Small RNA dynamics in cholinergic systems

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

Vorgelegt beim Fachbereich 14
der Johann Wolfgang von Goethe-Universität
in Frankfurt am Main

von
Sebastian Lobentanzer
aus Schlüchtern

Frankfurt 2019
(D30)

Small RNA dynamics in cholinergic systems

Abstract

Science still is very much in the discovery stage when it comes to transcriptional interactions, be it the long known workings of transcription factors or the recently discovered subtle fine-tuning of expression by small RNA, including microRNAs and transfer RNA fragments.

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.

Contents

This is the dedication.

»Ever tried. Ever failed. No matter.
Try again. Fail again.
Fail better.«

Simon Beckett

Acknowledgments

Thanks are due, for every scientist is not only standing on the shoulders of giants, but also on those of very real persons, without whom this dissertation would not have been possible. consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

Abbreviations

ACh acetylcholine

AD Alzheimer's Disease

Ago argonaute (protein)

API application programming interface

BD Bipolar Disorder

CAGE 5' cap analysis of gene expression

CNS central nervous system

DE differentially expressed

DMEM Dulbecco's modified eagle medium

FCS fetal calf serum

FDR false discovery ratio

GEO Gene Expression Omnibus (NCBI)

GO Gene Ontology

gp130 see IL6ST (gene)

LA-N-2 human neuroblastoma cell line (female)

LA-N-5 human neuroblastoma cell line (male)

miRNA microRNA

NCBI National Center for Biotechnology Information

PBS phosphate buffered saline

RT-qPCR real-time quantitative polymerase chain reaction

PD Parkinson's Disease

RIN RNA integrity number (RNA quality measure)

RISC RNA-induced silencing complex

RPMI1640 Roswell Park Memorial Institute medium

SCZ Schizophrenia

RNA-seq RNA sequencing

smRNA small non-coding RNA

SQL structured query language

TF transcription factor

tiRNA transfer RNA half

tRF transfer RNA fragment

tRNA transfer RNA

UTR untranslated region

vAChT vesicular acetylcholine transporter (from SLC18A3 gene)

Gene Symbols

ACHE acetylcholinesterase

AIF1 allograft inflammatory factor 1 (microglia marker protein)

CHAT choline acetyltransferase

CHRNA7 nicotinic acetylcholine receptor subunit $\alpha 7$

CNTF ciliary neurotrophic factor

CNTFR ciliary neurotrophic factor receptor (soluble)

GFAP glial fibrillary acidic protein (central astrocyte marker)

IL-6 interleukin 6

IL6R interleukin 6 receptor (soluble)

IL6ST interleukin 6 signal transducer (membrane bound; also known as gp130)

JAK janus kinase

LIF leukaemia inhibiting factor

LIFR leukaemia inhibiting factor receptor (soluble)

NGF nerve growth factor

OLIG1 oligodendrocyte transcription factor 1

RBFOX3 RNA-binding Fox-1 homolog 3 (neuronal marker gene; also known as NeuN)

SLC18A3 vesicular acetylcholine transporter (official gene symbol)

SST somatostatin

STAT signal transducer and activator of transcription

TYK tyrosine kinase

VIP vasoactive intestinal peptide

I know words. I have the best words.

Donald Trump

1 Introduction

1.1 Cholinergic Systems

Nary a process in the mammalian body can commence without participation of cholinergic systems. Acetylcholine (ACh) was chemically and pharmacologically described by Henry Dale more than 100 years ago[1]. A short time later, Otto Loewi published the first proof of signal transmission by small molecules: he transferred physiological solutions from electrically stimulated frog hearts to naive hearts and observed their reactions; the solution that provoked a parasympathetic response he proposed to contain a »vagus substance«[2]. Finally, in 1929, Henry Dale completed the picture by isolating acetylcholine from mammalian tissue and identifying it as the molecule responsible for the parasympathetic response[3]. Dale and Loewi's joint effort in »Discoveries in Chemical Transmission of Nerve Impulses« was rewarded with the »Nobel Prize in Physiology or Medicine« in 1936.

Although we have learned much about cholinergic systems in these past 100 years, our understanding of the mammalian nervous system still is fairly limited. Even when disregarding peripheral nervous systems, the complexity of cholinergic transmission is immense, and a myriad functions have been attributed to cholinergic circuits in the central nervous system (CNS). Central nervous projections of cholinergic fibres were extensively mapped by M. Marsel Mesulam and others in the 1980s[4], with a majority of long projection neurons originating in one of the eight cholinergic nuclei, Ch1-Ch8. While many of these anatomical structures have been filled with meaning by associations with both rudimentary as well as higher brain functions, there are still as many cholinergic pathways whose function is entirely unclear (Figure 1.1, from my first manuscript[5]).

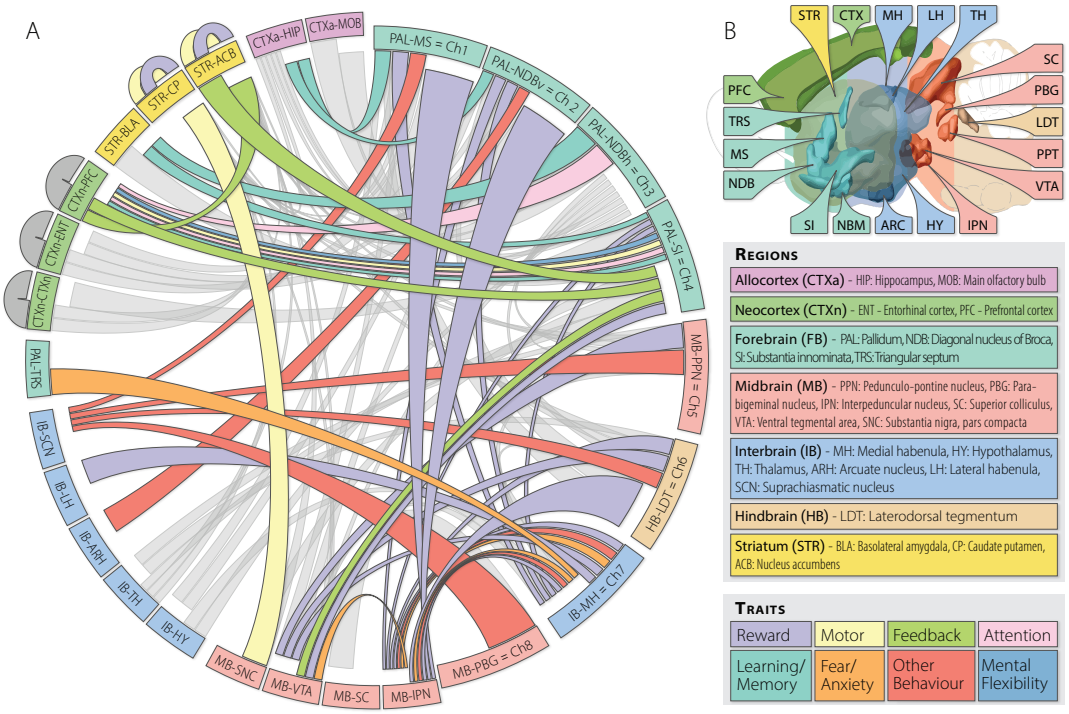

Figure 1.1: This is a figure that floats inline and here is its caption.

This holds particularly true for the only recently discovered cortical cholinergic interneurons, which, in comparison to their projecting counterparts, are very small and numerically vastly inferior to other neuron types in the cortex. Thus, their detection and analysis with current methods is challenging.

1.1.1 Cholinergic Aspects of Disease

Cholinergic systems are integral for a myriad physiological functions, and as such they are critically involved in aetiologies and phenotypes of a number of central and peripheral diseases. Of interest to this dissertation are the cholinergic aspects of degenerative and non-degenerative central nervous diseases (such as Alzheimer's Disease, Bipolar Disorder, Schizophrenia), ischemic conditions in stroke, and peripheral modulation of immune responses, particularly in the context of the aforementioned diseases.

Alzheimer's Disease

Cholinergic progression

Monotherapeutic approaches

Schizophrenia and Bipolar Disorder

Dirty therapeutics, multitarget

Stroke

Immunity

1.1.2 Neurokines

In comparison to the widely studied cholinergic projection neurons originating in the basal forebrain (Ch1-Ch4) that are known to depend on a retrograde survival signal by means of nerve growth factor (NGF), trophic influences on other cholinergic populations such as the cortical interneurons are unclear. NGF was described by Rita Levi-Montalcini in the 1950s as the first known instance of trophic peptides required for the survival of sympathetic ganglia[6], and the dependence of basal forebrain cholinergic neurons on retrograde NGF signalling was discovered in the 1980s[7].

A second group of trophic peptides with cholinergic implications are the so-called »neurokines«; the name results from the fact that this particular subgroup of cytokines has been associated with neuronal function in the central and peripheral nervous systems. Most prominently they include the ciliary neurotrophic factor (CNTF), leukaemia inhibiting factor (LIF), and interleukin 6 (IL-6), all of which coincidentally have been known under the acronym CDF. In the end of the 1980s, two groups of scientists (McManaman[8] and Rao[9]) independently identified proteins in extracts of muscle fibre that induced a differentiation of neurons towards a cholinergic type, and thus termed these proteins »choline acetyltransferase development factor« or »cholinergic differentiation factor« (both abbreviated CDF). Only later, through sequencing of the peptides, it became known that they had in fact discovered two distinct neurokines, LIF (Rao) and CNTF (McManaman, personal communication). IL-6, on the other hand, is abbreviated CDF for an entirely different reason: in this case it is short for »CTL (cytolytic T lymphocyte) differentiation factor«.

CNTF, LIF, and IL-6 convey their impact on neuronal activity through a partly redundant neurokine receptor pathway[10]. There are two basic types of neurokine receptors: soluble and transmembrane. The primary receptors for CNTF (CNTFR) and IL-6 (IL6R) are soluble proteins that are secreted into the extracellular space and, upon binding of a neurokine, bind to transmembrane receptor dimers on the cell surface. These transmembrane receptors are the LIF receptor (LIFR) and the »interleukin 6 signal transducer« (IL6ST, also known as gp130). Every neurokine has its preferred constellation of soluble and transmembrane receptors: CNTF binds to the soluble CNTF receptor and a dimer consisting of one gp130 and one LIFR protein; IL-6 binds to the soluble IL6R and a dimer of two units of gp130; LIF does not usually bind a soluble receptor but rather binds immediately to a dimer comprising one of each gp130 and LIFR; however, there is significant redundancy and crosstalk between those systems[11,12].

All receptor constellations result in a main effect of activation of the JAK/STAT cascade (Fig. 1.2). More specifically, neurokines can activate janus kinases (JAKs) 1 and 2 or the homologous tyrosine kinase (TYK) 2, and, successively, STAT (»signal transducer and activator of transcription«) isoforms

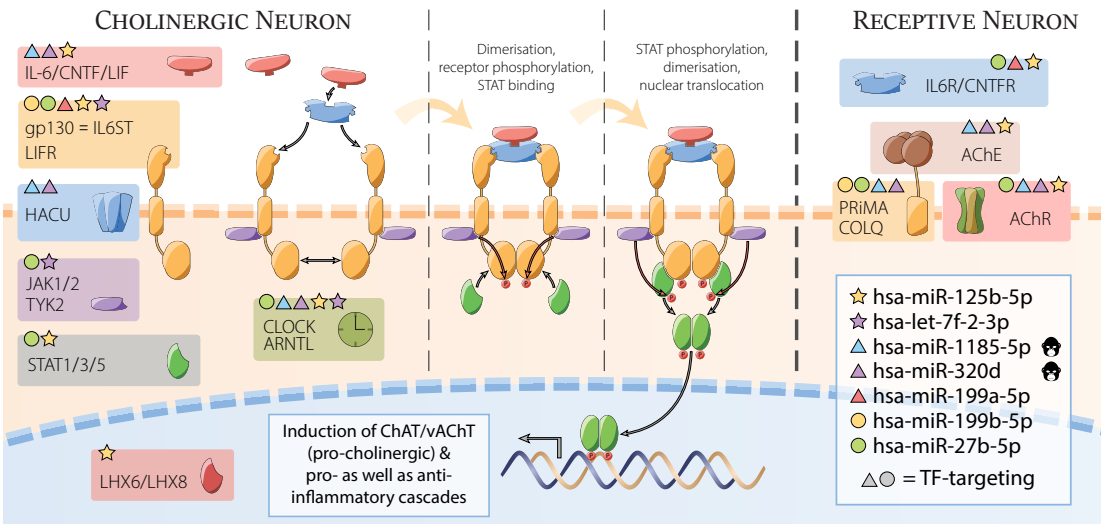

Figure 1.2: This is a figure that floats inline and here is its caption.

1, 3, 5A, and 5B, which then convey a multitude of cellular effects (e.g. in immunity or differentiation) through transcriptional activation. The STAT cascade is inherently self-limiting in that it usually leads to expression of transcription factors that serve as repressors of the STAT genes (XXX).

Neurokines, particularly IL-6(?), might serve as a link between the immunological and cholinergic aspects of physiological or disease processes. elaborate

1.2 Transcriptional Connectomics

The term »connectomics« is not strictly limited to one scientific discipline; it is frequently used when the studied matter is defined by complex relationships between interaction partners. The most frequent use outside of transcriptional matters is neuronal connectomics, i.e., the relationships and projections between brain regions. In this dissertation, connectomics generally refers to epi-transcriptional interaction, the processes surrounding protein-coding gene expression. For the sake of simplicity, in this dissertation all descriptions of genomics and transcriptomics matters, of genes and their small RNA regulators, are to be seen in the context of *Homo sapiens*, unless explicitly stated otherwise.

No matter their location, cholinergic neurons are defined by their ability to synthesise ACh and release it to neighbouring cells to a certain effect. To fulfil this task, two particular proteins are essential: the choline acetyltransferase (CHAT) to synthesise ACh from choline and acetyl-Coenzyme A, and the vesicular acetycholine transporter (vAChT, official gene symbol SLC18A3), which concentrates ACh in vesicles for later release. A notable genetic feature connects these two proteins beyond their functional association: the small *SLC18A3* gene (2420 nucleobases) sits inside the first intron of the *CHAT* gene and thus is already included in its primary transcript, and is subject to the *CHAT* promoter. However, oftentimes the (mature) transcript levels of *CHAT* and *SLC18A3* mRNA seem to be independently regulated; from the perspec-

tive of the organism, the possibility of differential regulation between these two genes makes sense. Since *SLC18A3* does not possess its own promoter, this differential regulation has to be conveyed epigenetically.

This dissertation deals in large parts with approaches aiming to decipher these interactions; and while its primary topic revolves around cholinergic systems, the methods described in the following are designed to be applicable to the entirety of the genome/epigenome. Four particular types of cellular actors are subjects of these methods and therefore will be briefly introduced: genes in the classical sense as the conveyors of cellular function by encoding for proteins; transcription factors (TFs), a subclass of protein coding genes that are able to regulate the expression of other genes; microRNAs (miRNAs), a class of small non-coding RNA (smRNA) that has been known for approximately two decades and is reasonably well described functionally and mechanistically; and transfer RNA fragments (tRFs), a second class of regulatory smRNA that has only recently been rediscovered and is significantly less well described regarding its functionality.

1.2.1 Transcription Factors

Transcription factors (TFs) were among the first intracellular regulatory mechanisms to be discovered (the earliest article referencing the term »transcription factor« in its title on PubMed was published in 1972). TFs commonly translocate from the cytosol into the nucleus upon activation (often by phosphorylation), where they bind specific DNA sequences that usually range in size from 6 to 12 nucleobases. The regions containing these binding sites (about 100 - 1000 bases in size) determine the effect upon binding, which can be one of two main modes: either a promoter, leading to an increased activity of transcription in the downstream vicinity of the binding site, or a repressor, having the opposite effect.

There exists a vast body of knowledge on TF-interactions with genes, mostly due to the long period of time since their discovery and the multitude of scientific publications, most often studying single TFs and their interactions with few genes, but cumulatively curated by several organisations. One of the currently largest curations of TF data, TRANSFAC, saw its original release in 1988. While these curation efforts can be extensive, they may present with serious bias towards particular TFs that might hold more scientific interest and thus are published far more frequently than others. Recently, comprehensive efforts have extended the available data significantly. Driven by the advent of RNA sequencing (RNA-seq), computational approaches have become able to not only comprehensively predict TF-gene interactions, but to do so in a highly tissue-specific manner (see Section 2.2.3). The human body is estimated to express up to 2600 distinct DNA-binding proteins, most of them presumed TFs [13], although other studies give lower estimates.

