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

Thesis advisor: Professor Jochen Klein Sebastian Lobentanzer

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

ACHE acetylcholinesterase

Ago argonaute

API application programming interface

CAGE 5' cap analysis of gene expression

CHAT choline acetyltransferase

CHRNA7 nicotinic acetylcholine receptor subunit $\alpha 7$

CNS central nervous system

CNTF ciliary neurotrophic factor

CNTFR ciliary neurotrophic factor receptor (soluble)

gp130 see IL6ST

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)

miRNA microRNA

NGF nerve growth factor

RISC RNA-induced silencing complex

SLC18A3 vesicular acetylcholine transporter (official gene symbol)

SQL structured query language

STAT signal transducer and activator of transcription

TF transcription factor

tiRNA transfer RNA half

tRF transfer RNA fragment

tRNA transfer RNA

TYK tyrosine kinase

UTR untranslated region

vAChT vesicular acetylcholine transporter

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 of 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]).

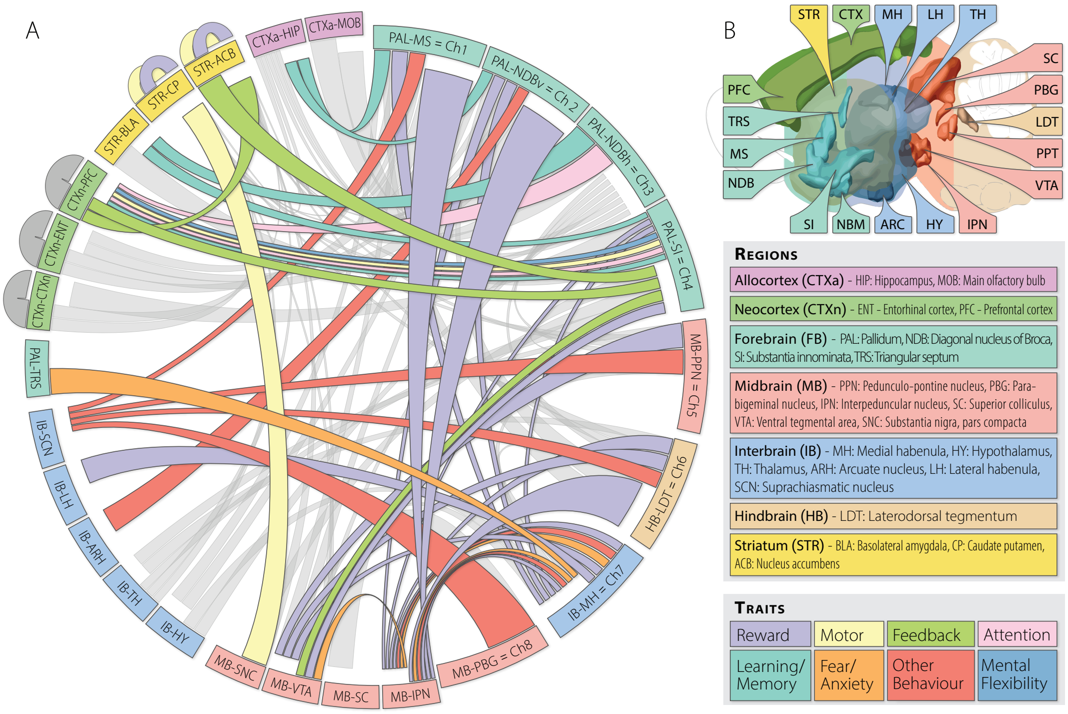
A
CTXa-HIP
CTXa-MOB
PAL-MS = Ch1
PAL-NDBv = Ch 2
PAL-NDBh = Ch3
PAL-SI = Ch4
MB-PPN = Ch5
HB-LDT = Ch6
IB-MH = Ch7
MB-PBG = Ch8
MB-IPN
MB-SC
MB-VTA
MB-SNC
IB-HY
IB-TH
IB-ARH
IB-LH
IB-SCN
PAL-TRS
CTXn-CTXn
CTXn-ENT
CTXn-PFC
STR-BLA
STR-CP
STR-ACB
B
STR
CTX
MH
LH
TH
SC
PBG
LDT
PPT
VTA
PFC
TRS
MS
NDB
SI
NBM
ARC
HY
IPN
Regions
Allocortex (CTXa) - HIP: Hippocampus, MOB: Main olfactory bulb
Neocortex (CTXn) - ENT - Entorhinal cortex, PFC - Prefrontal cortex
Forebrain (FB) - PAL: Pallidum, NDB: Diagonal nucleus of Broca, SI: Substantia innominata, TRS: Triangular septum
Midbrain (MB) - PPN: Pedunculo-pontine nucleus, PBG: Parabigeminal nucleus, IPN: Interpeduncular nucleus, SC: Superior colliculus, VTA: Ventral tegmental area, SNC: Substantia nigra, pars compacta
Interbrain (IB) - MH: Medial habenula, HY: Hypothalamus, TH: Thalamus, ARH: Arcuate nucleus, LH: Lateral habenula, SCN: Suprachiasmatic nucleus
Hindbrain (HB) - LDT: Laterodorsal tegmentum
Striatum (STR) - BLA: Basolateral amygdala, CP: Caudate putamen, ACB: Nucleus accumbens
Traits
Reward
Motor
Feedback
Attention
Learning/ Memory
Fear/ Anxiety
Other Behaviour
Mental Flexibility

