Thesis submitted for the degree of Doctor of Philosophy
at the University of Oxford

TRANSCRIPTIONAL BRAIN NETWORKS AND THEIR KEY REGULATORS ACROSS THE HUMAN LIFESPAN

Claudia Constanze Wehrspaun

MAGDALEN COLLEGE

OXFORD

MICHAELMAS TERM

2014

TRANSCRIPTIONAL BRAIN NETWORKS AND THEIR KEY REGULATORS ACROSS THE HUMAN LIFESPAN

Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford

Claudia Constanze Wehrspaun, Magdalen College, Michaelmas Term 2014

Abstract

The human brain's transcriptome undergoes substantial changes over the lifespan and shows characteristic patterns that reflect anatomical regions and cellular compositions. In this thesis, I applied combinations of network algorithms and tools from computational biology to analyse transcriptional networks and their key regulators in the human brain across space (brain regions) and time (the human life course). First, I identified an age-dependent transcriptional network enriched for microglial markers. The microglia network recapitulated haematopoietic master regulators that are crucial for early microglia development using data from the ageing human brain only. In the second project, I demonstrated that gene clusters linked to neurogenesis during fetal life show moderate to strong preservation in the human adult brain. In addition to temporal development, I analysed transcriptional network dynamics across the spatial axis and detected a network of ion channel/transporter genes that express their longest 3' untranslated regions (3'UTRs) exclusively in the brain. Enrichment for predicted miRNA response elements that are often shared among the ion channel/transporter genes, along with increased co-expression of this gene set, indicated that the extended 3'UTRs could serve as a hub for an endogenous competitive RNA network. I extended

the analysis to differentiate between brain regions and additional regulatory RNA elements, namely long noncoding (lnc) RNAs. I found that genes in the hypothalamus express a region-specific network to which are also associated co-expressed lncRNAs. Finally, I added global metrics to the analysis of local networks. The dynamics of global network metrics indicated strong coordination of expression across the lifespan compared to similar variance within age groups. This work shows that the transcriptome of human post-mortem brains at least partially preserves the network structure of cell types and functionally related genes, and how it may be dissected using suitable combinations of bioinformatic algorithms.

Table of Contents

Chapter 1 – Introduction 9

1.1 Transcriptomics 9

1.2 Quantifying gene expression 13

1.3 Gene regulatory networks 16

1.4 Human brain transcriptome 20

1.5 Functional modules in the brain's transcriptome 27

1.6 Research aims and thesis structure 30

Chapter 2 – Methods 32

2.1 Gene expression data from 'braincloud' 32

2.2 ChIP-Seq data 33

2.3 Computational tools 36

2.3.1 GAT 36

2.3.2 HOMER 37

2.3.3 CUFFLINKS 38

2.3.4 TargetScan 41

2.4 Functional enrichment 42

2.4.1 GO-terms 42

2.4.4 DAVID 43

2.4.5 GREAT 44

2.5 Network algorithms 46

2.5.1 WGCNA 46

2.5.2 ARACNE 52

Chapter 3 - Microglia recapitulate a haematopoietic master regulator network in the ageing human brain 55

3.1 Abstract 55

3.2 Introduction 56

3.3 Results 59

3.3.1 An age-dependent gene module in frontal cortex is preserved across independent data sets 59

3.3.2 Microglia surface receptors' expression levels reflect the module's age-dependency 67

3.3.2 A network of master regulators controls the microglia module 71

3.3.3 ChIP-Seq indicates causal role of master regulators for microglia module ... 78
3.4 Discussion ... 82
3.5 Methods ... 86
3.5.1 Data sets and expression quantification ... 86
3.5.2 Microarray gene expression ... 86
3.5.3 ChIP-Seq data ... 87
3.5.4 Transcription Factors ... 88
3.5.5 WGCNA ... 88
3.5.6 ARACNE ... 90
3.5.7 Master Regulator Analysis ... 91
3.5.8 Enrichment of functional categories and cell types ... 91
3.5.9 DAPPLE ... 92
3.5.10 Intersection of genomic intervals ... 92
3.6 Appendix ... 93
Chapter 4 – Preservation of transcriptional networksassociated with neurogenesis from fetal post-mortem brains during later life ... 106
4.1 Abstract ... 106
4.2 Introduction ... 108
4.3 Results ... 112
4.3.1 Cell cycle transcriptional modules in fetal brain reflect different cell stages ... 112
4.3.2 Neurogenesis transcription factors in developmental and adult stages include genes that regulate adult neurogenesis ... 119
4.3.3 Validation of neurogenesis TFs using ChIP-Seq data and motif enrichment ... 121
4.3.4 The transition of cell cycle transcriptional modules from fetal to adult life is characterized by age-dependent transcription factors ... 123
4.4 Discussion ... 131
4.5 Methods ... 136
4.5.1 ChIP-Seq data ... 136
4.5.2 Functional enrichment: GO-terms and MGI phenotypes ... 137
4.5.3 Enrichment of genomic regions ... 137
4.5.4 TF-binding motif enrichment ... 137
4.6 Appendix ... 139

Chapter 5 - Brain-expressed 3'UTR extensions strengthen miRNA cross-talk between ion channel/transporter encoding mRNAs 153

5.1 Abstract 153

5.2 Introduction 155

5.3 Results 158

5.3.1 Annotation and characterization of human genes 3' untranslated regions 158

5.3.2 Isoforms with long alternative 3'UTRs are often co-expressed and functionally related 162

5.3.3 Long alternative 3'UTRs strengthen miRNA mediated cross-talk among *ion channel/transporter genes* in the brain 165

5.4 Discussion 171

5.5 Methods 174

5.5.1 RNA sequencing data 174

5.5.2 Annotation of constitutive and alternative 3' untranslated regions 174

5.5.3 3'UTR expression 177

5.5.4 Lineage specific constraint 177

5.5.5 Functional enrichment 178

5.5.6 Expression correlation 178

5.5.7 Prediction of miRNA response elements 179

5.6 Appendix 181

Chapter 6 – Identification of candidate long noncoding RNAs acting as competitive regulatory elements in a transcriptional network linked to hormone activity 190

6.1 Abstract 190

6.2 Introduction 192

6.3 Results 195

6.3.1 Transcriptional network 195

6.3.2 Modules specific to brain regions 195

6.3.3 LncRNAs correlated with modules' eigengenes 201

6.3.4 Shared MREs between the hormone module genes' transcripts and the associated lncRNAs 208

6.3.5 Correlation between the antisense lncRNA's module membership and the median percentage of MREs it shares with the hormone module genes 214

6.4 Discussion 226

6.6 Methods 230

6.6.1 RNA-Sequencing data 230
6.6.2 Annotation of 3' untranslated regions 231
6.6.3 Expression of miRNAs 232
6.6.4 Annotation of long noncoding RNAs 232
6.6.5 Construction of transcriptional networks using WGCNA 234
6.6.6 Correlation of lncRNAs with the ME 235
6.6.7 Prediction of miRNA response elements 236
6.6.8 Density of miRNA response elements 236
6.6.9 Calculation of percentage shared MREs 236
6.7 Appendix 238
Chapter 7 – Global networks dynamics in the human brain's transcriptome across the human adult lifespan 254
7.1 Abstract 254
7.2 Introduction 256
7.3 Results 258
7.3.1 Transcriptional networks covering the lifespan show stronger coordination of expression than within age groups 258
7.3.2 Transcriptional networks from age groups but not covering the full lifespan are enriched for gene functional annotations 264
7.4 Discussion 283
7.5 Methods 287
7.5.1 Data 287
7.5.2 Network construction 287
7.5.3 Global network metrics 288
7.5.4 Null models 291
7.5.5 Functional enrichment 292
Chapter 8 – Conclusions and Outlook 293
Table of Figures 298
Table of Tables 307
References 308

Chapter 1 – Introduction

1.1 Transcriptomics

The transcriptome is the entirety of all transcripts in a cell, including protein-coding and non-coding sequences, with about three-quarters of the genome showing capability of being transcribed [1]. The transcriptional landscape is highly dynamic and varies among cell types, developmental stages, physiological conditions and different alleles [1-3]. Generating a comprehensive map of the transcriptome in a given cell or for an organism is challenged by the shortcomings of technologies for measuring transcript abundance and the need to integrate vast high-dimensional data sets. Traditionally, protein-coding genes have been annotated based on sequence conservation and the presence of open reading frames (ORFs) or from the hybridization of cDNA to genome tiling microarrays to measure transcriptional activity [4, 5]. The development of next generation sequencing allowed transcriptomes to be studied with a higher resolution and exons to be targeted on a nucleotide level [4]. However, despite the acceleration in sequencing capabilities, we still have a limited understanding of biological information that is encoded in DNA and RNA, especially concerning regulatory elements and their potential functional impact on the transcriptome. This is a crucial question as only about 1.5% of the human genome is composed of known protein-coding exons [6, 7], whilst the majority of genes in the genome is covered by processed (62.1%) or primary transcripts (74.7%) [1]. The Encyclopedia of DNA Elements (ENCODE) project, aiming at identifying functional elements, assessed that 80.4% of the human genome participates in biochemical events, and that 99% of bases lie within 1.7kb of such an event [8].

Elucidating the biological functions of coding and noncoding (nc) transcripts is the key to understanding gene regulatory networks and their dynamic changes during development and in health or disease. An ever growing number of ncRNA types has been identified, ranging from miRNAs (~22 nucleotide long microRNAs [9, 10]), piRNAs (~26-31 nucleotides long PIWI-interacting RNA [11, 12]), to lncRNAs (highly diverse group of long ncRNAs, arbitrarily defined as >200 nucleotides [6, 13, 14]). The transcript types discussed in this thesis are (*i*) messenger RNAs (mRNAs, the RNAs that are translated into proteins), (*ii*) lncRNAs, and (*iii*) miRNAs.

Protein sequence is encoded in mRNAs by 61 possible triplet codons for 20 naturally occurring amino acids, whereas genomically recoded, or in other words genomically modified, organisms can incorporate additional nonstandard amino acids [15, 16]. The diversity of proteins is increased by alternative splicing (AS) of pre-mRNA transcribed from one gene. During AS, multiple mRNAs are generated via the splicing of exons in different ways (Figure 1 [17]).

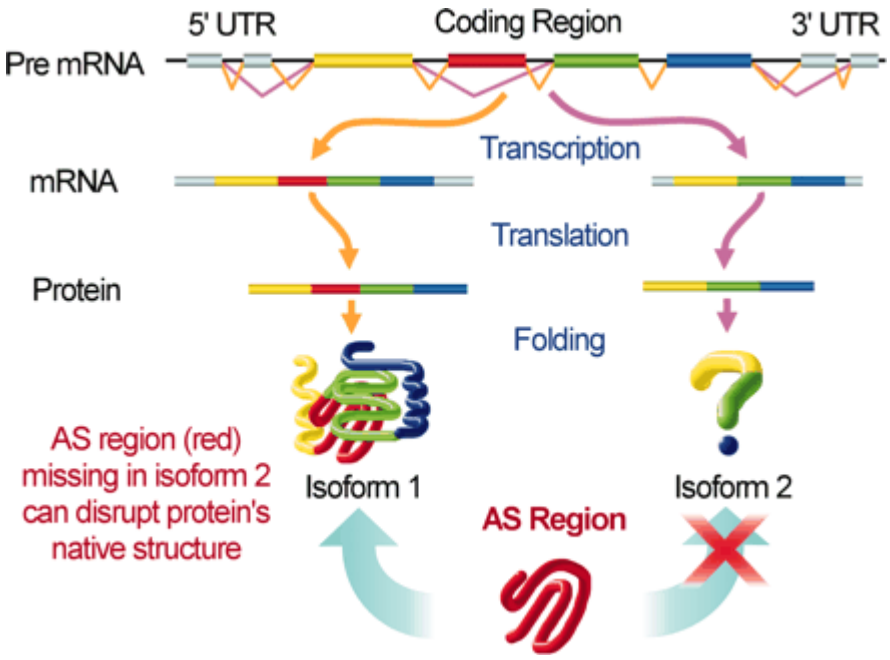


Figure 1. The genetic code is transcribed into pre-mRNA, consisting in 3'end and 5'end untranslated regions (UTRs), the protein coding regions consisting of expressed regions (exons, colored cylinders) and intragenic regions (introns, black lines). In the mature mRNA, introns are spliced out, and exclusion of exonic regions can lead to alternative isoforms (demarcated by orange and purple arrows). (In addition to the exclusion of exonic regions, alternative isoforms can be generated by the use of alternative exon boundaries, which is not displayed here). The terminal untranslated regions can regulate translation of mRNA into proteins. Here, only one isoform (orange arrows) includes the alternatively spliced region (red), leading to alternative protein products; Figure adapted from [17] (Copyright Romero 2006).

The vast majority of human genes undergo alternative splicing (~93% [18]), expressing an average of 10-12 isoforms per gene [1]. Different tissues show stronger variation in

splicing events than what is observed between different individuals [18]. Multiple mRNA isoforms can be produced by alternative splicing, alternative cleavage and polyadenylation (APA) and alternative promoter usage [18]. The most common way of generating alternative isoforms is skipping of an exon [19]. The resulting protein isoforms can range from related to distinct to opposing functions [18] and may have greatly varying 3-D structures [17]. Importantly, the abundance of an mRNA isoform is not necessarily a good proxy for abundance of the corresponding protein [20], due to post-transcriptional, post-translational and protein degradation regulation which together determine steady-state protein abundance [21].

Noncoding transcripts are not translated into proteins, but play crucial roles for regulating transcript abundance and processing. One important class of ncRNAs are microRNAs which regulate posttranscriptional repression of their target genes via binding to these genes' 3' untranslated regions (3'UTRs) [9]. Initially discovered in nematodes, miRNAs are now recognized as key regulators for diverse biological processes in mammals, during development [22, 23] or pathogenesis [24]. MiRNAs modulate gene expression by either blocking translation of their target or inducing its decay [25, 26]. It has been suggested that mRNAs, transcribed pseudogenes and lncRNAs 'talk' to each other by competing for the same miRNAs, and in this way potentially alter transcript abundance [27]. This competing endogenous RNA (ceRNA) activity could form a regulatory network for fine-tuning of transcript abundance [27] (discussed in detail in **Chapter 5** and **Chapter 6**).

1.2 Quantifying gene expression

The development of high-throughput methods for gene expression profiling made it possible to study the transcriptome on a genome-wide scale as opposed to quantifying single or few genes of interest. Different versions of microarray systems emerged during the 1990s [28-30]. In principle, microarrays use fluorescently labelled probes from mRNA which bind to cDNA targets. Fluorescence detection then allows expression abundance estimation [29]. A commonly used approach for transcriptome analyses applies two-colour microarrays. Here, mRNA isolated from two samples to be compared is reverse transcribed to cDNA and labelled using different fluorophores (usually for red and green wavelengths). Equal amounts of labelled cDNA targets from both conditions are then competitively hybridized to the same microarray. Gene expression is quantified by evaluating the ratio of the intensities of fluorescence emission at both wavelengths [30, 31]. Hybridization-based approaches have been expanded to custom microarrays which offer high resolution [5] and allow capturing of alternatively spliced isoforms [32]. However, these microarrays tend to be expensive and come with different disadvantages, such as the necessity of available sequence information and normalization of expression levels, providing a relative measure which is influenced by the choice of reference [33]. For these reasons, as well as the wider dynamic range, the possibility to determine splice variation and to identify new transcripts, transcriptome studies increasingly rely on sequencing-based approaches.

Historically, Sanger sequencing [34] was the method of choice for deriving DNA sequences before it was replaced for many large scale genome-wide studies by short sequencing tag-based methods [35, 36]. For example, serial analysis of gene expression (SAGE) obtains tags from 5'ends of mRNAs, or both terminal ends, allowing the

identification of transcriptional start sites (TSS) [35]. Similarly, cap analysis of gene expression (CAGE) derives tags from the initial 20 nucleotides from 5′ end mRNAs [36].

Ultra-high-throughput sequencing (RNA-Seq) of the transcriptome was first reported in 2008 [4, 37]. Basically, polyadenylated RNA is extracted and reverse transcribed to double-stranded cDNA using random hexamers or oligo(dT) primers. Next, the cDNA is fragmented and sequenced, for example by high-throughput Illumina sequencing, determining e.g. 35 base pairs (bp) of sequence from each fragment's end [4], or, depending on genome size, up to 300 bp maximum read lengths (Illumina Sequencing Systems). The reads – in other words a sequence of DNA bases – can be single end-reads that were generated from one end of the DNA fragment, or paired-end reads that were generated from both ends. Paired-end reads improve alignment and can provide information about splicing events to reconstruct alternative isoforms in RNA-Seq data [38, 39]. Recently alternative approaches have been developed for RNA-Seq library preparation, for example Ribo-Zero [40], which overcome some disadvantages of isolating poly(A) RNA, including the depletion of ribosomal RNA (rRNA) or selection of only the 3' most end if RNA is not intact [41]. RNA-Seq directly sequences cDNA without the need of cloning. However, in consequence no a priori information is available about the original transcripts from which reads were derived [42]. Short reads therefore need to be mapped back to the genome and/or transcripts assembled [43]. Different workflows are possible to achieve this, each associated with potential disadvantages [44, 45]. For example, read counts, or transcript abundance, may influence assembly depending on the data processing approach chosen [42] (Figure 2).

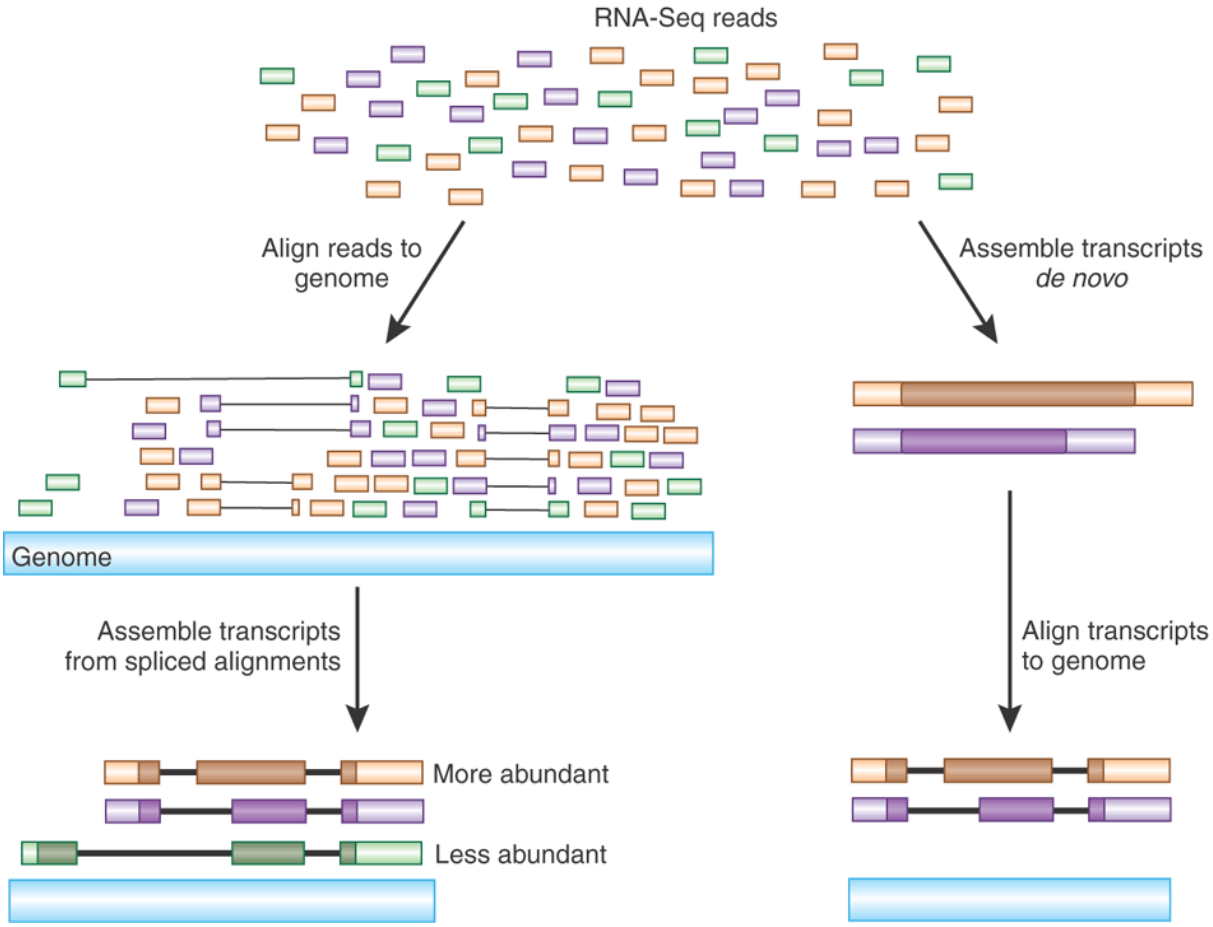


Figure 2. Comparison of alignments of reads to genome followed by transcript assembly (left) or transcript assembly followed by alignment of transcripts to genome (right). Reference [42] suggests that align-then-assemble (left) approaches are more sensitive because they can detect less abundant isoforms in contrast to assemble-then-align (right) methods. Figure reproduced from [42] with permission from the Nature Publishing Group, www.nature.com.

The value of RNA-Seq data for downstream analysis and experiments depends on certain quality control parameters. One important quality control parameter is the RNA integrity number (RIN), which reflects the integrity, or in turn the degradation, of the

sequenced RNA reads at the moment of RNA extraction from a given sample [46]. Controlling for the confounding effect of the distribution of RIN values across samples in an experiment is necessary, because naturally occurring rapid RNA degradation leads to shorter RNA fragments, which may confound the variance in the RNA abundance observed in an experiment [46].

1.3 Gene regulatory networks

Gene expression is orchestrated by multiple factors, from initiation of transcription by transcription factors (TFs) and co-factors to regulatory networks of various RNA types, including those described in **Chapter 1.2**. Transcription factors bind specific DNA-sequences, interact with other regulatory elements and in this way control the rate of transcription initiation [47] (Figure 3).

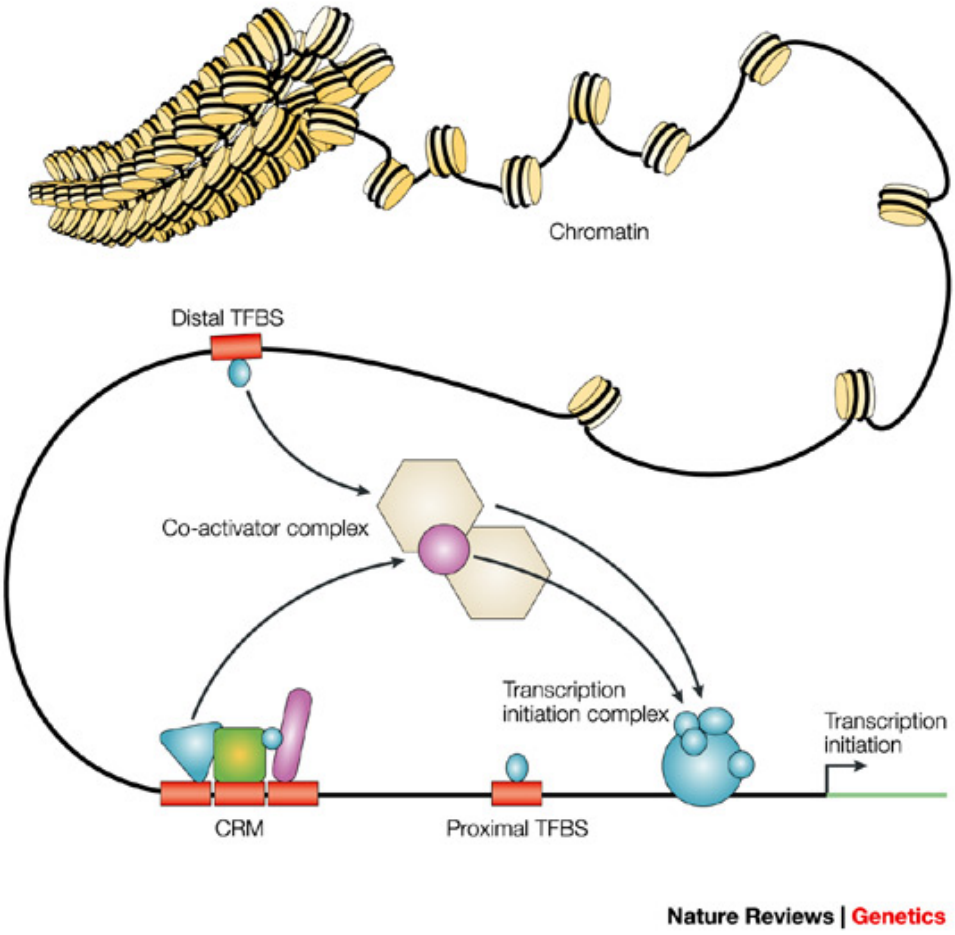


Figure 3. Transcription factors (TFs) target specific DNA sequences, called transcription factor binding sites (TFBS), which can be located proximal or distal to a transcription start site (TSS). Multiple TFs may form functional *cis*-regulatory modules (CRMs). The transcription initiation complex is further stabilized by interactions between bound TFs and cofactors. Figure reproduced from [47] with permission from the Nature Publishing Group, www.nature.com.

Analysis of single genes' expression patterns across time is limited to fold change or differential expression and cannot provide information about association between genes or the context of gene sets of interest [48]. For this reason, gene regulatory networks

(GRNs) which reflect temporal expression patterns of interacting genes are of interest. Here, gene regulatory networks are defined as co-expression networks with genes being nodes and the interconnection between all possible gene-pairs forming the network edges. Interconnection can be defined in various ways, such as binary connection (0 indicating unconnected and 1 connected), correlation of expression, weighted co-expression [49] or mutual information [50] (for details of measures and algorithms for network generation see **Chapter 2.3**). Measuring dynamic correlations between gene expression levels can provide information about regulatory activity between them [51].

Transcriptional regulatory networks governed by transcription factors have been derived for a variety of different species and systems [52-56]. The evolution of these regulatory networks has been linked to gene duplication [57], wiring new genes into networks via transposable elements [58, 59], and accumulation of single nucleotide polymorphisms (SNPs) in the centre of strain-specific transcription factor binding motifs [60]. Inference of gene regulatory networks based on transcription factor-target gene interactions allows the construction of directed networks and the subsequent identification of regulatory circuits and transcription factors acting as master regulators [50, 61].

The identification of transcription factors acting as master regulators can serve as a basis for building hierarchical regulatory models or selecting best predictors for phenotypes. In the following, I define master regulators as transcription factors which are specific for a defined gene set. More precisely, a transcription factor is termed a master regulator if the predicted regulon of the master regulator (in other words its target genes) shows significant overlap with the given gene set [61]. For example, the expression of master regulators for glioma-specific transcriptional networks predicts clinical outcome [61]. Master regulator analysis can be applied for dissecting disease pathways. For example, fibroblast growth factor receptor 2 (*FGFR2*) has been identified

as a risk factor for breast cancer in genome wide association studies (GWAS) [62]. The regulons of genes identified as master regulators acting upstream of *FGFR2* as well as genes regulated downstream by *FGFR2* itself are enriched for genes harbouring breast cancer risk loci [62].