1.2.2 microRNAs

The first endogenous »small RNA with antisense complementarity« was described in 1993 [14], but microRNAs (miRNAs) were only recognised as a distinct regulatory class of molecules in the early 2000s. They are typically between 18 and 22 nucleobase-long, single stranded RNA fragments, and their function is now largely undisputed: miRNAs serve as targeting molecules for a protein complex whose primary purpose is to repress translation of mRNA, and, in some cases, lead to mRNA degradation. The complex, therefore, is called RNA-induced silencing complex (RISC); central to its function is the family of argonaute (Ago) proteins, which can bind the mature miRNA and orient it for interaction with its targets. Guidance of RISC to the target mRNA is generally mediated via sequence complementarity between miRNA and the targeted mRNA. Specifically, a »seed« region, usually bases 2-8 on the miRNA, is mainly responsible for the interaction; in case of perfect complementarity of this seed to the mRNA sequence, the interaction is considered »canonical«.

In early miRNA research, the 3' untranslated region (UTR) of the mRNA was believed to contain most miRNA binding sites due to its greater accessibility (i.e., the lack of active ribosomes); however, cumulative recent reports suggest that binding inside the coding region of the mRNA is a regular occurrence(cite). The rules governing miRNA binding to target sequences show considerable flexibility; a recent study shows about 30% of analysed relationships to be of »non-canonical« nature(cite). In those cases, seed pairing with the mRNA is often imperfect. To ameliorate this loss of stability, compensation occurs typically by a secondary complementary structure after a small gap of non-complementary bases, leading to a »bridge«-type constellation. This flexibility has implica- FIGURE? tions in applications involving targeting algorithms; those that consider only the seed region are more prone to false negatives than models that consider, for instance, the free energy of the whole molecule (see Section 2.2.4).

Biogenesis

miRNAs, similar to coding genes, are transcribed from loci on the genome, many inside introns or even exons of coding genes [15]. The primary transcript (primary miRNA or pri-miRNA) typically contains a hairpin-like structure that usually results in a double-stranded molecule because of internal complementarity, and can contain up to six mature miRNAs. This hairpin structure is recognised by the DGCR8 protein (DiGeorge Syndrome Critical Region 8, in invertebrates called »Pasha«); the complex then associates with the RNA-cleaving protein »Drosha«, which removes bases on the opposite side of the hairpin, creating a miRNA precursor (or pre-miRNA), which is subsequently exported from the nucleus by the shuttle protein Exportin-5. In a final step in the cytosol, the ribonuclease »Dicer« removes the loop joining the 3' and 5' arms of the pre-miRNA, resulting in a duplex of mature miRNA, about 20 nucleotides long. Initially, it was thought to contain only one active

miRNA, resulting in a designation of »miRNA*« for the complementary strand (commonly, the strand with lower expression). However, this notion has been disproven, and to reflect the possibility of both strands performing miRNA functions, nomenclature has changed to specify the arm of the pre-miRNA from which the mature form originates (suffix »-3p« for the 3' arm, and »-5p« for the 5' arm).

miRNA genes, in the same way as protein coding genes, can be subject to promoters and repressors, adding another layer of expression control by TFs. However, these TF-miRNA relationships are far less well described than common coding gene interactions, because miRNAs due to their shortness are not amenable to many standard gene expression assay forms. Estimation of the number of distinct gene targets of any one miRNA varies widely; however, it is generally accepted to not be less than several dozen targets per miRNA, and up to thousands of genes per miRNA (although that estimate might be overenthusiastic).

Organisation and Curation

miRNAs are organised and curated by means of a periodically updated web-based platform, miRBase [16]. For *Homo sapiens*, miRBase v21 contains 2588 mature miRNAs from 1881 precursors. Evolutionarily, the miRNA repertoire has grown from rodents to primates, resulting in a number of primate-specific miRNAs that may convey additional function. miRNA nomenclature is organised [17] in a way that assigns evolutionarily conserved miRNAs the same designation (number) in all species in which they are expressed. In their full names, a prefix stating the organism of origin is added; for example, hsa-miR-125b-5p (for *Homo sapiens*) and mmu-miR-125b-5p (for *Mus musculus*) share the same sequence and most of their functionalities.

miRNAs are subcategorised in families (designated »mir« with lowercase »r«) by their genomic origin and phylogenetic homology aspects. As the annotation itself, family affiliations are in flux and change with each miRBase version. miRBase v21 lists 151 distinct miRNA families with 721 individual members in total. The remaining 1867 miRNAs do not (yet) belong to a larger family; the majority (80%) of those is newly discovered, as indicated by a 4-digit designation number.

Disease Association

miRNAs have been associated with a number of CNS diseases, including Alzheimer's Disease (AD), Parkinson's Disease (PD), Bipolar Disorder (BD), and Schizophrenia (SCZ). However, the largest contribution since their discovery by far has been made by cancer research; of the approximately 90 000 publications found on PubMed with the term miRNA, about 42 000 involve cancer (search term »miRNA AND cancer«). In comparison, »miRNA AND Alzheimer's Disease« results in about 600 hits, while a search for »miRNA AND Schizophrenia« yields just 363 publications (as of October 2019).

In AD, several groups of miRNAs have been found to show characteristic perturbations before the onset of symptoms, which makes them interesting biomarker candidates[18]. Some miRNAs have been extensively studied in a variety of contexts, most prominently hsa-miR-132-3p. Among its targets are several key neuronal regulators (e.g. FOXP2, FOXO3, P300, MeCP2), and it is in turn controlled by many pivotal neuronal elements (e.g. REST, ERK1/2, CREB); this presents an explanation for the many physiological and pathological situations that miR-132-3p has been found to play a role in. Its functions include the control of neuronal survival/apoptosis, migration and neurite extension, neuronal differentiation, and synaptic plasticity.

add to abbreviations?

miRNAs are able to fulfil their regulatory purpose in a context- and cell-type-dependent manner[19], such that the perturbation of one single miRNA might provide different functional outcomes in different tissues (e.g., glial cells and neurons), or different stages of disease. However, this »jack-of-all-trades« behaviour also poses significant problems in establishing miRNAs as pharmacological targets: In the case of antagonising of mimicking an existing miRNA, the amount of off-target effects would not only be enormous, the entire definition of an off-target effect would continuously change between tissues and during the course of the disease. For this reason, the design of custom oligonucleotides with limited capabilities might be preferable in the development of therapeutics based on RNA interference (See also Section 5.2).

Prediction?

1.2.3 Transfer RNA Fragments

Transfer RNA (tRNA) breakdown products have been known for decades, with first descriptions in the 1970s; back then, they were associated with a higher turnover of tRNA in cancer cells[20], and proposed as urine-based biomarkers for certain malignancies[21]. However, their genesis was attributed to random processes, and due to lacking molecular biology characterisation techniques, interest in those fragments quickly faded. It was not until recently that studies have shown tRNA to be a major source of stable expression of small noncoding RNA[22,23] in most mammalian tissues. Indeed, replicating the reports from the 1970s, tRNA breakdown products are the dominant form of small RNA in secreted fluids, such as urine and bile, and make up large parts of other bodily fluids as well[24]. They exist in two major forms: transfer RNA halves (tiRNAs), and the smaller transfer RNA fragments (tRFs). *from stroke paper* tiRNAs derive from either end of the tRNA, and are created by angiogenin cleavage at the anticodon loop[25,26]. Smaller fragments are derived from the 3' and 5' ends of the tRNA (3'-tRF/5'-tRF) or internal tRNA parts (i-tRF), respectively, and may incorporate into Ago protein complexes and act like miRNAs to suppress their targets[27,28].

However, there is considerable controversy about the generalisation of tRF functions, as distinct publications discover very different and sometimes opposing mechanisms of action for their respective fragments. An obvious assumption is the miRNA-like functionality, at least for those tRFs that are in the length range of miRNAs. There have been several instances of tRFs proven to act as

miRNA-like suppressors of translation in a RISC-associated manner[28], and of Dicer playing a large part in their biogenesis[22]. There are even instances of small RNA molecules previously mislabeled miRNAs that have been discovered to actually be tRNA-derived, such as miR-1280[29].

On the other hand, multiple groups have identified tRFs to function not in an antisense-complementary manner, but by homology aspects. A valine-derived tRF was found to regulate translation by competing with mRNA directly at the binding site at the initiation complex and thereby displacing the original mRNA, leading to its translational repression[30]. Others have found multiple classes of tRFs derived from glutamine, aspartate, glycine, and tyrosine tRNAs, that displace multiple oncogenic transcripts from an RNA-binding protein (YBX1), conveying tumor-suppressive activity[31]. Most counterintuitive is the recent finding of a tRF proven to bind to several ribosomal protein mRNAs and enhancing their translation, and, when specifically inhibited, leading to apoptosis in rapidly dividing cells[32].

There is no consistent nomenclature yet to describe and organise tRFs, which are by nature more heterogeneous than miRNAs; while only 61 mature tRNAs are required in a cell to achieve a one-to-one »codon→amino acid« translation, one tRNA molecule can be the origin of several hundred distinct tRF molecules. Additionally, the amount of human tRNA genes is estimated at 500-600[33], and there are many more pseudo-tRNA genes. To communicate the identity of individual tRFs, multiple approaches are common in current literature; most prominently, tRFs are tied to the parent tRNA and the amino acid carried by this tRNA. To illustrate: The 22-nucleotide-long LeuCAG3′ tRF (meaning: a fragment of 22 bases starting at the 3' end of the leucine-carrying tRNA with anticodon »CAG«) was shown to play an important role in regulating ribosome biogenesis[32]. Since there is no repository of the likes of miRBase yet, this approach can be cumbersome for replication purposes, and explicit statement of the exact sequence of each fragment is a must in publication. In fact, since the aforementioned paper does not mention the sequence explicitly, there exist 6 distinct possibilities of fragments fitting this description. While manageable on this small scale, this system prohibits efficient analysis of larger sets of tRFs that cannot be individually controlled. For this reason, the approach of Loher and colleagues[34] might be preferable: they propose the generation of a "license plate" based on the sequence of the fragment directly, composed of the prefix »tRF«, the length of the fragment, and a custom oligonucleotide string encoding (e.g., »B3« stands for »AAAGT«). This way, tRF names are unique and unmistakably linked to the sequence, nomenclature is species-independent, and tRNA origin can be quickly determined by sequence lookup.

Disease?

? Levels of tRFs may be modulated even more rapidly than levels of miRNAs, since tRNA molecules are very abundant in the cell and generation of mature tRFs requires only enzymatic degradation of tRNA but no de-novo transcription of the molecule in the nucleus (citation).

1.3 Nested Multimodal Transcriptional Interactions - The Need for Connectomics

The ultimate aim of transcriptional connectomics is the combination of all interacting cellular components in a model that satisfactorily explains our real-life observations and is able to predict the functional outcome of a modification of one of these players. Even in the simplified case of only studying the interactions between coding genes, TFs, miRNAs, and tRFs, the complexity of the required model exceeds our current capabilities by far. The more we know about the functioning of these intertwined systems, the more we understand how much there is still to learn.

For example, only recently has it become clear how complex transcriptional regulation by means of TFs really is, and, incidentally, the two systems studied foremost in this dissertation (nerve and immune cells) are the two most transcriptionally complex systems in any mammal. Through study of comprehensive genomic information of 394 tissue types in approximately 1000 human primary cell, tissue, and culture samples (from the FANTOM5 consortium) it was estimated that the mean number of active TFs towards any given gene is highest in immune (12 TFs per gene) and nervous cells (10 TFs per gene), and that any one TF in nervous and immune cells controls expression of a mean of 175 and 160 genes, respectively[35] (see also Section 2.2.3).

Similarly, it has been found that miRNAs, particularly in the nervous system, possess a much higher tissue specificity than coding genes, resulting in an expression landscape that varies widely between individual neuron types that are in close proximity in the brain. With the exception of single cell RNA-seq, no modern analysis method is capable of a resolution appropriate for accurate characterisation of these expression patterns, resulting in extinction of the signal of miRNAs that are not expressed consistently across cell types (similar to »housekeeping« genes) because of statistical interference. Very recent studies show that miRNA-gene co-expression networks are tightly linked to cell types in the nervous system, and that groups of miRs as functional modules associate with particular phenotypes in developmental and mature states[36]. This functional association with cell phenotype was found in quality comparable to the expression patterns of TFs, yet in quantity conveys smaller impact and thus is thought to be a fine-tuning mechanism, subtle and precise in purpose.

Another aspect of the tissue specificity of CNS-associated miRNAs is the high likelihood of underrepresentation or even non-discovery of those very specifically expressed miRNAs. Adding to the problem is the experimental bias towards rodent models when it comes to thorough studies of the CNS, where human or other primate samples are a rarity compared to rats or mice. Assessments of the numbers of yet unknown novel primate- and tissue specific miRNAs estimate their magnitude in the thousands[37], resulting in an effective doubling of currently known miRNAs.

These high numbers of potentially interacting players present computational challenges: If estimating the number of expressed genes in a human cell at 20 000 (and the number of TFs at a low 1000), this makes for an estimated minimum of 200 000 »real« interactions in the possible

$C = \frac{1000!}{10!(1000-10)!} \cdot 20\,000$, which practically equals infinity; this is without accounting for different tissue types or cell states (e.g., differentiation or disease). Similarly, the amount of mature miRNAs (2588 in miRBase v21) and their ability to target even more distinct transcripts than TFs with one single molecule present immense computational requirements for even listing all possible or actual relationships. An interaction table describing targeting of genes by miRNAs in one type of tissue has $2588 \cdot 20\,000 \approx 50\,000\,000$ individual fields.

Combining the different modes of transcriptional interaction presents additional challenges. A simple model system to visualise (in only one type of cell) the interaction of TFs targeting genes, and of miRNAs targeting genes as well as TFs, contains about 20 000 genes (a subset of which of the size of about 2000 are TFs), 2588 mature miRNAs, and a total of $2588 \cdot 20\,000 + 2000 \cdot 20\,000 \approx 90\,000\,000$ potential interactions. In standard application scenarios, such as the generation of an interaction network around a group of genes (e.g., the cholinergic genes), the processing requirements grow linearly with each added interaction partner, and exponentially with every regulatory layer that is added.

example of standard interaction gene x miR x TF

Practically, this information has to be provided, gathered, and integrated, which further multiplies the amount of storage and processing power required. miRWalk 2.0, a collection of miRNA interaction data, has collected 12 of the most popular miRNA-targeting prediction datasets, each of which has their strengths and weaknesses (see 2.2.4). Experimentally validated interactions (e.g. as collected in DIANA TarBase or miRTarBase) are gold standard, but far from comprehensive and strictly speaking only relevant for the cellular context in which the experiment was originally performed; there are also different evidence qualities to be accounted for, depending on the type of experiment performed. Ideally, all of these data are still accessible when performing the analysis, so a database created for this purpose should be able to incorporate all this information without any data loss while still remaining feasible in terms of computation time as well as space and working memory requirements.

This dissertation will first describe the creation of such a database and what has been learned during its various stages, and then go on to apply the database to different biological problems from real world experiments, such as the cholinergic differentiation of human male and female cultured neuronal cells, and the blood of stroke victims.

»Wir sehen in der Natur nie etwas als Einzelheit, sondern wir sehen alles in Verbindung mit etwas anderem, das vor ihm, neben ihm, hinter ihm, unter ihm und über ihm sich befindet.«

Johann Wolfgang von Goethe

2

miRNet: Creation of a Comprehensive Connectomics Database

The need for bioinformatic support in connectomics is immediately obvious from the sheer multitude of possible interactions between the participating factors. However, when I began working on this project (October 2015), there was no integrative database available for this purpose. Earlier that year, miRWalk 2.0 had been published, for the first time providing a relatively comprehensive source of predicted as well as experimentally validated miRNA targeting data[38] (see 1.2.2). One year later, Marbach's »regulatory circuits« were published[35], enabling analysis of comprehensive TF→gene relationships in 394 human tissues (see Section 1.2.1). These collections (as well as the data they were derived from) are the basis of the database further called *miRNet*, the development of which will be described in the following chapter.