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

Since cholinergic systems are integral for a myriad physiological functions, they are naturally 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

Schizophrenia and Bipolar Disorder

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. 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[10,11].

All receptor constellations result in a main effect of activation of the JAK/STAT cascade. 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 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.

find a place

FIGURE STAT

1.2 Transcriptional Connectomics

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 (2 420 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 perspective 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 subject 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 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 small non-coding RNA that has only recently been rediscovered and is significantly less well described regarding its functionality.

Where to put:

where goes this?

Distinguish neuronal connectomics from transcriptional connectomics

For the sake of simplicity, all descriptions of genomics and transcriptomics matters, genes, miRNAs, and tRFs in this dissertation are to be seen in the context of *Homo sapiens*, unless explicitly stated otherwise.

1.2.1 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 - 1 000 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, computational approaches have become able to not only comprehensively predict TF-gene interactions, but to do so in a highly tissue-specific manner (see 2.2.3). The human body is estimated to express up to 2 600 distinct DNA-binding proteins, most of them presumed TFs[12], although other studies give lower estimates.

1.2.2 microRNAs

The first endogenous »small RNA with antisense complementarity« was described in 1993[13], but 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. 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. 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 implications 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 2.2.4).

cite

cite

FIGURE?

miRNAs, similar to coding genes, are transcribed from loci on the genome, many inside introns or even exons of coding genes[14]. 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).

miRNAs are organised and curated by means of a periodically updated web-based platform, miRBase[15]. For *Homo sapiens*, miRBase v21 contains 2 588 mature miRNAs from XXX 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[16] 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.

Disease?

miRNA genes, in the same way as protein coding genes, can also 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 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). 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[17], and proposed as urine-based biomarkers for certain malignancies[18]. 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[19,20] 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[21]. 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[22,23]. 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[24,25].

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[25], and of Dicer playing a large part in their biogenesis[19]. There are even instances of small RNA molecules previously mislabeled miRNAs that have been discovered to actually be tRNA-derived, such as miR-1280[26].

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[27]. 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[28]. 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[29].

There is no consistent nomenclature yet to describe and organise tRFs, which are by nature more heterogeneous than miRNAs; considering their biogenesis, one tRNA molecule can be the origin of several hundred distinct tRF molecules. Multiple approaches are common in current literature, most prominently tRFs are tied to the parent tRNA and the amino acid coded for by this tRNA. For example, 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[29]. 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[30] 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).

