein". The goal of the gene transcription process is to make proteins, as proteins make up all the functions of the cells and ultimately our body as a whole. Although all cells contain the same DNA, the DNA is expressed differently. This leads to the big variation between cells in terms of morphology and function. A lot of steps and biomolecules are involved in all three stages of gene transcription, making cells of different phenotypes.

1.4.2 mRNA Modification as a Therapeutic Intervention in Psychiatric Diseases

The possibly overly simplified nature of today's available psychiatric medicines to treat diseases related to the PFC may become evident once we look at the cell as a whole. The cell is a highly complex environment which, when viewed in light of the gene number that exists in each cell, can be thought of as having at least 20.000 molecular dimensions. Therefore it can be argued that it makes little sense to attack a single gene with medicine consisting of a single molecule, and believe that it will be enough to cure or even halt a disease caused by dysregulation of the whole gene expression machinery. A study by Sul et al. (2009) showed that the transfer of astrocyte RNA into a neuron lead to the latter becoming more like an astrocyte. This challenges the idea of absolute phenotypes and opens up the possibility that phenotypes are dynamic and change according to the cell's transcriptome (Kim & Eberwine, 2010).

Studies of cytosolic mRNA and protein expression in healthy versus sick neurons provide the potential of designing mRNA using "healthy neuron mRNA" as a template. This set of mRNAs could somehow be transferred into the neurons that are out of balance and force the cell to build proteins from the healthy mRNA instead of the imbalanced mRNA produced by the cell itself.

1.4.3 The Dynamics of Gene Transcription

Gene transcription is usually described as a stochastic process that happens in bursts (Larsson et al., 2019). "Stochasticity" describes the cell "doing its own business", i.e. not being synchronized with other cells in terms of what genes are transcribed when (Elowitz et al., 2002).

"Transcriptional burst" means that there are times when the gene is being transcribed rapidly, followed by longer periods of "silence", referred to as no transcription. This is supported by many studies, including live-cell studies using Single-Molecule RNA Flourescent In-Situ Hybridization (smFISH) technique, with fluorescent probes bound to specific genes, as well as Single-Cell RNA Sequencing (scRNASeq), whose output data generate a significant amount of biological zeros. By “biological zeros” we mean that mRNA from the particular gene is not found in the analysed cell, i.e. it shows zero expression, even when we have ruled out the possibility of an error in the analysis (the latter is called a “non-biological zero”). Before the scRNASeq technique was developed, bulk RNA sequencing showed a “mean” of gene expression. However, once single cells were analysed, it became clear that a characteristic of the dataset produced shows mainly zero expression in most of the approx. 20.000 genes in every cell, meaning that most genes are “turned off” (not being transcribed) at any given time point (Jiang et al., 2022; Tunnacliffe and Chubb, 2020).

In this thesis, the widely accepted two-state model of gene transcription is used. This model states that genes can be either “switched on” (transcription is happening) or “switched off” (transcription is not happening). It is a simplified way of looking at an intriguingly complex process, which will be discussed in more detail in the "Discussion" part of our thesis.

1.4.4 Thresholds and Limiting Factors in Transcriptional Bursts

For the transcriptional burst, not to be confused with a neuronal burst, a possible threshold expression level has been identified: below it, only transcriptional burst frequency is modulated; above it only transcriptional burst size is modulated. This may happen due to a refractory period – at a certain point it is no longer possible to increase the frequency, and the only way to further increase expression is by increasing the rate of transcription or extending the duration of each burst. Episodic bursting could be a way for the cell to make the most out of its limited resources, such as transcription factors (Dar et al., 2012). It has been established that while there are about 20.000 genes, there are only a few dozen transcription factors. A theory of “transcription factories” has been described, where genes that are co-expressed or expressed at roughly the same time, enter a temporary cellular “organelle” consisting of groups of transcription factors (Sas-Nowosielska & Magalska, 2021).

1.4.5 Molecular Mechanisms that Govern Transcriptional Bursting

Studies suggest that some molecular mechanisms involved in transcription can affect every aspect of bursting (e.g. histone marks), whereas others predominantly influence the burst size (e.g. number and affinity of cis-regulatory elements, like promoters, enhancers, silencers) or the frequency (e.g. transcription factor availability) (Nicolas et al., 2017).

Additionally, it is possible that looking specifically at neurons adds another dimension to the process of gene transcription. Neurons are the most heterogenous cells in the body. There are many distinct phenotypes and subtypes have been identified for all of them, as revealed by immunostaining of phenotype markers, and their functions vary greatly depending on their anatomic placement and network properties. Neurons are specialized to change gene transcription responding to environmental input - stimulus driven gene transcription is widely studied in models of synaptic plasticity. A neuron is a highly complex environment, and it is likely that gene transcription is dependent on the cells' spatial awareness; for example, it may not matter only what kind of transcription factor (TF) is used, but also which direction the TF is coming from. The lack of proliferation in neurons that gives them a life-span function in the brain makes it likely that their transcription regulation mechanisms in regard to transcriptional dynamics are different from other cells that have a normal cell-cycle.

In this thesis, we try to challenge the artificial nature of growing a cell culture in a dish, flask etc. Neurons, more so than any other cells, need to be considered as part of a system, and cannot be studied well once removed from their natural environments, consisiting of other neurons, glial cells, vessels etc. Therefore, the methods used to study the neurons should be as close to an in-vivo situation as possible.

2. Method

2.1 Cell Harvesting and Cytosolic mRNA Extraction

73 male mice of the species C57BL/6N, whose ages varied between 27 and 40 days, were anesthetized and decapitated. The brain was removed and kept in ice-cold artificial cerebrospinal fluid (aCSF) containing sucralose for one hour. Then 300 µm slices were cut from the pre-frontal cortex, using a Vibratome. The slices were incubated in room-temperature aCSF for a minimum of one hour and then placed under a microscope, where the neuron was identified anatomically and electro-physiologically, using whole-cell patch-clamp method to measure the neuronal electric activity as well as extracting cytosol from the cells. For extraction of cytosol, a micropipette with a $\sim$ 1µm tip filled with RNAse-free cellular solution was used.

From the extracted cytosol sample, RNA was isolated and transcribed into cDNA using reverse transcriptase. Then the cDNA was amplified through two rounds of antisense-RNA (aRNA) amplification, a linear form of amplification, to preserve accuracy in mRNA copy number. The cDNA from each cell was then given its own unique barcode and amplified through three more rounds of aRNA amplification.