Since a large part of the scientific progress of this dissertation deals with practical problems of multimodal connectomics, I will begin by describing the infrastructure that makes effective computation of these problems possible. After this technical description of database structure and creation, I will explain the types and organisation of its content. The remainder of the chapter will then deal with the application of this infrastructure to real-world problems in transcriptional connectomics, and the statistical approaches suited to this special case.

2.1 Implementation

For any biological question to be asked in a bioinformatics setting, the effectiveness of the computational query determines the practicality of the approach. Because resources (i.e., processing power, storage, and working memory) are limited, the database that is queried should be organised in a way that facilitates retrieval of the desired information without excess processing of useless information. In the simplified case of only miRNAs interacting with genes in one direction (miRNA→gene), this means retrieval of only those interactions relevant for the queried genes or miRNAs.

Traditional table-based approaches (also known as relational databases) such as SQL (»Structured Query Language«) cannot provide such an implementation, since individual entries for genes and miRNAs (rows and columns) have to be accessed in their entirety, whether there is a connection between gene and miRNA (1) or not (0). Additionally, adding layers to these interactions (e.g., distinct prediction algorithms, tissues, or the interaction between TFs and genes) require the addition of entire tables the same size as the database, which is detrimental to effective use of space; and more complex queries also necessitate the transfer of information between those distinct tables (in SQL typically via a JOIN command), which claims additional working memory and processing time. Overall, the so-called »many-to-many« organisation of data does not lend itself to representation in a relational database.

Figure to explain tables?

The actual performance is determined by the processing power of the machine it is running on and several structural properties, such as organisation, indexing, monotony, and of course the size of the database; therefore, an estimation of processing time for queries is bound to be inaccurate. However, processing times typically do not vary on the scale of orders of magnitude, and thus general estimations can be made. Well optimised SQL databases with a size of 5 to 10 GB on disk usually require tens of minutes if not hours to complete one single complex query[39]; *miRNet* in its current form takes up approximately 15 GB of storage. Since one analysis typically consists of several hundreds (and, in the case of permutation analyses, several hundreds of thousands) of these queries, processing times in SQL implementation are too long to be practically useful. (It seems important to note that, as of 2018, SQL also offers a graph-based organisation in addition to the traditional, relational layout. These two are separate systems, and not to be confused. The advantages of Neo4j as explained in the following should be seen from the perspective of 2015, when the database was established, and when there was no graph-based SQL implementation.)

2.1.1 Neo4j: A Graph-Based Infrastructure

To query and display biological data that are organised in a network-like structure (many-to-many), a database that lends itself to the efficient processing and storage of network data is optimal. »Neo4j« utilises a database structure that is built on the save and recall of data points in »nodes« and »edges«, which represent entities (nodes) and relationships between those entities (edges); both nodes and

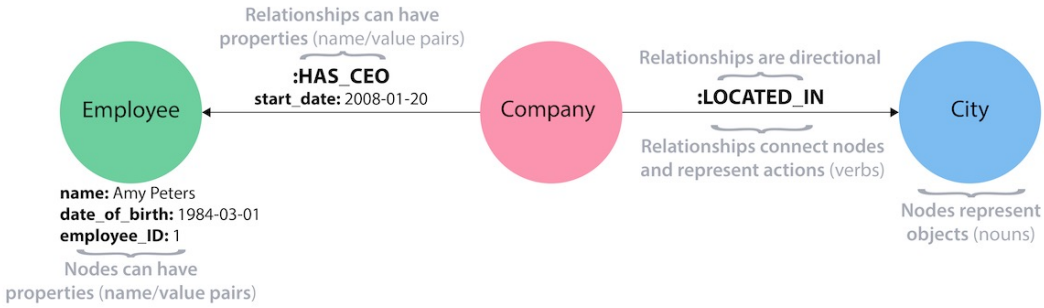

Figure 2.1: Organisation of a graph database. Will be replaced with own figure.

edges can have any number of attributes and a unique property called »type«, typically used to describe the class of the entry (such as »gene« or »miRNA«). This database organisation replicates the network-like structure of the biological data studied (Fig. 2.1). Neo4j combines the network-like data structure with an efficient indexing system for quickly finding the entries queried for, and then »walks« along the edges of the nodes that have been found, thus only searching and returning the data that is relevant to the current query. Theoretically, this makes the database more likely to be efficient in the setting of transcriptional interactions, an estimation that turned out to be true.

Depending on the input, these queries can also be rather large; however, the main pitfall of tabular databases such as SQL is circumvented: there is no need to process entire rows or columns of the table to make sure that the query is satisfied in its entirety. This is particularly useful in a setting of sparse information. To illustrate: Only 30 of the 2588 miRNAs target a specific gene, which is common; a relational database, after finding the index of the queried gene, would have to search 2588 fields for 1/0; the graph database, on the other hand, has to execute only 30 searches (or, more accurately, 30 »walks« along the edges connected to the indexed node). In practice, even in the very first prototype implementations, this accelerated standard-case computations approximately thousand-fold, and was even able to accommodate advanced approaches in situations that had been inaccessible in the tabular implementation.

2.1.2 High-throughput Database Generation

Neo4j provides several API (»application programming interface«) possibilities in implementation. For the purpose of entering large amounts of data into the database at once, the Java implementation is superior to the other forms in that it provides a batch processing mode via its `BatchInserter` class. I thus wrote a custom Java program for the purpose of creating an initial state of the database from the largest set of data, the complete miRWalk 2.0 content with 12 algorithms and validated interactions. The downloaded data was organised in a plain text based file format, with one text file

for each miRNA, totalling in size about 6 GB (for *H. sapiens*). The database was set up in a way that allows only one node for each individual miRNA and gene entered to avoid duplications, using the commands

- `createDeferredConstraint()`
- `assertPropertyIsUnique()`
- `createDeferredSchemaIndex()`

of the Neo4j Java package. This approach made sure to create only one node for each miRNA (type: MIR) and gene (type: GENE) in the data, which is essential for proper functioning of the database. Each of these nodes received several properties to store individual data, such as the various gene/miRNA identifiers, origin of data, and species.

Between those basic nodes, the batch insertion process created edges for each relationship that was found in the original data, assigning a type identifier to each edge detailing the origin of this interaction (type: name of the prediction algorithm or »VALIDATED« for experimental data). Thus, while the nodes for genes and miRNAs themselves are unique, an arbitrary number of relationships can exist between any two nodes, depending on how many interactions they share. aggregation here?

2.1.3 Maintenance and Quality Control

All additional datasets, such as the TF regulatory circuits or tRF targeting predictions, were entered into *miRNet* using the regular operation mode. Testing was also performed in regular operation, with manual as well as automated tests to assert the correct transfer of information from raw data to the graph database, and to avoid unpredictable behaviour. At times, conflicts had to be resolved manually, for instance when miRNA names conflicted between old »miRNA*« and new »3p/5p« notation; all manual edits are documented in the code, which was published alongside my first manuscript[5].

Except for the rapid import of large amounts of data in creation of a database, the Java implementation of Neo4j does not offer many advantages over the native R implementation, »RNeo4j«. Thus, after creation and a short period of experimentation with graphical user interfaces, I abandoned the Java program in favour of the more flexible R programming. While Java is an object-based programming language, whose benefits lie in extreme flexibility in regards to platform and purpose, high modularity, and speedy processing, R as a procedural language is the work horse of modern bioinformatics. Its procedural design (the division of data and functions that operate on that data) facilitates the transfer of approaches between distinct datasets, and the enormous vibrant community of data scientists using R provides a wealth of third party packages to tackle almost any bioinformatic task.

2.2 Materials

All materials used in the creation of *miRNet* have been acquired from resources that are non-commercial, web-available, and open-source (in the case of code). All properties and relationships derived from this data were entered into *miRNet* as either nodes, properties of nodes, edges, or properties of edges.

2.2.1 Gene Annotation

Even though »regular« protein coding genes have been known for a long time, there is no consensus yet about their nomenclature and organisation. Complicated by newly discovered functions and properties of phylogenetic nature, the scientific representation of the human genome is in constant flux. Several large organisations strive to provide a robust annotation of the human gene catalog, but also in many cases contradict one another. There are three nomenclature systems that are of high importance in modern genomics:

- The traditional naming system of acronyms (e.g. CHAT) and fantasy-names (such as »Sonic Hedgehog«), also occasionally called »gene symbol«, is still widely popular because of its accessibility to humans, but is also not particularly robust because of a high amount of synonyms with high confusion potential (see e.g. Section 1.1.2 on CDF) and instances of genes without names having to carry unwieldy systematic names.
- The American (National Center for Biotechnology Information (NCBI)), a branch of the National Institute of Health (NIH), curates and hosts a multitude of biological and medical data, and for the organisation of gene information uses its own systematic nomenclature termed »Entrez« ID. Entrez is a molecular biology database that integrates many aspects of biology and medicine in a gene-centered manner, and therefore Entrez IDs are useful to quickly connect a gene to its function, nucleotide sequence, or associated diseases. Entrez IDs are regular integers without additional characters.
- Akin to the NCBI effort, ENSEMBL is a project of the European Bioinformatics Institute (EBI) as part of the European Molecular Biology Laboratory (EMBL). Compared to the Entrez database, it is more focused on study and maintenance of the genome itself, and therefore has a more intricate nomenclature that allows for differentiation of, for example, genes and their various transcript isoforms (ENSEMBL IDs carry character prefixes for class identification, e.g., ENSG for genes, ENST for transcripts).

All of these are being used on a regular basis in many publications, and, often, they are used exclusively. As a result, the end user of the published data has to have access to all possible annotation forms, or, at least, a means to translate one into the other; often, this also introduces conflicts. For this reason, all ID types were entered into *miRNet* upon creation or during maintenance, for convenience and to minimise analysis prolongations due to conflict resolution.

2.2.2 microRNA Annotation

miRBase provides a consistent annotation for miRNAs. Due to their relatively recent discovery, there still are major changes from version to version; the syntax, however, is stable. In addition to the miRNA »names« that are composed of species, the string »miR«, pre-miRNA designation number, and strand origin (not in all cases!), such as »hsa-miR-125b-5p«, miRBase provides IDs for pre-miRNA molecules (also called ancestors) termed »MIID«, and IDs for mature miRNA molecules termed »MIMAT«. However, in practice, these are rarely used. Similarly, miRNA families are annotated using the »MIPF« ID.

2.2.3 Transcription Factor Targeting

The FANTOM5 project has applied 5' cap analysis of gene expression (CAGE) to a large number of human samples from diverse tissues to determine the accurate 5' ends of each transcript[40]. Knowledge of this fact enables accurate prediction of promoters likely to control a transcript's expression. Marbach and colleagues used this information in combination with detailed human gene expression data to derive a complex interaction network of TFs and genes (»regulatory circuits«), and in doing so aggregated samples with similar expression patterns and origins into 394 fictional tissues[35]. For every tissue, each TF was assigned transcriptional activities towards all genes that it supposedly targets (with the sum of all activities in any given tissue being 1). Marbach and colleagues have shown that the cumulative transcriptional activities towards any given gene correlate well with the actual gene expression in corresponding samples from an independent repository.

Even in its fifth iteration, FANTOM data is not entirely comprehensive, which came to my attention due to a cholinergic anomaly: The 5' CAGE peaks of the *CHAT* and *CHRNA7* (the nicotinic $\alpha 7$ receptor subunit) genes in raw FANTOM5 data do not pass the expression threshold, and therefore are not included in, e.g., Marbach's »regulatory circuits«. Both are critically important not only for neuronal cholinergic systems, but also for the non-neuronal aspect of immune processes. For instance, macrophages have been shown to produce ACh via CHAT(cite), and the $\alpha 7$ homomeric ACh receptor conveys direct immune suppression by its expression on monocytes(cite). Paradoxically, the CAGE peak of *SLC18A3*, which lies in the first intron of *CHAT*, crosses the threshold and therefore is included in the data. Unfortunately, I was not able to remedy these circumstances even upon personal communication with Daniel Marbach (author of »regulatory circuits«) and Hideya Kawaji of the FANTOM5 consortium, although the latter acknowledged the possibility of a gene annotation deficit leading to misattribution of the *CHAT* signal to *SLC18A3* due to the closeness of their 5' ends.

The entire collection of transcriptional activities in all tissues was downloaded from the project's web page[35], and neuronal and immune tissues were manually curated and entered into *miRNet*. The collected data comprises 33 neuronal tissues and 26 immune cell tissues (Appendix A), and 1 130 196

TF→gene relationships in total (not all 394 tissues were entered).

2.2.4 microRNA Interactions

The content of miRWalk 2.0 is freely available online[41]; however, there is no option of downloading the complete set. The targeting data thus was downloaded per miRNA using a custom crawler, with standard options for all 12 prediction algorithms (miRWalk, miRDB, PITA, MicroT4, miRMap, RNA22, miRanda, miRNAMap, RNAhybrid, miRBridge, PICTAR2, and TargetScan) in plain text format. For experimentally validated interactions, the main sources were DIANA TarBase[42] and miRTarBase[43], both of which offer complete download options. As of 2019, the 3.0 version of miRWalk allows complete species downloads; however, the developers have abandoned their third party algorithm plurality reducing the number of available alternatives from 12 to 4, which can be considered a significant disadvantage:

While sequence complementarity, particularly of the »seed«-region, is the primary paradigm of miRNA-mRNA interaction, prediction algorithms vary widely in their implementation, general purpose, and approach to interaction prediction (for a comprehensive review of approaches and rules, see[44]). A large group of available options utilise sequence conservation aspects to increase candidate viability (such as miRanda, PicTar, TargetScan, and microT4). Others, such as RNA22 and PITA, utilise biophysical aspects such as free energy of binding or the accessibility of target sites due to secondary RNA structures as prediction arguments. All of these approaches have their up- and downsides, e.g. considering their general precision and sensitivity, or their adequate prediction of particular cases, such as multiple site targeting. Thus, it has been proposed to use a combination of complementary approaches instead of only one algorithm per analysis[45]. For this reason, I might have preferred the 2.0 version of miRWalk, even if 3.0 had been available at the time.

One advantage of the collection of all data in a quickly accessible database is the opportunity to compare the different approaches to target prediction. A statistical evaluation of the collected interaction data from miRWalk 2.0 showed vast differences in general prediction quantity (Table 2.1) as well as prediction accuracy and sensitivity when compared to the validated subset of data (Table 2.2). Since the ground truth is not known, this is an additional argument for the combination of multiple algorithms instead of the use of a single set. Apart from RNAhybrid and miRBridge, all algorithms presented reasonable base hit frequencies and increases in the validated test set. While miRBridge already has the lowest positive frequency of all the algorithms, it is the only one to achieve a negative score in the validated test set. On the other hand, RNAhybrid has a vastly higher base hit frequency than the second highest scoring algorithm (by more than 300%), making it very likely to produce false positive results, and less valuable in the aggregation scoring system. The remaining 10 algorithms were included in *miRNet* targeting data. For ease of use, an additional relationship type was created from the aggregated single algorithm hits of any miRNA→gene relationship, with the sum of algorithms predicting the interaction as a score variable. This yields a theoretical score

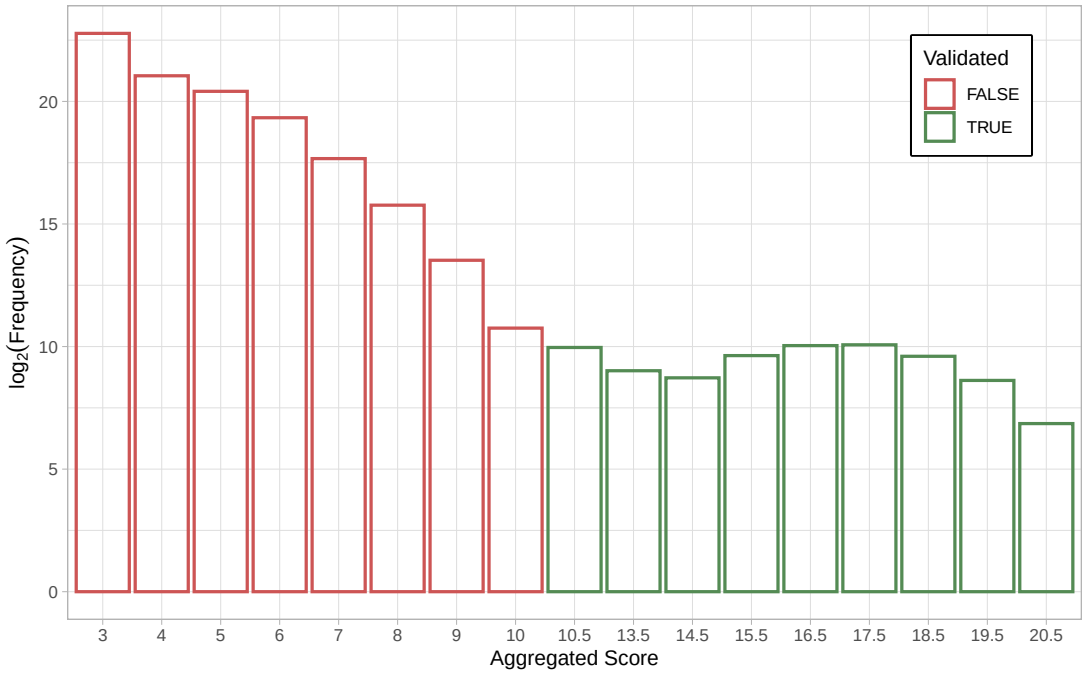

Figure 2.2: Histogram of miRNA→gene score distribution.

range from 3 to 10 (miRNA→gene relationships with only one or two hits were ignored for the sake of space). To account for experimentally validated interactions, each miRNA→gene relationship that was supported by strong evidence of interaction was modified by addition of 10.5 score points (a half point for quick identification of a validated relationship), extending the maximum score to 20.5 points. The resulting optimised graph contains 11 687 931 human miRNA→gene targeting relationships with a distinct score distribution (Fig. 2.2). In comparison, only 6146 miRNA→gene relationships are experimentally validated with »strong« evidence.