1.2.4 Nested Multimodal Transcriptional Interactions - The Need for Connectomics

..., multiple levels of obstacles have to be overcome. The ultimate aim of any such approach is the generation of a robust model for the studied phenomenon; the theoretical and practical hurdles to be surmounted to reach this goal are many. 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 it has 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 1 000 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), and that any one TF in nervous and immune cells controls expression of a mean of 175 and 160 genes, respectively[31] (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 sequencing, 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[32]. 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 for 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[33], 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 1 000), this makes for an estimated minimum of 200 000 »real« interactions in the possible $C = \frac{1\,000!}{10!(1\,000-10)!} \cdot 20\,000$, 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 (2 588 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 $2\,588 \cdot 20\,000 = 51\,760\,000$ individual fields.

Combining all aspects 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 2 000 are TFs), 2 588 mature miRNAs, and a total of $2\,588 \cdot 20\,000 + 2\,000 \cdot 20\,000 = 91\,760\,000$ 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 male and female cultured cells, or 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 bioinformatical 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[34] (see 1.2.2). One year later, Marbach's »regulatory circuits« were published[31], 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[35]; *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 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. 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.

Fig DB structure

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. For example: only 30 of the 2 588 miRNAs target a specific gene, which is common; a relational database, after finding the index of the queried gene, would have to search 2 588 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). 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

Java and R description? Where?

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.

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 know 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 and fantasy-names (e.g. CHAT), 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 and instances of genes without names having to carry unwieldy systematic names.

- The American 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.

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[36]. 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[31]. 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); and 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 $\alpha7$ 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 $\alpha7$ 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[31], and neuronal and immune tissues were entered into *miRNet*. The collected data comprises XX neuronal tissues and XX immune cell tissues (Appendix A), and XX TF-gene relationships in total.

2.2.4 microRNA Interactions

The content of miRWalk 2.0 is freely available online[37]; however, there is no option of downloading the complete set. The targeting data thus was downloaded per miRNA 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[38] and miRTarBase[39], 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:

mention custom crawler?

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[40]). 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 complemen-

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.

tary approaches instead of only one algorithm per analysis[41]. 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. Therefore, 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 range from 1 to 10. 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). The resulting optimised graph contains XX miRNA→gene targeting relationships with a distinct score distribution (Figure XX).

FIGURE: Histogram of score distributions?

TFs not the only CHAT anomaly

2.2.5 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 fragments[19,25], justifying an attempt to predict interactions in a comprehensive manner. Of the available options for nucleotide interaction prediction algorithms, TargetScan[42] seems particularly suited

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).

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[43], and through an additional algorithm allows the attribution of a score based on the branch length (on the species tree) of conserved targeting[44].

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 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 $p = \frac{800}{4^7} = 0.049$.