The structure of transcriptional networks can be analyzed on a global level or a local level. On the global level, one may ask the question how network measures differentiate between developmental stages or physiological conditions [52, 63]. For example, in a gene expression network of transcription factors and their target genes in yeast, topological statistics, such as degree (number of network edges directly connected to a node) and clustering coefficient (quantifying the abundance of 'triangles', defined as triplets of genes that are all connected by a network edge), which both reflect network interconnectivity, differed between cellular conditions [52]. Changes in interconnectivity indicate that transcription factors may act as transient hubs which regulate genes only in one condition but not in others [52]. On a local level, networks can be broken down into their basic building blocks, regulatory circuits or 'network motifs' [53, 54, 64]. Motifs in this context are defined as interconnection patterns which occur in the network more often than expected by chance. (The clustering coefficient described above is a specific case of motif quantification, namely the 'triangle' motif). These network motifs fulfil distinct biological functions such as governing temporal expression and regulating responses to external signals. For example a transcriptional network in *Escherichia coli* was dissected into three highly significant recurrent network motifs [53]: (*i*) Feed-forward loops in which one transcription factor regulates another transcription factor, so that both regulate a set of target genes. The feed-forward motif often occurs when systems rely on rapid responses to external signals. (*ii*) Single-input-module (SIM), defined as a set of target genes which are positively or negatively

regulated by the same transcription factor, whereby this transcription factor often shows autoregulatory activity. Within SIMs, the transcription factor may act as a master regulator for generating temporal expression patterns of its module via different gene activation thresholds. (*iii*) Dense overlapping regulons (DOR) refer to a dense layer of interactions among a set of transcription factors and a group of target genes [53]. Importantly, analysis of network motifs allows the generation of predictions which can be experimentally tested and validated [64].

1.4 Human brain transcriptome

The development of the human brain during evolution and across the lifespan shapes our intellectual abilities, consciousness, and all cognitive or emotional processes that define us as human. Identifying critical developmental stages and biological pathways which drive changes in the brain's organization and function at these points in time will improve our understanding of how the environment interacts with brain development, modulates brain plasticity, and potentially makes us susceptible to neurological or psychiatric diseases.

Research has aimed to dissect neuronal processes focusing on structure or activity in the human brain, analyzing morphology or biological markers in post-mortem brain tissue and linking phenotypes to the underlying genomic code [2, 3]. Multiple methods have been applied to define cohesive organizational units in the brain [3, 61]. Depending on the context and the research area, these units may be referred to as clusters, modules, networks or graphs [3, 53, 54]. They can be derived from coordinated brain activity during specific tasks [65], functional relatedness [3, 66] or various phenotypic patterns which evolve during the lifespan [67, 68]. Defining functionally related groups in the

transcriptional network of the brain is the focus of this thesis. In the following, I provide an overview of previous studies of the human brain transcriptome, its development from young to older age or in comparison to other species, and approaches to define transcriptional networks in the brain. For the sake of brevity, I will not discuss brain networks derived from different data types such as imaging or electrophysiology.

The human brain transcriptome undergoes major changes during the lifespan [2, 3, 69-74]. These changes do not evolve linearly from fetal to older age, but show turning points in magnitude and velocity depending on developmental stage [2], or brain area [72] and can be gender-specific [3, 69]. Most pronounced changes in gene expression happen before or around the age of birth, when expression patterns observed during fetal development are reversed during infancy (Figure 4) [2, 3].

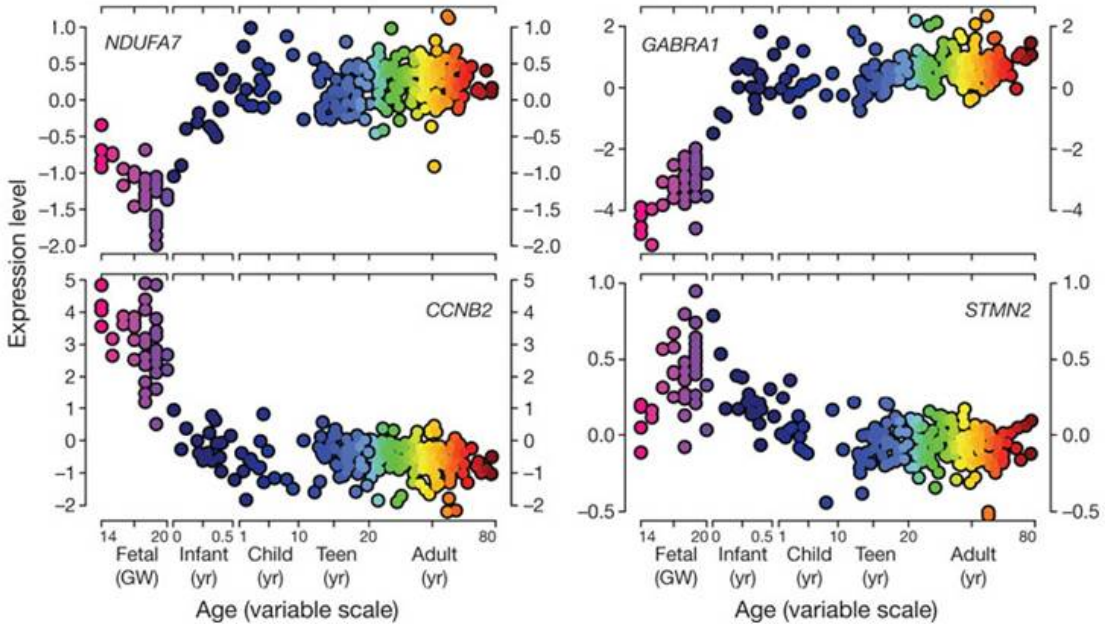


Figure 4. Illustration of gene expression patterns observed across the human lifespan using four individual genes. The gene expression level is shown on the y-axis and the age on the x-axis. The four genes illustrate strong expression changes before and around the age of birth, followed by more stable gene expression during adult age. Figure reproduced from [2] with permission from the Nature Publishing Group, www.nature.com.

A similar pattern occurs during later life around the age of 50 years, following the time of lowest fold change in expression between 20 to 40 years [2]. Three-quarters of the genes that showed expression changes during fetal life and infancy reversed their expression about the age of birth. More precisely, the expression of genes enriched for axonal function increased during fetal life and changed direction of expression around the age of birth, whereas genes enriched for ATP synthesis showed the opposite pattern. Genes whose expression increases in both stages were enriched for terms linked to the

synapse, compared to genes decreasing during this time which were enriched for cell division. Trajectories of gene expression in fetal life show a negative correlation with change rate after the age of 50 years [2]. This finding is in line with an age-dependent repression of neuronal genes linked to synaptic function [74], although no association was found with overall neuronal or synaptic loss. For example, the expression of genes linked to GABA-ergic inhibitory function was downregulated during increasing age [74]. Alternative splicing shows a similar pattern during the lifespan compared to gene expression changes. Namely, most alternative splicing events (70%) take place during development (0-14 years) compared to fewer events (30%) occurring during ageing (25-98 years) [70]. Corresponding to the gene expression changes during early development, that are mirrored later in life, splicing patterns are similar between infants and the elderly [70]. In addition to high rate of change during early life, inter-individual variance in gene expression is higher in children compared to adults. The genes showing this pattern tend to be annotated as genes involved in immune system processes, suggesting that their function may be linked to antigen regulation as well as developmental processes in the nervous system [75].

Apart from the time axis, the human brain's transcriptome evolves on a spatial axis, although it has been argued that the developmental stage is more relevant for the cellular transcriptome than regional identity [76]. Gene expression patterns display regional specificity which is highly conserved between individuals, depending on brain areas [3, 73, 77], as well as cell types [67, 78], for example neurons, oligodendrocytes, astrocytes and microglia [67, 73, 78]. In consequence, detected changes of expression levels may be caused either by different transcript abundance or variation in the relative abundance of cell types [79]. Furthermore, changes in cellularity across time might reflect epochs of human brain ageing [80]. The development of spatio-temporal

expression patterns in the human neocortex has been described as an hourglass shape: Major expression differences in interareal comparison were observed during the earliest phase in prenatal development, followed by a more stable preadolescent phase. The postadolescent phase is again characterized by increased differential expression although with a less defined gradient-like structure [71].

Spatial proximity of cortical regions and layers in the neocortex is accompanied by increased similarity in their transcriptome [73, 81], with enriched frontal lobe expression and specific features associated with primary sensorimotor cortices [73]. Expression patterns of genes showing variation between anatomical regions align to principal components thereby reflecting the distinct regions. For example, the three first principal components derived from 1000 genes with highest variance across 56 gyri in two healthy subjects differentiated (*i*) primary sensorimotor cortices, (*ii*) posterior orbital, paraolfactory and subcallosal gyri in frontal lobe, temporal lobe and primary visual cortex, and (*iii*) frontal cortex compared to temporal and occipital cortex [73] (Figure 5).

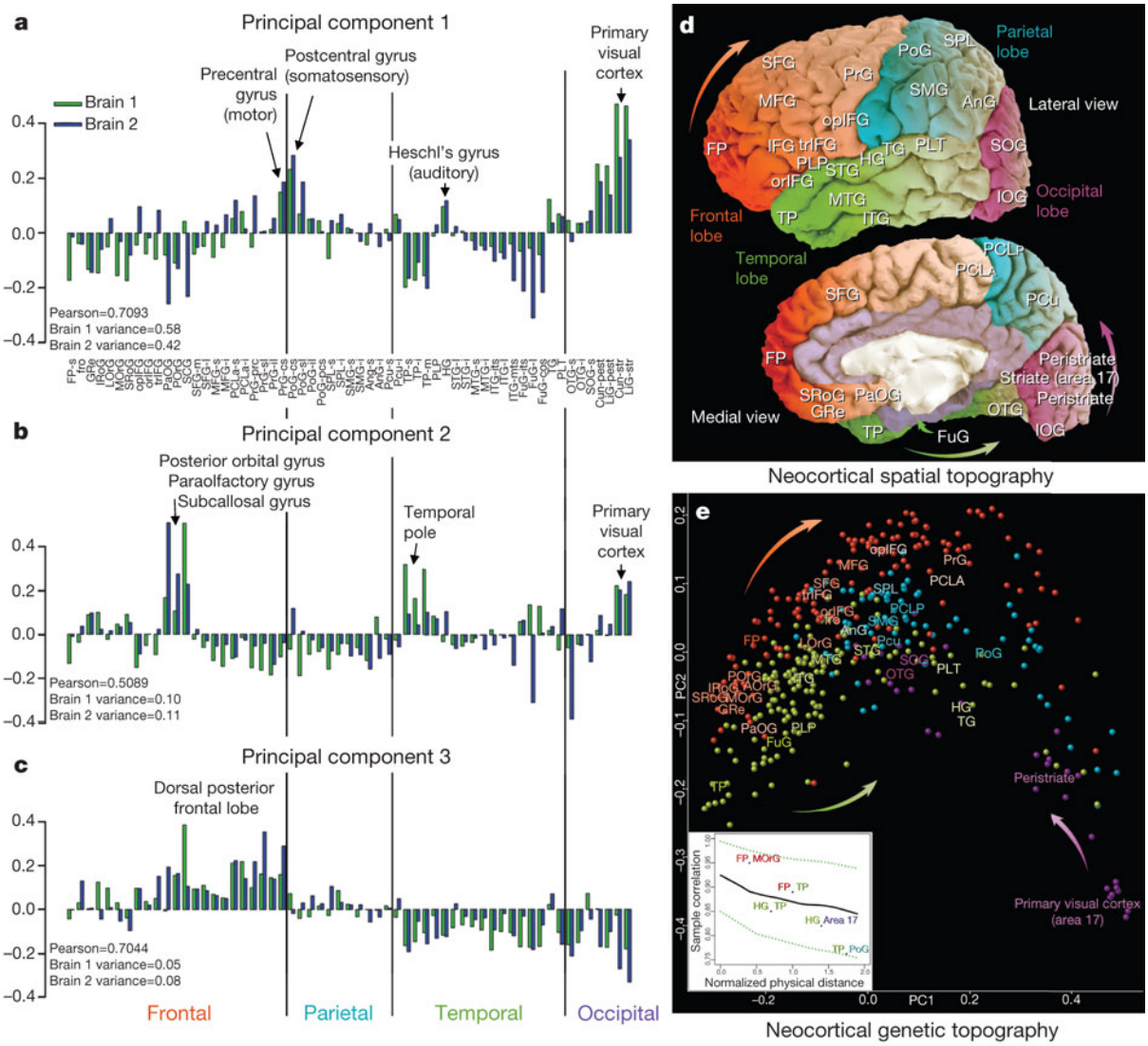


Figure 5. Gene expression patterns in neocortex reflect cortical organization. The first three principal components (a-c) were derived from 1000 genes with the highest variance across 57 cortical regions. All three components differentiate between distinct cortical divisions: (a) primary sensory and motor areas, (b) subdivisions in frontal, temporal and occipital lobe, and (c) caudal portion of the frontal lobe. Transcriptional profiles are mapped back onto the three dimensional brain topography (d-e). Gyri are labelled in a MRI representation of the brain (d). Multi-dimensional scaling allows mapping of gene expression profiles back into the

brain space, with distance reflecting the similarity in gene expression. Gene expression clusters reflect lobe and gyral positions as indicated by the arrows (e). Figure reproduced from [73] with permission from the Nature Publishing Group, www.nature.com.

Variation in the transcriptome in different brain areas is directly linked to the question of how the human brain differs from the brains of other species. Notably, not all regions of the brain's transcriptome show similar divergence across evolution. In particular, the frontal lobe displays a highly human-specific increase in complexity of gene expression patterns, in contrast to the caudate nucleus which is conserved between humans and other primates [82]. These findings are in line with the well known increase in brain size, number of neurons and proportion of high-level areas such as the prefrontal cortex (PFC) in humans [83-86], accompanied by a human-specific delay in the synaptic development compared to other species [87]. However, it is important to note that gene expression changes may – at least in part - evolve neutrally and cannot be directly translated into higher cognitive functions or their development [88].

The development of the human brain's transcriptome across time, in context of evolution or an individual's lifespan, raises the question as to what extent these changes are predetermined by genetic variations. Although a growing number of direct associations between genetic polymorphisms and gene or transcript abundance have been reported [89], no evidence has been found for an overall predictive value of genetic differences for the transcriptome as a whole [2]. Importantly, gene expression is a stochastic process and clonal cells exhibit substantial variations in their transcriptomes [90, 91] with mutations modulating the expression noise level [92]. Increasing age is associated with higher genome damage and higher stochasticity of gene expression. It

has been suggested that the DNA damage causes ageing by stochastic deregulation of the transcriptome [93]. Another potential consequence of a cell's non-genetic heterogeneity is stochasticity in cell fate decisions in stem cells [94]. For example, during haematopoiesis, the cell lineage choice is governed by the transition between metastable states of a slowly fluctuating transcriptome, accompanied by increased expression of lineage-specific transcription factors such as *PU.1* for the myeloid-prone cells and *GATA1* for erythroid-prone cells [94]. Transcription factors regulate gene expression of their targets and transmit information from a changing environment, using strength, amplitude and temporal dynamics of their signalling. Oscillatory transcription factor dynamics allow one transcription factor to direct multiple gene expression programs, whereby the cell faces a trade-off between fine-tuning differential gene expression using transcription factor dynamics, and higher gene expression noise when the complexity of dynamics of a single transcription factor increases [95]. Whether the source of correlated fluctuations in the transcriptome controlling lineage choice are specific master regulators or many parts of an interlocking genetic network remains poorly understood [96].

1.5 Functional modules in the brain's transcriptome

The brain's transcriptome is not restricted to global dynamics but is organized into subnetworks linked to various biological functions [3]. Interestingly, the representative expression trajectories of these subnetworks may relate to external traits, such as age, although the expression timecourses of individual genes may not [72]. Clustering based on pairwise gene expression can differentiate between brain areas from fetal age

onwards, with distinct areas showing differences in network cohesiveness [3]. When these clustering approaches were applied to the whole lifespan, age-dependency of modules was mostly driven by strong expression changes around the age of birth [3]. These modules were enriched for distinct biological functions and pathways. Namely, a module with high expression in early fetal life followed by decline during infancy was linked to neuronal differentiation and transcription factors, whereby a module with the opposite expression pattern was related to ion channels and neuroactive ligand–receptor interaction. Two modules showed opposite expression changes during ageing with drastic changes just after birth. The module with increased expression during ageing contained genes encoding membrane proteins, calcium signalling, synaptic transmission and neuroactive ligand–receptor interaction which are processes important during postnatal development. In comparison, the module decreasing in expression was enriched for zinc-finger protein and transcription factor genes [3].

Modules derived from co-expression not only differentiate between functionally related gene groups during development, but may also classify distinct cell types, including classes of neurons and glia [67], which emerge at different stages [80]. Importantly, expression changes depend on a given cell type, in addition to cellular composition - such as glial gene expression being modulated by the presence of neurons [97]. Another example is the improved predictive value of gene expression for synaptic density by genes related to oligodendrocytes [98]. Furthermore, associations between transcript abundance and neuroimaging measures delineating microstructure in white matter indicate that gene expression patterns are informative of the development of the brain's cellular scaffold [99]. Transcriptional clusters within distinct cell types may elucidate functional pathways which regulate neuroprotection or susceptibility to disease in an age-dependent manner. For example, the microglia transcriptome in mice revealed a

group of co-ordinated transcripts encoding proteins for sensing endogenous ligands and microbes, termed the microglia 'sensome' [66]. During ageing, expression of microglial genes linked to neuroprotection, microbe recognition and host defence were upregulated in contrast to transcripts associated with endogenous ligand recognition which were downregulated, suggesting that the neuroprotective priming state of microglia changes during the lifespan [66].

Functional modules in the human's brain transcriptome can be differentially regulated in health and diseases. For example, a transcriptional module derived from gene expression in human post-mortem brains and enriched for terms associated with activated microglia, astrocytes and immune system processes was found to be upregulated in autism patients [100]. In the context of ageing and neurodegeneration, a co-expression module linked to the immune system and microglia has been shown to be enriched for genes linked to Alzheimer's disease [101] and to be upregulated in patients with late-onset Alzheimer's disease [68]. During brain tumour growth, transcriptional modules are crucial for the transition from physiological to pathological states. The activation of regulatory modules and their master regulators are predictive for tumour aggressiveness and aberrant phenotypes [61].

In summary, constructing gene co-expression networks from the human brain transcriptome and analysing their organization in distinct subnetworks or modules not only improves our understanding of the regulation of physiological development during an organism's lifespan, but can also detect cohesive gene groups which encompass factors that are crucial for the initiation and the maintenance of disease traits.

1.6 Research aims and thesis structure

In this thesis, I analyse transcriptional networks in human post-mortem brains and their key regulators. The transcriptional networks were generated along the temporal axis, spanning the full human lifespan from fetal to older age (**Chapters 3 and 4**), and along the spatial axis, comparing coordination between genes and their regulatory elements between different brain areas (**Chapters 5 and 6**). In this way, I (*i*) identified causal regulatory factors for gene expression specific to distinct developmental periods, and (*ii*) detected transcriptional networks characterized by shared regulatory features in functionally related gene groups.

In **Chapter 2**, I describe data sets, computational tools and network algorithms which were used in multiple projects.

In **Chapter 3**, I provide evidence for the causal role of haematopoietic master regulators from fetal life for microglia homeostasis in the adult human brain. The transcription factors acting as master regulators for microglia were recapitulated using gene expression data from post-mortem human adult brains aged 12 years and older only.

In **Chapter 4**, I explore preservation of cell cycle gene modules from fetal age in post-mortem human adult brains. I found that two cell cycle modules show moderate to strong preservation in the adult age group, potentially linked to maintenance of a cellular micro-environment that allows initiation of neurogenesis or neural plasticity.

In **Chapter 5**, I describe the identification of a set of ion channel/transporter encoding genes that express their longest 3' untranslated regions exclusively in brain tissue. A significantly increased number of miRNAs shared between ion channel/transporter encoding genes suggests that their observed high co-expression is mediated by miRNA cross-talk.

In **Chapter 6**, I present brain area-specific transcriptional networks and long noncoding RNAs associated with these networks.

In **Chapter 7**, I compare global network metrics, characterising the age-dependent dynamics of the whole transcriptome in the brain, between different age groups.

In **Chapter 8**, I summarise the main findings presented in this thesis and outline perspectives for future research.

Chapter 2 – Methods

2.1 Gene expression data from 'braincloud'

The gene expression data set from the 'braincloud' study (GSE30272 [2, 3]) was used in **Chapter 3**, **Chapter 4**, and **Chapter 7** and is therefore described in further detail below. The complete gene expression data set contained 269 samples from the post-mortem human prefrontal cortex (PFC), spanning the whole lifespan from fetal development to ageing (80 years). No subject had severe neuropathology, or received a neurological or neuropsychiatric diagnosis during their lifetime. In addition, subjects with evidence of drug – or alcohol abuse were moved. The tissue homogenates included PFC grey matter (DLPFC, BA46/9, and corresponding regions in fetal brains). RNA was analyzed with custom 2-color microarrays from the National Human Genome Research Institute (NHGRI) microarray core facility. A pool of all samples was used as reference RNA. The RNA preparation, analysis and basic quality control followed the standard procedure as described in [2]. All samples with RNA integrity number (RIN) <5 were exluded. An age-appropriate linear fit was used to remove outliers (all probes >6 mean average deviations). Probes were removed if they showed nonspecific mapping (probes that were mapped to more than one expressed sequence), incorrect annotation, or contained polymorphisms with a minor allele frequency >0.01 in either the YRI (Yoruba in Ibadan, Nigeria) or CEU (Utah residents with ancestry from northern and western Europe) populations in Hapmap [2, 102]. A lower threshold for intensities was determined empirically at 5.3 on the $\log_2$ scale. All probes that showed intensities below this cutoff in half of the fetal or postnatal data points were removed. Additional background correction was applied to the $\log_2$ ratios (sample/reference) via normalization across mean $\log_2$ fluorescent intensities using the Loess correction [2,

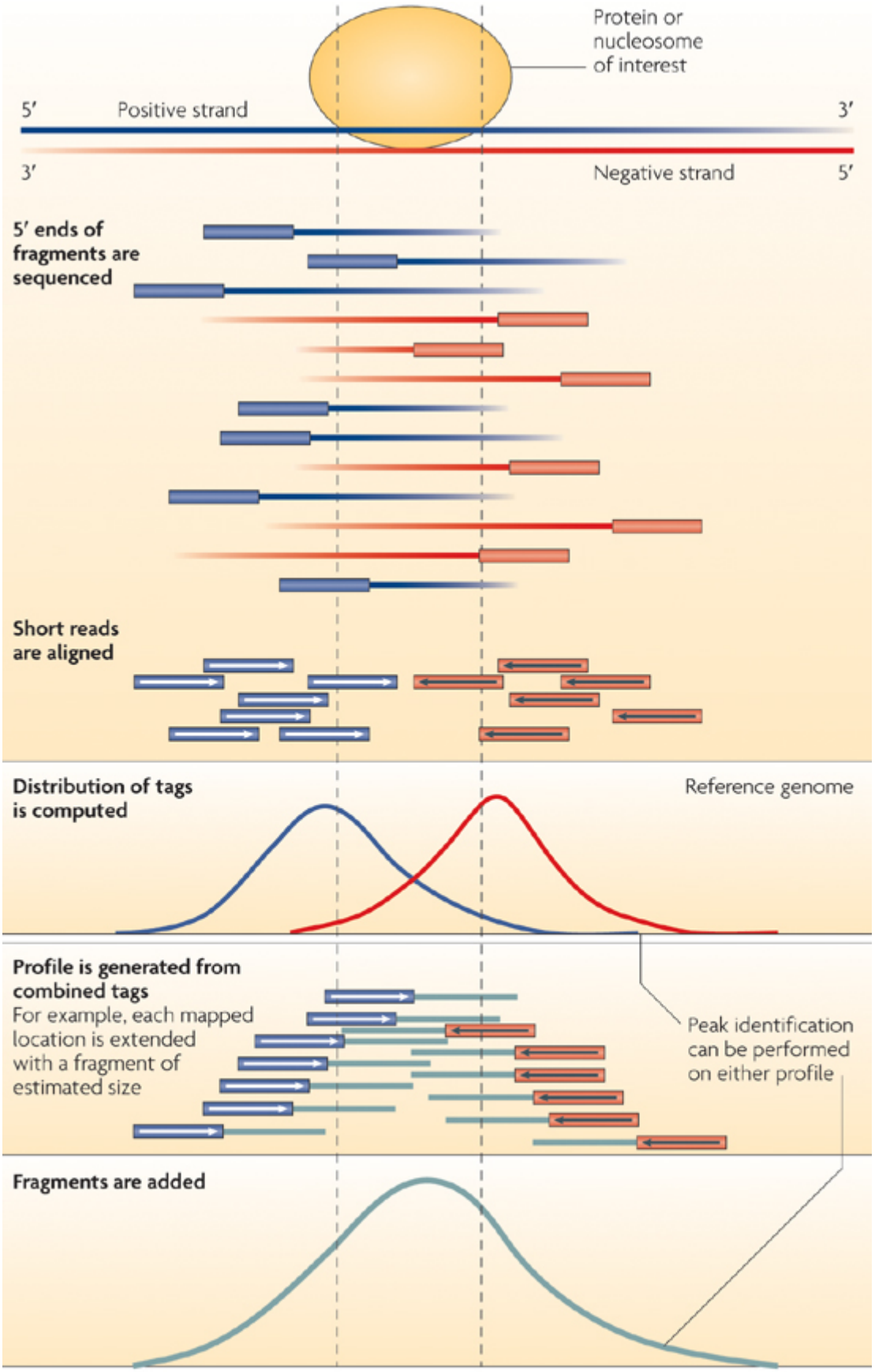
103]. Missing data were imputed [2]. The original 'braincloud' publication described a number of significant associations between individual single nucleotide polymorphisms (SNPs) and the expression of individual genes. However, there was no association between the genetic distance between individual humans and the similarity of their prefrontal transcriptional profiles, based on all pairwise comparisons between brain samples from African Americans and Caucasians [2]. This indicates that the observed age-dependent transcriptional dynamics are not driven by underlying genetic differences between individuals. The analysis was repeated considering (*i*) only SNPs which showed an association with the expression of specific genes and (*ii*) only considering individual races. In both cases, no association between global transcriptional distance and genetic distance was found. The authors concluded that the human genome produces a consistent molecular architecture across the human race [2].

2.2 ChIP-Seq data

Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) is a technique for mapping DNA-protein interactions [104, 105]. I used ChIP-Seq data to validate the computational prediction of transcription factor targets, as I shall discuss in **Chapter 3** and **Chapter 4**.

In the first step, ChIP uses antibodies to precipitate specific proteins, and thereby to enrich DNA fragments bound by these proteins. Next, these DNA fragments are sequenced and the resulting sequencing reads aligned to the genome. The enriched regions are referred to as 'peaks'. Meanwhile, a number of algorithms are available to call and annotate ChIP-Seq peaks, including HOMER for peak annotation which is described below (**Chapter 2.3.2**). The publicly available ChIP-Seq data used in

Chapter 3 and **Chapter 4** provided coordinates for called peaks, therefore, algorithms for peak calling will not be discussed in detail here. Briefly, 'peak callers' generate smoothed profiles of read distributions on both strands and merge them. Next, the peak is scored to assess whether it is accepted as statistically significant compared to data from a negative control sample. The ratio of observed signals in the sample compared to the control condition, and the absolute number of reads can be modelled using a Poisson model. The enriched regions are usually classified into different types, namely narrow, broad, or mixed, resulting in differently shaped peaks. The observed peak type is determined by the binding profile of the protein of interest, whereas narrow (or point-source) peaks often reflect transcription factor-binding in regulatory regions, in contrast to broad peaks that tend to reflect widely enriched proteins such as histone marks [104, 105].



Nature Reviews | Genetics

Figure 6. The DNA fragments enriched by ChIP are sequenced from 5'ends and aligned to plus and minus strands of the genome. The resulting two peaks, one on each strand, flank the binding region of the protein of interest. The distribution of the 'true' peak is approximated by extending reads on each strand in the appropriate direction. Figure reproduced from [104] with permission from the Nature Publishing Group, www.nature.com.

2.3 Computational tools

Publicly available computational and bioinformatics tools used repeatedly in the analyses are described in detail in this section. All custom scripts used during the analyses were written in python, matlab or R. Plots were created using R if not stated otherwise.