2.2 Single-Cell RNA-Sequencing with Illumina

First, cDNA is "barcoded", i.e. marked so that it can be identified which cell it came from. Then it was sequenced using next-generation sequencing (NGS). The Illumina NGS technique happens in 4 different stages. These stages are sample preparation, cluster generation, sequencing, and data analysis. This is a next generation sequencing technique were large numbers of sequencing cycles are performed, in this paper referred to as deep sequencing. In the first step, our isolated DNA gets fragmented into smaller pieces, these small pieces will then be introduced to oligonucleotides that bind to each side of the DNA fragment.

The fragments are then heated, in a way that they will be denatured. The fragments are then introduced to the flow cell and will bind to complementary binding sites. The clusters of DNA fragments are then amplified in a process called bridge amplification which will generate millions of copies of single-stranded DNA. This process is repeated until there are enough copies necessary for identification. The amplification happens on the flow cell. In the sequencing of the amplified single-stranded DNA (ssDNA), a primer is first added to the binding site of the oligonucleotide, complementary nucleotides are now added to the mix, containing specific fluorescent labels. The labels will then, after binding to the ssDNA be excited by a laser, and will emit a light, this light is then read, and given its complementary letter. Bioinformatic tools will now identify the sequence, and compare it to a reference genome, and with that overlay, the genome sequence can be deciphered (Illumina, 2016).

2.3 Normalization of Data

The exact volume of cytosol extracted from each cell is not known, therefore the data has been normalized using library size normalization. Ravasz et al. (2020) identified the 1000 genes that were most frequently expressed in all cells, i.e. the genes with the lowest amount of biological zeros. These cells were used as a reference for "pre-normalization", were a scaling factor was determined for each cell using the lowest root-mean-square deviation (RMSD) in relation to the average values of expressed genes in the 1000 genes. The values (i.e. the expression levels of genes) in the rest of the dataset were then "pre-normalized" using the calculated scaling factor. From the 1000 genes, the 500 that showed the lowest standard deviation were selected. Reference genes that showed more than a 10-fold difference in expression level between pyramidal cells and fast-spiking cells were filtered out. 409 genes were left and these were used to normalize the raw data. This was done by finding the multiplication factor for each cell that provided the lowest RSMD in the 409 reference genes compared with the averages of the non-zero values.

2.4 Data Processing in R

The data processing in this thesis was done in the R programming language, using RStudio; a free, open-source programming software designed especially for statistical analysis and bioinformatics.

2.4.1 R - Overview of the Beta-Poisson Model

The work done in R is based on the paper “Beta-Poisson Model for Single-Cell RNA-seq Data Analyses” by Vu et al. (2016). The authors present a modified Poisson model to infer transcriptional burst kinetics on static datasets. The model predicts that the expression values from cells where the gene is in ON-state is a data point on a Poisson curve - it is essentially a “translation” from static to dynamic data using a pseudo-time prediction.

The user can choose whether or not to use the dataset’s zero values to calculate the initial parameters. We chose to divide the dataset into pyramidal cells and fast-spiking cells, and analyze each dataset two times: once without taking the zero-values into account, and once with the zero-values taken into account. The modified Poisson model returns four parameters: the burst frequency (alpha) and burst size (beta) in addition to two scaling parameters; lambda1 and lambda2.

Three functions from the BPSC GitHub repository were used: estimateBPMatrix.R, getBPMCnullmatrix.R and getMCpval.R (Vu., et al 2016) repository can be found here: (https://github.com/nghiavtr/BPSC). The first function takes a single-cell RNA sequenced dataset as input, filters out only the genes that are expressed in more than 5 percent of the cells and calculates burst parameters for these genes. It also runs a statistical test (X2, “chi squared”). The second function takes the output from the first function as input and assigns a Monte Carlo p-value (MCp-value) to the parameters of each gene. This is done by simulating a large dataset (n=1000) using the same parameters, the results from the X2 test and the number of observations used in the model fitting. The final function returns the list of MCp-values generated by the second function.

2.4.2 R - The Filtering Process

To ensure the validity of our data, it was run through three filtering steps. The first step is the one built into the “estimateBPMatrix” function, where genes with less than 5% expression throughout the dataset were filtered out. From a starting point of 19683 genes, this left a total of 14480 (74%) genes for the pyramidal dataset and 12310 (63%) genes for the FSC dataset. The second step consisted of filtering out all the genes with Monte Carlo p-values <0.05. This left a total of 9790 (50%) genes for the pyramidal dataset and 9313 (47%) genes for the FSC dataset. The third and final filtering step was a comparison between remaining gene names in the two datasets: only names that could be found in both sets were kept. This left a total of 5758 (29%) genes. The last two steps were automated in R, to speed up the process and eliminate human error.

In addition to this, another group was made consisting only of genes with expression in at least 10 cells in each dataset.

2.4.3 R - The Statistical Process

A paired Kolmogorov-Smirnov test (KS test) was used to look for significant differences between the burst frequencies and burst sizes in the two "filtered" data sets. We preferred to use a KS test over a t-test because it does not require a normal distribution of the data; it is a non-parametric (i.e. it does not assume any specific type of distribution) test that, when using the "paired" mode, compares two samples and decide if they come from the same distribution or not.

2.4.4 R - The Visualization Process

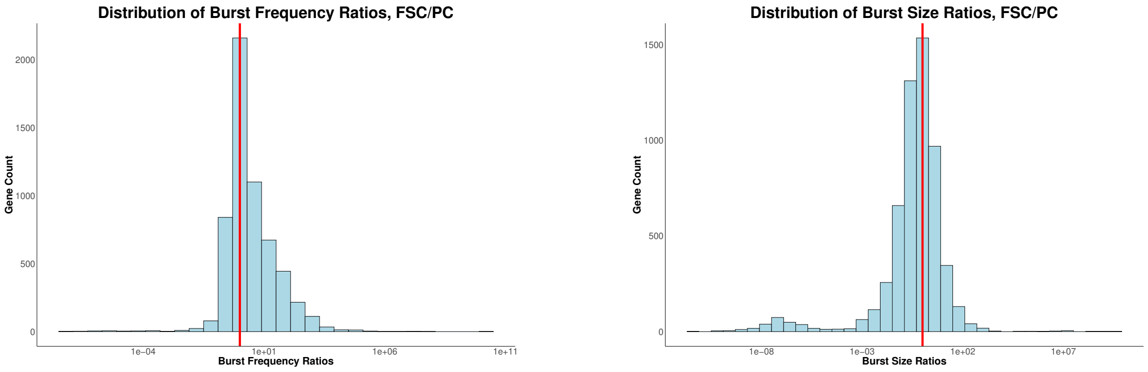
R was used to create histograms, density plots and scatter plots depicting our findings. For FSCs and PCs, burst frequency was plotted against burst size in the same scatter plot to visualize the differences in burst kinetics between the two neuron types.