Describe Targetscan process

2.3 Usage

2.3.1 Cypher Query Language

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

```
MATCH (gene:GENE {species: 'HSA'})
WHERE gene.name = 'CHAT'
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 done in real time on the whole genome and miRnome and only takes seconds for one iteration, a performance unimaginable in a relational database approach.

2.4 Statistical Approach to Transcriptional Connectomics

Permutation

This is some random quote to start off the chapter.

Firstname lastname

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, a 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.

Multicellular model prohibited by tissue specificity and novelty, animal model prohibited by lack of transferability, diseases have to be introduced up top

In-vivo situation unclear

Paucity of accurate data in mammals

Advent of single-cell sequencing

Sex differences prompted by disease distribution

3.1 Cortical Single-Cell Sequencing

3.2 The Cellular Model

SH-SY5Y

LA-N-2[8] LA-N-5[45]

3.2.1 Culture

3.2.2 Differentiation

3.3 Small RNA Sequencing and Differential Expression Analysis

3.3.1 microRNA Family Enrichment

3.4 Network Generation

3.5 The Cholinergic/Neurokine Interface

3.6 Application to Schizophrenia and Bipolar Disorder

Nulla facilisi. In vel sem. Morbi id urna in diam dignissim feugiat. Proin molestie tortor eu velit. Aliquam erat volutpat. Nullam ultrices, diam tempus vulputate egestas, eros pede varius leo.

Quoteauthor Lastname

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.

4.1 Background

4.2 Cohort

4.3 RNA Sequencing and Differential Expression Analysis

4.4 tRF Homology

4.5 WGCNA

4.6 Co-correlation

4.7 Networks

4.8 Direct Interaction

4.9 Feedforward Loops

This is some random quote to start off the chapter.

Firstname lastname

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.

Proin at eros non eros adipiscing mollis. Donec semper turpis sed diam. Sed consequat ligula nec tortor. Integer eget sem. Ut vitae enim eu est vehicula gravida. Morbi ipsum ipsum, porta nec, tempor id, auctor vitae, purus. Pellentesque neque. Nulla luctus erat vitae libero. Integer nec enim. Phasellus aliquam enim et tortor. Quisque aliquet, quam elementum condimentum feugiat, tellus odio consectetuer wisi, vel nonummy sem neque in elit. Curabitur eleifend wisi iaculis ipsum. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. In non velit non ligula laoreet ultrices. Praesent ultricies facilisis nisl. Vivamus luctus elit sit amet mi. Phasellus pellentesque, erat eget elementum volutpat, dolor nisl porta neque, vitae sodales ipsum nibh in ligula. Maecenas mattis pulvinar diam. Curabitur sed leo.

Nulla facilisi. In vel sem. Morbi id urna in diam dignissim feugiat. Proin molestie tortor eu velit. Aliquam erat volutpat. Nullam ultrices, diam tempus vulputate egestas, eros pede varius leo, sed imperdiet lectus est ornare odio. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Proin consectetuer velit in dui. Phasellus wisi purus, interdum vitae, rutrum accumsan, viverra in, velit. Sed enim risus, congue non, tristique in, commodo eu, metus. Aenean tortor mi, imperdiet id, gravida eu, posuere eu, felis. Mauris sollicitudin, turpis in hendrerit sodales, lectus ipsum pellentesque ligula, sit amet scelerisque urna nibh ut arcu. Aliquam in lacus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Nulla placerat aliquam wisi. Mauris viverra odio. Quisque fermentum pulvinar odio. Proin posuere est vitae ligula. Etiam euismod. Cras a eros.