2.2.5 Filtering of Aggregated Prediction Scores

Describe the analysis of prediction that resulted in selection of score cutoffs 6 or 7. elaborate

2.2.6 Identification of Cholinergic Controllers

Where does the CHAT miRNA anomaly fit? It is an early result, before cell culture. It even led to the selection of the cellular model. Maybe there (3.3.1)? The CHAT TF anomaly is described in this chapter, 2.2.3. Alternatively, dedicated section at the end of this chapter, applying all described methods (2.5)? TFs not the only CHAT anomaly

2.2.7 De-novo Prediction of tRF Targeting

Due to the recency of their (re-)discovery, no comprehensive interaction sources exist for transfer RNA fragments. There have been documented cases of miRNA-like behaviours of distinct RNA

algorithm	hit frequency
RNAHYBRID	71.62%
MIRMAP	19.90%
MIRWALK	19.74%
TARGETSCAN	16.33%
RNA22	12.34%
MICROT4	11.81%
MIRANDA	10.65%
PITA	4.90%
MIRDB	1.17%
MIRNAMAP	0.75%
PICTAR2	0.62%
MIRBRIDGE	0.15%

Table 2.1: Prediction algorithms ordered by the fraction of all possible interactions they predict as being real (positive rate). Different algorithms display a wide variation of hit rates in the entirety of predicted interactions between any miRNA and gene. Red: excluded from analysis.

algorithm	validated hit frequency	hit rate increase
PICTAR2	6.98%	1129.40%
MIRDB	9.80%	838.43%
MIRANDA	51.73%	485.94%
TARGETSCAN	70.63%	432.51%
MIRNAMAP	3.10%	410.95%
PITA	15.57%	317.20%
MICROT4	32.60%	276.10%
MIRMAP	53.86%	270.65%
MIRWALK	50.95%	258.15%
RNA22	22.51%	182.38%
RNAHYBRID	90.47%	126.32%
MIRBRIDGE	0.01%	0.00%

Table 2.2: Prediction algorithms ordered by their increase in true positive rate when considering only validated interactions. The hit rate increase when comparing experimentally validated interactions with the entire predicted data (Table 2.1) is also subject to strong variation. Hit rate increase is the increase of hit rate if only considering validated data as opposed to all predicted interactions. None of the studied algorithms unite a good precision (hit rate increase) and coverage (validated hit frequency).

fragments[22,28], justifying an attempt to predict interactions in a comprehensive manner. Of the available options for nucleotide interaction prediction algorithms, TargetScan[46] seems particularly suited for this task because it provides the option of evaluating the evolutionary conservation of target sites in the putatively targeted genes, thereby providing an additional layer of security: The sequence of 3' UTRs is evolutionarily less stable than the coding part of genes; thus, high conservation of the binding site might indicate evolutionary pressure to keep up the interaction with the fragment, making an actual function of the interaction more likely. TargetScan also presents with reasonable sensitivity and specificity as confirmed by an independent group[47], and through an additional algorithm allows the attribution of a score based on the branch length (on the species tree) of conserved targeting[48].

miRNA-like behaviour implies the existence of a region on the tRF similar to a miRNA »seed«, and TargetScan also expects a seed as input to its targeting algorithm. Since there has been no definitive answer to the question as to where the seed region in tRFs might be, it is safest to assume nothing and explore all possibilities, i.e., simulate every possible seed position for interaction discovery. For this purpose, all discovered sequences of tRFs (exceeding a base mean expression of 10 counts) were chopped into 7-base pieces (7mers), which is the lenght of miRNA seeds, and statistically improbable enough to appear in the genome at random; the average length of a human 3' UTR is 800 bases, so the probability of finding any 7mer randomly in any one 3' UTR is $p = \frac{800}{4^7} = 0.049$, which agrees with the 5% false discovery ratio (FDR) convention.

Describe Targetscan process

2.3 Usage

Neo4j uses a language (called »Cypher«) akin to SQL, which utilises keyphrases to issue commands, but combines it with a semi-graphical syntax to account for the graph-based layout of the data. In the following, I will describe its basic usage and the advantages it provides in the matter of transcriptional connectomics. The basic »finder« function (similar to SELECT in SQL) is called MATCH in Cypher, and, when combined with the semi-graphical syntax, can be used to identify nodes or more complex patterns in the database. The graphical syntax consists of two main building blocks that represent the basic types of data inside the database: nodes as regular brackets »`( )`« and edges between nodes as a construct of hyphens and box brackets, that can also have a direction indicated by the greater sign »`( )-[ ]->( )`«. To specify the elements to be found, attributes of nodes and/or edges can be filtered by using curly brackets in the node definition, or the WHERE clause. To be returned, elements need to be assigned arbitrary variable names:

Listing 2.1: MATCH

```
1 MATCH (gene:GENE {species: 'HSA'})
2 WHERE gene.name = 'CHAT'
3 RETURN gene
```

Query 2.1 identifies a node (arbitrarily designated »gene«) with type GENE (indicated by the colon), with attributes »species« (HSA, i.e. *H. sapiens*) and »name« (CHAT), and returns the node with all its attributes. Since the nodes of type GENE are restrained, there can only be one gene of species *H. sapiens* with this name in the database, and thus, only one data point will be returned. The graphical syntax further allows for pattern matching of, for instance, miRNA→gene relationships:

Listing 2.2: Patterns

```
MATCH (mir:MIR)-[rel:TARGETS]->(gene:GENE {species: 'HSA'})
WHERE gene.name = 'CHAT'
RETURN mir, rel, gene
```

Query 2.2, similar to query 2.1, starts by identifying the node of species HSA with the name CHAT, and proceeds to look for miRNA→gene relationship edges arriving at this node; the relationships have to be of the type TARGETS (the pre-aggregated score-based accumulation of targeting). As soon as no further edges are found, the process terminates and returns all found miRNAs (»mir«), relationships (»rel«), and genes (»gene«) in discrete form, including all their attributes, such as the ENSG and Entrez IDs, the MIMAT IDs for all found miRNAs, or the score value of their targeting relationship. In this query, since there is a constraint on genes, the only gene returned is *CHAT*. However, Cypher is not limited to filtering on unique attributes; it allows for query and return of as many data points as are needed. For example, if one is interested in all miRNA→gene interactions in the cholinergic system, the query might look as follows:

Listing 2.3: Filtering

```
MATCH (mir:MIR)-[rel:TARGETS]->(gene:GENE {species: 'HSA'})
WHERE gene.name IN {cholinergic_genes}
RETURN mir, rel, gene
```

The effectiveness of graph-based databases becomes clear in this approach: Query 2.3 is processed starting at a user-defined filter, the list of cholinergic genes as an input (containing *CHAT*, *SLC18A3*, cholinergic receptor genes, acetylcholinesterase, etc). In a first step, all nodes are found that fulfil the criteria: type GENE, from species *H. sapiens*, that are in the list of names given. Since the gene nodes are indexed, this only requires milliseconds. Then, through the connection of edges to these nodes, it finds all miRNA nodes that have a miRNA→gene relationship towards any of the cholinergic genes. By using the gene nodes as starting point, the query can end as soon as no other edges fulfilling these criteria are found on any of the nodes. In comparison, to satisfy this query in a relational database, the rows representing these cholinergic genes would have to be assessed in their entirety, not only in those columns that represent an extant relationship, thus prolonging execution.

The database then returns all miRNA→gene relationships in this set, representing the network of

cholinergic miRNA regulators, including all of their attributes. The advantages of graph-based data do not end there; say one wants to return only »master« regulators of cholinergic systems, defined as miRNAs that target at least 4 of the genes in the cholinergic set. In a relational database, this would have to be done post-hoc, by aggregation of relationships and removal of any results that do not exceed this threshold. This requires storage of the entire result in memory, and additional computational steps that can be very taxing depending on the size of the result table. In Cypher, this can be done during the query (code comments indicated by »//« explain single steps):

Listing 2.4: Two-stage Filtering

```
MATCH (gene:GENE {species: 'HSA'})
WHERE gene.name IN {cholinergic_genes}
WITH gene //the found genes are used as input for the second query
MATCH (mir:MIR)-[rel:TARGETS]->(gene)
WHERE count(rel) >= 4
RETURN mir, rel, gene
```

Query 2.4 essentially proceeds in the same way as query 2.3 in that it identifies the gene nodes filtered for and looks for the miRNAs connected to those nodes by TARGETS-type relationships; however, in the second step (which is performed per gene node as returned by the WITH clause), it returns only those patterns that have at least 4 incoming miRNA→gene relationships. Query 2.4 only requires little additional processing compared to query 2.3, and thus does not require nearly as much time as the post-hoc filtering required in a relational database query. This filtering can be applied in many stages, and in many forms, such as sums, averages, maximum and minimum, or other combinations of arithmetic and logical classifiers. Additionally, the patterns can be extended to represent complex relationships inside the graph. For instance, the following query 2.5 was used to find miRNAs that regulate any given gene in the database, and, simultaneously, affect TFs that are involved in regulation of this same gene (this type of interaction is called feedforward loop, see also Section 4.9).

Listing 2.5: Feedforward Loop Identification

```
MATCH (gene:GENE) //find gene
WHERE gene.id = ID //by identifier (Entrez)
WITH gene //use as input for next step
MATCH (tf:GENE {species: 'HSA', tf:TRUE})-[rel]->(gene)
//find TFs targeting that gene
WHERE type(rel) IN {tissue_types} //TFs only from specific tissues
//for instance, CNS cell types (Appendix A)
WITH gene, rel, tf //use as input for next step
```

```
MATCH (gene)<-[rel_m1:TARGETS]-(mir:MIR {species:
    'HSA'})-[rel_m2:TARGETS]->(tf)
//find miRNAs that target both gene and TF
WHERE rel_m1.score > 5 AND rel_m2.score > 5
//filter by minimum cumulative score
RETURN gene, tf, rel, type(r) AS tissue, mir, rel_m1, rel_m2
```

This analysis can be performed in real time, on the whole genome and miRnome, and merely takes seconds for one iteration, a performance unimaginable in a relational database approach; advanced statistical approaches such as permutation only become viable at this timescale.

2.4 Statistical Approach to Transcriptional Connectomics

The enormous amounts of data generated by modern molecular biology methods, such as RNA-seq and bioinformatics, present new challenges to statistical methodology. A major objective in the analysis of large datasets is a robust statistical representation of the distribution of this data. Traditionally used approaches such as Student's t-test are not automatically applicable to the intermediary results of these modern methods, because the premise of a normal distribution often does not hold, or has to be proven first. This section will describe the statistical problems encountered in the analysis of intermediary data produced by *miRNet*; the statistical properties of large count data directly generated by RNA-seq will be discussed in Section 3.4.3.

2.4.1 Permutation

The evaluation of comprehensive prediction datasets regarding miRNA→gene interactions on a genome scale is statistically challenging. Molecular interaction studies have explored only a minority of all possible targeting relationships, and as such, the ground truth of miRNA→gene interaction is unknown (see Section 2.2.4). Since there is no negative interaction data, validated interactions can only be defined in the positive space. Additionally, the various prediction algorithms also heavily diverge in their predictions, which leads to the question of how to approach the estimation of false discovery ratio (FDR) while simultaneously avoiding high false negative rates.

One possible approach that can aid in identification of the most pertinent effects in this case is random permutation. In this approach, the result of an analysis (e.g., a numeric targeting score of a miRNA→gene interaction, or a Spearman correlation between two gene sets) is compared to a null distribution that was generated from an iterative analysis similar to the initial one, but with randomised input (e.g., a group of miRNAs of the same size as the original set, randomly selected from all miRNAs, or the gene sets from the original analysis with randomly scrambled group affiliations). This permutation of the analysis is performed for a large number of times (usually between 10 000

and 1 000 000 iterations, depending on the context), and results in a distribution of possible outcomes that can be arranged from lowest to highest, often resulting in a normal (or »normal-like«) distribution, thus facilitating the estimation of confidence intervals, and, similarly, p-values for the »real« result.

A positive side-effect of performing such a permutation on a base collection of data, such as *miRNet*, is the automatic correction of inherent biases. For instance, should a particular gene by its genetic structure invite a large amount of false positive predictions as to the miRNA→gene interactions towards it, these will be present in the test as well as in the permutation comparison, and thus cancel out and yield a high p-value for this interaction.

2.4.2 Gene Set Enrichment

The objective of gene set enrichment is the identification of statistically over-represented entities in a dataset. The standard use case in biomedicine is the Gene Set Enrichment Analysis (GSEA), that is used to identify the most important classes of genes in large datasets, such as the ones produced by RNA-seq. Briefly, the analysis follows these steps: the studied genes have to be scored by a certain method, such as p-values from differential expression analysis, which enables the identification of a relevant subgroup, the test set (e.g., the 100 genes with lowest p-values). This test set is then compared to a background of genes (usually, all detected genes, or a large amount of genes from the entire dataset) by a statistical method fit to determine their enrichment in pre-defined categories. Often, ontological categories are used, such as the »biological process« type of Gene Ontology (GO), or KEGG pathways.

For each of these categories, the method tests for a representation of genes in the test set exceeding the frequency statistically expected by random sampling from the background of genes; thus enabling an estimation of the functionality these test set genes might inhabit in the process that is studied. Statistical approaches often employed in gene set enrichment are Kolmogorov-Smirnov statistics, permutations, or, more generally, hypergeometric tests such as Fisher's exact test. There are a wide variety of software solutions available for the implementation of gene set enrichment testing.

2.4.3 Leave-One-Out Approaches

Build networks with progressively less gene targets for all possible combinations iteratively

2.5 Identification of Cholinergic Regulators

Describe iterative Leave-One-Out (LOO) approach that led to the discovery of the CHAT miRNA anomaly here? Combines database structure, score filtering, LOO iteration

where goes CHAT miRNA anomaly?

There is no scientific study more vital to man than the study of his own brain. Our entire view of the universe depends on it.

Francis Crick

3

microRNA Dynamics in Cholinergic Differentiation of Human Neuronal Cells

Even though much has been achieved in the integrative study of miRNA control of gene expression, computational analysis of transcriptional interactions has not yet reached the level of sophistication needed for the accurate prediction of events inside mammalian cells. For this reason, combination of a bioinformatical assay with modern molecular biology methods can strengthen the message and reproducibility of any approach. The spectrum of processes worthy of study is as wide as modern biomedicine. Similarly, experimental models can span the entire repertoire available to a modern laboratory. The selection of a model adequate to the research question therefore is as important as diligent analysis and careful interpretation of results.