2.3.1 GAT

GAT (Genomic Association Tester) calculates the intersection of two sets of genomic intervals while controlling for different potential confounders such as varying G+C content or differences among chromosomes [106]. It assigns a p-value based on a randomization simulation procedure within the provided workspace – potentially the whole genome, or a pruned background, such as the genome after removal of gaps. More precisely, GAT takes the two sets of interest, for example regions annotated as genes and haplotypes of GWAS risk SNPs, and compares their overlap to randomly chosen sets of intervals across the workspace. Multiple testing correction is applied using a false discovery rate (FDR) procedure (Benjamini-Hochberg correction) that estimates the expected proportion of falsely rejected null hypotheses [106]. FDR is

defined as the expectation of Q, which reflects the proportion of erroneously rejected null hypotheses (Q=V(True null hypothesis is declared significant)/R(True or non-true null hypothesis is declared significant)). The q-value of the FDR is the equivalent of a p-value and reflects the minimum FDR at which the hypothesis may be called significant [107]. Here, the p-value corrected for FDR is referred to as Storey's q-value [108]. The FDR correction is less conservative than for example Bonferroni corrections that adjust the threshold for significance (α) by the number of performed tests [109]. More specifically, the Benjamini-Hochberg correction [107] controls for the expected FDR by (*i*) ranking the p-values, (*ii*) multiplying each p-value by the number of performed tests and (*iii*) dividing the product by the assigned rank.

2.3.2 HOMER

HOMER is a software suite for motif discovery and next generation sequencing analysis [110]. Here, part of the HOMER suite was used to annotate regions in the genome for ChIP-Seq peaks from publicly available data sets (see **Chapter 3** and **Chapter 4**). The input for annotations were ChIP-Seq peaks (if necessary lifted over to human genome built hg19 using the online liftOver tool from the UCSC browser), and the genome annotation file for hg19, using default settings with the "UCSC refGene" annotation. In the present analyses, I only used basic genomic annotations that are assigned by HOMER as the region occupied by the center of the given region, in this case ChIP-Seq peaks. Genomic regions considered were TSS (transcription start site, defined as -1kb to +100bp), TTS (transcription termination site, defined as -100bp to +1kb), coding exons, 5'UTR (untranslated regions) exons, 3'UTR exons, introns and intergenic regions.

2.3.3 CUFFLINKS

Transcripts in the Illumina body map [111, 112] and the GTEx data [113] sets (described in detail in **Chapter 5** and **Chapter 6** respectively) were assembled using Cufflinks [114]. Cufflinks takes aligned paired end RNA-Seq reads, sorted by the reference position, as input files. Here, cufflinks was run with a reference annotation (UCSC annotations) and allowing for the assembly of novel transcripts. Cufflinks uses approaches from graph theory to construct transcripts which best "explain" the given RNA-Seq data set [114]. The transcript abundance is estimated based on cDNA fragments which were sequenced and mapped back to the genes (or the transcripts). After normalization, the relative abundance of the RNA-Seq fragment counts can be used to quantify the expression. Cufflink outputs the transcripts' abundance as Fragments Per Kilobase of exon per Million fragments mapped (FPKM). Paired-end RNA-Seq provides two fragments per read (which is sequenced from both ends, see **Chapter 1.2**). The correct quantification of the transcript abundance depends on the correct assignment of the read to a given transcript, which must overcome the challenge of fragments potentially aligning to multiple isoforms of the same gene. Cufflinks applies numerical optimization for finding the unique maximum of a likelihood function, estimating the likelihood of the abundance of a set of transcripts given the observed set of fragments [114]. The Cufflinks workflow is shown in Figure 7 which is taken from [114].

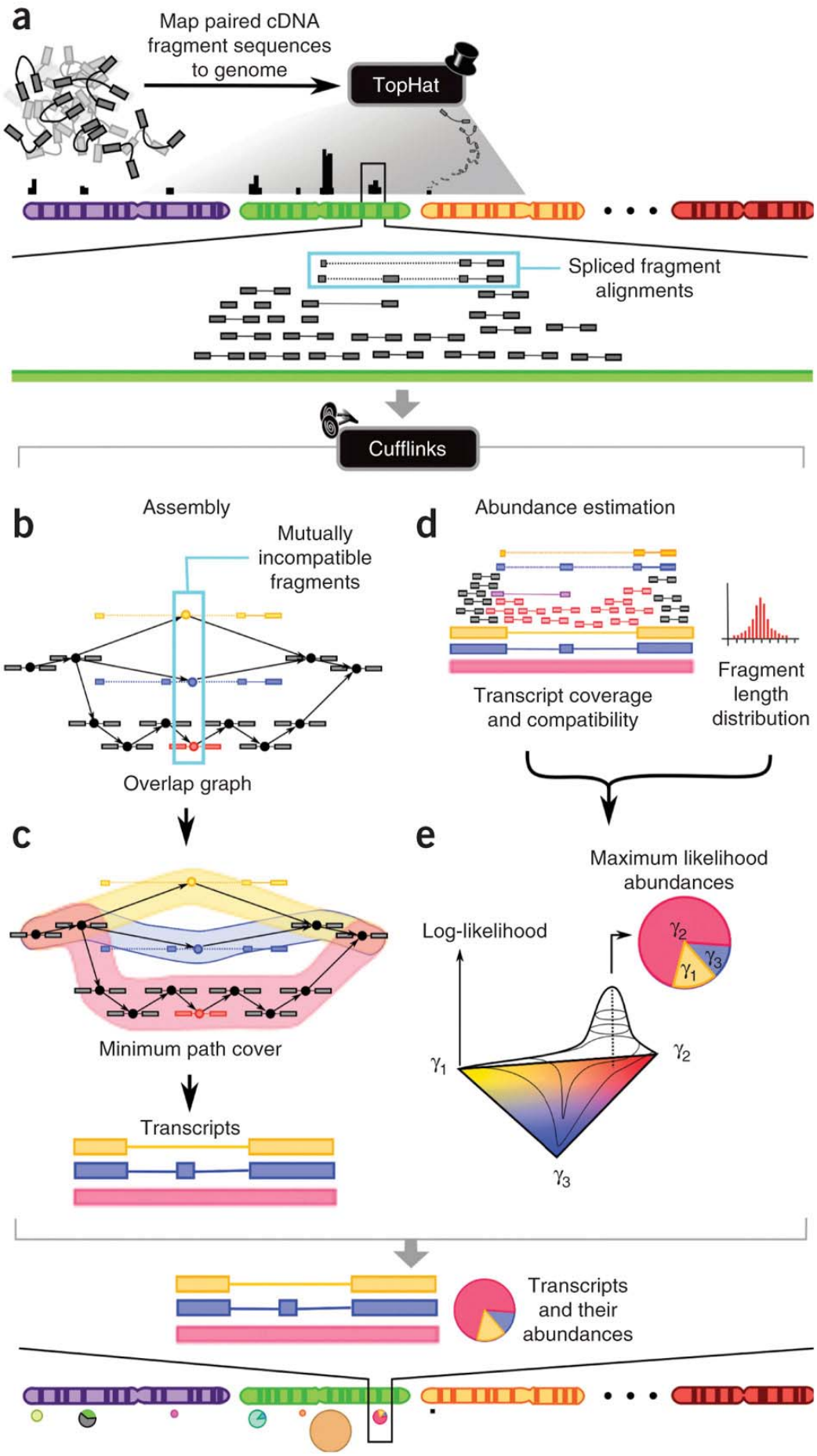

a
Map paired cDNA fragment sequences to genome
TopHat
Spliced fragment alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ1
γ2
γ3
Transcripts and their abundances

Figure 7. Flowchart for RNA-Sequencing (RNA-Seq) reads' alignment and assembly using Cufflinks [114]. A) Cufflinks' input consists of aligned cDNA fragment sequences. The software TopHat performs the required spliced alignments. B-C) Cufflinks defines paired-end RNA-Seq reads as single alignments and assembles the overlapping fragment alignments separately. D-E) The next steps estimate the abundance of the assembled transcripts. B) First, 'incompatible' fragments are identified which originate from differentially spliced mRNA isoforms. The 'compatible' fragments, with alignments overlapping in the genome, on the other hand are connected in an 'overlap graph'. In the example shown here, yellow, blue, and red fragments must have originated from different isoforms, whereas all remaining fragments could belong to any of the three shown paths. C) The isoforms are assembled from the overlap graph. Each path (three in the example), corresponds to fragments that could be merged into the same isoform. Cufflinks works based on Dilworth's Theorem, which states that the number of incompatible reads equals the minimum number of transcripts needed to 'explain' the fragments. More precisely, Dilworth's Theorem is used to produce a minimal set of paths that cover the fragments in the overlap graph, by identifying the maximum sets of reads that could not come from the same isoform. D) The transcript abundance is calculated after matching the fragments to their 'original' isoforms (illustrated here by colours). For example, the violet fragment could have come from the blue or the red isoforms, whereas the grey fragments could have come from any of the three isoforms. The probability of observing a fragment is modelled with a linear function based on the abundance of the transcript from which it could have originated. Cufflinks incorporates the distribution of the fragment lengths for the assignment of fragments to the isoforms. This is

important as the fragment assignment often implies different lengths, although only the ends of fragments are sequenced and therefore the lengths maybe unknown. This is illustrated here by the violet fragment, which could be longer if it came from the red instead of the blue isoform. E) The pie chart shows the isoform abundances that best explain the observed fragments, based on the numerical maximization of a likelihood function for all possible sets of the relative abundances of the red, yellow, and blue isoforms (γ_1,γ_2,γ_3) [114]. Figure reproduced from [114] with permission from the Nature Publishing Group.

2.3.4 TargetScan

TargetScan ([115] version 5.0 in **Chapter 5**, and version 6.0 in **Chapter 6**) was used to predict miRNA response elements (MREs) in 3' untranslated regions (3'UTRs, see **Chapter 5** and **Chapter 6**) and in long noncoding RNAs (lncRNAs, **Chapter 6**). TargetScan predicts MREs in the provided input regions (here 3'UTRs and lncRNAs) based on an input file containing the miRNA's name, the seven nucleotide long seed region sequence for this miRNA, and the species identification (here in all cases *Homo sapiens*). I used the default input file for the miRNAs and their seed regions provided by TargetScan [115]. More precisely, TargetScan searches for the presence of conserved 8mer and 7mer sites that match the seed region of a miRNA. 7mer refers to the given seven nucleotide long seed region sequence. 8mer is defined as the exact match to the mature miRNA followed by an 'A', based on the observation that the first position downstream of the seed match was often a highly conserved 'A' [115]. Here, I included all predicted targets and applied normalizations and correction for false positives as described in **Chapter 5** and **Chapter 6** respectively.

2.4 Functional enrichment

2.4.1 GO-terms

The Gene Ontology (GO) project [116] aims to provide consistent descriptions of gene products. It developed three structured vocabularies (ontologies) for gene products' associated biological processes (molecular events with a defined beginning and end), cellular components (parts of the cell or the extracellular environment), and molecular functions (gene products' activities on a molecular level). A gene product can be annotated with zero or multiple ontologies. Generally, the GO project recommends assigning the most detailed level of an ontology that correctly describes the biology of the gene product. In turn, a gene product automatically belongs to all parents of its assigned ontology [116].

The online REViGO tool [117] summarizes redundant GO-terms to facilitate interpretation and visualization of enrichment analysis. Briefly, REViGO uses a clustering algorithm based on semantic similarity measures to reduce an input list of GO-terms to a representative subset of terms. I used default settings as suggested by the authors [117] if not explicitly stated otherwise.

2.4.2 KEGG-pathways

KEGG (Kyoto Encyclopedia of Genes and Genomes [118]) is a database integrating information (genomic, chemical and functional) about the higher-level functions of the cell. More precisely, KEGG consists of 16 main databases categorized into genomic information about genes and proteins, chemical information about chemical substances, and functional data about interaction, reaction and relation networks. The KEGG molecular networks include pathway maps. The pathway maps represent known

molecular interaction and reaction networks for the metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases [118].

2.4.3 MGI-phenotypes

MGI (Mouse Genome Informatics [119]) is a database resource for genetic, genomic, and biological data of the laboratory mouse. It includes data for spontaneous, induced, and genetically-engineered mutations, and their associated phenotypes. These associations between mutations and phenotypes can then be used as a proxy for corresponding associations in humans, and thus potentially serve as disease models. In turn, human genes with an ortholog in mouse can be tested for enrichment of genes linked to specific phenotypes against a suitable background [119].

2.4.4 DAVID

The online DAVID tool (Database for Annotation, Visualization, and Integrated Discovery [120]) provides a data-mining environment for analysing large gene lists. After uploading the gene list of interest, the gene list can be compared against a suitable background of genes, for example testing for the enrichment of GO-terms (described in **Chapter 2.4.1**) in the gene list compared to the background. DAVID can perform different well-known statistical tests, for example Fisher's exact test (FET) that calculates exact p-values using contingency tables [120]. In addition, DAVID offers a selection of tests for p-value correction, including Bonferroni correction (see **Chapter 2.3.1**).

2.4.5 GREAT

GREAT (version 2.0.2, Genomic Regions Enrichment Annotations Tool [121]) tests for the functional enrichment of genes annotated with a specific ontology (for example GO-terms, see **Chapter 2.4.1**) in a gene list of interest compared to a suitable background. The enrichment analysis using GREAT requires an input file including the chromosome number as well as the start and the stop coordinates of the regions of interest (the so called BED file), which can be either coding or non-coding. The statistics provided by GREAT are calculated by associating the provided genomic regions to the nearby genes (described in detail below). This calculation includes two steps: First, a regulatory domain is assigned to each gene. I used the default settings given in the online GREAT tool, which lead to the assignment of a basal regulatory domain. More precisely, each gene is assigned a minimum distance upstream (5.0 kb) and downstream (1.0 kb) of the transcription start site (TSS), regardless of other nearby genes. The distal extension (up to 1000.0 kb) is applied in both directions, but only to a length that does not overlap the next nearest gene's basal regulatory domain. I used the option allowing inclusion of curated regulatory domains, which extends the assigned regulatory domains to known regulatory elements based on the experimental evidence reported in the literature. Next, each genomic region in the input file is associated with all genes whose regulatory domains it overlaps. GREAT supports hypergeometric tests (inferring proximal regulatory domains and calculating $P=Pr_{hyper}(k\geq$number of genes selected by proximal genomic regions with annotation $x \mid N=$number of genes in background, $K=$number of genes in background with annotation x, $n=$number of genes selected by proximal genomic regions)), and binomial tests (inferring distal regulatory domains and calculating $P=Pr_{binom}(k\geq$number of genomic regions with annotation $x \mid n=$number of total genomic regions, $p=$fraction of background with annotation x)) for comparing the

fraction of the genomic regions in the input associated with a given term from an ontology to the expectation from the background [121]. Multiple testing correction can be applied using the False Discovery Rate (FDR; see **Chapter 2.3.1**).

2.5 Network algorithms

2.5.1 WGCNA

Correlation networks can be used to find patterns in gene expression data as well as for detection of gene clusters, or in other words gene modules. Weighted correlation network analysis (WGCNA [122]) assembles modules of highly correlated genes and summarizes the expression of these genes using the module eigengene (ME), the first principal component of the gene expression across all genes within the given module. Subsequently, WGCNA calculates the network statistics such as the network connectivity e.g. for detecting hub genes. By default, WGCNA defines a co-expression network as an undirected network based on the pairwise absolute correlations between genes across time. Optionally, directed networks can be generated which preserve information about the negative pairwise correlations. Throughout this thesis, discussed co-expression networks are undirected if not explicitly stated otherwise. The analysis steps taken using the corresponding R package 'WGCNA' are described in the following, all according to reference [122].

(1) Network construction

The adjacency matrix a_{ij} [n x n] contains all information on the network nodes (row and column names) and the edges (matrix elements ij). In a pairwise correlation network each matrix element is a correlation coefficient between 0 and 1, indicating the connection strength between the nodes i and j. The adjacency matrix contains the absolute correlation values, the so called co-expression similarity s_{ij}.

$$s_{ij} = \left| cor\left(x_i, x_j\right) \right| \tag{1}$$

The edges in the network can be pruned down using different thresholding approaches. Hard thresholding would use one value τ and include only the edges passing this threshold $s_{ij} \geq \tau$. Alternatively, a soft threshold may be defined for the adjacency function which transforms the expression data to the adjacency network. Soft thresholding requires the definition of the thresholding parameters. The WGCNA package allows the use of the scale-free topology criterion to choose a soft thresholding parameter $\beta \geq 1$ for approximate scale free network topology. Briefly, the scale free network topology implies that the distribution of degrees of network nodes (the number of edges directly connected to a network node) follows a power law. (In turn, a network containing only nodes with the same degree would not be scale free).

Using β, the co-expression similarity is raised to a weighted network adjacency.

$$a_{ij} = s_{ij}^{\beta} \tag{2}$$

Equation (2) implies proportionality between the weighted adjacency a_{ij} between the two genes and their similarity on a logarithmic scale.

$$\log\left(a_{ij}\right) = \beta \cdot \log\left(s_{ij}\right) \tag{3}$$

(2) Module detection

The modules are densely interconnected gene sets [122]. WGCNA uses the topological overlap in an unsupervised hierarchical clustering approach for measuring the interconnectedness. The branches in the hierarchical clusters represent the module; the number of modules correspondingly depends on the choice of a branch cutting threshold. WGCNA provides a function for a dynamic tree cut method. The expression profile of a module can be summarized using different network statistics. The module eigengene (ME), in other words the first principal component of the gene expression

within the module, represents a weighted average expression profile. Another network statistic uses the intramodular connectivity to identify the most highly connected intramodular connectivity hub genes as the representatives for a given module, whereby the expression of the intramodular hub genes is highly correlated with the module eigengene [122]. WGCNA assigns each gene to exactly one module. It does not use "forced assignment", as it allows genes that cannot be assigned to a module to belong to a neutral group (termed "grey module"). The module membership (MM) offers a continuous measure to determine whether a node lies near the boundaries of a module or is between multiple modules. Basically, the MM reflects the correlation between a given gene's expression pattern and the ME of a given module.

$$K_{cor,i}^{(q)} := cor\left(x_i, E^{(q)}\right) \tag{4}$$

with x being the expression profile of a gene (or a network node) i, and $E^{(q)}$ the ME of the module q. The larger $\left|K_{cor,i}^{(q)}\right|$, the more similar the expression of the gene i is to the ME.

(3) Linking modules to external traits

Different possibilities exist to assess whether a gene module reflects biological information, for example, testing for the enrichment of functional terms, such as GO-term annotations (see **Chapter 2.4.1**). The representative expression profile of a module can be linked to an external trait T by assessing the correlation between the ME and the given trait T. In the same way, the correlation between each gene and an external trait can be calculated. Here, this measure is referred to as the Gene Significance (GS) which reflects the significance of a given's gene expression profile for the external trait.

$$GS_i = \left|cor\left(x_v, T\right)\right| \tag{5}$$

Correspondingly, a p-value can be assigned to the correlation between the expression profile x_i and the sample trait T. In this way, the GS is defined as a p-value based node significance measure (equation 6) as an alternative to using the absolute correlation (equation 5).

$$GS_i = -\log p_i \tag{6}$$

An extension of WGCNA asks the question whether the modules detected in one network are preserved and reproducible in another transcriptional network [123]. Intuitively, one could simply quantify whether a module from a test network contains the same genes as a module from a reference network. However, this would not provide information about the preserved connectivity between these genes. One possible approach to quantify the preservation is using a summary statistic which agglomerates different preservation statistics, namely $Z_{summary}$ [123].

$Z_{summary}$ (equation 7) summarizes four density preservation statistics and three connectivity based statistics (discussed in detail below; equations 7-20 [123]).

$$Z_{summary} = \frac{Z_{density} + Z_{connecitivity}}{2} \tag{7}$$

The authors [123] derived empirical p-values based on simulations and suggest strong evidence for preservation for $Z_{summary} > 10$, moderate evidence for $2 < Z_{summary} < 10$, compared to weak evidence $Z_{summary} < 2$. The composite statistic in equation (7) includes

$$Z_{density} = median(Z_{meanCor}, Z_{meanAdj}, Z_{propVarExpl}, Z_{meanKME}) \tag{8}$$

and

$$Z_{connectivity} = median(\mathrm{Z}_{cor.kIM}, \mathrm{Z}_{cor.kME}, \mathrm{Z}_{cor.cor}) \quad (9)$$

Here the Z_a statistic for statistic a is defined by equation (10), based on obs_a, the observed value for the statistic a, and the calculation of the mean μ_a and the standard deviation σ_a after randomly permuting the module labels in the test network n_{perm} times.

$$Z_a = \frac{obs_a - \mu_a}{\sigma_a} \quad (10)$$

$Z_{density}$ (equation 8) summarizes the Z statistics for the measures defined in equations 11-16.

$$meanCor^{[\mathrm{test}](\mathrm{q})} = mean\{\mathrm{vectorizeMatrix}(\mathrm{sign}(\mathrm{r}_{ij}^{[\mathrm{ref}](\mathrm{q})})\,\mathrm{r}_{ij}^{[\mathrm{test}](\mathrm{q})})\} \quad (11)$$

with $r_{ij}^{[\mathrm{ref}](\mathrm{q})}$ referring to the matrix elements

$$r_{ij}^{[\mathrm{ref}]} = cor(\mathrm{x}_i^{[\mathrm{ref}]}, \mathrm{x}_j^{[\mathrm{ref}]}) \quad (12)$$

after restricting the correlation matrix to the network nodes in the module q. The matrix elements in the test network are defined accordingly to the matrix elements in the reference network (equation 12).

$$meanAdj^{[\mathrm{test}](\mathrm{q})} = density^{[\mathrm{test}](\mathrm{q})} = mean(\mathrm{vectorizeMatrix}(\mathrm{A}^{[\mathrm{test}](\mathrm{q})})) \quad (13)$$

with the matrix $A=a_{ij}$ being turned into a vector of non-redundant components.

$$propVarExpl^{[\mathrm{test}](\mathrm{q})} = mean_{i \in \mathrm{M}_q}\{(\mathrm{kME}_i^{[\mathrm{test}](\mathrm{q})})^2\} \quad (14)$$

with the module eigennode $E^{(q)}$, in co-expression networks also referred to as the module eigengene (ME), or the first principal component. The module membership (MM) can then be quantified for the network node i in the module q:

$$kME_i^{(q)} = cor(\mathrm{x}_i, \mathrm{E}^{(q)}) \quad (15)$$

The mean sign-aware module membership is defined as:

$$meanKME^{[\text{test}](q)} = mean_{i \in \mathrm{M}_q} \{\text{sign}(\text{kME}_i^{[\text{ref}](q)})\, \text{kME}_i^{[\text{test}](q)}\} \quad (16)$$

$Z_{connectivity}$ (equation 9) summarizes the Z statistics for the measures defined in equations 17-20.

$$cor.kIM^{(q)} = cor(\text{kIM}^{[\text{ref}](q)}, \text{kIM}^{[\text{test}](q)}) \quad (17)$$

based on the intramodular connectivity of the network node i in the module q, defined as:

$$kIM_i^{(q)} = \sum_{j \in \mathrm{M}_q, j \neq i} a_{ij}^{(q)} \quad (18)$$

Equation (17) assumes that if the module q from a reference network is preserved in an independent network, the highly connected hubs in the first network will often be highly connected hubs in the second network as well. Thus, the correlation between the intramodular connectivities (equation 17) should be high in a highly preserved network.

Correspondingly, equation (19) is based on the assumption that variables with high module membership in one network will also show high module membership in the independent test network.

$$cor.kME^{(q)} = cor_{i \in \mathrm{M}_q}(\text{kME}_i^{[\text{ref}](q)}, \text{kME}_i^{[\text{test}](q)}) \quad (19)$$

The measure *cor.cor* (equation 20) aims to measure the connectivity patterns in the module q between the reference and the test network:

$$cor.cor^{(q)} = cor(\text{vectorizeMatrix}(\text{r}^{[\text{ref}](q)}), \text{vectorizeMatrix}(\text{r}^{[\text{test}](q)})) \quad (20)$$

2.5.2 ARACNE

The algorithm ARACNE uses mutual information (MI) to reconstruct transcriptional networks and allows the prediction of transcription factor targets [50, 124, 125]. Reverse engineering yields a scale-free hierarchical gene-gene co-expression network which may be driven by a comparably small number of major hub genes [124]. Importantly, ARACNE takes an unbiased approach which uses the full dynamic range of the data without assumptions of the underlying network topology [125]. ARACNE has been used in two of my projects to predict target genes of transcription factors, as I shall discuss in **Chapter 3** and **Chapter 4**.

The construction of a network from genome-wide expression data is based on the definition of the statistical dependencies between all variables, such as the expression timecourses. ARACNE uses the joint probability distribution (JPD) in the form:

$$P(\{g_i\}) = \frac{1}{Z}\exp[-\sum_{i}^{N}\phi_i(g_i) - \sum_{i,j}^{N}\phi_{ij}(g_i, g_j) - \sum_{i,j,k}^{N}\phi_{ijk}(g_i, g_j, g_k) - ...] \equiv e^{-H(\{g_i\})} \qquad (21)$$

with N being number of genes, Z the normalization factor (or the partition function), ϕ the potentials, and $H(\{g_i\})$ the Hamiltonian, defining the system's statistics [50]. The interaction between two variables – for example two genes - requires that the single potential which depends only on these given variables is nonzero. ARACNE identifies these irreducible dependencies and eliminates the remaining ones. In the simplest scenario, the genes could be independent of each other (equation 22). Here, the marginal probabilities $P(g_i)$ can be derived from the empirical observations and be used for estimating the first-order potentials.

$$H(\{g_i\}) = \sum \phi(g_i) \qquad (22)$$

With increasing sample size, the higher order marginals and their potentials may be integrated to restore the complete form of the JPD. The degree of interactions dictates the necessary sample size, whereby sample set size > 100 is sufficient for 2-way marginals [50]. On the pairwise interaction level, equation 21 incorporates the approximation:

$$H(\{\mathrm{g_i}\}) = \sum_i \phi_i(\mathrm{g_i}) + \sum_{i,j} \phi_{ij}(\mathrm{g}_i, \mathrm{g}_j) \tag{23}$$

With equation 23, all genes with potential $\phi_{ij} = 0$ are non-interacting, including both the genes that are independent, and the genes that share dependency via another gene. ARACNE estimates the pairwise mutual information (MI, equation 24) which becomes zero if and only if $P(\mathrm{g_i}, \mathrm{g_j}) = \mathrm{P(g_i)\,P(g_j)}$. The network is pruned based on eliminating interactions below a certain MI threshold derived from a chosen p-value, based on the null hypothesis of independence between two given genes. In the following, logarithms are natural.

$$I(\mathrm{x, y}) = \mathrm{S(x) + S(y) - S(x, y)} \tag{24}$$

where *S(t)* is the entropy of a discrete variable:

$$S(\mathrm{t}) = -\langle \log p(\mathrm{t}_i) \rangle = -\sum_i p(\mathrm{t}_i) \log p(\mathrm{t_i}) \tag{25}$$

and $p(\mathrm{t_i}) = \mathrm{Prob(t = t_i)}$ is the probability for each state – or value – of the variable. ARACNE estimates the MI using a Gaussian kernel as described in further detail in [50]. The redundant network edges are pruned down using the data processing inequality (DPI). The DPI tests each gene triplet, irrespectively of whether the edges in this triplet have been examined before, and removes the edges with the lowest MI. How many triplets are broken depends on the chosen DPI tolerance, ranging from 0, where

all triplets are broken on the weakest link to 100% which preserves all triplets; recommended values are between 0 -15% [125]. If a list of known transcription factors is provided as a prior, the putative transcription factor-target connections are protected in a way that the edges between a given transcription factor and its target cannot be eliminated in favour of the edges between two non-transcription factors [125].

Practically, a significant threshold can be introduced to report only the interactions which pass this given threshold. It can be inferred either directly as a discrete value for MI, or via a p-value threshold, for example 10^{-7} [50, 126] which is converted into the corresponding MI threshold by ARACNE. In this case, this would refer to a chance of 10^{-7} of reaching an MI score equal or greater than the calculated threshold by chance, given that a gene pair is independent [126]. The robustness of the network can be increased by building a consensus network based on bootstrap samples. Within each bootstrap step, the positions of the edges are randomly permuted while the total number is held constant. The statistical significance of an edge is then assessed against the null distribution of support for this edge across all the random permutations. If this p-value passes the defined threshold, the edge will be maintained in the consensus network [126].