Nunc auctor bibendum eros. Maecenas porta accumsan mauris. Etiam enim enim, elementum sed, bibendum quis, rhoncus non, metus. Fusce neque dolor, adipiscing sed, consectetuer et, lacinia sit amet, quam. Suspendisse wisi quam, consectetuer in, blandit sed, suscipit eu, eros. Etiam ligula enim, tempor ut, blandit nec, mollis eu, lectus. Nam cursus. Vivamus iaculis. Aenean risus purus, pharetra in, blandit quis, gravida a, turpis. Donec nisl. Aenean eget mi. Fusce mattis est id diam. Phasellus faucibus interdum sapien. Duis quis nunc. Sed enim.

Pellentesque vel dui sed orci faucibus iaculis. Suspendisse dictum magna id purus tincidunt rutrum. Nulla congue. Vivamus sit amet lorem posuere dui vulputate ornare. Phasellus mattis sollicitudin ligula. Duis dignissim felis et urna. Integer adipiscing congue metus. Nam pede. Etiam non wisi. Sed accumsan dolor ac augue. Pellentesque eget lectus. Aliquam nec dolor nec tellus ornare venenatis. Nullam blandit placerat sem. Curabitur quis ipsum. Mauris nisl tellus, aliquet eu, suscipit eu, ullamcorper quis, magna. Mauris elementum, pede at sodales vestibulum, nulla tortor congue massa, quis pellentesque odio dui id est. Cras faucibus augue.

Suspendisse vestibulum dignissim quam. Integer vel augue. Phasellus nulla purus, interdum ac, venenatis non, varius rutrum, leo. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Duis a eros. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Fusce magna mi, porttitor quis, convallis eget, sodales ac, urna. Phasellus luctus venenatis magna. Vivamus eget lacus. Nunc tincidunt convallis tortor. Duis eros mi, dictum vel, fringilla sit amet, fermentum id, sem. Phasellus nunc enim, faucibus ut, laoreet in, consequat id, metus. Vivamus dignissim. Cras lobortis tempor velit. Phasellus nec diam ac nisl lacinia tristique. Nullam nec metus id mi dictum dignissim. Nullam quis wisi non sem lobortis condimentum. Phasellus pulvinar, nulla non aliquam eleifend, tortor wisi scelerisque felis, in sollicitudin arcu ante lacinia leo.

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.

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

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

[3] H H Dale and H W Dudley. 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] M. M. Mesulam, E. J. Mufson, A. I. Levey, and B. H. Wainer. 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. ISSN 03064522. doi: 10.1016/0306-4522(84)90163-5.

[5] Sebastian Lobentanzer, Geula Hanin, Jochen Klein, and Hermona Soreq. Integrative Transcriptomics Reveals Sexually Dimorphic Control of the Cholinergic/Neurokine Interface in Schizophrenia and Bipolar Disorder. *CellReports*, pages 1–19, 2019. ISSN 2211-1247. doi: 10.1016/j.celrep.2019.09.017. URL https://doi.org/10.1016/j.celrep.2019.09.017.

[6] R Levi-Montalcini and B Booker. 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, mar 1960. ISSN 0027-8424. 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] F. Hefti. Nerve growth factor promotes survival of septal cholinergic neurons after fimbrial transections. *Journal of Neuroscience*, 6(8):2155–2162, 1986. ISSN 02706474.

[8] James L. McManaman, Frances G. Crawford, S. Scott Stewart, and Stanley H. Appel. 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] M S Rao, P H Patterson, and S C Landis. Multiple cholinergic differentiation factors are present in footpad extracts: comparison with known cholinergic factors. *Development (Cambridge, England)*, 116(3):731–44, nov 1992. ISSN 0950-1991. URL http://www.ncbi.nlm.nih.gov/pubmed/1289063.

[10] J. S. Rawlings. The JAK/STAT signaling pathway. *Journal of Cell Science*, 117(8):1281–1283, 2004. ISSN 0021-9533. doi: 10.1242/jcs.00963. URL http://jcs.biologists.org/cgi/doi/10.1242/jcs.00963.