3. Results

3.1 Possible Differences in Burst Frequency and Burst Size

To accurately compare the genes from the FSC dataset to the genes from the PC dataset, transcriptional burst frequency ratios and transcriptional burst sizes were calculated by dividing the FSC frequency or size by the PC frequency or size. From the calculations, the following was observed from the dataset of 5758 matched genes: 64.3% of FSC genes show higher (ratio $>$ 1) burst frequencies than their corresponding PC genes (Figure 5a, 6a). 24% of FSC showed a 10x or higher burst frequency than their corresponding PC genes. In comparison, 2.1% of PCs have a 10x or higher burst frequency than their corresponding FSC genes.

The KS test showed a significant difference between the transcriptional burst frequencies in FSCs as compared to PCs, and also a significant difference in transcriptional burst sizes when comparing the two neuron types.

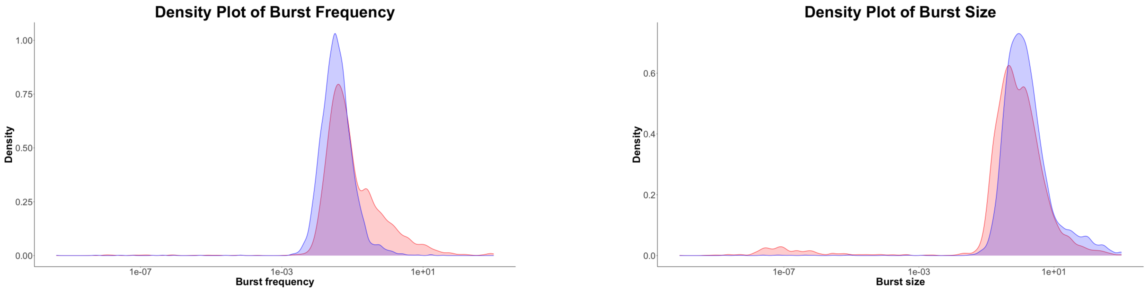


(a) Burst Frequency Ratios

(b) Burst Size Ratios

Figure 5: Distribution of burst frequency ratios (a) and burst size ratios (b) in the 5758 genes. The red line represents a ratio number of 1 - this is where we find the genes that show no difference in burst kinetics between the two datasets. Ratios were calculated by dividing the burst frequency or burst size of the gene in the FSC dataset by the burst frequency or burst size of the corresponding gene in the PC dataset.

60.7% of PC genes show a higher (ratio > 1) burst size than their corresponding FSC genes (Figure 5b, 6b). 21.1% of PC genes show a 10x or higher burst size than their corresponding FSC genes. In comparison, 8.1% of FSC genes show 10x or higher burst size than their corresponding PC genes.



(a)

(b)

Figure 6: Density plot of burst frequencies (a) and burst sizes (b) in the 5758 genes. FSCs are visualised in red and PCs in blue. These plots show the probability that a gene is found within a certain range. Plot a shows that there is a higher probability of finding a gene from the FSC dataset in a higher burst frequency range, whereas plot b shows that there is a lower probability of finding a gene from the FSC dataset in the higher burst size ranges.

A scatter plot of burst frequencies and burst sizes was made (Figure 7). Four clusters of genes were identified and named (Figure 8). Most notably, an overweight of FSC genes were seen in two distinct places in the plot - the upper right corner and the lower right corner. The upper right corner represents the genes with the highest burst sizes and burst frequencies, whereas the lower right corner represent genes with higher burst frequencies and lower burst sizes.

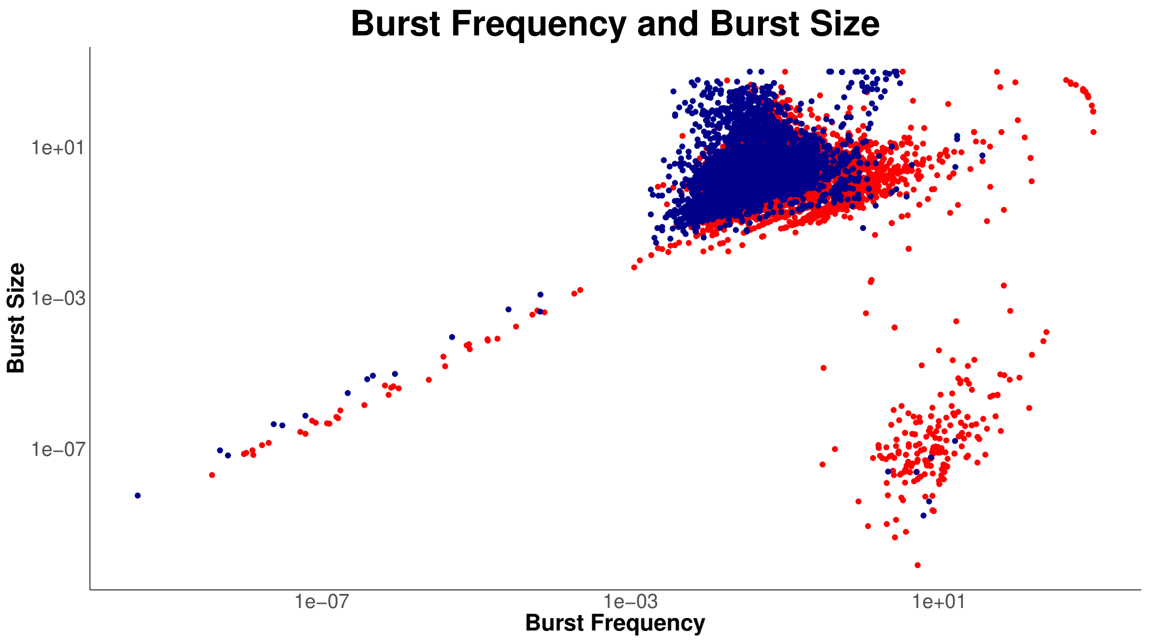


Figure 7: Distribution burst frequencies plotted against burst sizes in the 5758 "matched" genes, where the blue dots represent the PCs and the red dots represent the FSCs.