This chapter will discuss the current state of knowledge on brain transciptomics, generally and in the specific case of cholinergic neurons in the CNS, and then go on to explain the steps I undertook to elucidate small RNA processes in central cholinergic systems. First, my aim was to clarify co-expression patterns of central cholinergic neurons, which required analysis of transcriptome data in single-cell resolution. Based on this information, I selected two human models of cholinergic neuronal differentiation and established a differentiation protocol amenable to RNA extraction and successive molecular biology assays, most importantly, RNA-seq. The expression patterns so obtained were then used to perform bioinformatics analyses using the database introduced in Chapter 2, *miRNet*.

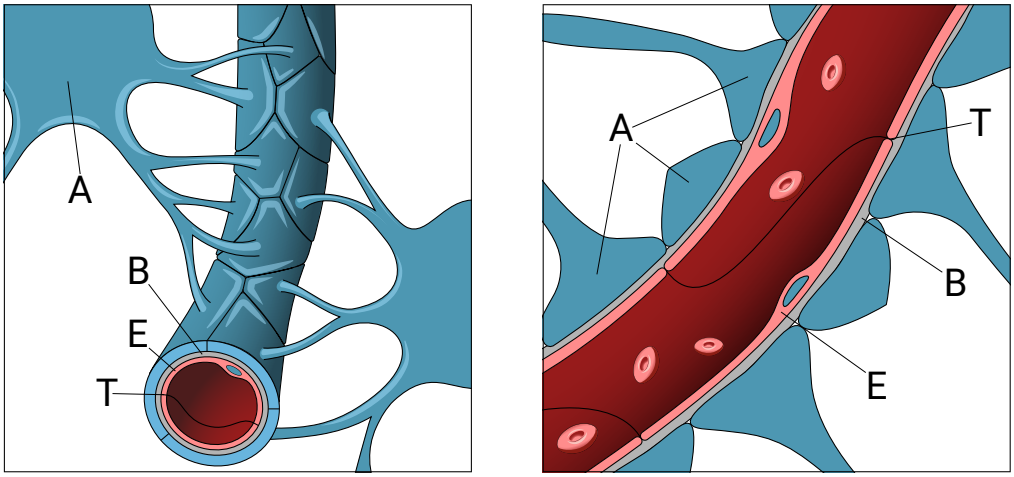

Figure 3.1: Schematic display of the blood-brain-barrier. The blood-brain-barrier surrounds virtually every capillary in the CNS. A: Astrocyte, B: Basal Membrane, E: Endothelial Cell, T: Tight Junction. Modified from Lobentanzer & Klein, 2019[51].

3.1 Neuronal Transcriptomes - Background

The mammalian brain requires a constant supply of oxygen and nutrients, because it does not provide storage for either. Though it only makes up approximately 2% of the entire human body mass, its energy expenditure is around 20% of the whole[49]. For this reason, each square millimetre of brain tissue (except for the ventricles) is infiltrated by hundreds of capillaries[50]. Since the blood-brain-barrier is essentially provided by supporting glia cells surrounding all capillaries from the »inside« (see Fig. 3.1, adapted from my second publication[51]), neurons constitute only a minority of brain tissues (but burn two thirds of its energy).

Before the advent of single-cell RNA-seq, studies that endeavoured to clarify the transcriptional profiles of neurons used microarray technology, which was recently succeeded by RNA-seq (also known as deep sequencing or next generation sequencing). For these methods, several cubic millimetres of brain tissue are required at the least; often, cubic centimetres are used. In contrast, the diameter of neuronal somata is usually in the micrometre range. Thus, the resolution of the method and the actual cellular resolution differ by a factor of approximately 1000. Additionally, even among the neuronal population, there is considerable heterogeneity and transcriptomic plurality; single brain regions rarely consist of less than 30 different neuron types, tightly packed next to each other, each with their own transcriptional identity[52,53,54,55]. Newest studies, deciphering the murine nervous system by sequencing of 500 000 individual cells, show that neuron diversity is very similar regardless of brain region[56]. These circumstances hold true for any mammal, and most of our knowledge stems from the analysis of our favourite research animal, the mouse. In humans, the complexity is only exacerbated; in fact, the elevation in CNS transcriptional complexity might be the reason for our superior cognitive abilities(cite).

Cholinergic neurons always constitute a minority in any neuronal population, sometimes to extremes. Most tissues are dominated by few neuron types, such as pyramidal cells in the cortex; the most common neurotransmitter types are GABAergic (inhibitory) and glutamatergic (excitatory), each with several subtypes. It is estimated that more than 80% of cortical neurons are excitatory, and more than 90% of synapses release glutamate[49]. There are two major cholinergic regions in the mammalian brain: The striatum is fairly well-populated with rather large cholinergic interneurons, and the basal forebrain holds a large amount of (smaller) cholinergic projection neurons (compare Fig. 1.1). However, in transcriptomic analyses, these tissues are seldom used, for lack of scientific interest, or because they are notoriously hard to access (the basal forebrain is small and deeply imbedded in the midbrain). The cortex, particularly the neocortex, is most often the tissue of choice in these studies, due to its scientific interest and accessibility. Though it contains only a minuscule amount of cholinergic interneurons whose identity still is a matter of debate, several of the recent single-cell RNA-seq approaches have independently identified cholinergic interneurons in cortical regions (see Fig. 3.2).

All of the above taken into consideration, several limitations apply when it comes to the selection of a cellular model for the cholinergic processes we aim to understand. A multicellular model is prohibited by the novelty of the subject; possibilities include *in vivo* or *ex vivo* experimentation on rodents or human (3D-)cell culture with multiple cell types. While the former certainly is closest to reality, our diseases of interest display a noticeable lack of transferability from lower mammals to human(cite). The latter, on the other hand, introduces a complexity not well suited to the level of knowledge we possess about the studied processes; in addition, these models are very new, and thus, too many variables would be unknown. For these reasons, I selected two mono-cultures of human neuronal cells for my experiments. I chose to introduce a second cellular model during the experimental phase, because some of the studied diseases display a clear sexual dimorphism; the first experiments were performed on the female-originated LA-N-2 cell line, and later, to explore sex-related differences, the male-originated LA-N-5 cells were added.

3.2 Cortical Single-Cell RNA Sequencing

To estimate the potential impact of the transcriptomes studied in our model on the diseases of interest, co-expression of the relevant genes has to be asserted. Similarly, if neurokines are to possess any relevance for cholinergic properties of central nervous cells, the cells in question would have to express molecular machinery required to receive neurokine signals. The advent of single-cell RNA-seq for the first time enables the resolution of gene expression on a cellular basis, and thus the disentangling of spatially close individual neuron types (and other, non-neuronal CNS cells); most of this information would be lost in RNA-seq performed on brain homogenate, even of a small biopsy. Differences in genes would be reduced to the universally expressed »housekeeping« genes, and in miRNAs, the

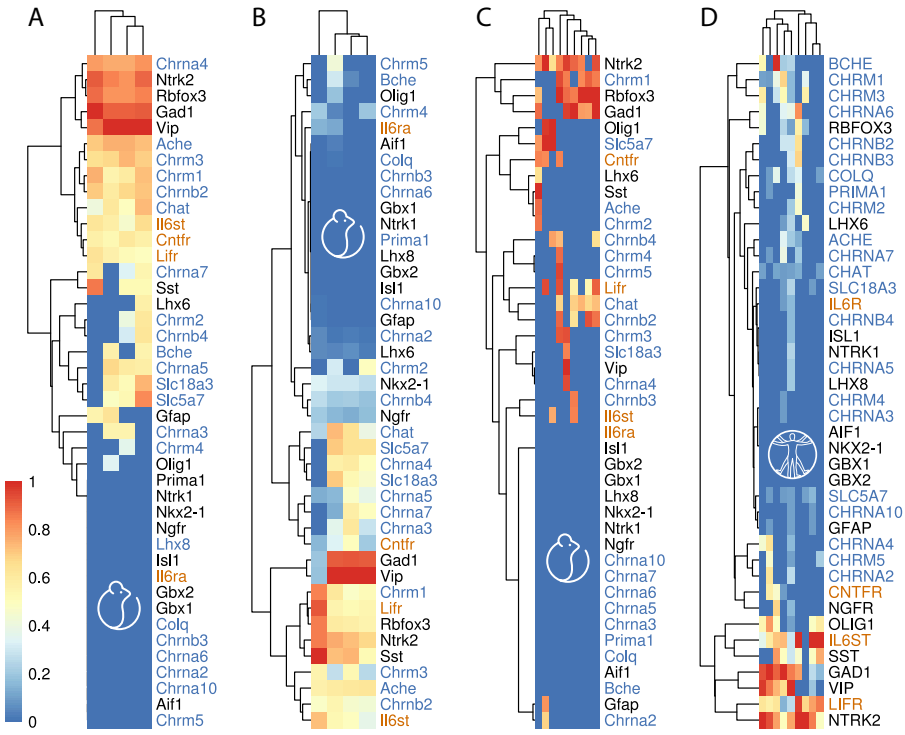

Figure 3.2: This is a figure that floats inline and here is its caption.

situation would be worse, in parallel to their even more tissue-specific expression.

For this reason, I analysed all publicly available single-cell RNA-seq datasets relevant to our questions, which provide a detailed tally of transcriptional subtypes in the CNS. More specifically, all studies that were available at the time focused on some subsection of the cortex (visual or somatosensory) or the hippocampus. Additionally, the data provided by those studies was in some cases pre-aggregated to represent groups of single neurons with similar transcriptomes (Fig. 3.2 A&B[53,54]); in other cases, every single neuron was represented (Fig. 3.2 C&D[52,55]).

An important quality-related parameter of a single-cell RNA-seq experiment is the sequencing depth achieved per individual sequenced cell. Some datasets I screened do not provide sufficient depth to resolve genes with medium expression, which includes our primary cholinergic markers *CHAT* and *SLC18A3*. The datasets which did provide adequate sequencing depth were filtered for their expression of these markers, and additionally characterised by their expression of common markers for cell types to be expected in the CNS. This resulted in a collection of transcriptomes for potentially cholinergic cells in the sampled brain regions, and allowed an assessment of the functional type and gene co-expression patterns in central cholinergic cells (Fig. 3.2).

Most cells identified as cholinergic by this definition expressed the general neuronal marker *RBFOX3*, also known by its trivial name NeuN, but not the microglial marker *AIF1*. Few cells (or clusters of cells) expressed non-neuronal markers such as *GFAP* (astrocytes) or *OLIG1* (oligodendrocytes), hinting at sparse non-neuronal cholinergic functions. In agreement with my findings, cells or clusters identified as cholinergic by the authors of the respective studies[53,54] (also by personal com-

munication) had been classified as interneurons and co-expressed a number of known phenotypic neuronal markers, such as *somatostatin (SST)* and *vasoactive intestinal peptide (VIP)*.

The identified cholinergic cells also revealed a constant co-expression with neurokine-related genes, particularly the transmembrane neurokine receptors LIFR and IL6ST, demonstrating a capacity to receive and process neurokine signals. In contrast, the high affinity receptor for NGF, *NTRK1*, is not co-expressed in mature (NeuN-positive) cholinergic neurons, fundamentally distinguishing these cells from the basal forebrain cholinergic projection neurons.

Clusters?

Permutation targeting analyses?

3.3 The Cellular Model

3.3.1 Requirements of a Cholinergic Neuron Cellular Model

Describe here the preliminary computational analyses leading to discovery of the CHAT miRNA anomaly, and thus the search for a CHAT-centric cell model?

where goes the preliminary analysis?

3.3.2 The SH-SY5Y Neuroblastoma Cell Line

A prominent example of human neuronal cell culture used in the identification and elucidation of cholinergic processes is the immortalised neuroblastoma cell line SH-SY5Y[57]. Derived from its parent line SK-N-SH, an adrenergic neuroblastoma[58], it expresses ample amounts of *ACHE*, and thus had become a work horse in many cholinergic fields, such as Alzheimer's Disease (which is treated with ACHE antagonists), pesticide development, and warfare(cite). However, in spite of its usefulness for processes involving *ACHE*, it turned out a less than optimal choice for the study of molecular events surrounding *CHAT* and *SLC18A3*, as it barely expresses both genes(cite), and cannot be coerced to elevate *CHAT* expression by the usual differentiation techniques (own experimentation, data not shown). Thus, for the questions asked in this chapter of the dissertation, SH-SY5Y does not qualify as adequate representation of a »cholinergic neuron«.

3.3.3 The LA-N Neuroblastoma Cell Lines

Following the elimination of SH-SY5Y as a suitable subject, I scoured the literature for candidates representing a cholinergic neuronal transcriptome, and found, among others, representatives of the LA-N neuroblastoma cell lines developed by R.C. Seeger around 1980[59,60]. Neuroblastoma is a form of neuronal cancer often affecting small children, and, consequentially, the two cell lines used in my experiments are immortalised biopsies of a 3 year old girl (LA-N-2[59]) and of a 4 month old boy (LA-N-5[60]). The decision to use LA-N-2 as my initial cellular model was influenced by three factors: it is well described in literature, although most studies had been published in the 1980s and 90s; it expresses a substantial amount of *CHAT* and *SLC18A3*; and it responds to neurokine-mediated differentiation by assuming a neuronal morphology accompanied by further elevation of *CHAT* and

SLC18A3 expression. LA-N-5 was not nearly as well described as LA-N-2, but later added to the experimental roster because of the complementary sex and hints towards cholinergic differentiation under retinoic acid[61].

3.3.4 Culture

LA-N cell culture is not as straightforward as many »go-to« human cell lines used in today's laboratories. Because LA-N-2 and LA-N-5 are very similar in this regard, they were treated similarly and I will describe the procedure for both. They have comparatively high duplication times, which can be lowered by using certain conditions that affect medium composition, nutrition, and CO_2 content. The cells were acquired at DSMZ (Braunschweig, Germany), which recommends keeping them in a 50:50 mixture of Dulbecco's modified eagle medium (DMEM) and Roswell Park Memorial Institute medium (RPMI1640), with 20% fetal calf serum (FCS) added. Sometimes, recommendations also suggest Leibovitz's L-15 medium, which is specifically designed for low CO_2 conditions, and others have suggested increased CO_2 levels inside the incubator. I found a combination of the DSMZ-recommended medium with 8% CO_2 atmosphere inside a 37°C incubator to accelerate growth to a degree that the cells could be split 1:3 to 1:4 in a weekly cycle. This protocol was used for all further experiments, which were performed between splits 2 to 8 after thawing of a batch from -80°C. All handling during maintenance and experimentation was performed under a laminar flow hood.

3.3.5 Differentiation

Neuronal differentiation of neuroblastoma cell lines has been performed in many instances, utilising a wide variety of differentiation agents such as the very general retinoic acid or 5-bromo-uracil, or very specific reagents, such as the neurokines IL-6 and CNTF(cite). LA-N cells have also been described to react to a selection of these substances; however, due to the elevated interest in neurokine mechanisms, I chose to go ahead with a reagent from that group. Somewhat arbitrarily, I used CNTF, because upon my inquiry, James McManaman revealed in personal communication that the »*CHAT* development factor« that he had discovered[8] was, in fact, CNTF, which had never been published. Additionally, of the neurokines used for differentiation purposes, CNTF is best described in literature and easily acquired in dried form from Merck (formerly SigmaAldrich, Darmstadt, Germany). CNTF was resuspended in pure water to a concentration of 25 $\mu g\,ml^{-1}$ and stored for experimentation in aliquots at -20°C.

LA-N cells are very sensitive to repeated temperature changes (or other handling-related disturbances), which resulted in increased amounts of apoptotic cells following repeated removal from the incubator after seeding or medium changes during the experiment. For this reason, I chose to only add the differentiation reagent once, 24h after initial seeding of the cells into the wells of a 12-well-plate, and avoid further disturbances until the time of lysis. For the maximum duration of my

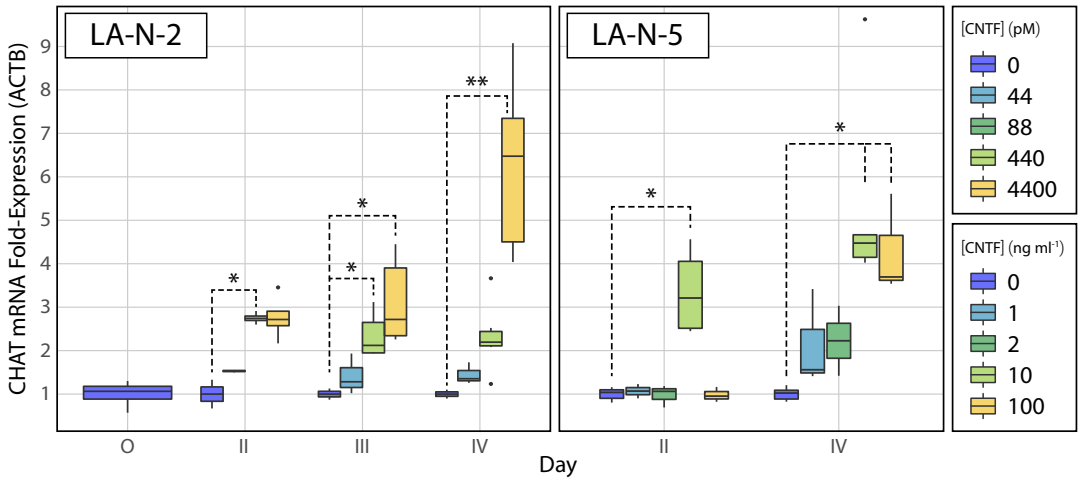

Figure 3.3: Time-dose curve of CNTF-mediated differentiation of LA-N-2 and LA-N-5.

experiments, 120h from seeding until lysis, the initially supplied medium was sufficient for survival.