[11] Neil M. Nathanson. Regulation of neurokine receptor signaling and trafficking. *Neurochemistry International*, 61(6):874–878, nov 2012. ISSN 01970186. doi: 10.1016/j.neuint.2012.01.018. URL https://linkinghub.elsevier.com/retrieve/pii/S0197018612000307.

[12] M Madan Babu, Nicholas M Luscombe, L Aravind, Mark Gerstein, and Sarah A Teichmann. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283–291, jun 2004. ISSN 0959440X. doi: 10.1016/j.sbi.2004.05.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X04000788.

[13] Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, dec 1993. ISSN 00928674. doi: 10.1016/0092-8674(93)90529-Y. URL https://linkinghub.elsevier.com/retrieve/pii/009286749390529Y.

[14] A. Rodriguez. Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Research*, 14(10a):1902–1910, sep 2004. ISSN 1088-9051. doi: 10.1101/gr.2722704. URL http://www.genome.org/cgi/doi/10.1101/gr.2722704.

[15] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162, jan 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1141. URL https://academic.oup.com/nar/article/47/D1/D155/5179337.

[16] Victor Ambros, Bonnie Bartel, David P Bartel, Christopher B Burge, James C Carrington, Xuemei Chen, Gideon Dreyfuss, Sean R Eddy, Sam Griffiths-Jones, Mhairi Marshall, Marjori Matzke, Gary Ruvkun, and Thomas Tuschl. A uniform system for microRNA annotation. *RNA (New York, N.Y.)*, 9(3):277–9, mar 2003. ISSN 1355-8382. doi: 10.1261/rna.2183803. URL http://www.ncbi.nlm.nih.gov/pubmed/12592000http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1370393.

[17] Ernest Borek, B S Baliga, Charles W Gehrke, C W Kuo, Sidney Belman, Walter Troll, and T Phillip Waalkes. 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.

[18] John Speer, Charles W Gehrke, Kenneth C Kuo, T Phillip Waalkes, and Ernest Borek. tRNA breakdown products as markers for cancer. *Cancer*, 44(6): 2120–2123, dec 1979. ISSN 0008-543X. 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.

[19] Christian Cole, Andrew Sobala, Cheng Lu, Shawn R Thatcher, Andrew Bowman, J. W.S. Brown, Pamela J Green, Geoffrey J Barton, and Gyorgy Hutvagner. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA*, 15(12):2147–2160, dec 2009. ISSN 1355-8382. 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.

[20] Yong Sun Lee, Yoshiyuki Shibata, Ankit Malhotra, and Anindya Dutta. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes & development*, 23(22):2639–49, nov 2009. ISSN 1549-5477. doi: 10.1101/gad.1837609. URL http://www.ncbi.nlm.nih.gov/pubmed/19933153http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2779758.

[21] Paula M Godoy, Nirav R Bhakta, Andrea J Barczak, Hakan Cakmak, Susan Fisher, Tippi C. MacKenzie, Tushar Patel, Richard W Price, James F Smith, Prescott G Woodruff, and David J Erle. Large Differences in Small RNA Composition Between Human Biofluids. *Cell Reports*, 25(5):1346–1358, oct 2018. ISSN 22111247. 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.

[22] Satoshi Yamasaki, Pavel Ivanov, Guo-fu Hu, and Paul Anderson. Angiogenin cleaves tRNA and promotes stress-induced translational repression. *The Journal of Cell Biology*, 185(1):35–42, apr 2009. ISSN 0021-9525. doi: 10.1083/JCB.200811106. URL http://jcb.rupress.org/content/185/1/35.long.

[23] Pavel Ivanov, Mohamed M. Emara, Judit Villen, Steven P. Gygi, and Paul Anderson. Angiogenin-Induced tRNA Fragments Inhibit Translation Initiation. *Molecular Cell*, 43(4):613–623, aug 2011. ISSN 10972765. 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.