Chapter 3 - Microglia recapitulate a haematopoietic master regulator network in the ageing human brain

3.1 Abstract

Microglia form the immune system of the brain. During organismal aging, they undergo major changes in morphology. Previous studies in cell cultures and animal models suggest altered activation states and cellular senescence in the aged brain. I found that microglia gene markers in the context of the post-mortem frontal cortex of healthy controls assemble into a transcriptional module in a gene co-expression network in three independent microarray data sets. These markers are predominantly for the M1 and M1/M2b activation phenotypes. The microglia module's representative expression pattern negatively correlated with age during the adult lifespan. The observed decrease in expression was more pronounced in microglia surface receptors for microglia/neuron crosstalk than in markers for activation state phenotypes. In addition to these receptors for exogenous signals, microglia are controlled by brain-expressed regulatory factors. I identified a subnetwork of transcription factors, including *RUNX1*, *IRF8*, *PU.1* and *TAL1*, acting as master regulators for the age-dependent microglia module. The causal impact of these master regulators on the microglia module was verified using publicly available ChIP-Seq data. Apart from the gene-expression level, interactions of key master regulators were preserved in a protein-protein interaction network. Importantly, these master regulators play crucial roles for haematopoiesis and myeloid cell-fate decisions during embryogenesis, whereas the present data indicates that they are also essential for regulating microglia homeostasis in the adult human brain.

3.2 Introduction

The human brain's transcriptome is highly plastic over its lifespan, drastically changing at specific developmental stages, such as at the time of birth, and by the end of infancy or childhood [2, 3], or in response to insults or injuries [127-129]. In any cell type alternative rates of transcription allow the adaptation to changing environments [83, 130] and therefore play a crucial role in healthy ageing or pathogenesis [131-136]. Distinct cell types and functionally related genes are expected to orchestrate their transcriptional patterns, which reflect regulatory crosstalk between interacting genes [67, 82, 122]. A previous network analysis of the human brain's transcriptome from fetal to old age detected transcriptional modules of coexpressed genes that are distinguished mostly by strong fold changes in expression before and after birth and during early childhood [3]. However, it is less clear how age affects gene modules during the more subtle gene expression changes in later life [2, 3]. Identification of age-dependent modules during adulthood could improve the understanding of periods that are critical for brain plasticity or susceptibility for late-onset diseases.

Gene expression levels are regulated in part by transcription factors (TFs). These recognize binding motifs in the vicinity of their target genes, and control temporal expression patterns [47, 137]. The complete set of genes regulated by a given transcription factor is defined as its 'regulon'. Transcription factors are highly diverse in their binding patterns, size of their regulons and consequentially their potential downstream influence [47]. Distinct gene modules are regulated by transcription factors that act as master regulators with significant overlap of their regulon and a module's genes [61]. The identification of master regulators in transcription factor cascades is a central issue for understanding the key components underlying transitions between

different physiological states [61, 138, 139]. For example *PU.1, RUNX1* and *TAL1* act as master regulators for haematopoiesis and microglia genesis during embryonic development [140]. Transcriptional modules have been identified in the human brain's transcriptome based on their link to Alzheimer's disease [68, 101]. However, it remains poorly understood which master regulators coordinate these modules' expression during healthy ageing. Importantly, elucidating regulatory patterns for microglia gene expression allows the identification of developmental stages that are critical to disease susceptibility.

Microglia form the innate immune system of the CNS [141, 142]. They play a fundamental role in healthy and pathological ageing [143, 144], undergoing major morphological changes with increasing age [145-147] while maintaining cell numbers [141, 148]. Transition between their phenotypical activation states, M1 and M2, has been linked to various disease pathways [149-151]. M1 reflects the classically activated stage and might contribute to neurotoxicity [150], induced by Interferon gamma (*IFNγ*), compared to a continuum of alternative activated M2 states whose focus is on tissue repair [152]. M2a cells are induced by *IL-4* or *IL-13*, M2b are induced by immune complexes and Toll-like receptor (*TLR*) agonists, and M2c are induced by *IL-10* or transforming growth factor (*TGF-β*) [151]. Animal studies suggest that during ageing M1-phenotypes increase, accompanied by a decrease in M2 [153]. Microglia are key factors for communication between the CNS and the peripheral immune system [141], by crosstalking to T-cells [151] and by modulating the permeability of the blood-brain barrier [154, 155]. During early development, microglia emerge from haematopoietic progenitors in the yolk sac [156]. Injuries may induce a second wave of haematopoiesis during adulthood and following infiltration of CNS by bone marrow-derived myeloid cells [152, 157] microglia from both origins share a dependence on critical transcription

factors, such as *PU.1* [157] which is expressed in the human adult brain [158]. The microglia response to insults is reflected by dynamic transitions between M1 and M2 phenotypes [159]. These transitions are mediated via intrinsic (e.g. *RUNX1*, *IRF8*, and *PU.1*) and extrinsic factors interacting with microglia receptors (e.g. *CD200*, *CX_3CR1*, and *TREM2*) [160].

I sought to identify transcription factor networks that act as master regulators for age-dependent modules within the context of the healthy human post-mortem brain. The intact brain environment, including all cell types, is essential for understanding the microglia regulatory network during ageing as it evolves in close communication with other cells, e.g. neurons [160], and the peripheral immune system [141]. To detect these key regulatory networks in the ageing brain, I applied an unbiased data driven approach which first detected modules whose representative expression pattern is correlated with ageing and then predicted transcription factors acting as master regulators for these modules. Combining two approaches for network analysis, I identified a microglia module, whose expression pattern was negatively correlated with age, especially for those genes that encode microglia surface receptors for neuron/microglia crosstalk. This microglia module in the healthy adult brain was regulated by a set of haematopoietic master regulators, including transcription factors known to act as endogenous factors regulating microglia activation or cell differentiation during embryonic development [160], such as *RUNX1*, *IRF8*, *PU.1* and *TAL1*. This indicates that these master regulators play key roles not only during earliest development, but also in microglia cell homeostasis in the human adult brain.

3.3 Results

3.3.1 An age-dependent gene module in frontal cortex is preserved across independent data sets

I used three independent microarray gene expression data sets from healthy human post-mortem frontal cortex tissue (Figure 8A-B) to generate transcriptional networks across the lifespan using 'Weighted correlation network analysis' (WGCNA [122], described in detail in **Chapter 2.5.1**). Genes whose expression is distinctive of cell types and/or functionally related genes are expected to show higher coordination of expression and thus can be assembled by this analysis into gene modules [122]. Age-dependent modules were detected using an unbiased approach, focusing only on modules whose representative expression pattern, the module's eigengene (ME) [122], was significantly correlated with age (α=0.05) indeed at higher significance levels than potential confounding variables, namely RNA integrity number (RIN, described in **Chapter 1.2**), pH and post mortem interval (PMI). One module only fulfilled this criterion in all three data sets and showed strong enrichment for gene annotations linked to the immune system in the gene ontology (GO) data base (Table 1) for all three data sets used (Figure 8C). The main themes of enrichment, all associated with the immune system, were replicated using MGI-phenotypes (all ontologies used for testing for functional enrichment are described in **Chapter 2.4**).

Table 1. Functional enrichment for GO-terms for genes contributing to the age-dependent module in all three data sets combined (*n*=426 genes), calculated using a hypergeometric test in GREAT (described in Chapter 2.4.5 [121]). P-values were corrected for multiple testing by applying false discovery rate (FDR Q-Val, see Chapter 2.3.1). 'Fold' indicates fold enrichment. Categories highlighted in yellow are non-redundant GO-terms for the given modules, summarized by REViGO [117] (see Chapter 2.4.1).

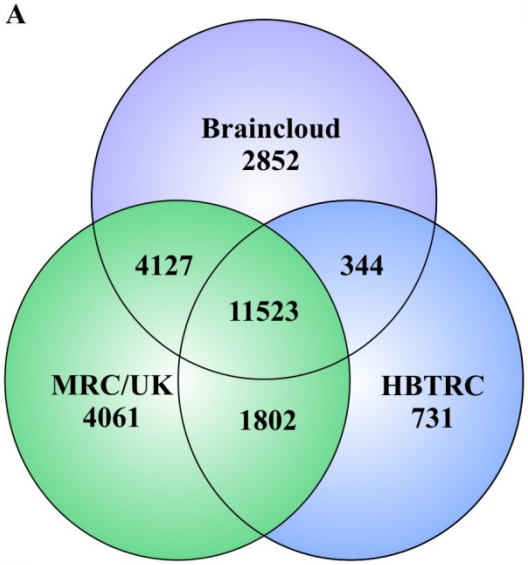
Term Name	Hypergeometric test	
GO Biological Processes	FDR Q-Val	Fold
immune response	4.4E-22	4.9
defense response	9.9E-21	4.5
immune system process	4.8E-19	3.4
regulation of immune response	4.7E-16	5.4
activation of immune response	5.7E-16	7.8
positive regulation of immune response	5.6E-16	6.8
positive regulation of immune system process	3.6E-14	5.0
regulation of immune system process	1.8E-13	3.8
innate immune response	7.8E-13	5.3
response to wounding	3.0E-12	3.3
response to stress	3.6E-12	2.2
inflammatory response	1.2E-09	5.2
immune response-activating signal transduction	3.4E-09	6.7
adaptive immune response	5.7E-09	9.6
immune response-regulating signaling pathway	6.8E-09	6.4
immune effector process	7.8E-09	6.8
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	1.2E-08	36.6
regulation of response to stimulus	1.2E-08	2.1
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	2.0E-08	9.6
positive regulation of response to stimulus	2.5E-08	2.9
immune response-activating cell surface receptor signaling pathway	2.8E-07	7.9
immune response-regulating cell surface receptor signaling pathway	6.6E-07	7.4
immunoglobulin mediated immune response	1.8E-06	11.0
B cell mediated immunity	3.2E-06	10.3
antigen receptor-mediated signaling pathway	9.4E-06	7.3

humoral immune response	1.3E-05	8.9
lymphocyte mediated immunity	2.9E-05	8.3
positive regulation of cell activation	4.6E-05	4.6
leukocyte mediated immunity	5.7E-05	6.8
regulation of cytokine production	6.9E-05	3.8
positive regulation of leukocyte activation	1.5E-04	4.5
cellular response to cytokine stimulus	2.7E-04	4.0
regulation of cell activation	3.0E-04	3.6
complement activation	3.1E-04	12.6
regulation of interleukin-8 production	4.3E-04	11.9
regulation of leukocyte activation	5.7E-04	3.6
positive regulation of cytokine production	5.6E-04	4.9
positive regulation of lymphocyte activation	8.5E-04	4.4
antigen processing and presentation of peptide antigen via MHC class II	9.1E-04	33.6
regulation of lymphocyte activation	1.1E-03	3.8
positive regulation of T cell activation	1.8E-03	4.7
regulation of body fluid levels	2.9E-03	2.6
protein activation cascade	3.3E-03	8.5
response to cytokine stimulus	3.7E-03	3.1
regulation of defense response	3.8E-03	3.1
wound healing	3.7E-03	2.5
positive regulation of multicellular organismal process	3.8E-03	2.8
cellular response to mechanical stimulus	3.7E-03	8.2
positive regulation of defense response	3.7E-03	4.0
cell activation	3.7E-03	2.6
regulation of T cell activation	4.4E-03	3.9
leukocyte activation	4.5E-03	3.3
complement activation, classical pathway	4.5E-03	13.6
microglial cell activation involved in immune response	4.6E-03	43.1
cytokine-mediated signaling pathway	5.3E-03	3.8
T cell receptor signaling pathway	5.6E-03	6.3
complement activation, alternative pathway	6.6E-03	19.2
regulation of interleukin-6 production	6.7E-03	7.3
humoral immune response mediated by circulating immunoglobulin	7.5E-03	12.0
regulation of response to external stimulus	9.6E-03	3.2
cellular response to molecule of bacterial origin	9.9E-03	6.8
T cell costimulation	1.0E-02	6.7
icosanoid biosynthetic process	1.0E-02	8.3
regulation of innate immune response	1.0E-02	3.8
positive regulation of tumor necrosis factor production	1.1E-02	10.9
prostanoid metabolic process	1.2E-02	10.7
blood coagulation	1.2E-02	2.6

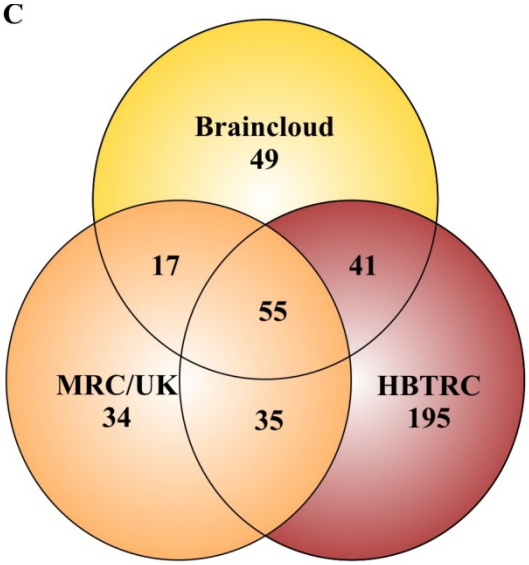
leukotriene production involved in inflammatory response	1.2E-02	30.2
antigen processing and presentation of exogenous peptide antigen via MHC class II	1.2E-02	30.2
negative regulation of granulocyte differentiation	1.2E-02	30.2
regulation of response to stress	1.3E-02	2.3
hemostasis	1.3E-02	2.5
unsaturated fatty acid biosynthetic process	1.4E-02	7.6
positive regulation of innate immune response	1.5E-02	4.2
activation of innate immune response	1.4E-02	4.7
positive regulation of B cell differentiation	1.5E-02	27.5
leukocyte migration	1.5E-02	3.8
B cell receptor signaling pathway	1.8E-02	9.5
positive regulation of interleukin-8 production	2.4E-02	13.0
cellular response to biotic stimulus	2.4E-02	5.6
regulation of actin polymerization or depolymerization	3.1E-02	5.4
microglial cell activation	3.1E-02	21.6
antigen processing and presentation of exogenous peptide antigen	3.1E-02	21.6
positive regulation of interleukin-8 biosynthetic process	3.1E-02	21.6
antigen processing and presentation	3.1E-02	4.7
leukocyte chemotaxis	3.2E-02	6.4
positive regulation of production of molecular mediator of immune response	3.5E-02	11.5
macrophage activation involved in immune response	3.6E-02	20.1
inositol trisphosphate metabolic process	3.6E-02	20.1
regulation of actin filament length	3.7E-02	5.2
cellular response to lipopolysaccharide	4.0E-02	6.1
myeloid cell activation involved in immune response	4.2E-02	10.9
regulation of interferon-beta production	4.2E-02	10.9
regulation of tumor necrosis factor production	4.3E-02	6.0
icosanoid metabolic process	4.5E-02	5.9
leukocyte activation involved in immune response	4.7E-02	5.9
positive regulation of type IIa hypersensitivity	4.8E-02	50.3
detection of triacyl bacterial lipopeptide	4.8E-02	50.3
detection of diacyl bacterial lipopeptide	4.8E-02	50.3
detection of bacterial lipopeptide	4.8E-02	50.3
cellular response to diacyl bacterial lipopeptide	4.8E-02	50.3
cellular response to triacyl bacterial lipopeptide	4.8E-02	50.3
antigen processing and presentation of exogenous antigen	4.6E-02	17.8
GO Molecular Function		
receptor activity	2.2E-04	2.1
transmembrane signaling receptor activity	5.4E-03	2.3
signaling receptor activity	2.0E-02	2.1

GO Cellular Components		
MHC class II protein complex	5.5E-06	33.6
complement component C1 complex	5.4E-04	75.5
MHC protein complex	1.2E-03	12.1
lysosomal membrane	2.9E-02	4.7
lysosome	3.5E-02	3.0
Fc-epsilon receptor I complex	4.4E-02	50.3
Toll-like receptor 1-Toll-like receptor 2 protein complex	4.4E-02	50.3
perinuclear endoplasmic reticulum	4.4E-02	50.3
NADPH oxidase complex	4.4E-02	16.8

A



C



B

Data set	Subjects	Age [years]	Probes
Braincloud	184	12.9-78.2	18846
MRC/UK	97	16-83	21513
HBTRC	100	50-95	14400

Figure 8. Venn diagram showing the overlap of genes in the three microarray gene expression data sets and in the microglia module. A) Overlap of genes from the three microarray gene expression data sets braincloud, Medical Research Council (MRC/UK) and Harvard Brain Tissue Resource Center (HBTRC). B) Demographics and coverage of the three data sets. C) Venn diagram showing the microglia modules and their gene overlaps.

The ME was negatively correlated with age at lower p-values than with confounding variables in each data set (Figure 9 and Figure 10).

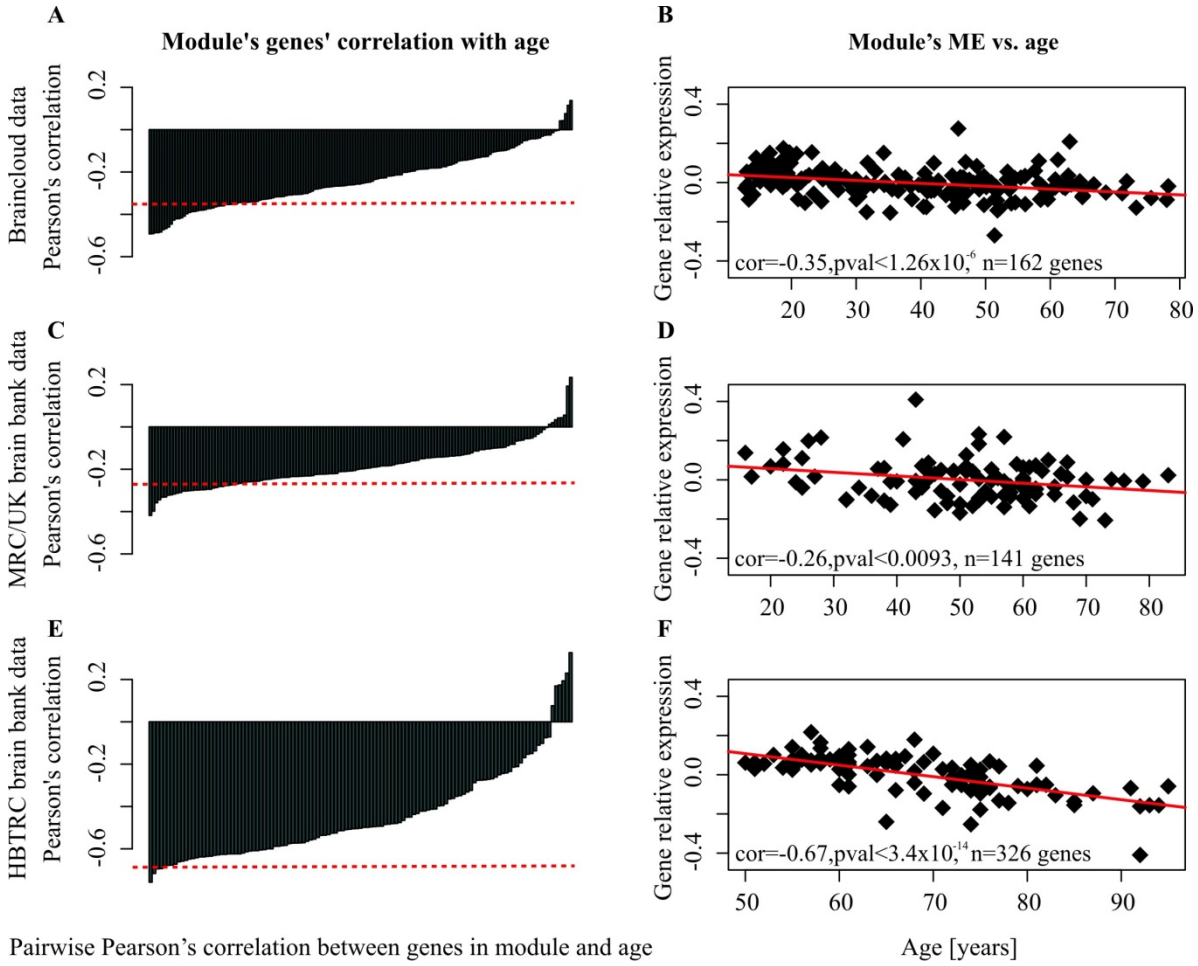


Figure 9. Microglia modules' ME's [gene relative expression: $\log_2$(sample/reference)] negative correlation with age. A) Pearson's correlation of all *n*=162 genes in the microglia module with age, derived from braincloud data ranked by increasing correlations. The red dashed line indicates the Pearson's correlation for the microglia module's ME with age (Pearson's correlation=-0.35). B) showing a negative slope of a linear model fit with the microglia module's ME's expression as the dependent variable and age [years] as the independent variable. C and D display the same metrics for the network derived from MRC/UK data (*n*=141 genes in the microglia module); E and F correspondingly display the same

metrics for the network derived from HBTRC data (*n*=326 genes in the microglia module).

I validated the negative correlation between ME and age after matching samples for the confounding variables. 'Matched' in this context refers to the testing for correlation between the ME and age only for those samples that share the same value for a given confounding variable. Figure 10 illustrates an example of Pearson's correlation coefficients between ME and age (in years) for constant RIN-values in the braincloud data. Clearly, for the majority of RIN values the negative correlation between ME and age (Figure 9B) is replicated (Figure 10). Similar results were found for pH-values, and when Spearman's correlation was calculated instead of Pearson's correlation (not displayed).

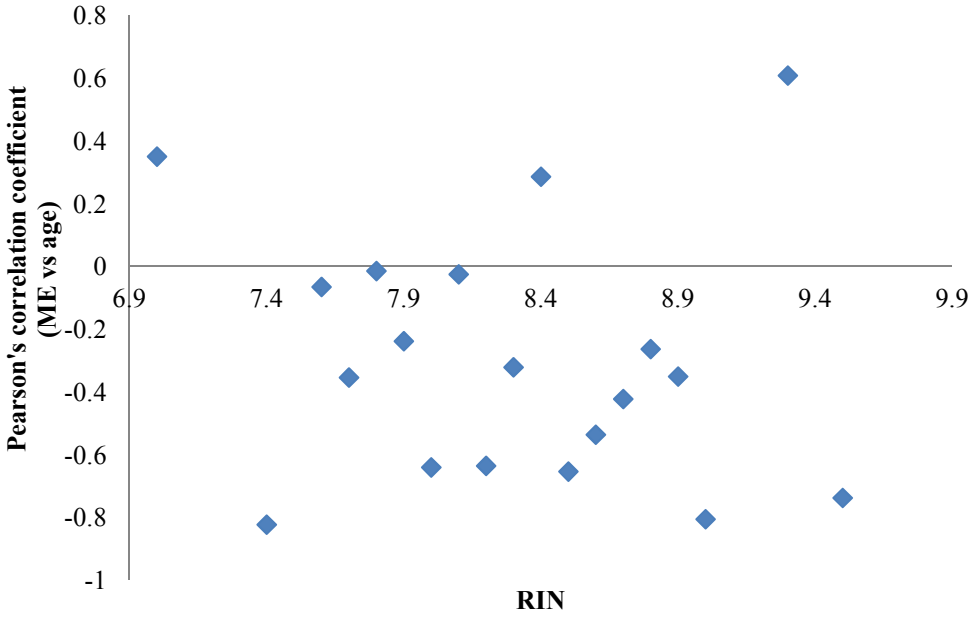


Figure 10. Scatterplot of Pearson's correlation coefficient (y-axis) for the correlation between module eigengene (ME) and age [years] at constant RIN values (x-axis).

3.3.2 Microglia surface receptors' expression levels reflect the module's age-dependency

To determine whether a specific cell type can explain this age-dependent module, through variation of either its population size or its gene expression repertoire, I sought for known gene markers that specify different CNS cell types [101, 142, 161, 162] (Supplementary Table S3.1) among the genes that are present in the module in all three data sets combined (n=426 genes). The module displayed a highly significant 20-fold enrichment for microglial markers compared to random gene sets of identical size (40 of 162 microglia markers, compared with a median of 2 expected by chance, *empirical p-value*<0.001); moreover, it contained only one of the n=29 markers specific for neurons, oligodendrocytes, endothelial cells or astrocytes. The module contained 4 out of 6 gene

markers (*P2RY12*, *GPR34*, *C1QA* and *PROS1*) that have been validated as being expressed in human adult microglia but not in human blood-derived monocytes [162]. This "microglia module" thus principally reflects gene expression variation across the human brain life course in microglia, and not in other CNS cell types, and is not derived from blood contamination.

I next considered whether the module reflected genes that are highly expressed in particular microglial subtypes, more specifically 62 genes that encode known markers for M1 or M2 phenotypes (Supplementary Tables S3.1). The combined microglia module predominantly contained marker genes for M1 or M1/M2b phenotypes. Nevertheless, these markers did not display a negative correlation with aging (Figure 11 shows results for MRC/UK data in detail and Figure 12 shows summarized results for the same metrics for braincloud and HBTRC data). To test whether the markers for activation phenotypes (M1/M2) in the microglia module were representative for all expressed markers, I calculated the mean correlation of all expressed markers for specific microglia phenotypes, regardless of module-assignment (Supplementary Figure S3.1 for MRC/UK data and Supplementary Figure S3.2 for braincloud data). Only the markers for the anti-inflammatory phenotype (M2a) showed a significant decrease of gene expression in both data sets across the lifespan (α=0.05, uncorrected p-values). A decrease in expression of M2a-markers, combined with relatively stable expression of M1 markers is in line with previous studies indicating a M1-skewed microglial activation during ageing [163]. However, in contrast to the M1 markers, the M2a markers were distributed across different modules in both the MRC/UK and the braincloud data. This indicates that expression of M2a markers becomes lower with increasing age, but in a more stochastic way than the expression of the genes assembled in the age-dependent microglia module.