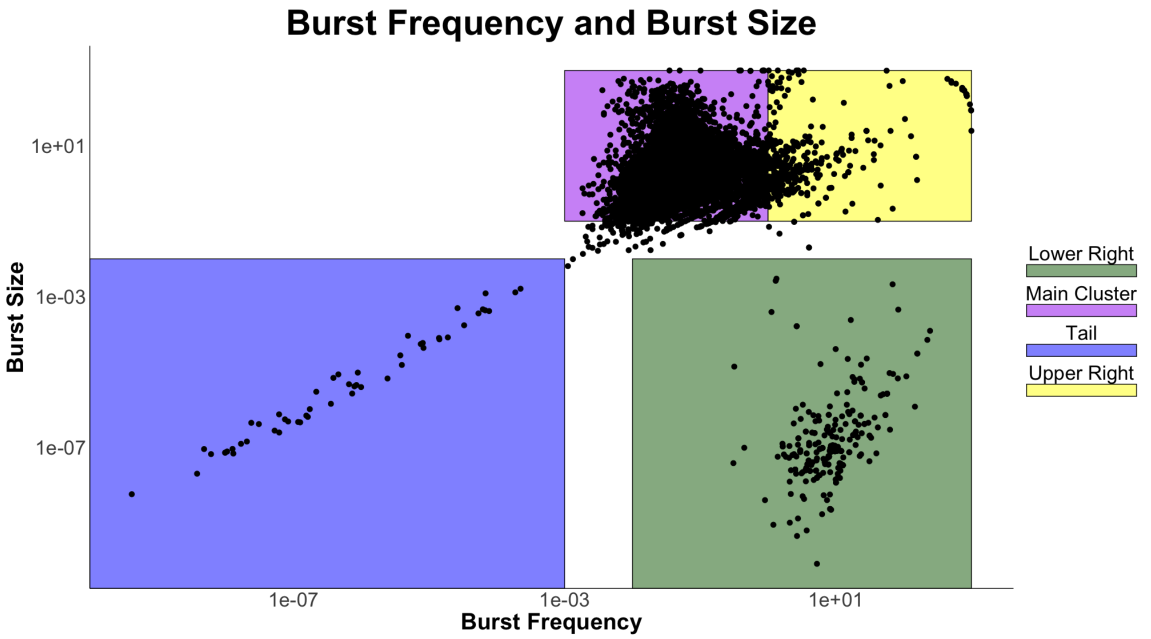


Figure 8: Clusters from Figure 7 were identified and marked by coloured boxes. The blue box represents the "Tail", the purple box the "Main cluster", the yellow box is the "Upper right" cluster and the green box is the "Lower right" cluster.

3.2 Taking On-state frequencies into account

The on-state frequency, i.e. the number of cells in the dataset that showed any expression, was calculated for all genes. The on-state frequency can be viewed in much the same way as sample size: lower on-state frequency compares to smaller sample size, because there are fewer datapoints to base the calculations on. In our project, this means that there are less points to fit a Poisson curve template on.

Significant differences in the mean on-state frequency between the four clusters were found. In the main cluster, the on-state frequency was higher than in all the other groups (Figure 9, 10).

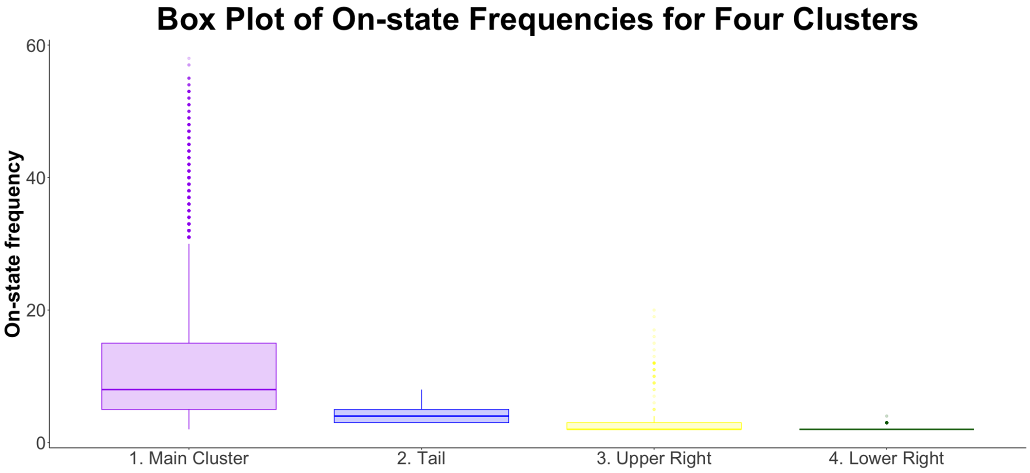
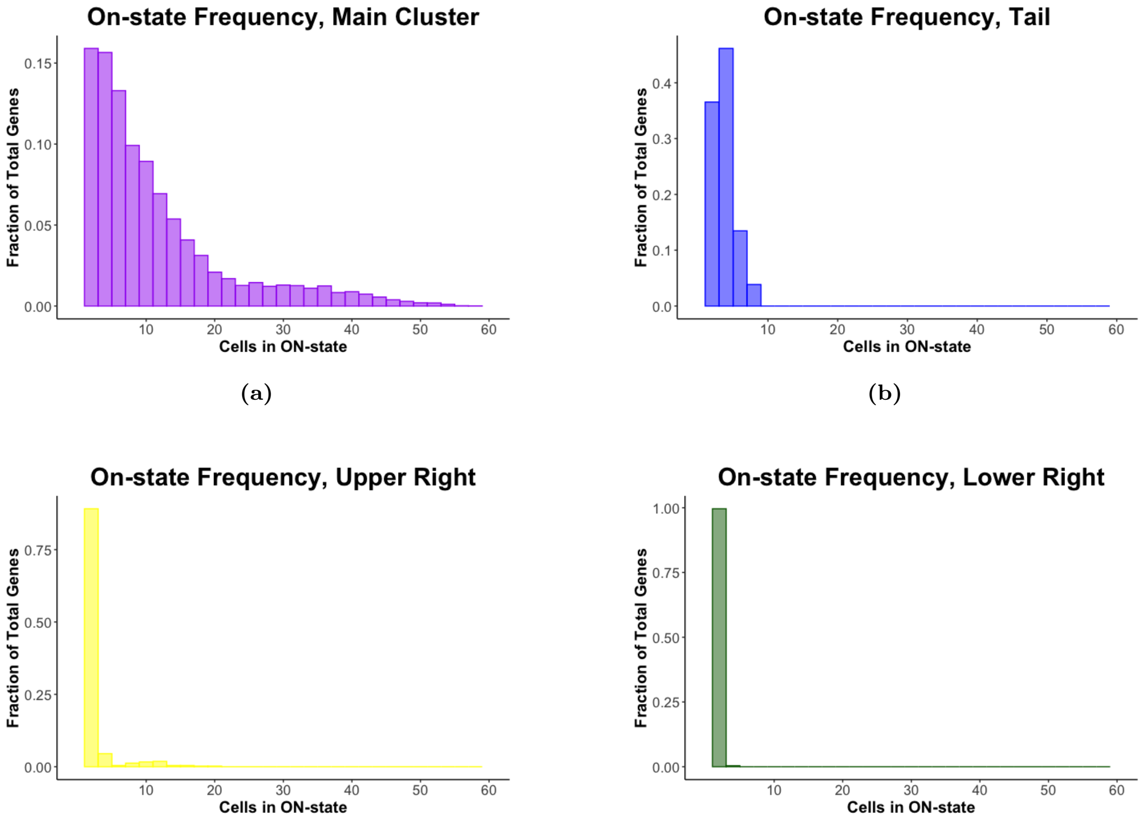


Figure 9: Box plot of on-state frequencies for the four clusters in Figure 8



(a) **(b)** **(c)** **(d)**

Figure 10: Histograms showing ON-state frequencies for the different clusters seen in Figure 8
.