Differentiation was performed in regular growth medium without changes in FCS content, and CNTF was added to the medium after an initial growth period of 24h. Cells were seeded into 12-well plates at approximately 200 000 cells/well, with 1 ml of growth medium. To determine the optimal amount of CNTF for differentiation, time-dose curves were determined for both cell lines in a range from 1 ng ml^{-1} to 100 ng ml^{-1}. Here, I discovered the first pharmacological difference between LA-N-2 and LA-N-5: the maximum of their cholinergic response to neurokine stimulation (i.e., an elevation in *CHAT* and *SLC18A3* transcription) occurs at different concentrations of CNTF. While LA-N-2 cells respond most strongly to 100 ng ml^{-1}, LA-N-5 cells show an »inverted u«-type dose response with a maximum around 10 ng ml^{-1} CNTF (Fig. 3.3). James McManaman, who studied LA-N differentiation thoroughly in the 1990s[62], believes both lines to respond in an »inverted u«-type manner (personal communication); thus, I assume that the LA-N-2 response would also diminish at CNTF concentrations significantly higher than 100 ng ml^{-1}. I also suspect that CNTF concentrations could be significantly lowered by removal of the high amount of FCS in the medium, however, that would likely require the use of a special serum-free medium, which would have to be established up front, and might have other, unforeseen consequences. Regardless, CNTF concentrations around 100 ng ml^{-1} (i.e., pico- to nano-molar) still are well within the physiological range of concentrations that the mammalian brain is able to reach by paracrine secretion via, e.g., astrocytes[63].

To study the small RNA dynamics following CNTF exposure of LA-N-2 and LA-N-5, I selected 4 time points at which the experiment was stopped and the cells were quickly lysed *in situ* to preserve total RNA in that state: for the quick, immediate-early-like phase, at 30 and 60 minutes after the addition of CNTF, and, for the long-term effects of differentiation, at 48 and 96 hours after the addition of CNTF (Fig. 3.4, from my first publication[5]). Each time point was controlled by a pseudo-treated (using pure water) culture from the same batch that had been seeded at the same time as the experimental group. In the final series used for the parallel sequencing of LA-N-2 and LA-N-5, all

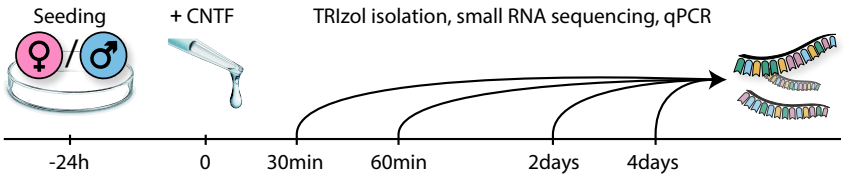

Figure 3.4: LA-N-2 / LA-N-5 Differentiation Timeline.

experiments were carried out in quadruplicates. add early/late

3.3.6 RNA Isolation

Total RNA was isolated using TRIzol (ThermoFisher Scientific), essentially as suggested by the manufacturer, with slight changes to the protocol to enrich small RNA species. The cells, growing in a monolayer in 12-well-plates, were cleared of medium, washed two times with 500 µl of cell culture grade phosphate buffered saline (PBS) (Gibco), and immediately suspended in 1 ml of TRIzol, pipetting up and down until visibly dissolved. After incubation for 5 minutes at room temperature, the samples were stored in -20°C for short periods of time until RNA isolation.

TRIzol-suspended lysates (1 ml) were added to RNA-separation centrifuge tubes (PhaseMaker Tubes, ThermoFisher Scientific), adding 200 µl of pure chloroform and mixing vigorously for 15 seconds. After two minutes, the mixture was centrifuged at 12 000 g and 4°C for 15 minutes, and the upper, watery phase containing the RNA was extracted. This was mixed with approximately 2 parts of pure ethanol and incubated for 10 minutes at room temperature to precipitate the RNA. The precipitate was spun at 12 000 g and 4°C for another 10 minutes, and the supernatant discarded. The pellet was washed with 85% ethanol (vortexed briefly) and centrifuged again for 5 minutes at 7500 g and 4°C.

After the final centrifugation step, the samples were transferred to the laminar flow hood, and air dried after removal of most of the supernatant via micropipettors. The pellet was allowed to dry almost until completion and resuspended in 30 µl to 50 µl pure RNase-free water. RNA concentration was measured at a Nanodrop 2000 instrument (ThermoFisher Scientific) and samples were diluted to a uniform concentration of 100 ng μl^{-1}. Finally, RNA samples were aliquoted according to later purpose and stored at -80°C.

RNA quality was determined by analysis on a 2100 Bioanalyzer instrument (Agilent) using a nano chip and 1 µl of sample; RNA integrity number (RIN) was near optimal for all samples (> 9).

3.4 Small RNA Sequencing and Differential Expression Analysis

For the detection and analysis of small RNA species, RNA-seq is the current gold standard method. It allows the mapping of a comprehensive transcriptome and thus is vastly superior to

small scale and consecutive methods such as real-time quantitative polymerase chain reaction (RT-qPCR), and even the larger scale microarrays. Microarrays, while also potentially allowing a "snapshot" of entire transcriptomes, are limited by the predetermined sequences on the chip. RNA-seq, on the other hand, is not biased towards any structural property of the sample; this is particularly important in the analysis of small RNA species, since their sequences are very variable (tRFs) and still not completely catalogued (miRNAs). Assuming an adequate sequencing depth ($\gtrsim$ 1 M reads/sample), RNA-seq allows a comparison of all expressed small RNA species at once, which is immensely helpful when dealing with processes on the combinatorial scale of miRNA regulation.

3.4.1 Sequencing

For small RNA sequencing, the aliquoted samples were shipped on dry ice to the cooperating institute at the Hebrew University of Jerusalem, the Silberman Institute of Molecular Biology, the laboratory of Prof. Hermona Soreq. 600 ng of total RNA per sample were prepared for sequencing using the NEBNext Small RNA Library Prep Set for Illumina (New England BioLabs). The libraries were multiplexed with coloured barcodes, allowing for sequencing of all 48 samples on one chip. Briefly, this includes ligation of sequencing adapters to both 3' and 5' ends of all (single-stranded) RNA fragments in the sample, followed by 12-15 cycles of reverse transcription to form the RNA library. Ligated and amplified libraries were then size selected via gel electrophoresis on a 6% Polyacrylamide gel. The band representing small RNA species on the gel was excised and prepared for loading onto the sequencing chip. After loading, the chip was sequenced in a NextSeq 550 series instrument (Illumina) with a read length of 80 bases, single-ended.

Sequencing quality was determined by analysis of the raw reads using the fastqc software(cite).

describe quality parameters

examples?

3.4.2 Sequence Alignment

Raw reads were adapter-trimmed and quality filtered using the flexbar software[64] with parameters

```
-a adapters.fa -q TAIL -qf sanger -qw 4
-min-read-length 16 -n 1 --zip-output GZ
```

The sequence used in the *adapters.fa* file, as recommended by the manufacturer, was

AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

For the alignment of miRNA sequences, parts of the miRExpress 2.0[65] pipeline were used according to the documentation. First, a lookup table for the current miRBase version 21 was created as per the instructions of the authors. The alignment was then performed using the commands *Raw_data_parse*, *statistics_reads*, *alignmentSIMD*, and *analysis*; *Trim_adaptor* was skipped because the adapters had already been trimmed in the quality filtering step. Additionally, since miRExpress

describe what miRExpress does?

is not accepting of sequences of any length, the raw data was length filtered to include only reads up to a length of 25 bases before input into miRExpress. Thus, raw reads were aligned to the miRnome provided by miRBase v21, yielding count tables of mature miRNAs and miRNA precursors for each sample. In total, 1913 mature miRNAs from miRBase v21 were discovered in the data.

3.4.3 Differential Expression Analysis - R/DESeq2

To determine the effect and dynamics of CNTF-mediated differentiation of LA-N-2 and LA-N-5, the expression state of each measured time point was compared to the respective control using the established R package DESeq2[66]. DESeq2 determines differential expression in count-based data by application of a linear regression model to a negative binomial distribution based on a fitted mean μ and a gene-specific dispersion value α (for gene i and sample j). The mean is derived using a sample-specific »size factor«, s, and a parameter q proportional to the expected true concentration of RNA fragments in the sample. The DESeq2 differential expression pipeline is composed of the following commands:

- `estimateSizeFactors()` (to estimate s_j)
- `estimateDispersion()` (to estimate α_i)
- `nbinomWaldTest()` (application of a generalised linear model to determine log-fold changes and statistics via the Wald test, using $\mu_{ij} = s_j q_{ij}$ and $log_2(q_{ij}) = x_j \beta_i$)

The Wald test, named after Abraham Wald, is an approach to hypothesis testing that measures the distance between the tested unrestricted estimate and the null hypothesis, using the precision as a weighting factor. The larger the distance between tested values and the null, the more likely the measured values are »true«. RNA-seq data can be modelled using binomial distributions[67], such as the Poisson distribution, and the difference between two Poisson means (e.g., »treated« vs »control«) can be tested by generalised linear models based on the distributions themselves, Fisher's exact test, or the likelihood ratio test. However, comparative analysis has shown that the Wald test on log-transformed data provides statistical power superior to these other methods[68], particularly in lowly expressed fragments. The design formula for the linear regression was a simple combination of condition and time point and applied to LA-N-2 and LA-N-5 separately:

$$y \sim condition\_time$$

To reduce the noise introduced by the high variance in low-count genes while preserving large, »real« differences, the authors propose the »shrinkage« of log-fold changes to avoid arbitrary low-cut filtering at a predefined expression (count) value. Multiple variants are available; for miRNA data, the adaptive algorithm »apeglm«[69] (adaptive t prior shrinkage estimator) yielded sensible results (see Fig. 3.5).

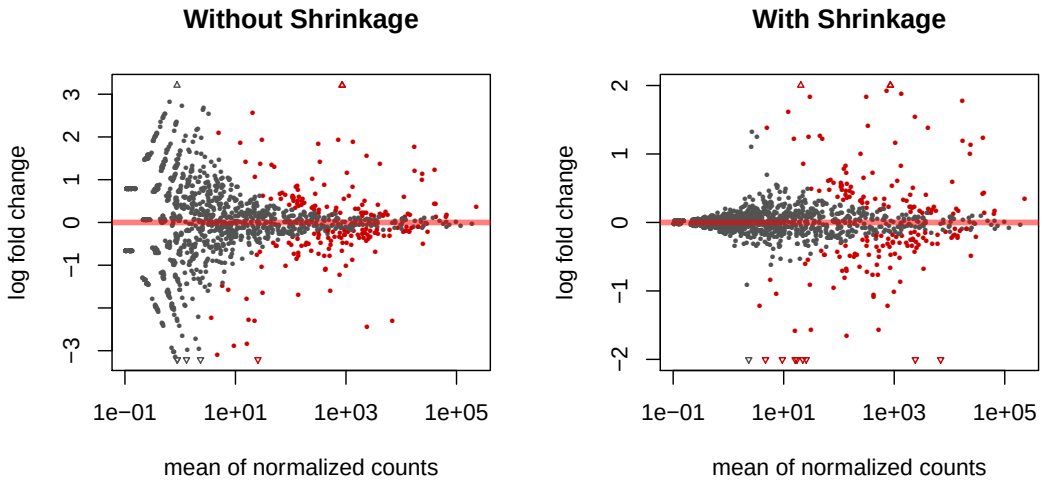

Figure 3.5: MD Plot Shrinkage Comparison. A mean-difference plot (MD Plot) is a plot of log-intensity ratios (differences, »M-values«) versus log-intensity averages (means, »A-values«); it is synonymous with »MA Plot«. The DESeq2 function plotMD shows the fold changes attributable to a given variable over the mean of normalised counts for all the samples in the data set. Points will be coloured red if the adjusted p value is less than 0.1. Points which fall out of the window are plotted as open triangles pointing either up or down. The left plot is generated from the standard linear model, the plot on the right is corrected by the »apeglm« algorithm [69] to reduce noise in the low-count fragments.

3.4.4 microRNA Dynamics in CNTF-mediated Cholinergic Differentiation of LA-N-2 and LA-N-5

Differential expression analysis performed in this manner yielded 490 differentially expressed (DE) miRNAs across all groups, with characteristic distributions between cell lines and time points. The raw data and processed counts were deposited to NCBI Gene Expression Omnibus (GEO), accession GSE132951. An earlier sequencing experiment (deposited as GSE120520), which was similar in principle, but only comprised three biological replicates and only LA-N-2, reproduced 80% of DE miRNAs in the newer experiment. Considering the general reproducibility of RNA-seq and the lower replicate number, 80% is an excellent substantiation of the result.

Differential Expression in Both Cell Lines

114 mature miRNAs were detected as DE in both cell lines, with some changes similar in both, while others were inverted (Fig. 3.6). In both cases, however, count-change values (see Box 1) correlated highly between the two cell lines (similar: 76 miRNAs, Spearman's $\rho = 0.9066$, p < 2.2E-16; inverted: 38 miRNAs, $\rho = 0.9294$, p < 2.2E-16).

Box 1: The count-change metric.

The frequently used log-fold change metric is not ideally suited for assessing the potential effect of expression changes for individual miRNAs because it does not reflect mean expression levels. To

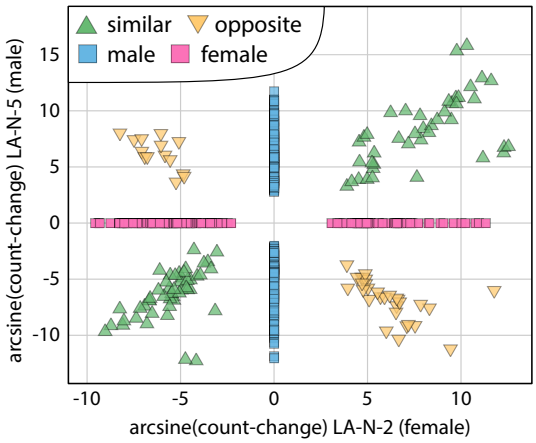

Figure 3.6: LA-N-2 / LA-N-5 Count-change Correlation.

determine the absolute change in expression, I introduced the count-change metric, a combination of base mean expression and log-fold change, to weigh DE miRNAs against one another. The count-change is defined as follows:

$$CC = (BM \cdot 2^{LFC}) - BM$$

CC: count-change, BM: base mean expression, LFC: log-2-fold-change.