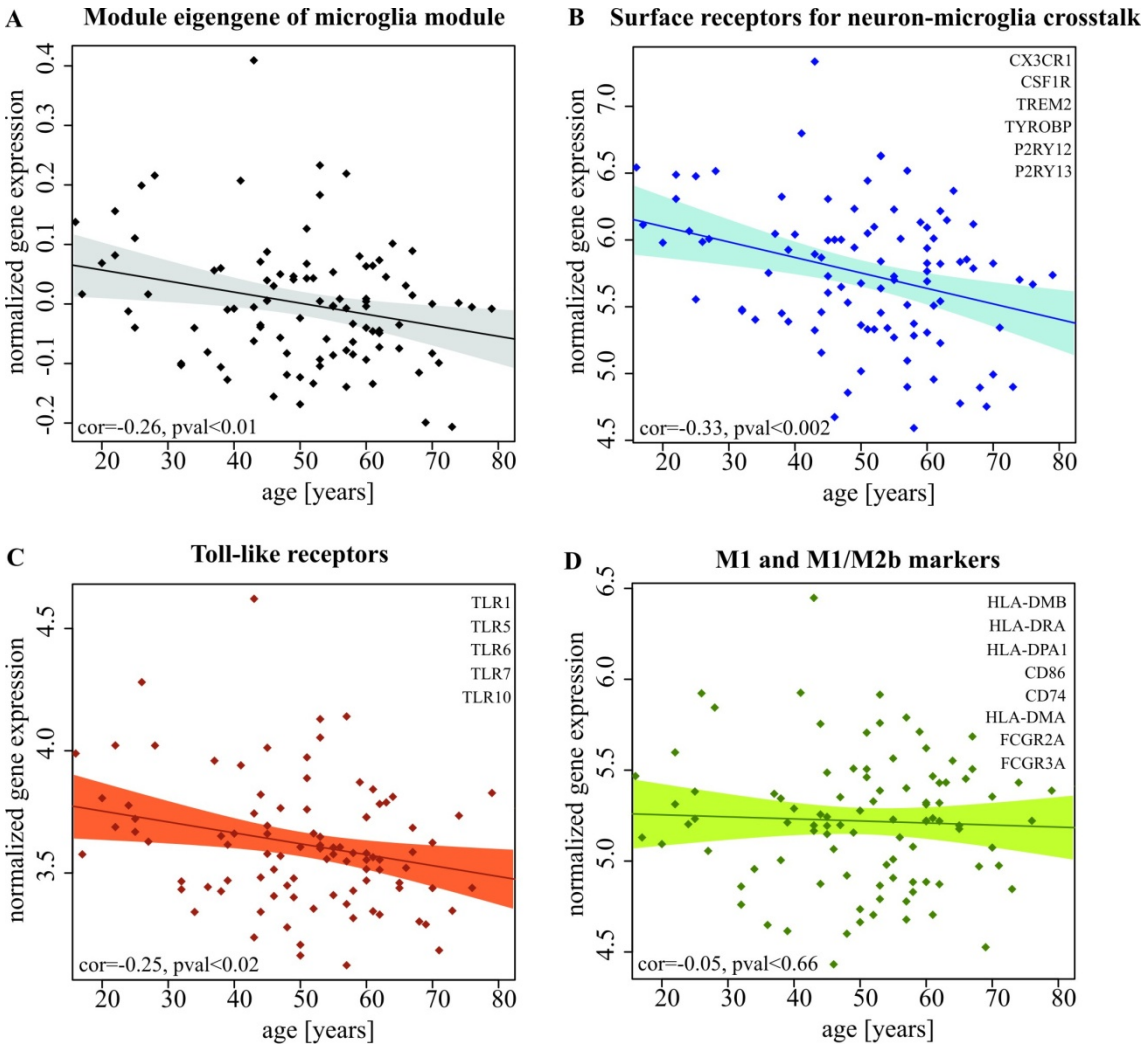


Figure 11. Microglia surface receptors' and microglia M1 and M1/M2b phenotype markers' expression correlation with age for MRC/UK data. The regression model fit and confidence interval for normalized gene expression versus age is for A) microglia module's ME (shown in grey), B) surface receptors for neuron-microglia crosstalk (blue), C) Toll-like receptors (red), and D) M1 and M1/M2b activational phenotype markers (green). The x-axis shows age and the y-axis shows normalized gene expression of the microglia module's ME. Legends in B-D indicate all respective marker genes that are expressed in the MRC/UK data set and therefore were included in the regression. Pearson's correlation coefficient of the mean

expression of the expressed marker genes (B-D) or module ME versus age (A) is shown in the bottom-left corner of each plot.

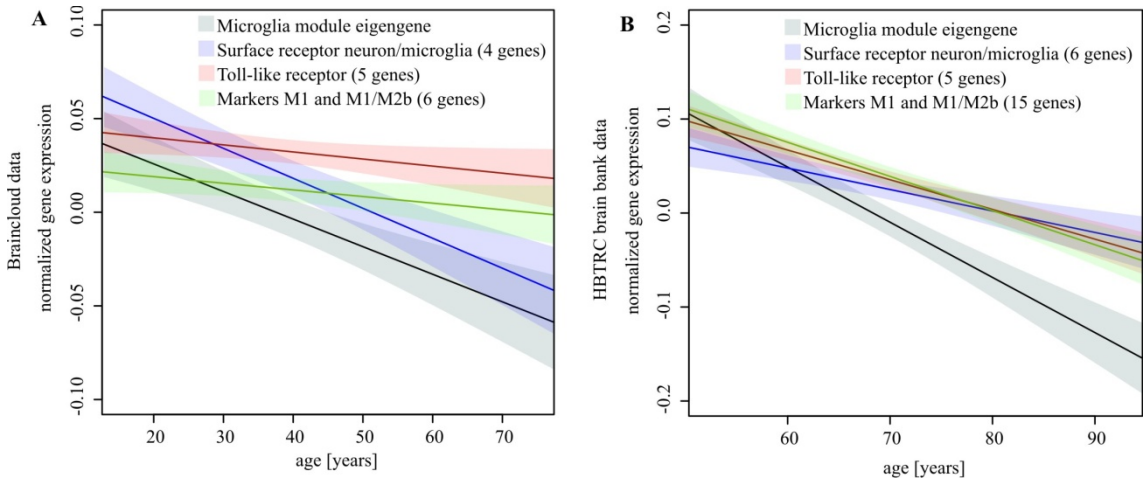


Figure 12. Microglia surface receptors' and microglia M1 and M1/M2b phenotype markers' expression correlation with age for (A) braincloud data and (B) HBTRC brain bank data. The regression model fit and confidence interval for average gene expression versus age are for the microglia module's ME (grey), surface receptors for neuron-microglia crosstalk (blue), Toll-like receptors (red) or M1 and M1/M2b activational phenotype markers (green). The x-axis shows age and the y-axis shows normalized gene expression of the microglia module's ME.

Figure 11 and Figure 12A indicate that microglia surface receptors for microglia-neuron crosstalk and toll-like receptors (TLRs) for communication with pathogen-associated molecular patterns (PAMPs) [160] tend to show a negative correlation between gene expression and age. This is in line with previous findings showing decreased expression of the microglia-specific surface receptor *CX_3CR1* in the aged brain [164, 165]. This

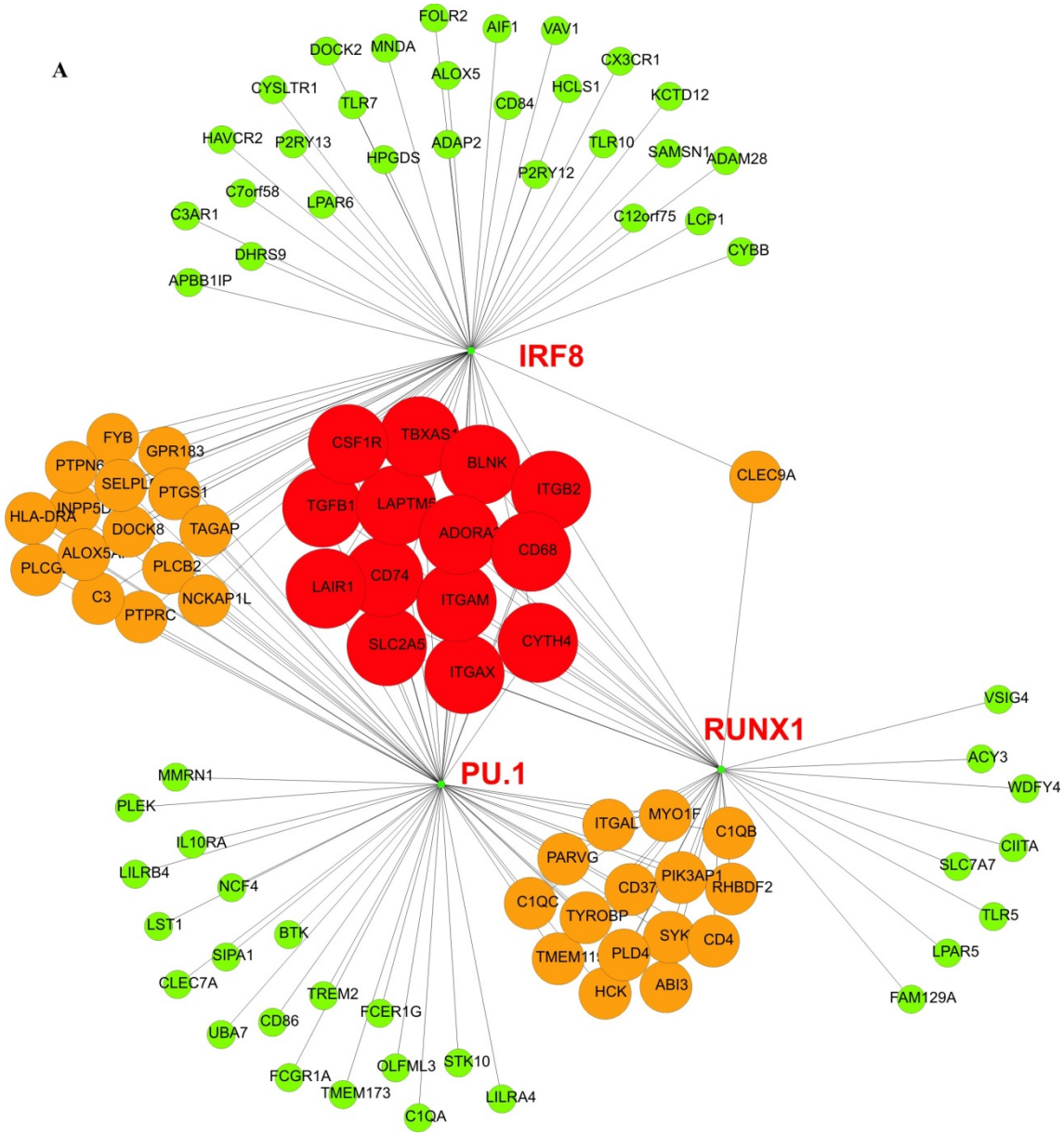
trend was less pronounced however, in the data set of subjects aged 50 years and older (HBTRC, Figure 12B), and was not observed when only considering subjects 50 years and older in either MRC/UK or braincloud data sets (results not shown). This indicates that expression of these receptors decreases over the early adult lifespan but stays on a relatively constant level within the oldest age group. Compared to receptors for neuron-derived signals, the microglia module contained a smaller number of receptors for glioma cell secreted factors [166], namely *IL10RA* in the MRC/UK data set (Pearson's correlation with age=-0.22, *p-value*<0.03) and *IL10RA* (Pearson's correlation with age=-0.60, *p-value*<7.98×10^{-11}) as well as *IL4R* (Pearson's correlation with age=-0.65, *p-value*<1.80×10^{-13}) in the HBTRC data set. These gene expression timecourses indicate that the age-dependency of the microglia module may be driven by expression changes of intrinsic factors that regulate the reception of exogenous signals for microglia activation, especially in microglia-neuron crosstalk [160, 166].

3.3.2 A network of master regulators controls the microglia module

To detect microglia-specific transcription factors in the healthy adult human brain, which I shall refer to as "master regulators" (MRs [61]), transcription factor targets were predicted by reconstructing a transcriptional network using ARACNE [50, 126]. Master regulators for the microglia module were identified based on significant overlap of a transcription factor's predicted regulon with the microglia module (Fisher's exact test (FET), $\alpha<0.05$, corrected for 1655 transcription factors in the ARACNE-derived transcriptional network). Importantly, using this approach, transcription factors can be identified that are not part of the microglia module themselves, for example because they function as upstream regulators. *N*=42, 23 and 64 master regulators were identified in the three data sets, resulting in *n*=107 master regulators in all three sets combined (Supplementary Table S3.2 and Supplementary Table S3.4). Five master regulators –

TAL1, PLEK, RUNX1, IRF8 and *ZFP36L2* – were replicated in all three data sets. Twelve more were identified in at least two data sets (*MAF, FLI1, KLF6, HHEX, LYL1, PARP12, BNC2, PRDM4, IFI16, ELK3, FOS* and *NFATC2*).

The top-ranked master regulators include transcription factors that are known to be expressed in microglia [158, 167] and act in experimentally validated interactions: For example RUNX1 and PU.1 form a negative feedback loop for regulating *PU.1* transcript abundance and subsequently regulate myeloid cell fate [160, 168], and PU.1 and IRF8 function as interaction partners in microglia [167]. RUNX1 also directly interacts with TAL1 and FLI1 [140], whereby PU.1 in turn mediates TAL1 activity [169]. The interdependence among the master regulators is further illustrated by the strong overlap of their regulons (Figure 13). Almost all predicted target genes that are regulated by at least two of the master regulators RUNX1, PU.1 and IRF8 belong to the microglia module (Figure 13). More specifically, all 14 genes that are regulated by all three master regulators, RUNX1, PU.1 and IRF8, belong to the microglia module (Figure 13A). This is in contrast to the majority of genes which are regulated by only one of these three regulators, this gene set tends not to be part of the microglia module (Figure 13B). In addition to the trivial GO-enrichment for regulatory functions of transcription factors, 107 master regulators were also significantly enriched in KEGG-pathways "Acute myeloid leukaemia" (p-value=2.5×10^{-3}, Bonferroni corrected) and "Pathways in cancer" (p-value=6.9×10^{-3}, Bonferroni corrected) terms. These findings likely reflect their crucial roles in myeloid and haematopoietic cell fate decisions [140].


A
IRF8
PU.1
RUNX1
FOLR2
AIF1
VAV1
DOCK2
MNDA
CX3CR1
ALOX5
CYSLTR1
TLR7
CD84
HCLS1
KCTD12
HAVCR2
P2RY13
HPGDS
ADAP2
TLR10
SAMSN1
ADAM28
P2RY12
C7orf58
LPAR6
C3AR1
C12orf75
LCP1
DHRS9
CYBB
APBB1IP
FYB
GPR183
CSF1R
TBXAS1
BLNK
CLEC9A
PTPN6
SELPLG
PTGS1
ITGB2
TGFB1
LAPTM5
HLA-DRA
INPP5D
ADORA2A
DOCK8
TAGAP
CD68
ALOX5AP
PLCG2
PLCB2
CD74
LAIR1
C3
ITGAM
NCKAP1L
PTPRC
SLC2A5
CYTH4
ITGAX
VSIG4
ACY3
WDFY4
MMRN1
PLEK
ITGAL
MYO1F
C1QB
IL10RA
PARVG
CIITA
SLC7A7
LILRB4
CD37
PIK3AP1
RHBDF2
NCF4
C1QC
TLR5
LST1
TYROBP
BTK
SYK
CD4
SIPA1
TMEM119
PLD4
LPAR5
CLEC7A
HCK
ABI3
TREM2
FAM129A
FCER1G
CD86
UBA7
FCGR1A
OLFML3
STK10
TMEM173
LILRA4
C1QA

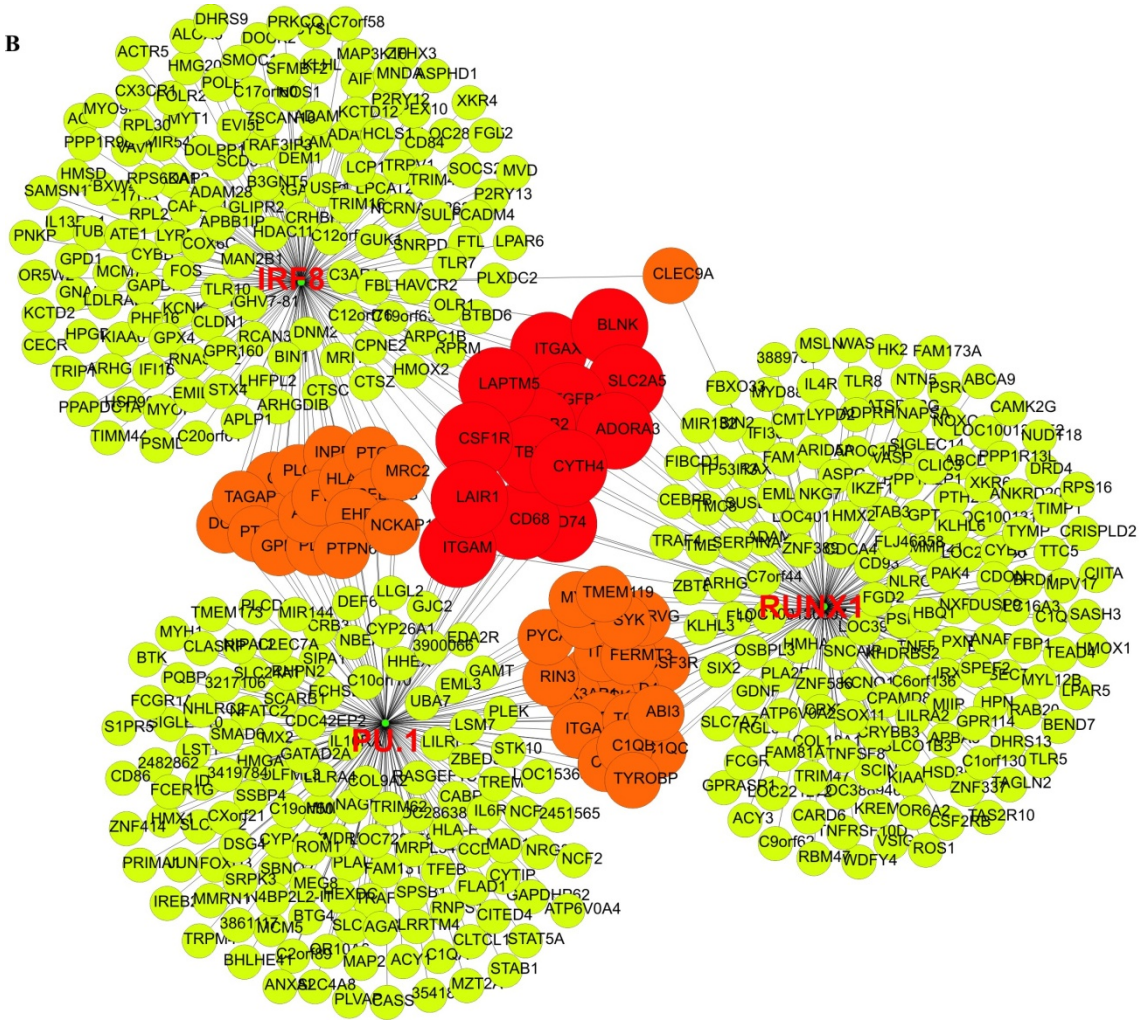

B
IRF8
PU.1
RUNX1
CLEC9A
BLNK
ITGAX
SLC2A5
LAPTM5
CSF1R
ADORA3
CYTH4
LAIR1
CD68
ITGAM
MRC2
TAGAP
NCKAP1
PTPN6
TMEM119
RIN3
ABI3
ITGAL
C1QB
C1QC
TYROBP
FERMT3
PYCARD
GAMT
EML3
UBA7
PLEK
LSM7
STK10
ZBED3
LOC153622
IL6R
NCF2
TFEB
CYTIP
FLAD1
GAPDHP62
ATP6V0A4
CITED4
STAT5A
CLTCL1
STAB1
MAP2
ACY1
C1QA
MZT2A
PLVAP
CASS4
ANXA2
C2orf88
BHLHE41
TRPM2
IREB2
SRPK3
MEG8
SBNO2
PRIMA1
JUNFOS
DSG4
CYP27A1
ROMO1
MRPL54
HLA-E
C19orf21
SSBP4
CXorf21
SLC25A37
ZNF414
HMOX1
CD86
FCER1G
SIGLEC5
FCGR1A
S1PR5
PQBP1
BTK
MYH10
TMEM173
PLCD3
MIR144
DEF6
LLGL2
GJC2
CYP26A1
EDA2R
NBEA
CRB3
SIPA1
RAPGEF2
EC7A
PACSIN2
CLASP1
SLC2A3
C10orf54
ECHS1
SCARB1
FCGR1A
NHLRC2
ATP2C2
CDC42EP3
SMAD6
MX2
IL1R2
LILRA5
GATAD2A
HMGA1
LST1
2482862
3419784
LILRA4
COL9A2
RASGEF1C
TREM1
CABP5
TRIM69
NAGK
2228638
2451565
MAD1L1
NRG1
CCDC72
MRPL54
ROMO1
PLAC8
FAM129A
SPSB1
RNPS1
CITED4
AGAP2
LRRTM4
SLC35A1
MCM5
3861197
BTG4
C2orf88
BHLHE41

Figure 13. Overlap of master regulator's (*RUNX1*, *PU.1* and *IRF8*) predicted target genes in the microglia module (A) compared to genes from whole data set (MRC/UK) (B). A) Predicted regulons (in other word target genes) of *RUNX1, PU.1,* and *IRF8* in the MRC/UK gene expression data for predicted regulons which are part of the microglia module. Size and colour of nodes reflect indegree of nodes. More specifically, small and green nodes have a small number of incoming edges, and, by contrast, large and red nodes have a high number of incoming edges. Yellow nodes are intermediate. Here, the possible range of indegree goes from zero (a master regulator which is not regulated by another master regulator and therefore has no incoming edges) to two (an incoming edge from both other master regulators.) For simplicity, only edges that connect at least one of the three master regulators (*RUNX1*, *PU.1* and *IRF8*) with another gene are displayed, excluding edges between target genes. B) Same measures including all predicted target genes. It is of note that *n*=14 genes that are regulated by all three master regulators are present in both A and B.

Next I asked whether transcriptional interactions, as derived above via network analysis, are also manifested on the protein level. I tested whether the 107 master regulators engage in protein-protein interactions. For this, I used DAPPLE (Disease Association Protein-Protein Link Evaluator) to search for significant physical connectivity, based on protein-protein interactions that have been reported in the literature (Figure 14) [170]. DAPPLE assesses the significance of interactions in the input gene sets based on node-label permutations [170]. I found that 40 of 107 master regulators reached a corrected p-value<0.05 (Supplementary Table S3.3), including three of the five master regulators that were replicated in every data set (TAL1, RUNX1, IRF8) and the top-ranked master

regulator in the MRC/UK data set (PU.1). Protein-protein network statistics were highly significant, showing a higher number of direct connections between master regulators than expected by chance (n=48 direct interactions between 36 proteins in direct network, p-$value$<0.001), higher average direct connectivity of master regulators (or direct binding degree; $mean$=2.7, p-$value$<0.001), as well as indirect connectivity ($mean$=50.7, p-$value$<0.001). The protein-protein interaction network derived by DAPPLE indicates that interactions between the master regulators of the microglia module act not only on the transcriptional but also on the protein level.

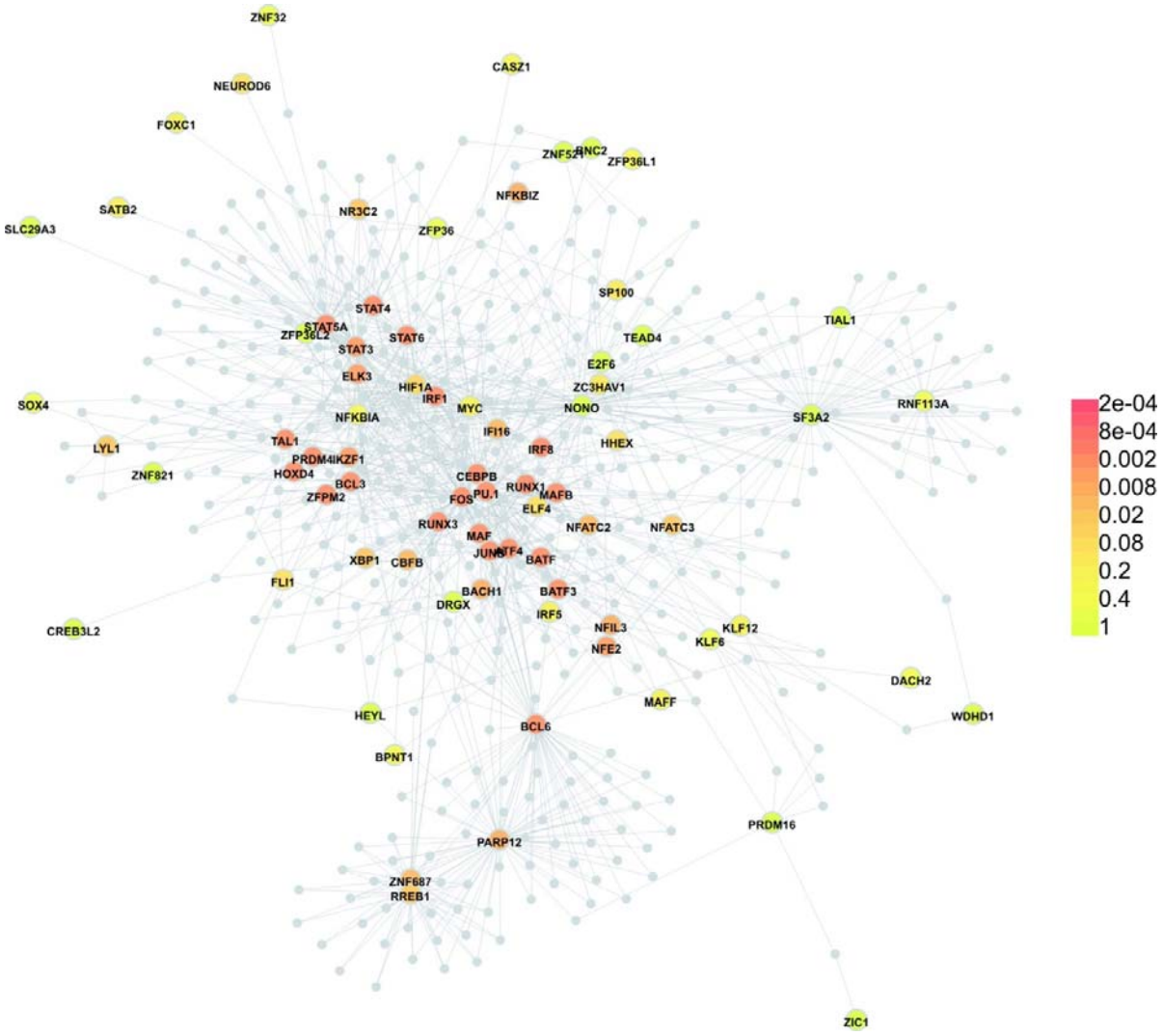


Figure 14. Protein-protein interactions are recapitulated in a master regulator (MR) network. A protein-protein network, derived from direct and indirect interactions among proteins encoded by MRs' genes. Colours are DAPPLE p-values (see legend), calculated from 1000 within-degree node-label permutations. The network was constructed on the basis of physical protein-protein interactions that have been reported in the literature [170].

3.3.3 ChIP-Seq indicates causal role of master regulators for microglia module

The predicted target genes of the master regulators could reflect either indirect (non-causal) coordination between genes in the transcriptional network, or direct (causal) interactions of transcription factors and their target genes. To distinguish these possibilities, I verified that the master regulators' targets in the microglia module predicted by ARACNE were causal using publicly available ChIP-Seq data for RUNX1, IRF8 and TAL1, transcription factors that were replicated in all three data sets, and PU.1, which was top-ranked in the MRC/UK set, but not replicated (Figure 15). For each of the four master regulators, I tested for significant overlap between their binding sites as indicated by their ChIP-Seq peaks, and the basal regulatory domains (5kb upstream and 1kb downstream to each gene's transcription start site (TSS) [121]) of their predicted target genes in the combined microglia module (n=406 genes available in ENSEMBL bioMart). The overlap between ChIP-Seq peaks and combined microglia module was compared to a background of 19045 basal regulatory domains of all genes available in ENSEMBL bioMart and expressed in at least one of the three microarray gene expression data sets. As a negative control, I tested for overlap between ChIP-Seq peaks of a randomly chosen TF which was not identified as a master regulator (NANOG) and basal regulatory domains of all expressed genes from the combined microglia module.

PU.1 is known to bind near the promoter regions of its target genes [171]. ChIP-Seq peaks for PU.1 binding annotated within TSS-promoter regions in three available cell lines from the ENCODE data (GM12878, GM12891 and K562) [137] showed significant overlap with the genes in the combined microglia module (Figure 15A; all significant *q-values*<0.002) in each of the three cell lines. Enrichment of PU.1 binding sites in the combined microglia module validates PU.1 as correctly identified master

regulator although it was only detected in one data set. Correspondingly, ChIP-Seq peaks for RUNX1, IRF8 and TAL1 indicated significant overlap of these master regulators' binding sites with the basal regulatory domains of the genes in the combined microglia module (Figure 15). In contrast, ChIP-Seq peaks for a randomly chosen TF (NANOG), which was not identified as a master regulator, did not significantly overlap with the basal regulatory domains of the genes in the combined microglia module (*p-value*>0.05).