In the "Main Cluster" the mean on-state frequency (OSF) was 11.92 cells in ON-state per gene, while the median was 8, meaning that at least half of the genes in the data set had 8 or more cells that showed expression. In the "Tail", the mean OSF was 4.21 and the median 4. In the "Upper Right" cluster, the mean OSF was 2.9 and the median 2. In the "Lower Right" cluster, the mean OSF was 2.14 and the median was 2, meaning that at least half of the genes in the data set showed expression in two cells only. When comparing the total number of matched genes, i.e. the scatter plot as a whole (n=5758 genes), the FSCs had a mean OSF of 6.15 and a median of 5 (standard deviation (SD)=4.18), whereas the PC had a mean OSF of 17.2 and a median of 14 (SD=11.70).

Since the PC dataset consisted of 59 cells in total and the FSC dataset consisted of 25 cells in total, we wanted to test if the transcriptional burst frequencies of the

PC genes changed if only 25 randomly selected cells from the dataset were used in the calculations. This gave us the scatterplot in Figure 11, where the same pattern that was previously only seen in FSCs has appeared also in PCs.

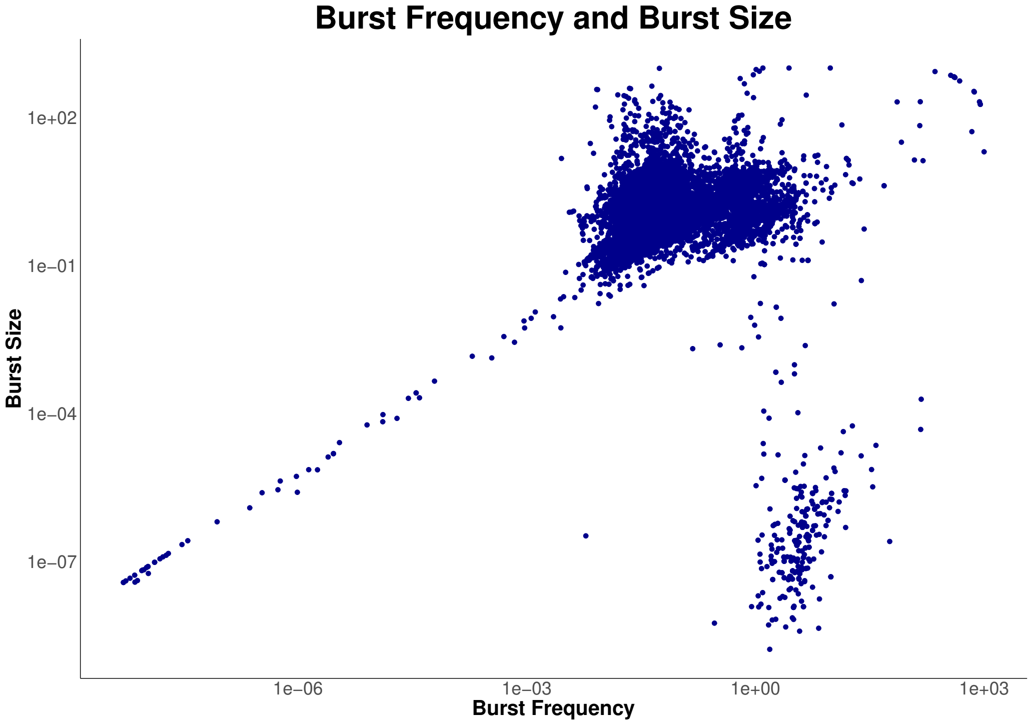


Figure 11: A scatterplot of only genes from the PC dataset (shown in blue dots), after randomly selecting out 25 cells to use for calculations. The distribution is similar to that of the FSCs in Figure 7.

To control for the differences in OSF found within and between groups, we decided to make a scatterplot containing only genes that 1) could be found in both datasets ("matched" genes) and 2) were expressed in a at least 10 cells i each dataset (OSF≥10). This is shown in Figure 12. Again, ratios of burst frequencies (alpha ratios) and burst size (beta ratios) were calculated (Figure 13). Then, SD was calculated for both alpha and beta ratios and red lines marking 3 times SD (3SD) were added to the plot. The genes outside of 3SD were then extracted and added to a table (Figure 14).

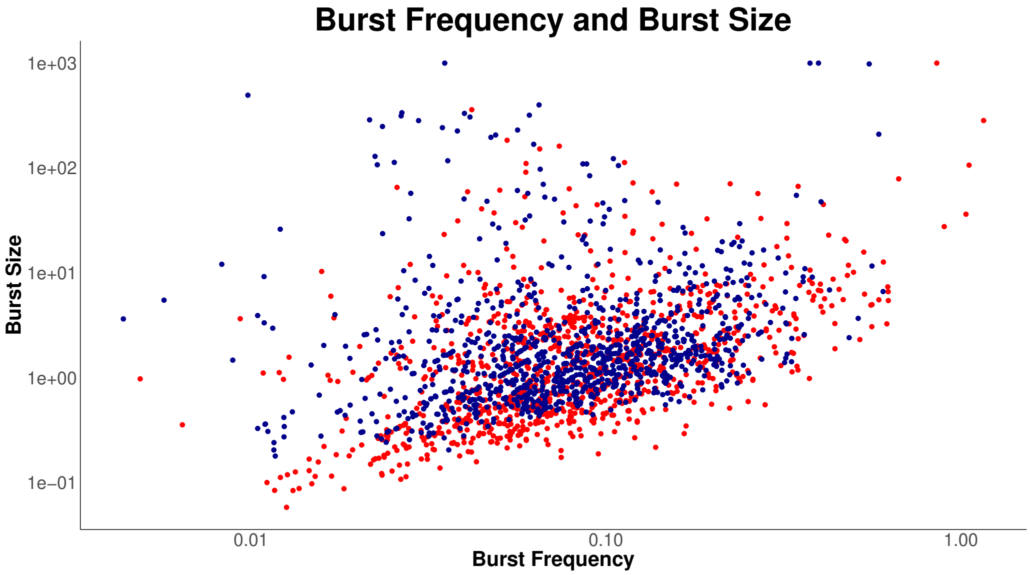


Figure 12: Burst kinetics of genes expressed in 10 or more cells in each dataset, 1056 genes in total. Again, the FSCs are marked with red and the PCs are marked with blue. The "outliers" seen in Figure 7 and Figure 8 have disappeared from this plot, i.e. there are no clusters outside of the "main" cluster.