Differential Expression Along The Timeline

For consistency, from hereon out, time points 30 minutes and 60 minutes will be termed »early«, while 2 days and 4 days will be referred to as »late«. Differential expression was detected in all groups, lending credibility to the rapid changes in expression needed for a miRNA response of the »immediate-early« type. However, the response to long-term CNTF stimulation was larger in miRNA numbers as well as effect sizes (Fig. 3.7 A&B). Of all early perturbed miRNAs, only 3 and 13 miRNAs were exclusively perturbed immediate-early-like in LA-N-2 and LA-N-5, respectively; all others were still DE after 2 and/or 4 days. In LA-N-2, the late time points at 2 and, particularly, 4 days showed the greatest perturbation; in LA-N-5, the picture was more complex (Fig. 3.7 C&D). However, generally, there were large similarities as well as exclusivities between the time points 2 and 4 days in both cell lines. When comparing early and late time points between LA-N-2 and LA-N-5 directly, similarly complex patterns emerged (Fig. 3.7 E&F). Particularly at late time points (Fig. 3.7 F), every possible combination of overlap exists. 24 miRNAs were DE in all late conditions; 107 miRNAs were DE only in LA-N-2, and 269 miRNAs were DE only in LA-N-5.

something special?

compare most important targets early/late

Differential Expression Between LA-N-2 and LA-N-5

While there was considerable intersection in DE miRNAs between the cell lines, a substantial amount of miRNAs was only DE in one of the two lines. Generally, response to CNTF was higher in the male-

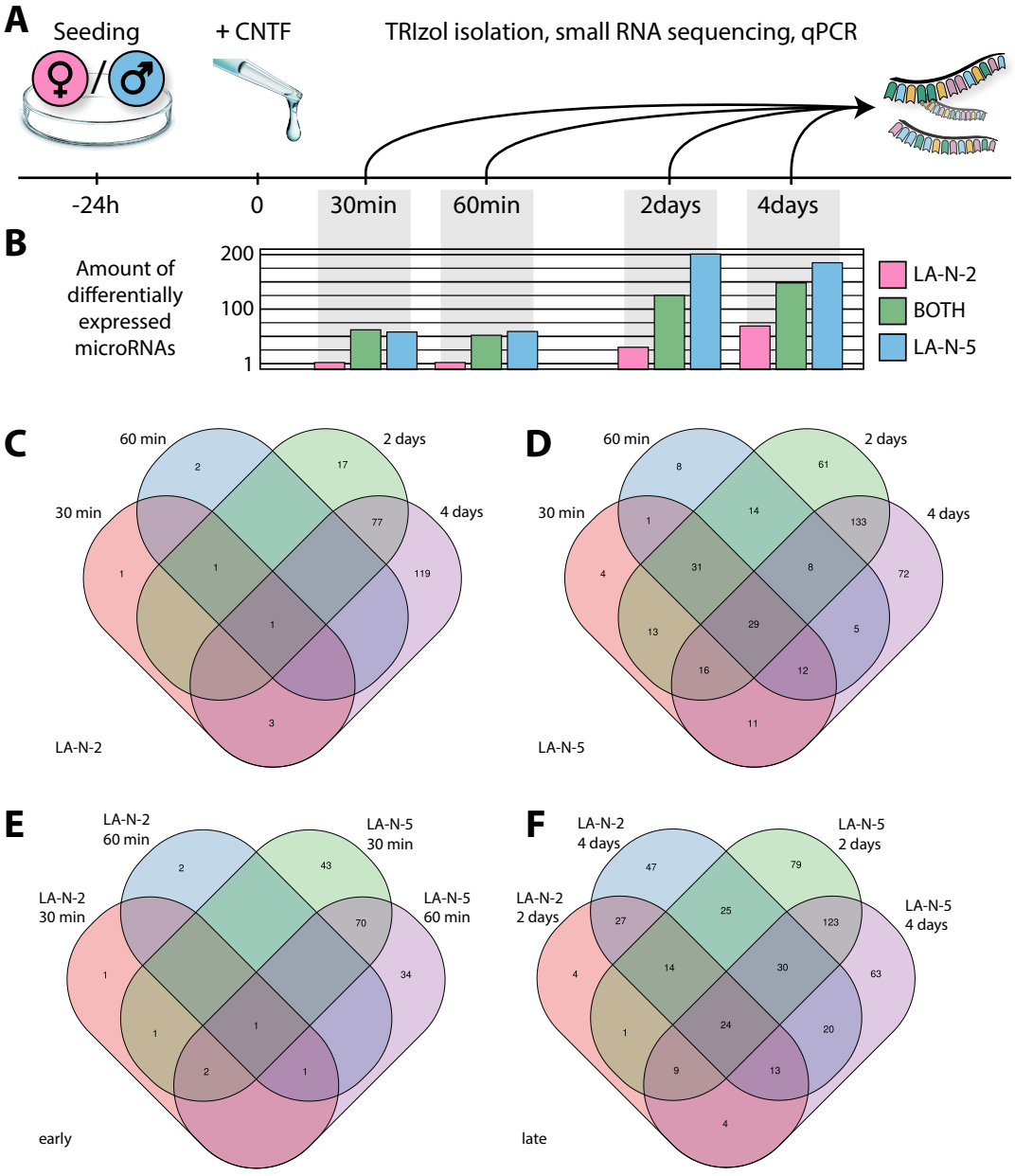

Figure 3.7: LA-N-2 / LA-N-5 Timeline and Differential Expression.

Figure 3.8: DE miRNAs between LA-N-2 and LA-N-5.

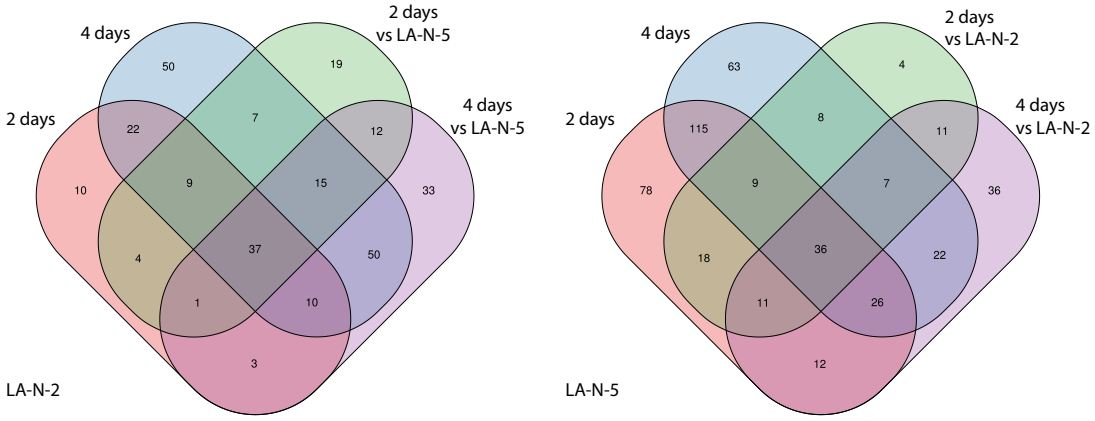

Figure 3.9: LA-N-2 / LA-N-5 DE miRNAs late time-points vs in-between.

originated LA-N-5 cells; however, there were also miRNAs found DE only in the female LA-N-2 (compare Fig. 3.7). Thus, not all of the differences in miRNA expression can be attributed to a higher sensitivity in LA-N-5. Similarly, LA-N-5 shows a »non-significant trend« toward higher count-change values (mean of absolute count-change across all DE time points, 20,907 versus 3,066, Welch two-sample t test, p = 0.08).

To determine the influence of genotype on the differentiating effect of CNTF, I employed a statistical interaction design in the DESeq2 Wald test. Briefly, by including an interaction term in the linear regression formula, we can statistically isolate the effect of the condition (CNTF or control at each time point) between the two genotypes:

$$y \sim condition + genotype + condition : genotype$$

Using the interaction term *condition* : *genotype*, I filtered out the miRNAs that reacted significantly different to CNTF stimulation in LA-N-5 compared to LA-N-2 at each time point.

combine

intersect with detected miRs in CNTF vs control experiments, 2 venns

To categorise and systematise the sexual dimorphism of CNTF differentiation of LA-N cells, I performed gene set enrichment of miRNA families in the differential expression datasets.

3.4.5 microRNA Family Enrichment

Of the 151 miRNA families listed in miRBase v21, members of 71 families are DE in LA-N-2 and LA-N-5. To test for the enrichment of male, female, and ubiquitously DE miRNAs in these families, I performed gene set enrichment based on Fisher's exact test for each of the families. I found 5 families to be enriched in both male and female cells, and 12 families enriched in only one of the two cell lines (Fig. 3.10, left side). The size range of enriched families was substantial, from small families with only 4 mature members to extensive families with dozens of mature miRNAs.

Gene Targeting of Enriched Families

Using *miRNet*, the targets of all individual miRNAs were determined. Of note, the amount of family members in any miRNA family did not correlate with the absolute amount of targets predicted. Rather, the influence of individual miRNAs was the main factor determining the size of the gene target network. However, those families that were enriched in only one cell line presented with significantly smaller target sets than those that were found DE in both (mean targeted genes per miRNA 217 versus 378, Welch two-sample t test, $p = 0.001$). Relative to family size, 4 of the enriched families targeted less genes than all others (Fig. 3.10, right side): mir-10 ($p = 0.016$), mir-192 ($p = 0.042$), mir-379 ($p = 0.011$), and mir-515 ($p < 0.001$). Hypothetically, the spectrum of target amounts might correlate with the degree of functional specification of distinct miRNA families: on one end, broadly acting families such as let-7 with sex-independent function, on the other, families with a narrow target profile, such as mir-10, whose restricted function can associate with sex-specific effects.

3.5 Network Generation

3.6 The Cholinergic/Neurokine Interface

3.7 Application to Schizophrenia and Bipolar Disorder

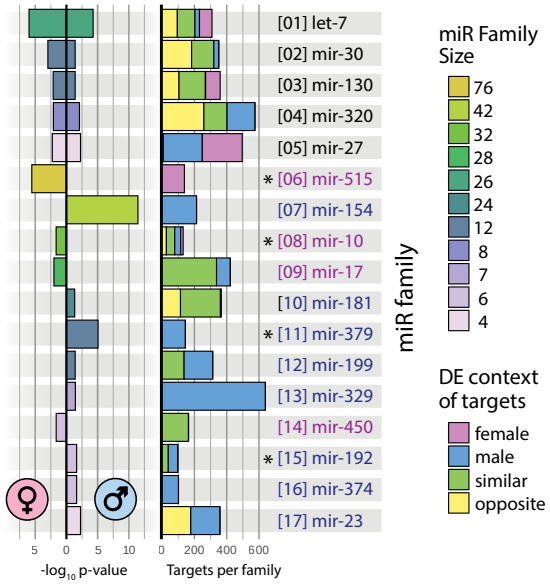

Figure 3.10: miRNA Families Enriched in Differential Expression.

I realized, "Oh my gosh! I'm having a stroke!" And the next thing my brain says to me is, Wow! This is so cool! How many brain scientists have the opportunity to study their own brain from the inside out?"

Jill Bolte Taylor

4

Dynamics Between Small and Large RNA in the Blood of Stroke Victims

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

If the human brain were so simple that we could understand it, we would be so simple that we couldn't.

Emerson M. Pugh

5 Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

5.1 Methods

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.

5.2 Small RNA Therapeutics and Pharmacology

Extant approaches, methods, diseases, PCSK9, asthma, using small RNA antisense as substitute for single-target small molecules, reduce off-target effects, side effects of a different kind

Transcriptomics as basis for selection and design of antisense therapy, combinatorial, compare dirty drugs from psychiatric disorders, serendipity impossible, determinant is the sequence as opposed to functional groups that can be iteratively modified (only 4 building blocks)

6

Conclusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum tortor quam, feugiat vitae, ultricies eget, tempor sit amet, ante. Donec eu libero sit amet quam egestas semper. Aenean ultricies mi vitae est. Mauris placerat eleifend leo. Quisque sit amet est et sapien ullamcorper pharetra. Vestibulum erat wisi, condimentum sed, commodo vitae, ornare sit amet, wisi. Aenean fermentum, elit eget tincidunt condimentum, eros ipsum rutrum orci, sagittis tempus lacus enim ac dui. Donec non enim in turpis pulvinar facilisis. Ut felis.

Cras sed ante. Phasellus in massa. Curabitur dolor eros, gravida et, hendrerit ac, cursus non, massa. Aliquam lorem. In hac habitasse platea dictumst. Cras eu mauris. Quisque lacus. Donec ipsum. Nullam vitae sem at nunc pharetra ultricies. Vivamus elit eros, ullamcorper a, adipiscing sit amet, porttitor ut, nibh. Maecenas adipiscing mollis massa. Nunc ut dui eget nulla venenatis aliquet. Sed luctus posuere justo. Cras vehicula varius turpis. Vivamus eros metus, tristique sit amet, molestie dignissim, malesuada et, urna.

Cras dictum. Maecenas ut turpis. In vitae erat ac orci dignissim eleifend. Nunc quis justo. Sed vel ipsum in purus tincidunt pharetra. Sed pulvinar, felis id consectetuer malesuada, enim nisl mattis elit, a facilisis tortor nibh quis leo. Sed augue lacus, pretium vitae, molestie eget, rhoncus quis, elit. Donec in augue. Fusce orci wisi, ornare id, mollis vel, lacinia vel, massa.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.

References

[1] Dale H H. THE ACTION OF CERTAIN ESTERS AND ETHERS OF CHOLINE, AND THEIR RELATION TO MUSCARINE. *Journal of Pharmacology and Experimental Therapeutics*, 6(2) (1914).

[2] Loewi O. Über humorale Übertragbarkeit der Herznervenwirkung. *Pflügers Arch. Ges. Physiol.*, 189:239–242 (1921).

[3] Dale H H & Dudley H W. THE PRESENCE OF HISTAMINE AND ACETYLCHOLINE IN THE SPLEEN OF THE OX AND THE HORSE. *J. Physiol.*, 68:97 (1929).
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1402860/pdf/jphysiol01676-0019.pdf

[4] Mesulam M M, Mufson E J, Levey A I, & Wainer B H. Atlas of cholinergic neurons in the forebrain and upper brainstem of the macaque based on monoclonal choline acetyltransferase immunohistochemistry and acetylcholinesterase histochemistry. *Neuroscience*, 12(3):669–686 (1984). doi:10.1016/0306-4522(84)90163-5.

[5] Lobentanzer S, Hanin G, Klein J, & Soreq H. Integrative Transcriptomics Reveals Sexually Dimorphic Control of the Cholinergic/Neurokine Interface in Schizophrenia and Bipolar Disorder. *CellReports*, pp. 1–19 (2019). doi:10.1016/j.celrep.2019.09.017.
URL https://doi.org/10.1016/j.celrep.2019.09.017

[6] Levi-Montalcini R & Booker B. Destruction of the sympathetic ganglia in mammals by an antiserum to a nerve-growth protein. *Proceedings of the National Academy of Sciences*, 46(3):384–391 (1960). doi:10.1073/pnas.46.3.384.
URL http://www.ncbi.nlm.nih.gov/pubmed/16578497http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC222845http://www.pnas.org/cgi/doi/10.1073/pnas.46.3.384

[7] Hefti F. Nerve growth factor promotes survival of septal cholinergic neurons after fimbrial transections. *Journal of Neuroscience*, 6(8):2155–2162 (1986).

[8] McManaman J L, Crawford F G, Stewart S S, & Appel S H. Purification of a Skeletal Muscle Polypeptide Which Stimulates Choline Acetyltransferase Activity in Cultured Spinal Cord Neurons. *Journal of Biological Chemistry*, 263(12):5890–5897 (1988).

[9] Rao M S, Patterson P H, & Landis S C. Multiple cholinergic differentiation factors are present in footpad extracts: comparison with known cholinergic factors. *Development (Cambridge, England)*, 116(3):731–44 (1992).
URL http://www.ncbi.nlm.nih.gov/pubmed/1289063

[10] Berger M. *Psychische Erkrankungen* (2014).