[24] Alexander Maxwell Burroughs, Yoshinari Ando, Michiel Laurens de Hoon, Yasuhiro Tomaru, Harukazu Suzuki, Yoshihide Hayashizaki, and Carsten Olivier Daub. 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, jan 2011. ISSN 1547-6286. doi: 10.4161/rna.8.1.14300. URL http://www.tandfonline.com/doi/abs/10.4161/rna.8.1.14300.

[25] Pankaj Kumar, Jordan Anaya, Suresh B Mudunuri, and Anindya Dutta. 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, dec 2014. ISSN 1741-7007. 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.

[26] Bingqing Huang, Huipeng Yang, Xixi Cheng, Dan Wang, Shuyu Fu, Wencui Shen, Qi Zhang, Lijuan Zhang, Zhenyi Xue, Yan Li, Yurong Da, Qing Yang, Zesong Li, Li Liu, Liang Qiao, Ying Kong, Zhi Yao, Peng Zhao, Min Li, and Rongxin Zhang. tRF/miR-1280 Suppresses Stem Cell–like Cells and Metastasis in Colorectal Cancer. *Cancer Research*, 77(12):3194–3206, jun 2017. ISSN 0008-5472. 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.

[27] Jennifer Gebetsberger, Leander Wyss, Anna M Mleczko, Julia Reuther, and Norbert Polacek. A tRNA-derived fragment competes with mRNA for ribosome binding and regulates translation during stress. *RNA Biology*, 14(10):1364–1373, oct 2017. ISSN 1547-6286. 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.

[28] Hani Goodarzi, Xuhang Liu, Hoang C.B. Nguyen, Steven Zhang, Lisa Fish, and Sohail F. Tavazoie. Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. *Cell*, 161(4):790–802, may 2015. ISSN 00928674. 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.

[29] Hak Kyun Kim, Gabriele Fuchs, Shengchun Wang, Wei Wei, Yue Zhang, Hyesuk Park, Biswajoy Roy-Chaudhuri, Pan Li, Jianpeng Xu, Kirk Chu, Feijie Zhang, Mei-Sze Chua, Samuel So, Qiangfeng Cliff Zhang, Peter Sarnow, and Mark A. Kay. A transfer-RNA-derived small RNA regulates ribosome biogenesis. *Nature*, 552(7683):57, nov 2017. ISSN 0028-0836. doi: 10.1038/nature25005. URL http://www.nature.com/doifinder/10.1038/nature25005.

[30] Phillipe Loher, Aristeidis G Telonis, and Isidore Rigoutsos. MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Scientific Reports*, 7(1):41184, mar 2017. ISSN 2045-2322. doi: 10.1038/srep41184. URL http://dx.doi.org/10.1038/srep41184http://www.nature.com/articles/srep41184.

[31] Daniel Marbach, David Lamparter, Gerald Quon, Manolis Kellis, Zoltán Kutalik, and Sven Bergmann. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, 13(4):366–370, apr 2016. ISSN 1548-7091. doi: 10.1038/nmeth.3799. URL http://www.ncbi.nlm.nih.gov/pubmed/26950747http://www.nature.com/articles/nmeth.3799http://regulatorycircuits.org.

[32] Tomasz J Nowakowski, Neha Rani, Mahdi Golkaram, Hongjun R Zhou, Beatriz Alvarado, Kylie Huch, Jay A West, Anne Leyrat, Alex A Pollen, Arnold R Kriegstein, Linda R Petzold, and Kenneth S Kosik. Regulation of cell-type-specific transcriptomes by microRNA networks during human brain development. *Nature Neuroscience*, 21(12):1784–1792, dec 2018. ISSN 1097-6256. 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.

[33] Eric Londin, Phillipe Loher, Aristeidis G Telonis, Kevin Quann, Peter Clark, Yi Jing, Eleftheria Hatzimichael, Yohei Kirino, Shozo Honda, Michelle Lally, Bharat Ramratnam, Clay E. S. Comstock, Karen E. Knudsen, Leonard Gomella, George L. Spaeth, Lisa Hark, L. Jay Katz, Agnieszka Witkiewicz, Abdolmohamad Rostami, Sergio A. Jimenez, Michael A. Hollingsworth, Jen Jen Yeh, Chad A. Shaw, Steven E. McKenzie, Paul Bray, Peter T Nelson, Simona Zupo, Katrien Van Roosbroeck, Michael J Keating, George A Calin, Charles Yeo, Masaya Jimbo, Joseph Cozzitorto, Jonathan R. Brody, Kathleen Delgrosso, John S. Mattick, Paolo Fortina, and Isidore Rigoutsos. 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, mar 2015. ISSN 0027-8424. doi: 10.1073/pnas.1420955112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1420955112.