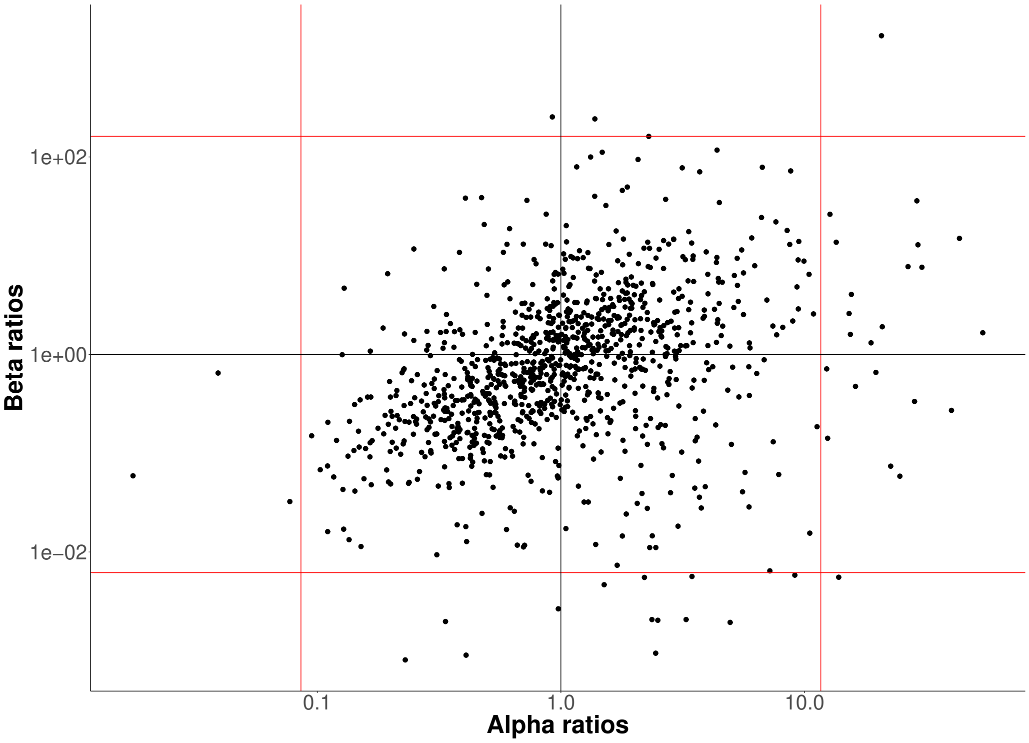
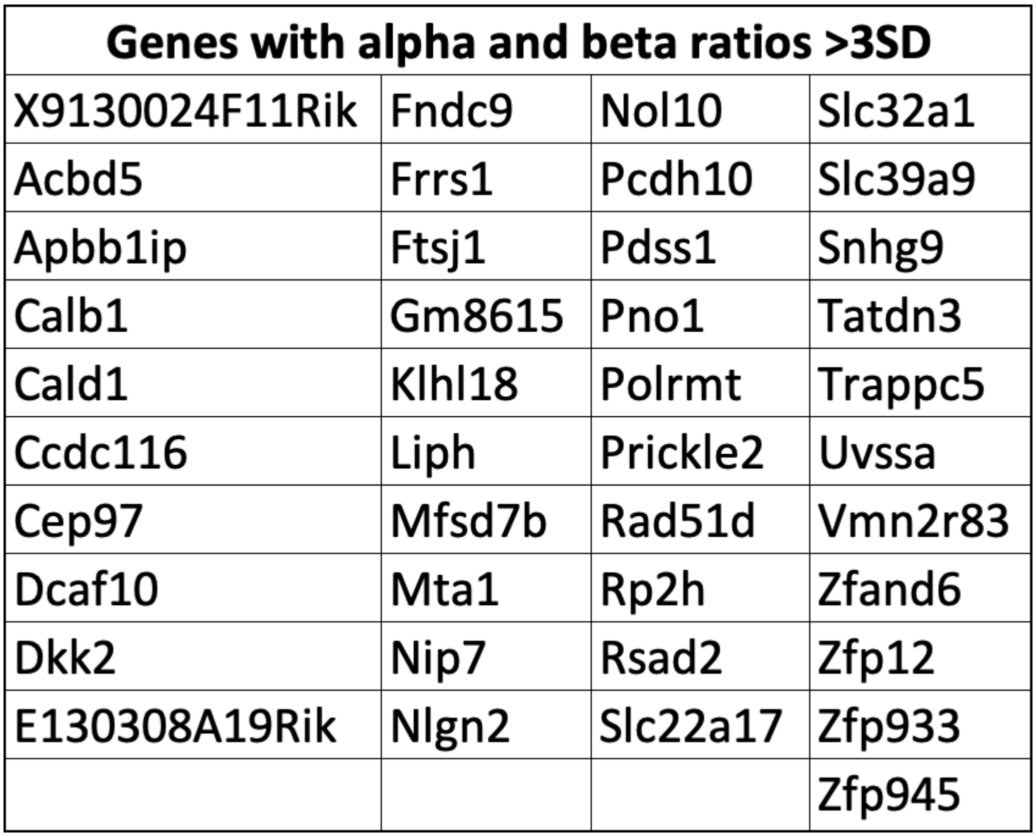


Figure 13: Alpha and beta ratios from the 1056 genes in Figure 12



Genes with alpha and beta ratios >3SD			
X9130024F11Rik	Fndc9	Nol10	Slc32a1
Acbd5	Frrs1	Pcdh10	Slc39a9
Apbb1ip	Ftsj1	Pdss1	Snhg9
Calb1	Gm8615	Pno1	Tatdn3
Cald1	Klhl18	Polrmt	Trappc5
Ccdc116	Liph	Prickle2	Uvssa
Cep97	Mfsd7b	Rad51d	Vmn2r83
Dcaf10	Mta1	Rp2h	Zfand6
Dkk2	Nip7	Rsad2	Zfp12
E130308A19Rik	Nlgn2	Slc22a17	Zfp933
			Zfp945

Figure 14: Genes located outside of the red lines (3SD) in Figure 13.

4. Discussion

4.1 Genes of Interested

In the genes that showed a significant difference (at least 3SD), there are some that may be of interest, such as the Calb1 gene which it believed to control the entry of calcium into the cell upon stimulation of glutamate receptors and Pcdh10, which is thought to play a role in establishing cell-cell connections in the brain. Unfortunately this discovery was made very late and so we did not have the time to investigate the genes or their properties further.