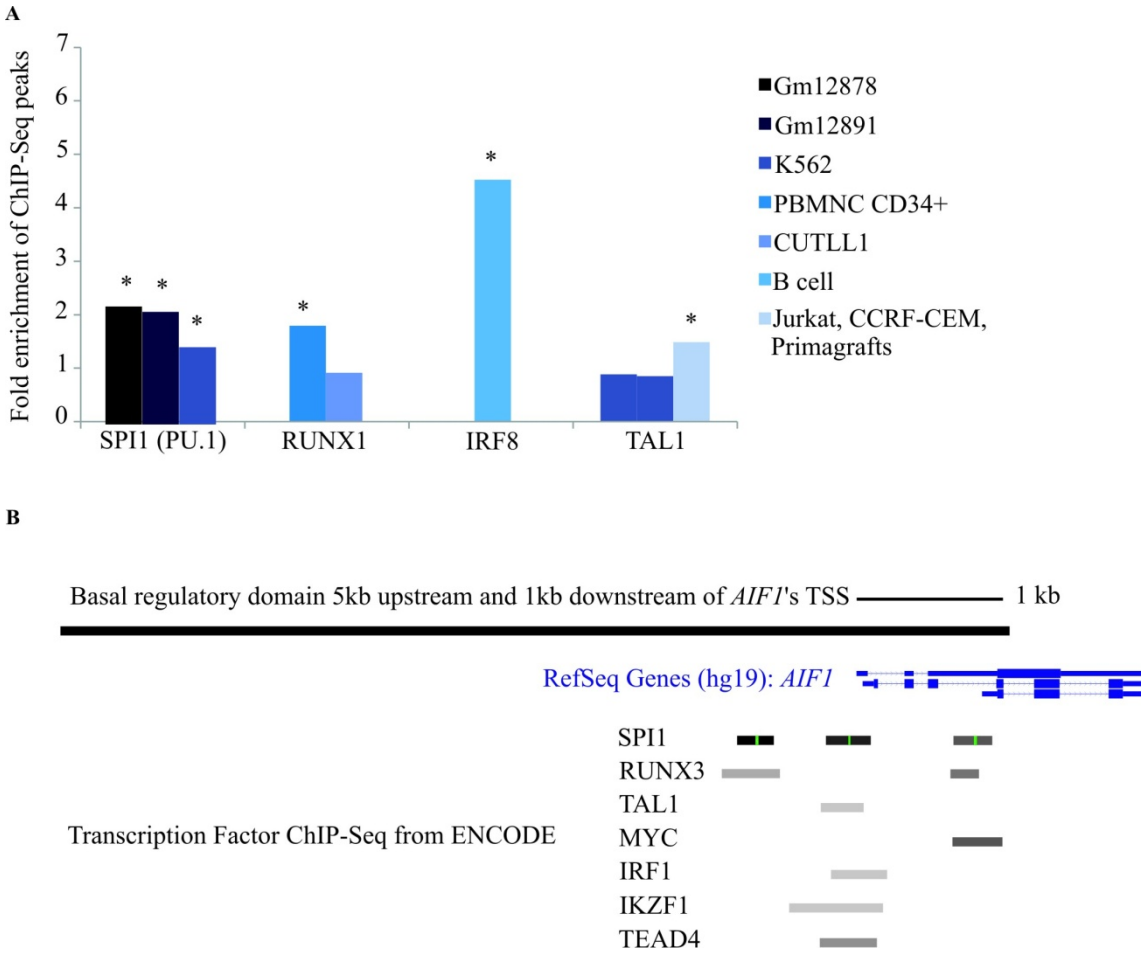


Figure 15. ChIP-Seq peaks for SPI1 (PU.1), RUNX1, IRF8 and TAL1 in genes in the combined microglia module. A) Fold enrichment for SPI1 (PU.1) ChIP-Seq peaks in all annotated basal regulatory domains of all genes in the combined microglia module (*n*=406 assigned to the microglia module in all three data sets combined). Stars demarcate significant (*p-value*<0.05) enrichment. Legend indicates the cell lines that were used for each master regulator. B) Illustration of master regulator bindings sites overlapping the basal regulatory domain of *AIF1* taken from UCSC genome browser.

The significant overlap of TF binding sites as detected by ChIP-Seq and the combined microglia module indicates a causal role for the identified haematopoietic master regulators in the expression of their target genes. As assembly of the microglia module as well as prediction of microglia-specific master regulators was based on gene expression changes across the human adult lifespan, the present data show that the identified master regulators play a key role in microglia homeostasis not only during embryogenesis but also in the ageing brain.

3.4 Discussion

The regulation of microglia homeostasis in the healthy adult human brain is poorly understood [160]. Previous studies have elucidated how microglia emerge from the yolk sac during embryogenesis, and are dependent on the orchestrated functions of transcription factors, including RUNX1, PU.1 and IRF8 [156, 167]. In response to insult or injury, microglia maintain the ability to self-renew [172] or monocytes may be recruited to differentiate into microglia from the bone marrow [152, 157]. Studies that have mostly been performed in cell lines or animal models provided evidence that key transcription factors are also necessary for regulating the transition between microglia activational phenotypes, namely for the transition between pro-inflammatory (M1) and anti-inflammatory (M2) states [160, 168, 173, 174]. However, it remains unclear what role these key regulators play in the homeostasis of microglia in the healthy adult human brain [160].

Here I show that a gene co-expression module, which is highly enriched for microglia markers – predominantly M1 and M2b phenotypes, is controlled by a network of master regulators that also act as key players during haematopoiesis and myeloid cell lineage determination. The microglia module included genes encoding receptors for exogenous signals for microglia activation: *CX3CR1* transmits "resting" signals, *TREM2* and *TYROBP* initiates phagocytosis, purinergic receptors (*P2RY12, P2RY13, P2RY6, P2RY5*) signal neuronal injury and *CSF1R* induces cell survival or proliferation [160]. More specifically, genes encoding microglia surface receptors for microglia-neuron crosstalk or toll-like receptors showed negative correlation of expression with increasing age. The predicted master regulators were not all members of the microglia module, and included subsets of endogenous transcription factors (e.g. *RUNX1, PU.1*

and *IRF8*) that are known to regulate microglia activation [160], and are required for differentiation during embryogenic development [140, 167]. These transcription factors act in close interactions with highly overlapping binding sites. For example, an IRF8 and PU.1 heterodimer exhibits stronger binding activity than IRF8 alone [175, 176]. Importantly, my analysis demonstrates that interacting haematopoietic master regulators from early development are recapitulated from adult human brain tissue. This indicates that a number of the key factors acting during early microglia development – including *RUNX1, PU.1* and *IRF8* – also play a crucial role for homeostasis of adult microglia.

Previous work has provided evidence for a cohesive immune-microglia module that is present in the human brain. These studies compared different neuroanatomical subdivisions [67, 68, 73, 101], or searched for candidate genes for susceptibility to neurodegenerative disease [68, 101]. For example, *TREM2* is a predisposing factor for late-onset Alzheimer's disease (LOAD) [101] and was also present in the microglia module in all three data sets in the current study. However, although *TREM2* showed highly significant module membership, in other words correlation with the module eigengene, it was not ranked in the top 10% of microglia module genes with highest module membership across the three data sets and would not have been selected as a central hub gene in an objective, data driven approach. It has to be noted however that I focused on the frontal cortex, whereas in previous work *TREM2* was identified as a major hub in different brain regions [101]. Furthermore, previous studies investigating various areas in human post-mortem brains from 55 subjects aged 20-99 years (hippocampus, entorhinal cortex, superior-frontal gyrus, and postcentral gyrus) reported up-regulation of genes related to inflammation and immune components [69, 177]. These results could indicate regional specificity for immune-related transcriptional regulation in the adult human brain. Other studies did focus on comparing brain areas,

but post-mortem brains from healthy controls and disease cases. In one of these studies, *TYROBP* was identified as a key regulator based on differential expression between control compared to late onset Alzheimer's disease (LOAD) post-mortem brains and regulatory strength, highlighting its potential role in disease status [68].

The microglia module was identified in three independent data sets based on the negative correlation of gene expression with ageing.

Changes in transcript abundance can be caused by changes in transcription rate in the cell, or by changes in relative or absolute cell numbers of a specific cell type [79]. Previous lines of work have shown that – although the relative number of glia cells compared to neurons increases with age [178] – the cell number of microglia stays stable during the adult lifetime [141, 148]. Furthermore, as the microglia module predominantly contained markers for the proinflammatory (M1) phenotype, a decrease in expression during ageing cannot be explained by a shift towards proinflammatory phenotype during ageing [144, 153].

An alternative explanation for the observed decrease in expression levels of the microglia module is lower transcription rate for genes regulating microglia activation. It has been demonstrated that expression of receptor *CX3CR1* is diminished in the aged brain [164, 165]. Along these lines morphological changes in microglia during older age [145-147] potentially reflect dystrophy and cell senescence rather than increased activation, which could explain a downregulation of endogenous receptor levels including for *CX3CR1* [164].

The microglia module in adult human brains recapitulates master regulators for microglia development during embryogenesis [160]. This indicates that the identified master regulators maintain a causal role for microglia homeostasis during ageing. One

possible role the master regulators could fulfil for microglia in the adult brains is regulation of replicative senescence. For example, given *PU.1*'s role in preventing excessive haematopoietic stem cell division [179], this known master regulator for early microglia development potentially also mediates microglia replicative senescence in normal ageing. Accordingly *RUNX1* and *TAL1* are directly linked to cell growth rates [180, 181]. In line with this basic notion, both microglia senescence during ageing [182] and expression levels of *PU.1* [183], *RUNX1* [184] and *TAL1* [185] are affected by stress [183]. Therefore, while *PU.1* sustains its levels via autoregulation [179], it could prevent excessive proliferation of M1 microglia in response to accumulating oxidative stress during ageing in coordination with other master regulators.

3.5 Methods

3.5.1 Data sets and expression quantification

Data sets, statistical tests and online tools that, in addition, are used in other research chapters are described in detail in **Chapter 2**.

3.5.2 Microarray gene expression

Microarray gene expression data from healthy human controls from three independent data sets were used to generate transcriptional networks. In all cases, I averaged multiple probes available for the same annotated gene.

- Braincloud [2]. All microarray expression data from the prefrontal cortex were quantified per gene ($\log_2$(sample/reference)) as relative gene expression and processed as described previously [2]. I included *n*=184 subjects, aged 12.9-78.2. If multiple probes were available for one gene, I calculated the average expression for this gene, resulting in a total of *n*=18846 available genes (GSE30272). The representative expression patterns of the transcriptional modules were tested for correlation with the available confounding variables RIN, pH, and post-mortem interval (PMI). Braincloud data are described in detail in **Chapter 2.1**.
- MRC Sudden Death Brain and Tissue Bank [186]. If multiple probes were available for one gene, I calculated the average expression for this gene, resulting in a total of *n*=21513 probes from frontal cortex of *n*=97 subjects, aged between 16 to 83 years (available from GSE46706) [187]. Background correction and array pre-processing were performed using RMA quantile normalization. Quantile normalization is a part of the expression summary

measure robust multi-array average (RMA) and aims to normalize a batch of arrays for allowing meaningful comparisons. More precisely, quantile normalization is applied to make the distribution of probe intensities the same for all arrays. It maps probe levels from all *n* arrays in a way that an *n*-dimensional quantile plot follows a *n*-dimensional identity line (y=x with slope=1, see **Chapter 3.5.2**) [188]. The representative expression patterns of the transcriptional modules were tested for correlation with the available confounding variable PMI.

- Harvard brain tissue resource center (HBTRC) [68]. Gene expression data from frontal cortex of *n*=100 healthy subjects between 50 to 95 years were included (GSE44770). *N*=14400 probes with annotated gene names available in ENSEMBL genes were used out of *n*=30305 available probes. Expression was quantified as $\log_2$(sample/reference) as relative gene expression.

3.5.3 ChIP-Seq data

I used publicly available ChIP-Seq data to validate transcription factor targets (details of ChIP-Seq data are described in **Chapter 2.2**). If available, I annotated peaks from cell lines provided by ENCODE [137]; alternatively I searched for ChIP-Seq data for master regulators described in the literature. The publicly available data provided called ChIP-Seq peaks (see **Chapter 2.2**). All peaks were annotated using HOMER (v4.3 [189], described in **Chapter 2.3.2**).

- For PU.1, I used broad ChIP-Seq peaks from three cell lines in the ENCODE data: GM12878 (B-lymphocyte, lymphoblastoid; 42935 peaks), GM12891 (B-lymphocyte, lymphoblastoid; 48830 peaks), and K562 (leukaemia; 28677 peaks).

- RUNX1 peaks were annotated from peripheral blood mononuclear cells CD34+ and transferred from build hg18 to hg19 using the online liftOver tool (PBMNC CD34+; sample GSM722708 [190]; 14536 peaks), and from cell line CUTLL1 (T lymphoblastic leukemia/lymphoma cell line; sample GSM1252931; 28156 peaks).
- IRF8 ChiP-Seq peaks were retrieved from human germinal center B cells (LY1 cell lines). I included only peaks in IRF8-target genes which were replicated across all cell lines used in the original study (183 peaks) [175]. Peaks were transferred from built hg18 to hg19 using online liftOver tool (part of the UCSC Genome Browser).
- ChiP-Seq Peaks from ENCODE data from two replicates K562 were also used for TAL1 (26255 and 54782 narrow peaks). In addition, I used genomic regions occupied by TAL1 in Jurkat, CCRF-CEM, and Primagrafts (Prima 2 and Prima 5) [191] (19844 occupied regions). The coordinates of the occupied genomic regions were transferred from genome build hg18 to hg19 using the online liftOver tool.
- For NANOG, I used narrow peaks (n=5473) from cell line H1-hESC (human embryonic stem cells) available in the ENCODE data.

3.5.4 Transcription Factors

N=2001 gene names for transcription factors from the literature [192, 193] were used as input for ARACNE (described in detail in **Chapter 2.5.2**).

3.5.5 WGCNA

Co-expression networks were built using step-by-step network construction and module detection as implemented in WGCNA (Weighted correlation network analysis [122],

described in detail in **Chapter 2.5.1**). Briefly, in the first step a pairwise Pearson's correlation matrix of all genes across all timepoints was calculated. This correlation matrix was then converted into an adjacency matrix. The soft thresholding power was chosen to reach a scale-free topology fit index indicating approximate scale-free topology, as suggested by the authors [122]. Next, the adjacency matrix was transformed to a Topological Overlap Matrix (TOM), thereby reducing spurious associations, and a dissimilarity matrix was calculated, which was then used to produce a hierarchical tree. The minimum cluster size was arbitrarily set to ten genes. The standard method for branch cutting, namely Dynamic Tree Cut, was applied [122]. This function returns gene modules, each of which is assigned a colour. To reduce redundant information, modules with highly similar expression profiles were merged. This merging procedure was applied using a height cut of 0.25 (corresponding to a correlation of 0.75). Merging was based on the comparison of the module eigengenes (MEs) which provides a summary expression metric. These MEs were also correlated with any external quantitative trait to analyse the dependency between the given trait and the module's expression. Here, I used age and three potential confounding variables (RIN, pH and PMI). More specifically, I selected modules whose ME was correlated with age at more significant p-values than for available confounding variables. For each available confounding variable, I further tested for Pearson's correlation between age and ME after matching the samples for the available confounding variables. The matching procedure consisted in first selecting samples with the same value for a confounding variable, and subsequently testing for correlation of the ME using these samples only. The transcriptional networks were undirected, or in other words did not contain information on edge orientation. Similar results were found for directed networks.

3.5.6 ARACNE

ARACNE uses mutual information (MI) to reconstruct transcriptional networks and allows the prediction of transcription factor targets [50, 124, 125] (see **Chapter 2.5.2**). Reverse engineering yields a scale-free hierarchical gene-gene co-expression network [124]. Briefly, 'scale-free' refers to a network with a small number of major hub genes (see **Chapter 2.5.1**). Importantly, ARACNE takes an unbiased approach which uses the full dynamic range of the data without assuming the underlying network topology [125]. The construction of a network from genome-wide expression data is based on the definition of statistical dependencies between all variables, such as expression timecourses. ARACNE estimates MI using a Gaussian kernel as described in further detail in [50]. Redundant network edges are pruned using data processing inequality (DPI). DPI tests each gene triplet, irrespectively of whether edges in this triplet have been examined before, and removes edges with the lowest MI. How many triplets are broken depends on the chosen DPI tolerance, ranging from 0, where all triplets are broken on the weakest link, to 1 which preserves all triplets; recommended values are between 0 and 0.15 [125]; here I applied 0.1. If a list of known transcription factors is provided as a prior, putative transcription factor-target connections are protected in a way that edges between a given transcription factor and its target cannot be eliminated in favour of edges between two non-transcription factors [125]. I provided a list of n=2001 transcription factors previously described in [192, 193]. I used a p-value of 10^{-7}, which is converted to an MI threshold internally by ARACNE, to report only MI values above the 10^{-7} level of significance; referring to a probability of 10^{-7} to obtain an MI score ≥ calculated threshold by chance if a gene is independent [126]. As suggested by the authors [126], I generated a consensus bootstrapping network using n=100 sampling steps.

3.5.7 Master Regulator Analysis

To identify transcription factors (TFs) acting as master regulators (MRs), I predicted the regulon (in other words the target genes) of each TF using ARACNE. Next, I used Fisher's Exact Test (FET) to calculate whether the intersection of the microglia module with each TFs' regulon was larger than expected by chance (α=0.05, corrected in each data set for the number of TFs that were included in the transcriptional network constructed by ARACNE). More precisely, the contingency table for the FET contained genes in the module (yes/no) versus genes in the TF's regulon (yes/no). TFs whose regulon shared a higher number of genes with the microglia module than expected by chance were termed 'master regulators' for the microglia module.

3.5.8 Enrichment of functional categories and cell types

Enrichment was assessed using GO-terms, MGI-phenotypes and KEGG-pathways comparing the genes in the microglia module against a background of all genes in the WGCNA network using the functional classification tool Database for Annotation, Visualization, and Integrated Discovery (DAVID [120]) and Genomic Regions Enrichment of Annotations Tool (GREAT version 2.0.2 [121]). All ontologies and online tools used are described in detail in **Chapter 2.4**. Briefly, GREAT assigns "basal regulatory domains" as defined by [121] to the genomic regions of interest, spanning 5 kb upstream to 1 kb downstream of the transcription start site of the nearest gene relative to the regions provided in the input, plus a distal extension of maximal 1000 kb or up to the next nearest gene's basal regulatory domain in both directions. I included GO-terms and KEGG-pathways from analysis using GREAT with a significant q-value (α=0.05, hypergeometric p-value, corrected for multiple testing using false discovery rate (FDR)) and KEGG-pathways in DAVID with a significant p-value (α=0.05,

Bonferroni corrected). The underlying cell population of the age-dependent gene co-expression module was tested by comparing the number of expressed microglial markers in the module combined for all three data sets to the number of expressed microglial markers in *n*=1000 random gene sets of the same sample set size using the genes available in all three microarrays combined as a background. For testing enrichment of the named ontologies for master regulators (*n*=107, united for all three data sets), I used all genes with annotated gene names and genomic regions in all three data sets as a background (*n*=18661 genomic regions).

3.5.9 DAPPLE

DAPPLE (Disease Association Protein-Protein Link Evaluator) takes as input a list of genes (here 107 master regulators) and searches for physical connectivity between proteins based on known protein-protein interactions from the literature. DAPPLE assesses p-values for direct and indirect interactions using *n*=1000 within-degree node-label permutations [170]. "Degree" of a node refers to the number of edges directly connected to the node. An empirical p-value is calculated for each node as the ratio of number of random permutations which yield a degree for the node which is as high as or higher than the observed node degree, divided by the total number of permutations [170].

3.5.10 Intersection of genomic intervals

GAT (Genomic Association Tester) [106] is described in detail in **Chapter 2.3.1**.

3.6 Appendix

Table S3.1. Gene names for marker genes indicating a given cell type or microglia phenotype [101, 142, 161, 162].

Gene	cell type/phenotype
FCGR2A	M1
FCGR2B	M1
FCGR3B	M1
CXCL10	M1
CCL15	M1
CCL19	M1
CCL8	M1
FCGR3A	M1
CXCL11	M1
CXCL9	M1
IL15	M1
IL12A	M1
IL12B	M1
IL1A	M1
IL23A	M1
NOS2	M1
CD74	M1 and M2b
CD86	M1 and M2b
HLA-DMA	M1 and M2b
HLA-DMB	M1 and M2b
HLA-DPA1	M1 and M2b
HLA-DPB1	M1 and M2b
HLA-DQA2	M1 and M2b
HLA-DRA	M1 and M2b
HLA-DOB	M1 and M2b
HLA-DQB1	M1 and M2b
IL6	M1 and M2b
HLA-DQB2	M1 and M2b
HLA-DRB1	M1 and M2b
HLA-DRB3	M1 and M2b
HLA-DRB4	M1 and M2b
HLA-DRB5	M1 and M2b
TNF	M1 and M2b
CXCL13	M1 and M2c
CCL20	M1 or M2b
IGF1	M2a
PDGFB	M2a
PDGFC	M2a

TGFB1	M2a
CCL13	M2a
CCL14	M2a
CCL17	M2a
CCL22	M2a
CCL23	M2a
CCL24	M2a
CCL26	M2a
FN1	M2a
IL1RN	M2a
PDGFA	M2a microglia_oligodendrocyte
RETNLB	M2a
IL10	M2a and M2b
MSR1	M2a and M2c
MRC1	M2a and M2c
ARG1	M2a or M2c
CCL18	M2a or M2c
CD163	M2a, M2b and M2c
CCL1	M2b
CXCL2	M2b
CXCL3	M2b
SPHK1	M2b
IL1B	M2b
CCL16	M2c
CD68	microglia
PTPRC	microglia
P2RY12	microglia unique
TMEM119	microglia unique
OLFML3	microglia unique
P2RY13	microglia unique
CX3CR1	microglia unique
GPR34	microglia unique
HEXB	microglia unique
RHOB	microglia unique
JUN	microglia unique
RAB3IL1	microglia unique
SERPINE2	microglia unique
CCL2	microglia unique
FCRLS	microglia unique
SCOC	microglia unique
SIGLECH	microglia unique
SLC2A5	microglia unique
LRRC3	microglia unique
PLXDC2	microglia unique
USP2	microglia unique
CTSF	microglia unique

CTTNBP2NL	microglia unique
TGFBR1	microglia unique
ATP8A2	microglia unique
LGMN	microglia unique
SLCO2B1	microglia unique
MAFB	microglia unique
EGR1	microglia unique
BHLHE41	microglia unique
FCGR1A	microglia unique
HPGDS	microglia unique
CTSD	microglia unique
HSPA1A	microglia unique
LAG3	microglia unique
CSF1R	microglia unique
ADAMTS1	microglia unique
F11R	microglia unique
GOLM1	microglia unique
NUAK1	microglia unique
CRYBB1	microglia unique
LTC4S	microglia unique
ASPH	microglia_neuron
BASP1	microglia_neuron
CD34	microglia_neuron
TANC2	microglia_neuron
SLC7A8	microglia_neuron
SYNGR1	microglia_neuron
RGMB	microglia_neuron
NPNT	microglia_neuron
MEF2A	microglia_neuron
RTN4RL1	microglia_neuron
MEF2C	microglia_neuron
KHDRBS3	microglia_neuron
MYO1B	microglia_neuron
RTN1	microglia_neuron
SALL1	microglia_astrocyte
LTGB5	microglia_astrocyte
CST3	microglia_astrocyte
SLC1A3	microglia_astrocyte
GPR56	microglia_astrocyte
MERTK	microglia_astrocyte
CABLES1	microglia_astrocyte
SALL3	microglia_astrocyte
ETV5	microglia_astrocyte
CHST7	microglia_astrocyte
EYA4	microglia_astrocyte
ARHGAP5	microglia_astrocyte

RNF180	microglia_astrocyte
PMP22	microglia_astrocyte
TMEM144	microglia_astrocyte
NPL	microglia_astrocyte
TRIM47	microglia_astrocyte
TLR3	microglia_astrocyte
LTGA6	microglia_astrocyte
ABCA1	microglia_astrocyte
TMEM100	microglia_astrocyte
ARHGAP12	microglia_astrocyte
FADS1	microglia_astrocyte
SPIRE1	microglia_astrocyte
TSPAN7	microglia_astrocyte
JAM2	microglia_astrocyte
LRRC8A	microglia_astrocyte
IL18	microglia_astrocyte
C1QC	microglia_oligodendrocyte
C1QB	microglia_oligodendrocyte
C1QA	microglia_oligodendrocyte
CTSS	microglia_oligodendrocyte
TIMP2	microglia_oligodendrocyte
BIN1	microglia_oligodendrocyte
FRMD4B	microglia_oligodendrocyte
TMCC3	microglia_oligodendrocyte
CHN2	microglia_oligodendrocyte
GAB1	microglia_oligodendrocyte
SPSB1	microglia_oligodendrocyte
SLC12A2	microglia_oligodendrocyte
SCARB2	microglia_oligodendrocyte
RAP1GDS1	microglia_oligodendrocyte
PAK1	microglia_oligodendrocyte
EPN2	microglia_oligodendrocyte
GLT1	astrocyte
CX30	astrocyte
GFAP	astrocyte
GLAST	astrocyte
AQP4	astrocyte
ALDOC	astrocyte
FGFR3	astrocyte
NG2	oligodendrocyte
SOX10	oligodendrocyte
CX47	oligodendrocyte
MBP	oligodendrocyte
MOG	oligodendrocyte
UGT8A	oligodendrocyte

GAL3ST1	oligodendrocyte
MOBP	oligodendrocyte
MAG	oligodendrocyte
MAL	oligodendrocyte
NEFL	neuron
GABRA1	neuron
SYT1	neuron
KCC2	neuron
SNAP25	neuron
KCNQ2	neuron
SV2B	neuron
OCLN	endothelial
VWF	endothelial
TIE2	endothelial
CLDN5	endothelial
PECAM1	endothelial

Table S3.2. Transcription factors identified as master regulators for the microglia module when results for all three data sets were combined.

Master regulators
ARNTL2
ATF4
BACH1
BATF
BATF3
BCL3
BCL6
BNC2
BPNT1
CASZ1
CBFB
CEBPB
CEBPD
CREB3L2
DACH2
DLX6
DOT1L
DRGX
DUSP12
E2F6
ELF4
ELK3
FLI1
FOS
FOXC1
HEYL
HHEX
HIF1A
HLA-DQB2
HLX
HOXD4
IFI16
IKZF1
INF2
IRF1
IRF5
IRF8
JUNB
KLF12

KLF6
LYL1
MAF
MAFB
MAFF
MYC
NEUROD6
NEUROG2
NFATC2
NFATC3
NFE2
NFIL3
NFKBIA
NFKBIZ
NONO
NR3C2
PARP12
PLEK
PRDM16
PRDM4
RNASE2
RNF113A
RREB1
RUNX1
RUNX3
SATB2
SF3A2
SIX4
SLC29A3
SLC4A10
SOX11
SOX15
SOX4
SP100
SPI1 (PU.1)
STAT3
STAT4
STAT5A
STAT6
TAL1
TBX15
TBX3
TEAD4
TFDP3

TIAL1
U2AF1L4
WDHD1
XBP1
ZC3HAV1
ZFP36
ZFP36L1
ZFP36L2
ZFPM2
ZIC1
ZNF14
ZNF189
ZNF248
ZNF32
ZNF366
ZNF425
ZNF521
ZNF526
ZNF628
ZNF687
ZNF696
ZNF776
ZNF821
ZSCAN16

Table S3.3. Master regulators in the protein-protein-interaction network.

ATF4
BACH1
BATF
BATF3
BCL3
BCL6
CBFB
CEBPB
ELK3
FOS
HIF1A
HOXD4
IFI16
IKZF1
IRF1
IRF8
JUNB
LYL1
MAF
MAFB
NFATC2
NFATC3
NFE2
NFIL3
NFKBIZ
NR3C2
PARP12
PRDM4
RREB1
RUNX1
RUNX3
SPI1
STAT3
STAT4
STAT5A
STAT6
TAL1
XBP1
ZFPM2
ZNF687

Table S3.4. Transcription factors identified as master regulators (MR) for the microglia module for all three data sets separately, ranked by Fisher's exact test's (FET) p-values.