[34] Harsh Dweep and Norbert Gretz. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nature Methods*, 12(8):697–697, 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3485. URL http://www.nature.com/doifinder/10.1038/nmeth.3485.

[35] Surajit Chaudhuri, Vivek Narasayya, and Ravi Ramamurthy. Estimating Progress of Execution for SQL Queries, jun 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.

[36] Chung-chau Hon, Jordan A Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen J L Rackham, Julian Gough, Elena Denisenko, Sebastian Schmeier, Thomas M Poulsen, Jessica Severin, Marina Lizio, Hideya Kawaji, Takeya Kasukawa, Masayoshi Itoh, A Maxwell Burroughs, Shohei Noma, Sarah Djebali, Tanvir Alam, Soichi Kojima, Yukio Nakamura, Harukazu Suzuki, Carsten O Daub, Michiel J L De Hoon, Erik Arner, and Long Rnas. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature Publishing Group*, 543(7644):199–204, 2017. ISSN 0028-0836. doi: 10.1038/nature21374. URL http://dx.doi.org/10.1038/nature21374.

[37] H Dweep and N Gretz. miRWalk2 web page.

[38] Dimitra Karagkouni, Maria D Paraskevopoulou, Serafeim Chatzopoulos, Ioannis S Vlachos, Spyros Tastsoglou, Ilias Kanellos, Dimitris Papadimitriou, Ioannis Kavakiotis, Sofia Maniou, Giorgos Skoufos, Thanasis Vergoulis, Theodore Dalamagas, and Artemis G Hatzigeorgiou. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Research*, 46(D1):D239–D245, jan 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1141. URL http://academic.oup.com/nar/article/46/D1/D239/4634010.

[39] Chih-Hung Chou, Sirjana Shrestha, Chi-Dung Yang, Nai-Wen Chang, Yu-Ling Lin, Kuang-Wen Liao, Wei-Chi Huang, Ting-Hsuan Sun, Siang-Jyun Tu, Wei-Hsiang Lee, Men-Yee Chiew, Chun-San Tai, Ting-Yen Wei, Tzi-Ren Tsai, Hsin-Tzu Huang, Chung-Yu Wang, Hsin-Yi Wu, Shu-Yi Ho, Pin-Rong Chen, Cheng-Hsun Chuang, Pei-Jung Hsieh, Yi-Shin Wu, Wen-Liang Chen, Meng-Ju Li, Yu-Chun Wu, Xin-Yi Huang, Fung Ling Ng, Waradee Buddhakosai, Pei-Chun Huang, Kuan-Chun Lan, Chia-Yen Huang, Shun-Long Weng, Yeong-Nan Cheng, Chao Liang, Wen-Lian Hsu, and Hsien-Da Huang. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1):D296–D302, jan 2018. ISSN 0305-1048. 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.

[40] Dong Yue, Hui Liu, and Yufei Huang. Survey of Computational Algorithms for MicroRNA Target Prediction. *Current Genomics*, 10(7):478–492, nov 2009. ISSN 13892029. 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.

[41] T M Witkos, E Koscianska, and W J Krzyzosiak. Practical Aspects of microRNA Target Prediction. *Current molecular medicine*, 11(2):93–109, 2011. ISSN 15665240. doi: 10.2174/156652411794859250.

[42] Robin C Friedman, Kyle K.-H. Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, oct 2009. ISSN 1088-9051. 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.

[43] Panagiotis Alexiou, Manolis Maragkakis, Giorgos L. Papadopoulos, Martin Reczko, and Artemis G. Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25(23):3049–3055, dec 2009. ISSN 1460-2059. 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.

[44] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4, aug 2015. ISSN 2050-084X. doi: 10.7554/eLife.05005. URL https://elifesciences.org/articles/05005.

[45] David P Hill and Kent a Robertson. 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, aug 1997. ISSN 01653806. doi: 10.1016/S0165-3806(97)00076-X. URL http://www.ncbi.nlm.nih.gov/pubmed/9298234https://linkinghub.elsevier.com/retrieve/pii/S016538069700076X.

A

Transcription Factor Regulatory Circuits - Tissue Types