4.2 The Beta-Poisson Model for Single-Cell RNA-seq Data

While the Beta-Poisson model has some advantages in its relative ease of use and accessibility, it lacks some crucial functions needed for accurate inference of burst kinetics on scRNA-seq datasets. The mathematical model we used here is just that - a model. It is simplified and does not contain all the possible parameters that would influence gene transcription. The model forces the Poisson distribution, and fitting a curve to the dataset will eliminate a portion of the data, because some of the data may fit better and other parts of the data will not fit as well. The authors claim better fitting than other models that were popular at the time (2016), however these models were designed for bulk RNA-seq. data and not single-cell RNA-seq. data, so the comparison may be unfair. In addition, the authors claim that the model can be used in samples with as few as 25 cells. This is contrary to our findings, which show that with 25 cells and a filtering of only 5% (i.e., genes with expression in only 2/25 cells are accepted by the model and used in calculations), the model is inaccurate and shows characteristic clusters that disappear once we select out genes with expression in 10 or more cells in each dataset. In recent years, or even recent months, multiple mathematical models to infer burst kinetics on scRNA-seq. datasets have been developed that may have better fitting than the

Beta-Poisson model.

4.3 The Two-State Model of Transcription

The two-state model of transcription (ie. genes are either in ON state or in OFF state) that is implicit in the Beta-Poisson model fails to address the fact that gene-state switching is not a one-step process – multiple steps are involved, including but not limited to chromatin opening, recruiting transcription factors and transcription initiation (S. Luo et al., 2023). The two-state model has even been described as "largely unsatisfactory" by Tunnacliffe and Hubb (2020). mRNA from a gene may be found in the cytosol without the gene being "switched on" at that exact moment, and due to the "snapshot" nature of our dataset (i.e. we have a pseudo-time axis), we do not know whether the amount of mRNA is increasing or decreasing at the time of cytosol extraction.

4.4 Memorylessness and Randomness

A major drawback of the Beta-Poisson model is that it assumes Markovian dynamics, or "memorylessness", in gene transcription; meaning that what happens in the present is not influenced by what happened in the past, and the future is not influenced by what happens in the presence. We know that this is not true in gene transcription, because there are many molecular mechanisms that determine how and when transcription will happen. It would be more appropriate to use a statistical model that takes prior knowledge into account and modifies the output accordingly, such as Bayesian statistics, which is often used in machine learning algorithms.

4.5 mRNA versus Expressed Protein

Because the proteins are not measured directly, but indirectly through cytosolic mRNA, we do not know if our data is truly valid. While there is a positive correlation between the amount and type of mRNA in the cytosol and the amount and type of encoded proteins in a cell, it is not clearly established how much of the cytosolic mRNA is translated into protein (Liu et al., 2016). mRNAs produced from different genes seem to show different cellular dynamics, depending on factors like their total length, the size of their poly-A tail and the number of translational initiation sites. Studying cytosolic mRNA is therefore not equivalent to studying expressed proteins, however it is the only option we have at the moment where neurons can be identified electrophysically before analysis. The amount of protein from all the genes in a cell is simply too small to measure directly, and proteins cannot be amplified in the same way as mRNAs.

4.5.1 Mouse Brains versus Human Brains

As the introduction of this thesis stresses the importance of finding new "druggable targets" (i.e. molecular structures in the brain that can be targeted with medicine) for psychiatric diseases in humans and the data we are working on is derived from mouse brains, we find it necessary to mention the limitations of mouse studies. Apart from the obvious difference in size, the rodent brain and the human brain share no common homologes between each other within the cortex. For example, there is no such thing as a dlPFC in the mouse brain - some scientists would even say that mice have no PFC at all. Some of the most striking differences between mice and human brains are how the genes react and respond to drugs. Specifically in serotonin, where the genes that sense serotonin differ between mice and humans, resulting in a different reaction to the neurotransmitter As serotonin receptors are a common druggable target used for depression, mouse models using this neurotransmitter will not show the same reaction as it will in humans (Hodge et al., 2019). Because of this, results from experiments done on mouse brains cannot be generalized to humans, but they can help aid our basic understanding of psychiatric disease.

4.6 Data Processing

4.6.1 Normalization and Filtering

Our data is not normalized in the way that the model suggests it should be, because this type of normalization (FPKM: Fragments per kilobase per million, CPM: Counts per million) is not suitable for between-sample comparisons (L. Ravasz et al., 2020). The question remains whether or not this matters when our main focus is to look at the differences between the two cell types. After all, the ratios between them are what matter - but because we have limited knowledge of exactly how the Beta-Poisson model works, we cannot rule out the possibility that normalization of data plays a part in the final results.

Another problem is the statistical significance of all the "lost" genes in the dataset. The filtering steps that were necessary to validate and compare the two datasets led to a huge loss of data, and we do not know whether what we are left with is representative for our dataset.

4.6.2 Unresolved Issues in R

When running the program in R, we consistently got the following error messages when calculating the Monte Carlo p-values (using the getBPMCnullmatrix function):

```
In ppois(x2, m): NaNs produced

In ppois(x1, m): NaNs produced
```

Where "ppois" is a function that gives the cumulative Poisson distribution and "NaN" is "Not a Number". This is a warning message that would show up if the lambda parameter is invalid. Even after extensive troubleshooting, we were unable to figure out the reason for this error message, and we have simply not had enough time to fix it. Therefore, we cannot be completely sure that our data is valid.

5. Conclusion

5.1 True Differences in Transcriptional Burst Kinetics?

Our analysis shows a significant difference between the transcriptional burst kinetics in fast-spiking cells compared to pyramidal cells in 41 genes.

We believe that this idea is worth exploring further. We suggest that this dataset should be re-analyzed using newer mathematical models for inferring burst kinetics on static single-cell RNA seq. datasets, such as the generalized telegraph model developed by S. Luo et al. (2023), which introduces some non-Markovian parameters, or other Bayesian models such as the beta-gamma-Poisson model developed by X. Luo et al. (2022).

5.2 The Usefulness of Mathematical Models

The mathematical models used to infer burst kinetics on scRNA-seq. datasets are continuously improving, and although they are not and never will be perfect, they are useful in bridging the gap between live single-cell studies of gene transcription and static scRNA-seq datasets. Models like the Beta-Poisson model used in this thesis will in the future create more accurate predictions and ultimately lead to better personalised medicines.

In addition to this, our hope is that bioinformatics work similar to what we have done here can ensure that scientists get more out of each dataset generated, which in turn could lead to fewer experiments on animals.