Braincloud		**MRC/UK**		**HBTRC**	
MR	**FET p-value**	**MR**	**FET p-value**	**MR**	**FET p-value**
TAL1	9.93E-133	*SPI1 (PU.1)*	9.54E-99	*PLEK*	1.07E-216
MAF	1.38E-130	*IRF8*	2.00E-89	*ELF4*	2.24E-130
PLEK	2.42E-124	*SP100*	8.32E-55	*BATF*	2.11E-122
FLI1	1.63E-74	*RUNX1*	7.42E-48	*IRF8*	4.89E-93
KLF6	6.17E-70	*PLEK*	4.00E-37	*IRF5*	4.85E-70
HLA-DQB2	1.70E-68	*ZFP36L2*	7.84E-34	*HIF1A*	1.23E-63
RUNX1	2.98E-52	*IKZF1*	1.04E-30	*RUNX3*	1.10E-61
HHEX	2.45E-48	*MAFB*	9.49E-21	*HHEX*	8.48E-59
SOX15	2.23E-36	*IFI16*	7.73E-18	*BCL6*	4.46E-57
ZFPM2	2.49E-36	*TAL1*	1.63E-17	*NFE2*	2.08E-50
STAT4	8.10E-30	*ZSCAN16*	2.55E-13	*ZFP36*	1.13E-49
DLX6	9.96E-30	*ELK3*	5.89E-12	*CEBPB*	1.55E-47
ZNF14	7.10E-28	*DRGX*	1.96E-11	*STAT3*	6.53E-41
DACH2	2.88E-25	*PARP12*	1.19E-10	*CEBPD*	1.61E-40
TIAL1	2.75E-24	*ZNF526*	1.50E-10	*TEAD4*	2.36E-38
SATB2	2.49E-22	*SLC29A3*	4.19E-09	*RUNX1*	1.11E-36
KLF12	8.97E-21	*FOS*	1.48E-08	*NFKBIZ*	3.38E-36
IRF8	1.95E-19	*DOT1L*	5.38E-07	*STAT5A*	3.38E-36
TBX3	2.91E-18	*INF2*	6.94E-07	*MYC*	5.47E-33
DUSP12	1.21E-17	*SLC4A10*	7.40E-07	*IFI16*	1.08E-30
HLX	1.48E-16	*SF3A2*	8.61E-07	*MAF*	2.22E-30
ZNF366	6.94E-15	*NFATC2*	2.85E-06	*BCL3*	1.70E-29
FOXC1	3.54E-14	*RNASE2*	9.47E-06	*KLF6*	2.45E-25
RNF113A	1.59E-13			*SIX4*	1.08E-23
ZNF687	2.12E-13			*BACH1*	2.84E-23
HOXD4	9.49E-13			*U2AF1L4*	1.32E-22
ZNF628	2.17E-12			*NEUROG2*	2.30E-21
LYL1	3.02E-12			*RREB1*	1.84E-18
ZC3HAV1	9.48E-11			*LYL1*	3.31E-17
BATF3	1.75E-10			*MAFF*	1.26E-15
ZNF821	4.86E-10			*ZNF32*	1.68E-15
ZNF248	2.18E-09			*FOS*	4.38E-15
PARP12	2.18E-08			*ARNTL2*	1.09E-14
ATF4	4.09E-08			*IRF1*	1.78E-14
CBFB	2.86E-07			*CASZ1*	1.31E-13

ZNF189	7.40E-07
NR3C2	1.73E-06
NFATC3	1.96E-06
BNC2	3.06E-06
ZFP36L2	3.43E-06
ZNF776	6.07E-06
PRDM4	2.48E-05

BNC2	1.02E-12
NFIL3	1.44E-12
NFATC2	2.21E-12
BPNT1	4.20E-12
WDHD1	7.08E-12
TFDP3	2.18E-11
NFKBIA	1.74E-10
NEUROD6	1.87E-10
XBP1	3.52E-10
TAL1	3.94E-10
ZNF425	1.24E-09
ZNF521	3.52E-09
ZFP36L1	8.06E-09
HEYL	1.00E-08
TBX15	1.51E-08
ELK3	2.19E-08
STAT6	3.05E-08
SOX11	1.44E-07
PRDM16	2.04E-07
E2F6	3.05E-07
ZNF696	1.06E-06
JUNB	1.32E-06
ZIC1	1.71E-06
PRDM4	2.71E-06
NONO	7.29E-06
CREB3L2	7.93E-06
ZFP36L2	1.56E-05
SOX4	2.63E-05
FLI1	3.21E-05

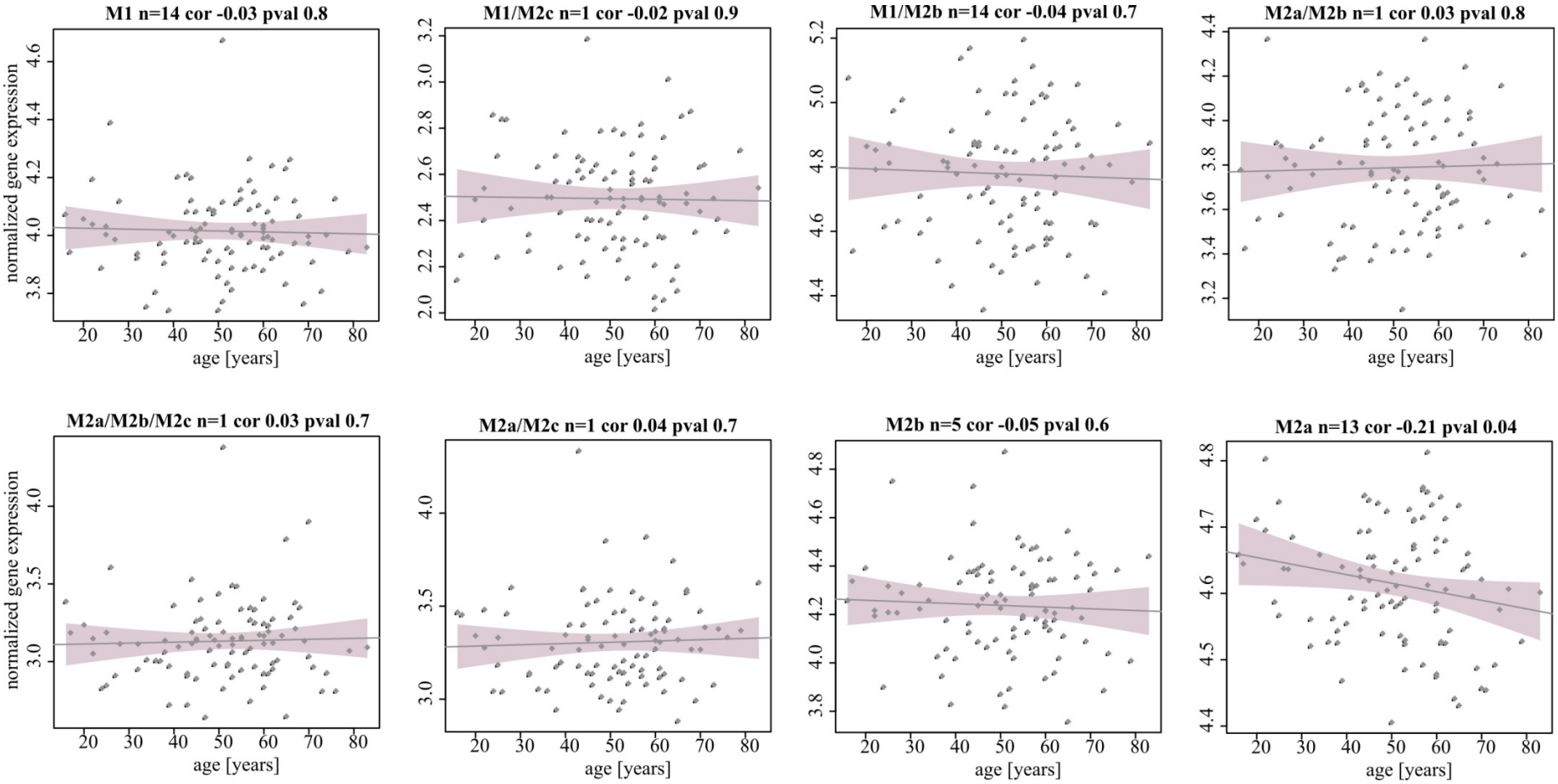


Figure S3.1. Pearson's correlation (indicated as "cor" and p-value as "pval") of age with average expression of all marker genes (number of marker genes indicated as "n") for specific microglia activation phenotypes (indicated in plots' titles) in the MRC/UK data.

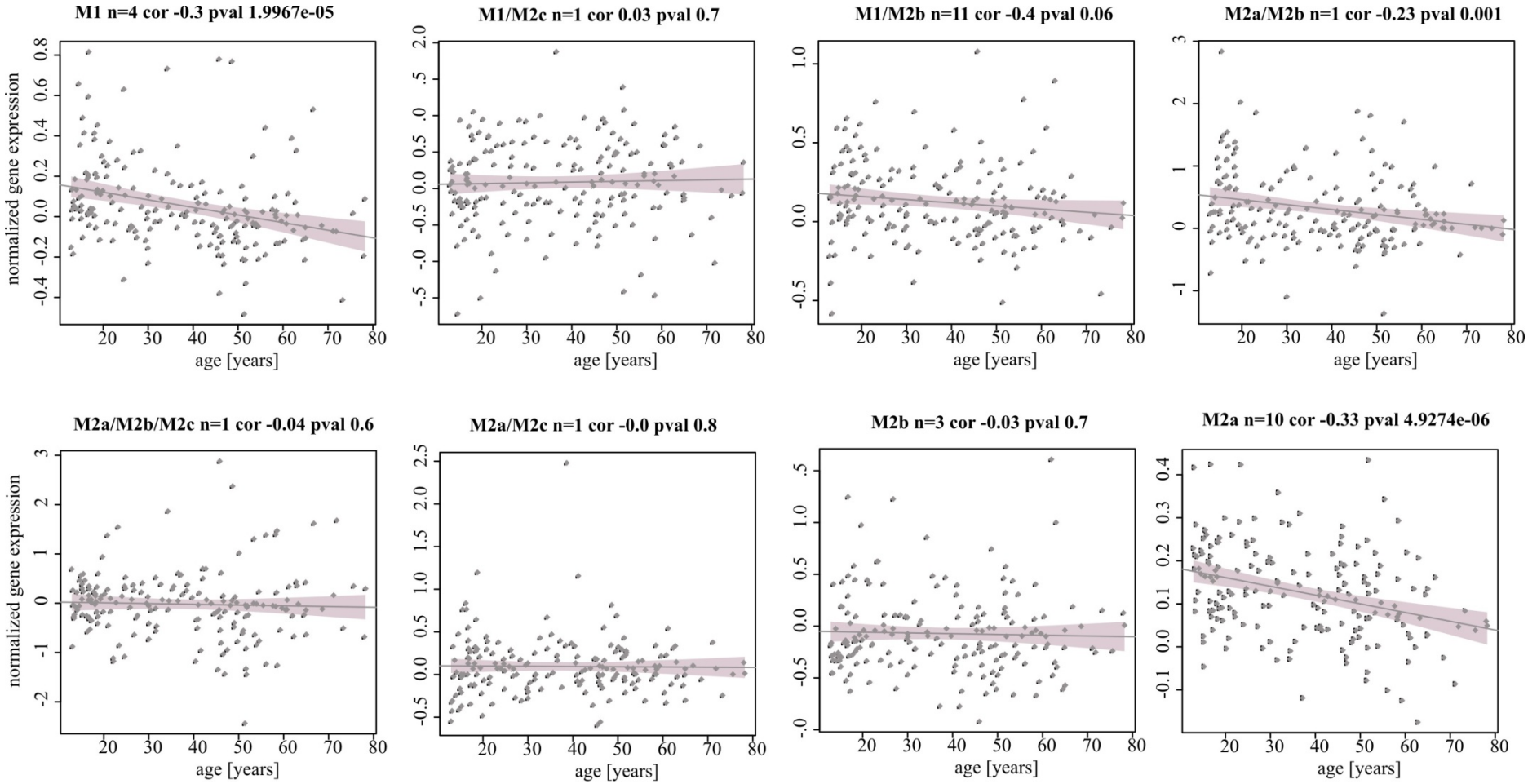


Figure S3.2. Pearson's correlation (indicated as "cor" and p-value as "pval")) of age with average expression of all marker genes (number of marker genes indicated as "n") for specific microglia activation phenotypes (indicated in plots' titles) in braincloud data.

Chapter 4 – Preservation of transcriptional networks associated with neurogenesis from fetal post-mortem brains during later life

4.1 Abstract

The discovery of neurogenic niches proved that neurogenesis continues during brain ageing, instead of the central nervous system (CNS) being cellularly static after early development. Recently, an increasing number of brain regions have been linked to neurogenic processes [194] outside of the well-characterized neurogenic niches, including cortex [195, 196], amygdala [197-199], hypothalamus [200, 201], striatum [202, 203], and substantia nigra [204, 205]. Transplantation of niche-astrocyte-produced signals might also induce neurogenesis outside of known neurogenic zones [206]. However, neurogenesis in adult brains occurs only at low levels or in response to aversive stimuli, such as brain injury or neurodegeneration, instead of spontaneous, continuous neurogenesis [194, 207]. Previous studies showed that the adult CNS appears to partially recapitulate molecular requirements of neuronal development during early life [208]. Maintenance of a molecular and cellular microenvironment that is supportive of neurogenesis might be a necessary precondition for a brain tissue's ability to react to injury with neurogenesis and plastic changes.

I hypothesized that such a neurogenic microenvironment would imply preservation of fetal neurogenesis transcriptional networks and its Transcription Factor (TF)-target gene interactions in adult brains. To test this hypothesis, I identified cell cycle gene expression modules in human fetal brains, as well as TFs that are specific for these modules. I tested for the preservation of the cell cycle modules and their TFs in the

prefrontal cortex of human adult post-mortem brains. I found that cell cycle modules from fetal life show moderate-to-strong preservation in human adult post-mortem brains, and are regulated by neurogenesis TFs whose activities are recapitulated from fetal age (e.g. *SOX2, SOX9, EMX2*) or are specific to adult age (e.g. *FOXO1*). The observed preservation of transcriptional interactions required for neurogenesis in adult human brains could indicate that a cellular microenvironment is maintained that allows neurogenesis in response to injury in the prefrontal cortex, outside of well-characterized neurogenic niches.

4.2 Introduction

The transcriptome of the human fetal brain shows stronger age-specific expression dynamics relative to any other developmental stage over the lifespan, including the highest rate of expression fold change and highest intersubject correlation [2, 3]. Moreover, specific fetal gene expression changes are reversed after birth or later in life [2]. During fetal brain development, the majority of cortical neurons are generated by neurogenesis, migrate to the neocortex and form the essential neural network. In contrast, neurogenesis in adult life has initially been observed only within distinct anatomical regions, namely the subependymal zone of lateral ventricles and the subgranular zone of the dentate gyrus in hippocampus [194]. More recently, the presence of neuroblasts and neurogenesis in the adult human brain has been reported in brain regions outside of these two 'classic' neurogenic niches, for example in the striatum [209]. However, adult neurogenesis outside of subventricular zone and hippocampus remains controversial and is potentially restricted to very low levels or to pathological conditions [194, 207]. Known heterogeneity in adult CNS plasticity contributes to this controversy, such as progenitor cell proliferation as opposed to true gliogenic/neurogenic processes, defined as a series of subsequent steps leading to new neurons with distinct functions [207, 210].

The production of new cells across the lifespan relies on the cell cycle regulatory machinery [194]. The cell cycle spans four phases leading to cell division, the first gap phase (G1), DNA replication in synthetic phase (S), second gap phase (G2) and mitosis (M), followed by cell cycle exit, migration and finally neuron maturation. Cell cycle progression is coordinated by a set of cell cycle regulators, which are functionally redundant and dependent on cell type and niche [194]. The cellular environment for

neurogenesis differs greatly between fetal and adult brains. Whilst early development is characterized by cell proliferation, the aging brain consist of mature and functionally committed cells [194]. Most cells in the adult brain maintain a quiescent differentiated state (G0) and may only reenter the cell cycle following mitotic signals in response to non-physiological conditions, such as neurodegeneration or brain injury [194]. Regulation of the cell cycle progression or exit in physiological conditions, as well as in reaction to pathological states during adult neurogenesis, remains poorly understood [194].

Early patterning of the neocortex depends on coordinated expression of TFs [211, 212], such as *PAX6* and *EMX2* [212]. High levels of *PAX6*, combined with low levels of *EMX2*, induce progenitors to develop into neurons that are appropriate for the motor cortex [212]. Co-expression of TFs in progenitor cells has been suggested as an indicator of cell-fate: For example, *SOX2* and *TBR2* (*EOMES*) expression in the human OSVZ (outer subventricular zone), potentially generates excitatory neurons, whereas *MASH1* (*ASCL1*) expressing cells are committed to generating inhibitory neurons [213]. *SOX2* expression is sufficient to generate proliferative neuroblasts from resident astrocytes in the adult and aged mouse brain [214]. The same TFs can play crucial roles for both embryonic and adult neurogenesis. For example, members of the *FOXO* gene family, including *FOXO1*, are expressed in adult neurogenic zones and are essential for proliferation and differentiation during embryonic and adult neurogenesis [215]. Both *FOXO1* and *SOX2* are mostly expressed in glia (such as oligodendrocytes, astrocytes and microglia) [216]. In the adult brain, *FOXO1* also controls cell death, by regulating cell death genes such as *BAX* [215]. The hypothesis has been proposed that *FOXO* family genes support neurogenic niches, microenvironments that allow proliferation and

differentiation of neural stem cells in the hippocampus and cerebral cortex across the lifespan [217].

Here I asked the question whether cell cycle transcriptional networks and their regulatory TF-target interactions in fetal life are preserved in the aging human brain outside of known neurogenic niches, namely in the prefrontal cortex. The cellular environment in fetal brains is tuned for cellular proliferation, in contrast to adult brains, which are mostly characterized by cellular quiescence [194]. Previous studies suggest that the adult CNS partially recapitulates the cellular requirements for proliferation of the developing brain, which presents the possibility of inducing neuronal differentiation outside of currently known adult neurogenic niches (e.g. [208]). For this reason, I sought to investigate what part of the cell cycle transcriptional network is maintained during adult life, which might possibly provide for a microenvironment with neurogenic potential, either *de novo* or following cell transplantation. More specifically, I identified transcriptional gene modules linked to the cell cycle during fetal life and predicted "neurogenesis TFs" that were specific for these gene sets for either fetal or adult age. I found that two cell cycle-associated modules detected in fetal brains showed moderate-to-strong preservation in adult brains. TFs that have been previously identified as being crucial for neurogenesis regulated significantly large subsets of genes in the cell cycle modules during either fetal (e.g. *FOXM1*) or adult life (e.g. *FOXO1*) but not in both. A third set of TFs, including *SOX2, SOX9, EMX2* and *PAX6*, were detected in both age groups but changed the majority of their target genes from fetal to adult life, indicating a shift in TF function during brain ageing. Preserved coordination between transcripts and TF-target interactions could reflect the maintenance of a microenvironment well after development thereby allowing the induction of neurogenesis, for example in response to

injury, in the ageing human prefrontal cortex outside of well-characterized neurogenic zones.

4.3 Results

4.3.1 Cell cycle transcriptional modules in fetal brain reflect different cell stages

I used whole genome microarray gene expression data from human post-mortem brains [2] (see **Chapter 2.1**) to derive co-expression modules in transcriptional networks using WGCNA [122] (see **Chapter 2.5.1**). Previous studies have shown that expression changes across the lifespan are driven predominantly by transcriptional dynamics occurring shortly before and after birth [2, 3]. For that reason, I asked whether fetal brain-specific transcriptional subnetworks linked to cell cycle are maintained in the prefrontal cortex during adult life, outside of known adult neurogenic niches. More specifically, I used transcriptional profiles from 38 fetal human brains to derive network modules which I then compared to modules from adult human transcriptional networks from four age intervals, namely 12-20 years, 20-35 years, 35-50 years and 50 years and older, each containing n=18846 probes from 38 subjects (matched in sample size to the fetal brain data set). Two modules in the fetal transcriptional network showed significant enrichment for GO-terms linked to "cell cycle" (Supplementary Tables S4.1 and S4.2) and were strongly negatively correlated with age (Pearson's correlation=-0.7, $p\text{-}value<2.3\text{x}10^{-6}$ in one module and Pearson's correlation=-0.5, $p\text{-}value<2.0\text{x}10^{-3}$ in the second module; Figure 16).

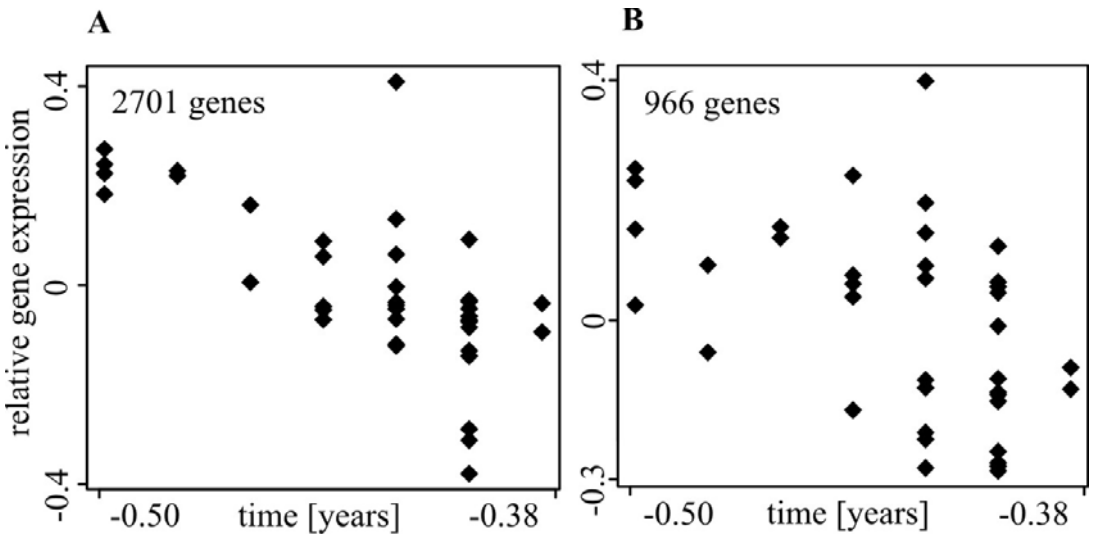


Figure 16. **Scatterplot of module eigengenes (ME, the first principal component) of each of the two modules enriched for GO-terms linked to "cell cycle". The y-axis indicates relative gene expression and the x-axis indicates fetal age. Age of the fetal brains is given in years, relative to date of birth. The total number of genes in each module is indicated in the plot, more precisely *n*=2701 genes in one module (A) and *n*=966 genes in the other module (B).**

No module enriched for cell cycle genes was detected for any of the four older networks. However, both cell cycle modules from fetal age showed evidence for being preserved in gene expression data from 184 subjects aged 12 years and older, based on composite preservation statistics implemented in WGCNA ($Z_{summary}$) [123]. Summary statistics $Z_{summary}>10$ indicate the strong preservation of a module and $2>Z_{summary}>10$ indicates moderate preservation. Here, one fetal cell cycle module (n=2701 genes, Figure 16A) showed strong evidence for preservation ($Z_{summary}$=13), whereas the other module (n=966 genes, Figure 16B) was moderately preserved in the adult data ($Z_{summary}$=8.7).

Expression of markers for cell stages (S phase, G1/S, G2, G2/M, or M/G1 [218]) indicated that one fetal module included more markers covering the cell cycle from G1 to G2 (M/G1, G1/S, S phase, G2, hereafter referred to as the "Interphases module"; Figure 17), whereas the second module was significantly enriched for the GO-term "M phase" (58 genes contributing to enrichment) and included at least twice as many markers for G2/M compared to other stages (referred to as the "G2/M module"; Figure 17). More precisely, both modules showed a higher total number of expressed cell stage markers than expected (*empirical p-value*<0.05), based on 1000 random samples of gene sets each matched in sample size to the two cell cycle modules (*n*=2701 genes in the Interphases module and *n*=966 genes in the G2/M module, Figure 17). More precisely, the Interphases module was significantly enriched for G1/S and G2 markers; compared to markers for S phase that showed *empirical p-value*=0.08 (Figure 16). Each of the five marker types was significantly enriched in the G2/M module, and yet G2/M markers were by far the most abundant (Figure 17). These enrichments for markers linked to different cell stages indicates that during fetal life, cellular progression during interphases compared to mitosis is differentially regulated.

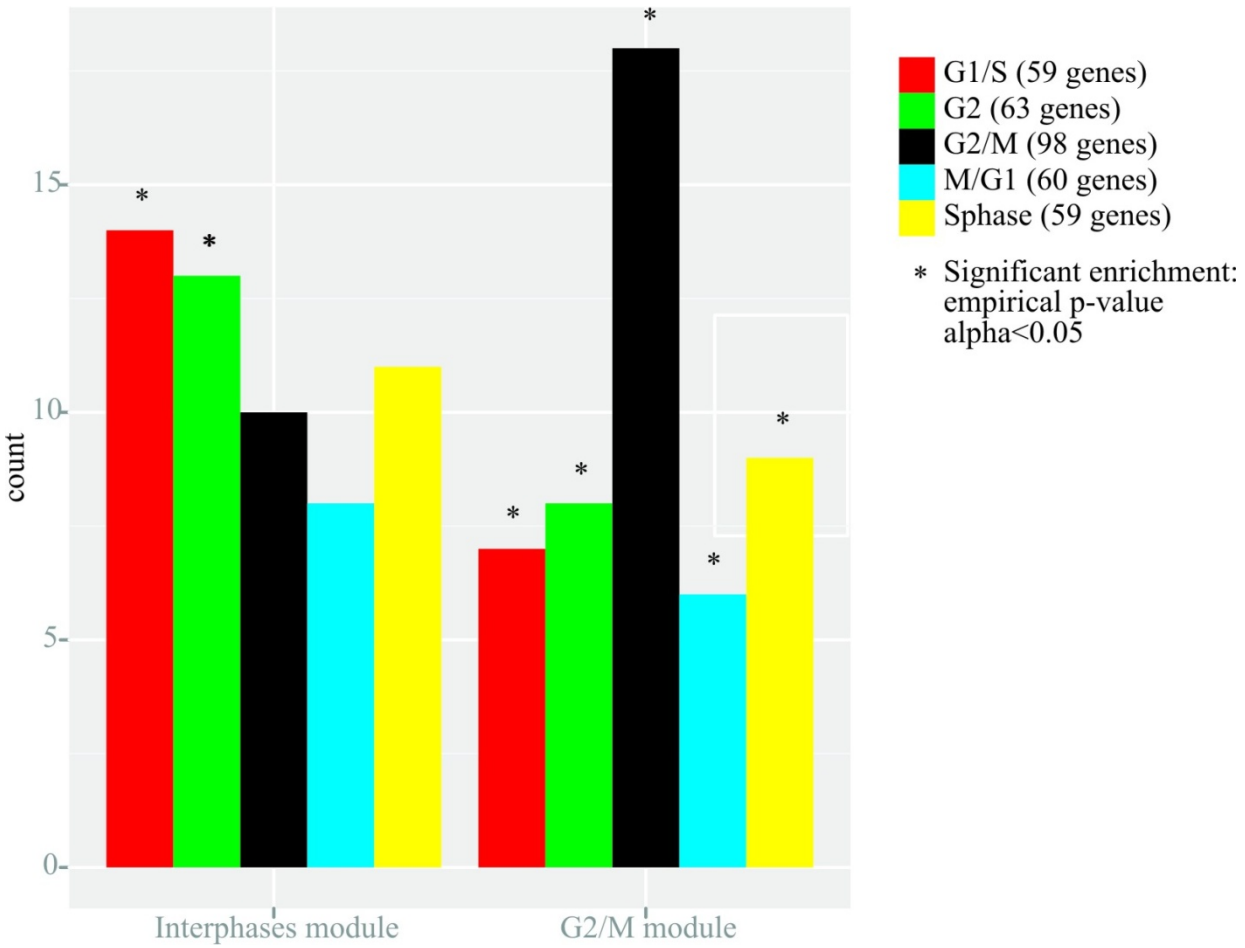


Figure 17. Enrichment of marker genes for cell cycle stages in two modules enriched for GO-terms linked to "cell cycle". Hereafter, the first module is referred to as the "Interphases module" (left) and the other module is referred to as the "G2/M module" (right). Cell cycle stages referred to using colours are indicated in the legend, including the total numbers of marker genes available for each stage [218].