[11] Rawlings J S. The JAK/STAT signaling pathway. *Journal of Cell Science*, 117(8):1281–1283 (2004). doi:10.1242/jcs.00963.
URL http://jcs.biologists.org/cgi/doi/10.1242/jcs.00963

[12] Nathanson N M. Regulation of neurokine receptor signaling and trafficking. *Neurochemistry International*, 61(6):874–878 (2012). doi:10.1016/j.neuint.2012.01.018.
URL https://linkinghub.elsevier.com/retrieve/pii/S0197018612000307

[13] Babu M M, Luscombe N M, Aravind L, Gerstein M, & Teichmann S A. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283–291 (2004). doi:10.1016/j.sbi.2004.05.004.
URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X04000788

[14] Lee R C, Feinbaum R L, & Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854 (1993). doi:10.1016/0092-8674(93)90529-Y.
URL https://linkinghub.elsevier.com/retrieve/pii/009286749390529Y

[15] Rodriguez A. Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Research*, 14(10a):1902–1910 (2004). doi:10.1101/gr.2722704.
URL http://www.genome.org/cgi/doi/10.1101/gr.2722704

[16] Kozomara A, Birgaoanu M, & Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162 (2019). doi:10.1093/nar/gky1141.
URL https://academic.oup.com/nar/article/47/D1/D155/5179337

[17] Ambros V, Bartel B, Bartel D P *et al.* A uniform system for microRNA annotation. *RNA (New York, N.Y.)*, 9(3):277–9 (2003). doi:10.1261/rna.2183803.
URL http://www.ncbi.nlm.nih.gov/pubmed/12592000http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1370393

[18] Salta E & De Strooper B. microRNA-132: a key noncoding RNA operating in the cellular phase of Alzheimer's disease. *The FASEB Journal*, 31(2):424–433 (2017). doi:10.1096/fj.201601308.
URL http://www.fasebj.org/doi/10.1096/fj.201601308

[19] Lu L F, Gasteiger G, Yu I S *et al.* A Single miRNA-mRNA Interaction Affects the Immune Response in a Context- and Cell-Type-Specific Manner. *Immunity*, 43(1):52–64 (2015). doi:10.1016/j.immuni.2015.04.022.
URL http://www.ncbi.nlm.nih.gov/pubmed/26163372http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4529747https://linkinghub.elsevier.com/retrieve/pii/S1074761315002551

[20] Borek E, Baliga B S, Gehrke C W, Kuo C W, Belman S, Troll W, & Waalkes T P. High Turnover Rate of Transfer RNA in Tumor Tissue. *CANCER RESEARCH*, 37:3362–3366 (1977).
URL https://cancerres.aacrjournals.org/content/37/9/3362.full-text.pdf

[21] Speer J, Gehrke C W, Kuo K C, Waalkes T P, & Borek E. tRNA breakdown products as markers for cancer. *Cancer*, 44(6):2120–2123 (1979). doi:10.1002/1097-0142(197912)44:6<2120::AID-CNCR2820440623>3.0.CO;2-6.
URL http://www.ncbi.nlm.nih.gov/pubmed/509391http://doi.wiley.com/10.1002/1097-0142{%}28197912{%}2944{%}3A6{%}3C2120{%}3A{%}3AAID-CNCR2820440623{%}3E3.0.CO{%}3B2-6

[22] Cole C, Sobala A, Lu C, Thatcher S R, Bowman A, Brown J W, Green P J, Barton G J, & Hutvagner G. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA*, 15(12):2147–2160 (2009). doi:10.1261/rna.1738409.
URL http://www.ncbi.nlm.nih.gov/pubmed/19850906http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2779667http://rnajournal.cshlp.org/cgi/doi/10.1261/rna.1738409

[23] Lee Y S, Shibata Y, Malhotra A, & Dutta A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes & development*, 23(22):2639–49 (2009). doi:10.1101/gad.1837609.
URL http://www.ncbi.nlm.nih.gov/pubmed/19933153http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2779758

[24] Godoy P M, Bhakta N R, Barczak A J *et al.* Large Differences in Small RNA Composition Between Human Biofluids. *Cell Reports*, 25(5):1346–1358 (2018). doi:10.1016/j.celrep.2018.10.014.
URL https://doi.org/10.1016/j.celrep.2018.10.014https://linkinghub.elsevier.com/retrieve/pii/S2211124718315778

[25] Yamasaki S, Ivanov P, Hu G f, & Anderson P. Angiogenin cleaves tRNA and promotes stress-induced translational repression. *The Journal of Cell Biology*, 185(1):35–42 (2009). doi:10.

1083/JCB.200811106.
URL http://jcb.rupress.org/content/185/1/35.long

[26] Ivanov P, Emara M M, Villen J, Gygi S P, & Anderson P. Angiogenin-Induced tRNA Fragments Inhibit Translation Initiation. *Molecular Cell*, 43(4):613–623 (2011). doi:10.1016/j.molcel.2011.06.022.
URL https://www.sciencedirect.com/science/article/pii/S1097276511005247?via{%}3Dihubhttps://linkinghub.elsevier.com/retrieve/pii/S1097276511005247

[27] Burroughs A M, Ando Y, de Hoon M L, Tomaru Y, Suzuki H, Hayashizaki Y, & Daub C O. Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. *RNA Biology*, 8(1):158–177 (2011). doi:10.4161/rna.8.1.14300.
URL http://www.tandfonline.com/doi/abs/10.4161/rna.8.1.14300

[28] Kumar P, Anaya J, Mudunuri S B, & Dutta A. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biology*, 12(1):78 (2014). doi:10.1186/s12915-014-0078-0.
URL http://www.ncbi.nlm.nih.gov/pubmed/25270025http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4203973http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-014-0078-0

[29] Huang B, Yang H, Cheng X *et al.* tRF/miR-1280 Suppresses Stem Cell–like Cells and Metastasis in Colorectal Cancer. *Cancer Research*, 77(12):3194–3206 (2017). doi:10.1158/0008-5472.CAN-16-3146.
URL http://cancerres.aacrjournals.org/http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-16-3146

[30] Gebetsberger J, Wyss L, Mleczko A M, Reuther J, & Polacek N. A tRNA-derived fragment competes with mRNA for ribosome binding and regulates translation during stress. *RNA Biology*, 14(10):1364–1373 (2017). doi:10.1080/15476286.2016.1257470.
URL https://www.tandfonline.com/action/journalInformation?journalCode=krnb20https://www.tandfonline.com/doi/full/10.1080/15476286.2016.1257470

[31] Goodarzi H, Liu X, Nguyen H C, Zhang S, Fish L, & Tavazoie S F. Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. *Cell*, 161(4):790–802 (2015). doi:10.1016/j.cell.2015.02.053.
URL https://www.sciencedirect.com/science/article/pii/

S0092867415003189?via{%}3Dihubhttps://linkinghub.elsevier.com/retrieve/pii/S0092867415003189

[32] Kim H K, Fuchs G, Wang S *et al.* A transfer-RNA-derived small RNA regulates ribosome biogenesis. *Nature*, 552(7683):57 (2017). doi:10.1038/nature25005.
URL http://www.nature.com/doifinder/10.1038/nature25005

[33] Parisien M, Wang X, & Pan T. Diversity of human tRNA genes from the 1000-genomes project. *RNA Biology*, 10(12):1853–1867 (2013). doi:10.4161/rna.27361.
URL http://www.ncbi.nlm.nih.gov/pubmed/24448271http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3917988http://www.tandfonline.com/doi/abs/10.4161/rna.27361

[34] Loher P, Telonis A G, & Rigoutsos I. MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Scientific Reports*, 7(1):41184 (2017). doi:10.1038/srep41184.
URL http://dx.doi.org/10.1038/srep41184http://www.nature.com/articles/srep41184

[35] Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, & Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, 13(4):366–370 (2016). doi:10.1038/nmeth.3799.
URL http://www.ncbi.nlm.nih.gov/pubmed/26950747http://www.nature.com/articles/nmeth.3799http://regulatorycircuits.org

[36] Nowakowski T J, Rani N, Golkaram M *et al.* Regulation of cell-type-specific transcriptomes by microRNA networks during human brain development. *Nature Neuroscience*, 21(12):1784–1792 (2018). doi:10.1038/s41593-018-0265-3.
URL http://www.ncbi.nlm.nih.gov/pubmed/30455455http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6312854http://www.nature.com/articles/s41593-018-0265-3

[37] Londin E, Loher P, Telonis A G *et al.* Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proceedings of the National Academy of Sciences*, 112(10):E1106–E1115 (2015). doi:10.1073/pnas.1420955112.
URL http://www.pnas.org/lookup/doi/10.1073/pnas.1420955112

[38] Dweep H & Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nature Methods*, 12(8):697–697 (2015). doi:10.1038/nmeth.3485.
URL http://www.nature.com/doifinder/10.1038/nmeth.3485

[39] Chaudhuri S, Narasayya V, & Ramamurthy R. Estimating Progress of Execution for SQL Queries (2004).
URL https://www.microsoft.com/en-us/research/publication/estimating-progress-of-execution-for-sql-queries/?from=https{%}3A{%}2F{%}2Fresearch.microsoft.com{%}2Fapps{%}2Fpubs{%}2F{%}3Fid{%}3D76556

[40] Hon C c, Ramilowski J A, Harshbarger J *et al.* An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature Publishing Group*, 543(7644):199–204 (2017). doi:10.1038/nature21374.
URL http://dx.doi.org/10.1038/nature21374

[41] Dweep H & Gretz N. miRWalk2 web page.

[42] Karagkouni D, Paraskevopoulou M D, Chatzopoulos S *et al.* DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Research*, 46(D1):D239–D245 (2018). doi:10.1093/nar/gkx1141.
URL http://academic.oup.com/nar/article/46/D1/D239/4634010

[43] Chou C H, Shrestha S, Yang C D *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1):D296–D302 (2018). doi:10.1093/nar/gkx1067.
URL http://www.ncbi.nlm.nih.gov/pubmed/29126174http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5753222http://academic.oup.com/nar/article/46/D1/D296/4595852

[44] Yue D, Liu H, & Huang Y. Survey of Computational Algorithms for MicroRNA Target Prediction. *Current Genomics*, 10(7):478–492 (2009). doi:10.2174/138920209789208219.
URL http://www.ncbi.nlm.nih.gov/pubmed/20436875http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2808675http://www.eurekaselect.com/openurl/content.php?genre=article{&}issn=1389-2029{&}volume=10{&}issue=7{&}spage=478

[45] Witkos T M, Koscianska E, & Krzyzosiak W J. Practical Aspects of microRNA Target Prediction. *Current molecular medicine*, 11(2):93–109 (2011). doi:10.2174/156652411794859250.

[46] Friedman R C, Farh K K H, Burge C B, & Bartel D P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105 (2009). doi:10.1101/gr.082701.108.
URL http://www.ncbi.nlm.nih.gov/pubmed/18955434http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2612969http://genome.cshlp.org/cgi/doi/10.1101/gr.082701.108

[47] Alexiou P, Maragkakis M, Papadopoulos G L, Reczko M, & Hatzigeorgiou A G. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25(23):3049–3055 (2009). doi:10.1093/bioinformatics/btp565.
URL http://www.ncbi.nlm.nih.gov/pubmed/19789267https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp565

[48] Agarwal V, Bell G W, Nam J W, & Bartel D P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4 (2015). doi:10.7554/eLife.05005.
URL https://elifesciences.org/articles/05005

[49] Raichle M E & Gusnard D A. Appraising the brain's energy budget. *Proceedings of the National Academy of Sciences*, 99(16):10237–10239 (2002). doi:10.1073/pnas.172399499.
URL http://www.pnas.org/cgi/doi/10.1073/pnas.172399499

[50] Bohn K A, Adkins C E, Mittapalli R K, Terrell-Hall T B, Mohammad A S, Shah N, Dolan E L, Nounou M I, & Lockman P R. Semi-automated rapid quantification of brain vessel density utilizing fluorescent microscopy. *Journal of Neuroscience Methods*, 270:124–131 (2016). doi:10.1016/j.jneumeth.2016.06.012.
URL http://www.ncbi.nlm.nih.gov/pubmed/27321229http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4981522https://linkinghub.elsevier.com/retrieve/pii/S0165027016301339

[51] Lobentanzer S & Klein J. Zentrales und Peripheres Nervensystem. In Wichmann & Fromme, (Editors) *Handbuch für Umweltmedizin*, chapter 11. Erg. lfg. edition (2019).

[52] Darmanis S, Sloan S A, Zhang Y, Enge M, Caneda C, Shuer L M, Hayden Gephart M G, Barres B A, & Quake S R. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):201507125 (2015). doi:10.1073/pnas.1507125112.
URL http://www.pnas.org/content/112/23/7285.abstract

[53] Zeisel a, Manchado a B M, Codeluppi S *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–42 (2015). doi:10.1126/science.aaa1934.
URL http://science.sciencemag.org.docelec.univ-lyon1.fr/content/347/6226/1138.abstract

[54] Tasic B, Menon V, Nguyen T N T *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, advance on(January):1–37 (2016). doi:10.1038/nn.4216.
URL http://dx.doi.org/10.1038/nn.4216

[55] Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta J J, Hession C, Zhang F, & Regev A. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*, 353(6302):925–928 (2016). doi:10.1126/science.aad7038.
URL http://www.sciencemag.org/lookup/doi/10.1126/science.aad7038

[56] Zeisel A, Hochgerner H, Lönnerberg P *et al.* Molecular Architecture of the Mouse Nervous System. *Cell*, 174(4):999–1014.e22 (2018). doi:10.1016/j.cell.2018.06.021.
URL http://www.ncbi.nlm.nih.gov/pubmed/30096314http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6086934https://linkinghub.elsevier.com/retrieve/pii/S009286741830789X

[57] Biedler J L, Roffler-Tarlov S, Schachner M, & Freedman L S. Multiple Neurotransmitter Synthesis by Human Neuroblastoma Cell Lines and Clones. *Cancer Res.*, 38(11_Part_1):3751–3757 (1978).
URL http://cancerres.aacrjournals.org/content/38/11{_}Part{_}1/3751.short

[58] Biedler J L, Helson L, & Spengler B A. Morphology and Growth, Tumorigenicity, and Cytogenetics of Human Neuroblastoma Cells in Continuous Culture. *Cancer Research*, 33(11):2643–2652 (1973).

[59] Seeger R C, Rayner S A, Laug W E, Neustein H B, & Benedict W F. Morphology, Growth, Chromosomal Pattern, and Fibrinolytic Activity of two New Human Neuroblastoma Cell Lines. *Cancer Research*, 37(5):1364–1371. (1977).

[60] Seeger R C, Danon Y L, Rayner S A, & Hoover F. Definition of a Thy-1 determinant on human neuroblastoma, glioma, sarcoma, and teratoma cells with a monoclonal antibody. *Journal of immunology (Baltimore, Md. : 1950)*, 128(2):983–9 (1982).
URL http://www.ncbi.nlm.nih.gov/pubmed/6172518

[61] Hill D P & Robertson K a. Characterization of the cholinergic neuronal differentiation of the human neuroblastoma cell line LA-N-5 after treatment with retinoic acid. *Developmental Brain Research*, 102(1):53–67 (1997). doi:10.1016/S0165-3806(97)00076-X.
URL http://www.ncbi.nlm.nih.gov/pubmed/9298234https://linkinghub.elsevier.com/retrieve/pii/S016538069700076X

[62] McManaman J L & Crawford F G. Skeletal Muscle Proteins Stimulate Cholinergic Differentiation of Human Neuroblastoma Cells. *Journal of neurochemistry*, pp. 258–266 (1991).

[63] Sun M, Liu H, Min S, Wang H, & Wang X. Ciliary neurotrophic factor-treated astrocyte-conditioned medium increases the intracellular free calcium concentration in rat cortical

neurons. *Biomedical Reports*, 4(4):417–420 (2016). doi:10.3892/br.2016.602.
URL https://www.spandidos-publications.com/https://www.spandidos-publications.com/10.3892/br.2016.602

[64] Roehr J T, Dieterich C, & Reinert K. Flexbar 3.0 – SIMD and multicore parallelization. *Bioinformatics*, 33(18):2941–2942 (2017). doi:10.1093/bioinformatics/btx330.
URL https://academic.oup.com/bioinformatics/article/33/18/2941/3852078

[65] Wang W C, Lin F M, Chang W C, Lin K Y, Huang H D, & Lin N S. miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics*, 10:328 (2009). doi:10.1186/1471-2105-10-328.
URL https://www.ncbi.nlm.nih.gov/pubmed/19821977

[66] Love M I, Huber W, & Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550 (2014). doi:10.1186/s13059-014-0550-8.
URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8

[67] Bullard J H, Purdom E, Hansen K D, & Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94 (2010). doi:10.1186/1471-2105-11-94.
URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-94

[68] Chen Z, Liu J, Ng H, Nadarajah S, Kaufman H L, Yang J Y, & Deng Y. Statistical methods on detecting differentially expressed genes for RNA-seq data. *BMC Systems Biology*, 5(Suppl 3):S1 (2011). doi:10.1186/1752-0509-5-S3-S1.
URL http://www.ncbi.nlm.nih.gov/pubmed/22784615http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3287564http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-5-S3-S1

[69] Zhu A, Ibrahim J G, & Love M I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, 35(12):2084–2092 (2019). doi:10.1093/bioinformatics/bty895.
URL https://academic.oup.com/bioinformatics/article/35/12/2084/5159452

A

Transcription Factor Regulatory Circuits - Tissue Types