Bibliography

Arnsten, A. F. T. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature Reviews Neuroscience*, *10*(6), 410–422. https://doi.org/10.1038/nrn2648

Bekkers, J. M. (2011). Pyramidal neurons. *Current Biology*, *21*(24), R975. https://doi.org/10.1016/j.cub.2011.10.037

Britannica, E. (n.d.). *Nerve cells*. Britannica ImageQuest. https://quest-eb-com.galanga.hvl.no/images/132_1285561

Cahass. (2006). *Pyramidal cell*. Wikimedia Commons. https://commons.wikimedia.org/w/index.php?curid=651365

Crabtree, G. W., & Gogos, J. A. (2014). Synaptic plasticity, neural circuits, and the emerging role of altered short-term information processing in schizophrenia. *Frontiers in Synaptic Neuroscience*, *6*. https://doi.org/10.3389/fnsyn.2014.00028

Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., Simpson, M. L., & Weinberger, L. S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, *109*(43), 17454–17459. https://doi.org/10.1073/pnas.1213530109

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, *297*(5584), 1183–1186. https://doi.org/10.1126/science.1070919

Ferguson, B. R., & Gao, W.-J. (2018). PV interneurons: Critical regulators of e/i balance for prefrontal cortex-dependent behavior and psychiatric disorders. *Frontiers in Neural Circuits*, *12*. https://doi.org/10.3389/fncir.2018.00037

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349. https://doi.org/10.1152/jn.1989.61.2.331

Hebb, D. (1949). *The organization of behavior* (First). John Wiley Sons.

Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Johansen, N., Penn, O., Yao, Z., Eggermont, J., Höllt, T., Levi, B. P., Shehata, S. I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., … Lein, E. S. (2019). Conserved cell types with divergent features in human versus mouse cortex. *Nature*, *573*(7772), 61–68. https://doi.org/10.1038/s41586-019-1506-7

Hu, H., Gan, J., & Jonas, P. (2014). Fast-spiking, parvalbumin sup/sup GABAergic interneurons: From cellular design to microcircuit function. *Science*, *345*(6196). https://doi.org/10.1126/science.1255263

Illumina. (2016). Illumina sequencing by synthesis.

Jiang, R., Sun, T., Song, D., & Li, J. J. (2022). Statistics or biology: The zero-inflation controversy about scRNA-seq data. *Genome Biology*, *23*(1). https://doi.org/10.1186/s13059-022-02601-5

Kim, J., & Eberwine, J. (2010). RNA: State memory and mediator of cellular phenotype. *Trends in Cell Biology*, *20*(6), 311–318. https://doi.org/10.1016/j.tcb.2010.03.003

Knopman, D. S., Amieva, H., Petersen, R. C., Chételat, G., Holtzman, D. M., Hyman, B. T., Nixon, R. A., & Jones, D. T. (2021). Alzheimer disease. *Nature Reviews Disease Primers*, *7*(1). https://doi.org/10.1038/s41572-021-00269-y

Kolk, S. M., & Rakic, P. (2021). Development of prefrontal cortex. *Neuropsychopharmacology*, *47*(1), 41–57. https://doi.org/10.1038/s41386-021-01137-9

Larsson, A. J. M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O. R., Reinius, B., Segerstolpe, Å., Rivera, C. M., Ren, B., & Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. *Nature*, *565*(7738), 251–254. https://doi.org/10.1038/s41586-018-0836-1

Liu, Y., Beyer, A., & Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell*, *165*(3), 535–550. https://doi.org/10.1016/j.cell.2016.03.014

Luo, S., Zhang, Z., Wang, Z., Yang, X., Chen, X., Zhou, T., & Zhang, J. (2023). Inferring transcriptional bursting kinetics from single-cell snapshot data using a generalized telegraph model. *Royal Society Open Science*, *10*(4). https://doi.org/10.1098/rsos.221057

Luo, X., Qin, F., Xiao, F., & Cai, G. (2022). BISC: Accurate inference of transcriptional bursting kinetics from single-cell transcriptomic data. *Briefings in Bioinformatics*, *23*(6). https://doi.org/10.1093/bib/bbac464

Nicolas, D., Phillips, N. E., & Naef, F. (2017). What shapes eukaryotic transcriptional bursting? *Molecular BioSystems*, *13*(7), 1280–1290. https://doi.org/10.1039/c7mb00154a

Patel, K. R., Cherian, J., Gohil, K., & Atkinson, D. (2014). Schizophrenia: Overview and treatment options. *P T*, *39*(9), 638–645.

Piller, C. (2022). Blots on a field? *Science*, *377*(6604), 358–363. https://doi.org/10.1126/science.add9993

Ravasz. (2020a). *The canonical network of pfc containing inputs (i1–i12) and outputs (o1–o2) of fs and pyr cell.* https://doi.org/10.1093/cercor/bhaa195

Ravasz. (2020b). *The canonical network of pfc containing inputs (i1–i12) and outputs (o1–o2) of fs and pyr cell.* https://doi.org/10.1093/cercor/bhaa195

Ravasz, L., Kékesi, K. A., Mittli, D., Todorov, M. I., Borhegyi, Z., Ercsey-Ravasz, M., Tyukodi, B., Wang, J., Bártfai, T., Eberwine, J., & Juhász, G. (2020). Cell surface protein mRNAs show differential transcription in pyramidal and fast-spiking cells as revealed by single-cell sequencing. *Cerebral Cortex*, *31*(2), 731–745. https://doi.org/10.1093/cercor/bhaa195

Sas-Nowosielska, H., & Magalska, A. (2021). Long noncoding RNAs—crucial players organizing the landscape of the neuronal nucleus. *International Journal of Molecular Sciences*, *22*(7), 3478. https://doi.org/10.3390/ijms22073478

Steward, M. M., Sridhar, A., & Meyer, J. S. (2012). Neural regeneration. In *Current topics in microbiology and immunology* (pp. 163–191). Springer Berlin Heidelberg. https://doi.org/10.1007/82_2012_302

Sul, J.-Y., Wu, C.-w. K., Zeng, F., Jochems, J., Lee, M. T., Kim, T. K., Peritz, T., Buckley, P., Cappelleri, D. J., Maronski, M., Kim, M., Kumar, V., Meaney, D., Kim, J., & Eberwine, J. (2009). Transcriptome transfer produces a predictable cellular phenotype. *Proceedings of the National Academy of Sciences*, *106*(18), 7624–7629. https://doi.org/10.1073/pnas.0902161106

Tunnacliffe, E., & Chubb, J. R. (2020). What is a transcriptional burst? *Trends in Genetics*, *36*(4), 288–297. https://doi.org/10.1016/j.tig.2020.01.003

Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., & Pawitan, Y. (2016). Beta-poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, *32*(14), 2128–2135. https://doi.org/10.1093/bioinformatics/btw202