The expression of cell stage marker genes across the lifespan showed that the strong negative correlation with fetal age was not maintained during later life (Figure 18 and Figure 19 for all expressed cell stage markers in braincloud microarray gene expression data set). These negative correlations in fetal brain data are consistent with the diminution of neurogenesis during development and its cessation after early prenatal life. Adult brains are no longer 'poised' for cell proliferation, and neurogenesis outside of neurogenic niches is restricted to low levels or specific circumstances [194, 207]. For these reasons, I interpret these data as implying that transcriptional modules are no longer sufficiently cohesive to be detected *de novo* within adult age categories, despite composite preservation statistics showing moderate-to-strong preservation of the two modules, which is an indication that expression coordination between these genes is partially maintained during ageing.

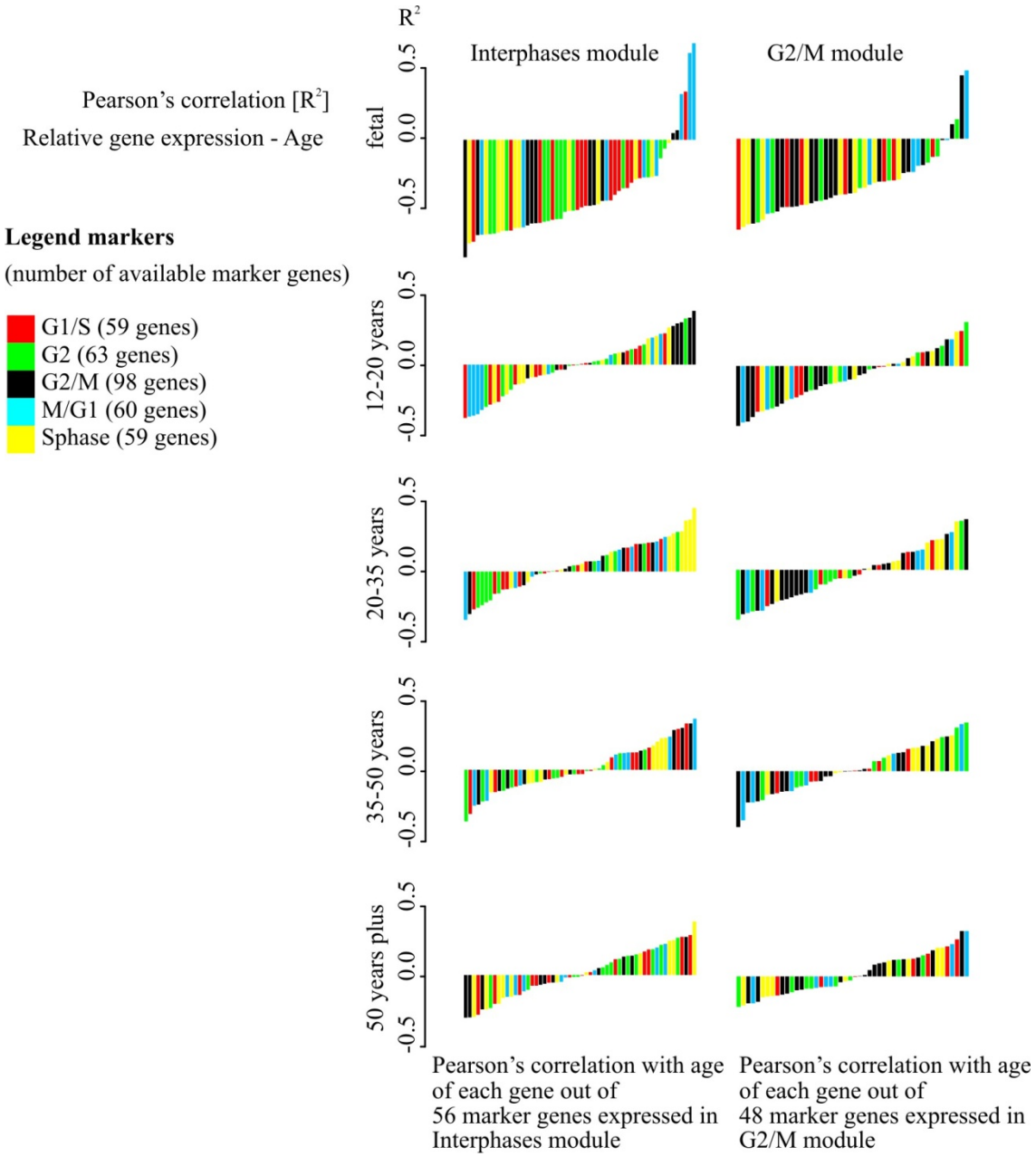


Figure 18. Barplot of Pearson's correlation (Pearson's correlation coefficient [-1...1] on the y-axis) of each cell cycle stage marker gene (x-axis) expressed in Interphases or G2/M module with age. Each bar is the Pearson's correlation of a given expressed marker gene with age, over the age bin indicated on the y-axis. Cell cycle stages referred to using colours are indicated in the legend, including total numbers of available marker genes [218].

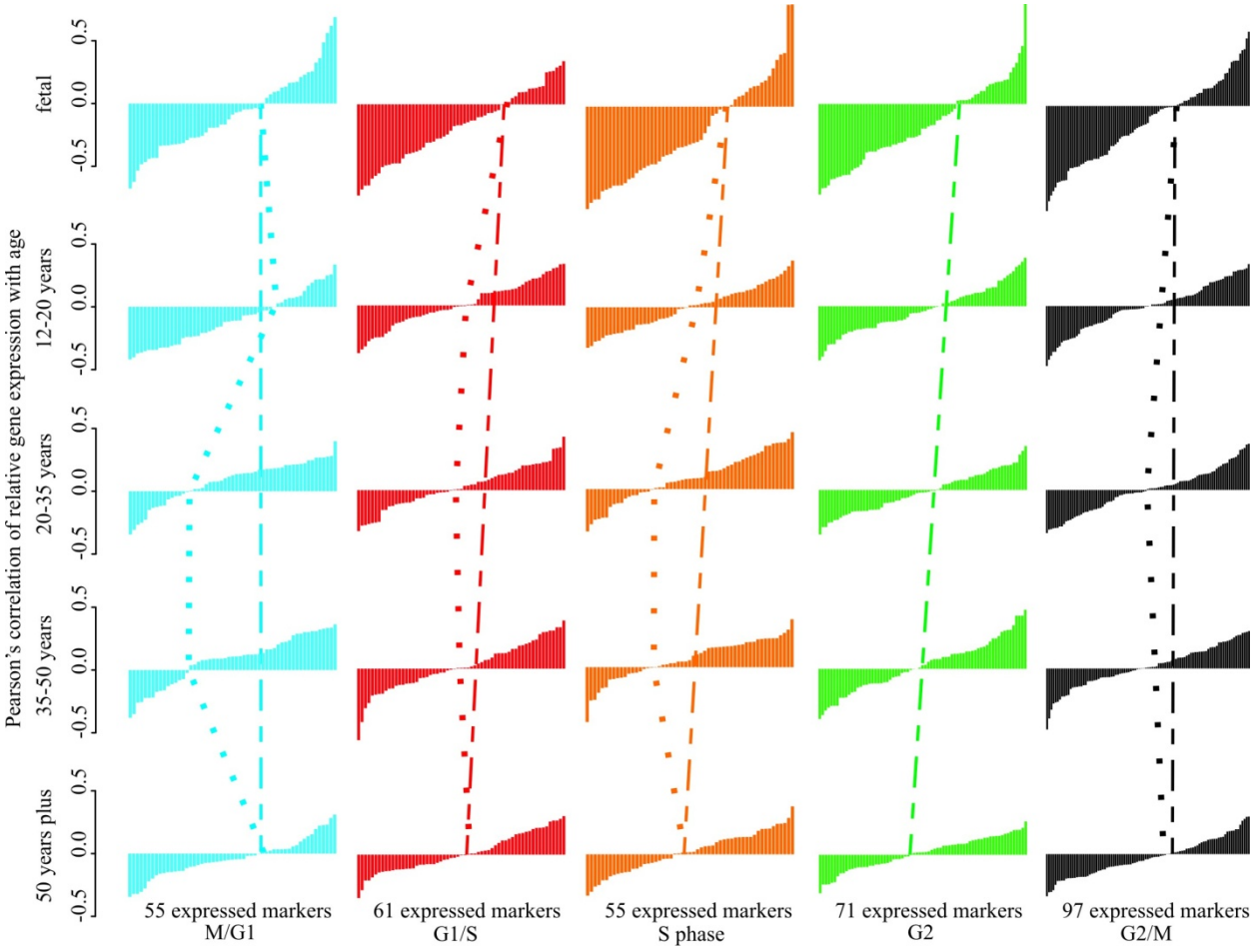


Figure 19. Barplot of Pearson's correlation (Pearson's correlation coefficient [-1...1] on the y-axis) with age of each cell cycle stage marker gene (x-axis) expressed in the complete microarray gene expression data set used to build the transcriptional network. Each bar is the Pearson's correlation of a given expressed marker gene with age over the age bin indicated on the y-axis. Cell cycle stages referred to using colours are indicated in the legend [218].

Prompted by the moderate-to-strong preservation of the cell cycle modules from fetal brains in gene expression data obtained from human adults, I next aimed to predict which TFs function as specific regulators for the two cell cycle modules during fetal life, and whether these TF-target interactions are subsequently maintained during brain ageing. Although cell cycle modules were not detected *de novo* in the adult transcriptional networks, the preservation of TF-target interactions could play a role for how these modules' genes respond to mitogenic signals to reenter cell cycle or undergo cell cycle progression instead of apoptotic cell death [194].

4.3.2 Neurogenesis transcription factors in developmental and adult stages include genes that regulate adult neurogenesis

I sought to identify "neurogenesis TFs" that regulate genes present in the Interphases (*n*=2701 genes) and G2/M modules (*n*=966 genes). For that purpose, I predicted TF-target interactions for genes in the modules separately for fetal and adult life using ARACNE [50, 126]. "Neurogenesis TFs" were defined as those TFs whose regulon (all predicted target genes) contained significantly more genes from either of the two cell cycle modules than expected by chance (Fisher's exact test, α=0.05, corrected for 1556 TFs tested). I used a data-driven approach and considered all genes which were assigned to the "cell cycle"-modules in fetal life without restricting the gene sets to for example known cell stage markers. In this way, I aimed to test for preservation of gene co-expression of all genes linked to cell cycle and neurogenesis during fetal life, regardless of specific known cell cycle pathways.

During fetal life, I identified 167 neurogenesis TFs for the Interphases module compared to 96 for the G2/M module (Supplementary Table S4.3 and S4.4). Fifteen

such TFs were detected for both modules (*RAD51, FOXN2, EMX2, TCF19, MXD3, GLI3, ZNF367, DNMT3A, AKNA, ARNTL2, SCRT2, THRA, NR3C2, LARP6, DACH1*).

Separately, over adult life (184 subjects aged 12 years and older), I identified 22 and 2 TFs whose predicted regulons overlapped significantly with genes from Interphases or G2/M modules in the adult age category, respectively (Supplementary Table S4.3 and S4.5). For the Interphases module, 15 out of the 22 TFs were replicates from fetal age (*SOX2, EMX2, SOX9, HES6, HOMEZ, TRPS1, ELAVL2, DNAJC1, ZFHX4, ZNF107, FOXN2, KLF15, SOX21, GLI3, RFX2*) compared to 7 TFs detected in adult life only (*FOXO1, TIPARP, HIST1H1C, ZIC1, TSC22D4, ONECUT2, ZNF146*). The adult G2/M module yielded only 2 TFs (*FOXO1, ZNF114*) that were absent from the fetal module.

FOXO1 was identified as a regulator for both cell cycle modules during adult life stages, but not during fetal life. *FOXO1* is known to play multiple roles during adult neurogenesis and cell death [215, 217] (Supplementary Table S4.5). Cells entering the cell cycle during adulthood potentially lack the factors needed for cell cycle progression and consequently undergo apoptosis instead of dividing [194, 219]. In that regard, genes in both cell cycle modules potentially shift from controlling proliferation to being apoptosis regulators during ageing, obviously in line with the basic notion of decreased neurogenesis in aged brains. The observation of some TFs being replicated in fetal and adult age (e.g. *SOX2* or *EMX2*) compared to TFs specific to adult age (e.g. *FOXO1*) leads to the question whether replicated TFs maintain the same target genes across the lifespan, and in turn whether TFs specific to adult age take over the regulon of TFs specific to fetal age (e.g. *FOXM1*).

4.3.3 Validation of neurogenesis TFs using ChIP-Seq data and motif enrichment

I validated TF-target gene interactions predicted by ARACNE using publicly available ChIP-Seq data, separately for fetal and adult life stages. More precisely, I tested for significant overlap between ChIP-Seq peaks and basal regulatory domains in the gene set of interest, consisting of the intersection between predicted target genes for a given TF and the genes in Interphases or G2/M modules. Basal regulatory domains were defined as basal extensions ranging from 5kb upstream to 1kb downstream of the transcription start site (TSS) of the annotated gene [121]. Only genes with an annotated genomic region and an associated gene name available in ENSEMBL BioMart were considered. The background consisted of basal regulatory domains, defined as for genes from the two modules, for all brain-expressed genes that were initially used to build the networks. I hypothesized that basal regulatory domains from predicted TF-targets in the cell cycle-modules show significantly greater overlap with ChIP-Seq peaks for the given TF than for all brain-expressed genes.

To validate TF-target interactions during fetal life, I used ChIP-Seq data available in ENCODE for a TF with a known role in neurogenesis, namely *FOXM1* [220]. I used ChIP-Seq peaks from lymphoblastoid cell line Gm12878 for *FOXM1*, which was identified as a G2/M module-specific TF. For *FOXM1*, I found significant enrichment of ChIP-Seq peaks in basal regulatory domains of the predicted target genes (165 genes) during fetal life (1.5-fold enrichment, *qvalue*=0.001 in replicate 1 and 1.5-fold enrichment, *qvalue*=0.001 in replicate 2).

During adult life, *FOXO1* was the only TF predicted as being specific for both Interphases- and G2/M modules. In contrast, *FOXO1* was not predicted to be specific

for either of the two modules during fetal age. I used ChIP-Seq peaks for *FOXO1* in CD8+ cells to validate the predicted *FOXO1*-target genes during adulthood. The basal regulatory domains of predicted *FOXO1*-target genes in both Interphases- and G2/M modules (111 genes) showed moderate overlap with the ChIP-Seq peaks. However, the fold-increase was not sufficiently strong to be significant (1.1-fold enrichment, *qvalue*=0.4 in replicate 1 and 1.7-fold enrichment, *qvalue*=0.1 in replicate 2). As described above, the ChIP-Seq data came from publicly available sources (see **Chapter 4.5.1**); it has to be noted that the results are therefore influenced by a lower number of available peaks for FOXO1 and a different underlying cell type (see **Chapter 4.5.1**).

In addition to testing for the enrichment of ChIP-Seq peaks, I tested for the enrichment of known *FOXO1*-binding motifs in the basal regulatory domains (5kb upstream and 1kb downstream of the transcription start site (TSS) of the annotated gene [121]) of predicted *FOXO1*-target genes against a background of basal regulatory domains from all brain-expressed genes using the MEME Suite (version 4.10.0 [221]). I used IUPAC codes for two *FOXO1*-motifs in vertebrates (NRWAAACAAN and GNNTTGTTTACNTT), downloaded from the TRANSFAC data base [222] (see **Chapter 4.5.4**). No *FOXO1*-motifs specific to humans were available. Motif enrichment in the sequences of the basal regulatory domains of *FOXO1*-target genes was tested using the online AME tool (Analysis of Motif Enrichment) from the MEME suite [223] (see **Chapter 4.5.4**). There was no significant enrichment of the two *FOXO1*-motifs found in the predicted *FOXO1*-target genes.

The enrichment of ChIP-Seq peaks within promoter-TSS in predicted target genes of *FOXM1* indicates a causal role of this neurogenesis TF for regulating the expression of its predicted targets. As would be expected, the observed signal was stronger during

fetal compared to older age – and only reached significance in the younger group. Clearly, the more subtle expression changes during later life could introduce a higher error rate, including false positives in the predicted target genes or false negatives, not included in the prediction. Furthermore – as described above and in **Chapter 4.4** – I used publicly available immortalized cell lines from non-brain tissues which can obviously serve only as a proxy for the age-dependent changes observed in human brain tissue.

4.3.4 The transition of cell cycle transcriptional modules from fetal to adult life is characterized by age-dependent transcription factors

Next I was interested in the transition from neurogenesis TFs during fetal life, which is characterized by the development of the majority of cortical neurons and strongest expression changes [2], to neurogenesis TFs during ageing. The regulation of neuronal plasticity in the aging human brain is still poorly understood, especially outside of neurogenic niches [194]. However, preservation of transcriptional cell cycle networks and their TF-target interactions could impact on the potential for neurogenesis in parenchyma either *de novo* or following injury or transplantation. I sought to test whether TFs regulating the cell cycle modules in fetal and adult age maintained the same regulon, and, correspondingly, whether adult-specific TFs replaced fetal-specific TFs that regulate expression of a given gene set.

The predicted regulon (defined as all predicted TF-target genes) of TFs replicated in fetal and adult life showed a surprisingly small overlap. For example, *SOX2*'s regulon increased from 198 genes during fetal age to 291 genes in the transcriptional networks derived from adult brains and yet only 16 genes were part of both young and old regulons (Figure19A; *ACSS1, ARSE, BCAN, C1orf59, CCDC102A, EPHX2, KCNN3,*

KIAA1244, METRN, MTMR9, NBEA, NKAIN4, PHGDH, PTPRO, TMED1, TTYH1). Similarly, *EMX2*'s regulon decreased from 346 genes during fetal age to 295 genes, with only 20 genes being replicated in the older group. *FOXO1* was not cell-cycle modules specific in fetal brains, but became the only TF that was replicated in both the Interphases and the G2/M modules during adulthood. The regulon of *FOXO1* markedly increased in size during ageing (12-78 years), going up from 38 to 356 genes Figure 20B). None of the initial 38 genes were preserved in the adult *FOXO1* regulon. *N*=15 genes formed part of the *FOXM1* regulon during fetal age and became part of the *FOXO1* regulon in the adult group (Figure 20B; *TPM3, NECAP1, ARL6IP6, ACTR10, KIF11, ATP6V1E1, GSTA4, AMPH, RP2, TPRG1L, ATP6AP2, MAP1A, GDA, EMX2, CHST14*). The *FOXM1* regulon itself preserved only 4 genes from fetal to older age (*C15orf42, IGF2BP3, CCNB2, BUB1*). Three other genes were part of the *FOXM1* regulon during fetal life, but became part of the *SOX2* regulon during ageing (*MAMDC2, DERA, MXD3*). The transition from a TF's regulon during fetal life to another regulon later in life within the genes present in the two fetal cell cycle modules, for example as observed for *SOX2* (Figure 20A), indicates that the same TFs may fulfil multiple roles in the regulation of cell cycle-related genes during different developmental stages.

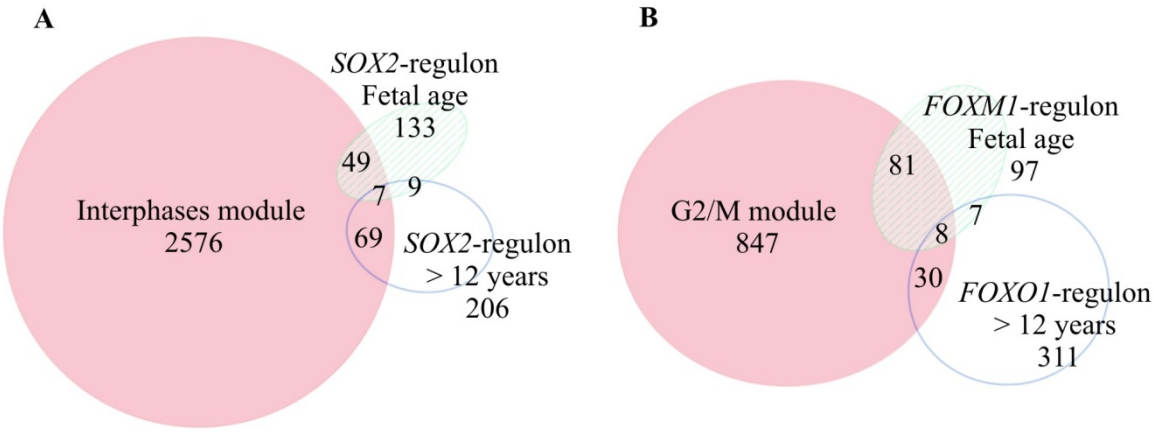


Figure 20. Venn diagrams indicating the overlap of the Interphases module's genes with predicted *SOX2* target genes during fetal or adult life (A), and G2/M module's genes with *FOXM1*-predicted target genes during fetal life compared to *FOXO1*-predicted target genes during adult life (B).

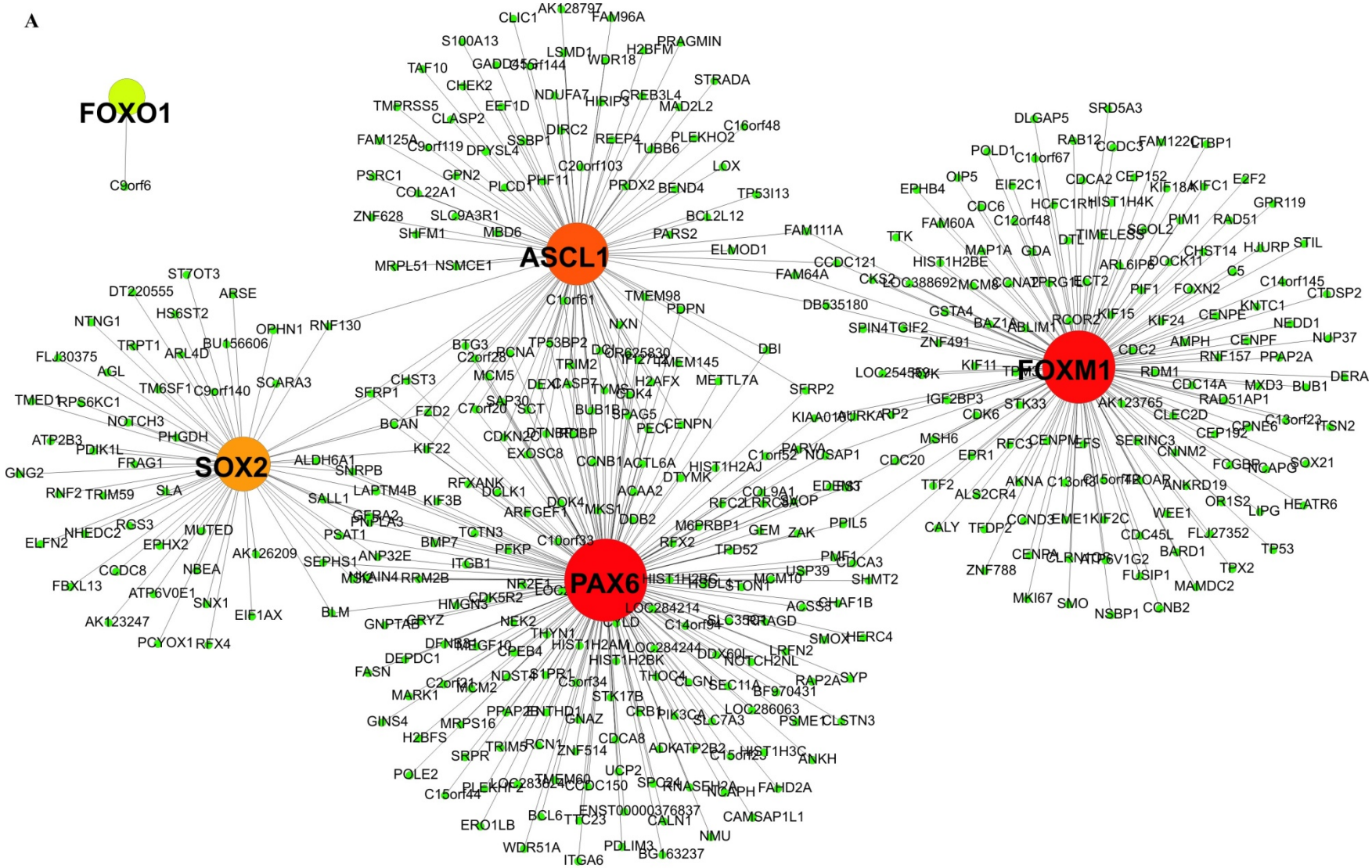
Clearly, major changes ensue from moving from embryonic neurogenesis, building up the CNS, to the mature brain, which is characterized by quiescent cells and low level heterogeneous neuro- and gliogenic processes. This heterogeneity derives from multiple sources, including the location and nature of progenitor cells, the existence of well characterized or atypical neurogenic niches, developmental stage, complete or incomplete neurogenesis and the differentiation of low level spontaneous, continuous neurogenesis from neurogenesis in response to injury [207]. However, despite this diversity of neurogenic processes, the preservation of cell cycle modules and cell cycle module-specific TFs (e.g. *SOX2, SOX9, EMX2*), as well as the regulation of adult cell cycle modules by TFs that we know to play crucial roles for neurogenesis (e.g. *FOXO1*; Supplementary Tables S4.3, S4.4, and S4.5), indicate that regulatory patterns deriving

from the fetal neurogenic microenvironment are partially maintained in the ageing prefrontal cortex. This is in line with previous studies suggesting that some cellular requirements for neurogenesis during development are recapitulated in the adult CNS [208], which raises the possibility of neuronal differentiation being induced outside of known neurogenic niches [208].

I identified 23 TFs as specific to Interphases – or G2/M module during adult life. These 23 TFs have documented roles in regulating neurogenesis, cell proliferation and other developmental pathways (Supplementary Table S4.5). Both modules when combined contained 3667 genes, of these 861 genes were regulated by at least one of the 23 neurogenesis TFs and up to 13 TFs targeted the same gene (*TULP3*). To test whether these 861 genes shared biological functions, I tested for functional enrichment of MGI and GO-terms against a background of all brain-expressed genes that were originally used to build the transcriptional network. All ontologies used are described in detail in **Chapter 2.4**. The top-ranked MGI term "abnormal neuronal precursor proliferation" (*hyper geometric p-value*=2.0×10^{-4}) and the GO-term "neuron differentiation" (*p-value*=4.4×10^{-4}) indicate a possible role of this gene set in neuronal proliferation, although no term survived the stringent multiple testing correction required.

Figure 21 illustrates examples of TFs identified in fetal life only (Figure 21A, *FOXM1*, *ASCL1*), TFs identified in both age groups (Figure 21A and Figure 21B, *SOX2*, *PAX6*), and of the one TF that was detected in the adult group only (Figure B, *FOXO1*); including only target genes which were part of the Interphases - or G2/M modules. Although the majority of *SOX2*'s target genes changed from fetal to adult age, *SOX2* shared a high number of target genes especially with *PAX6* at both stages. *PAX6* is – as is *SOX2* – a TF expressed in immature neurons [224] and was nominally significant as a

neurogenesis TF in the fetal G2/M module (*p-value*=7.0×10^{-4}, α=0.05, corrected for 1556 TFs tested) and again nominally significant for the adult Interphases module (*p-value*=5.0×10^{-5}, α=0.05). During fetal age, *PAX6* shared target genes with other TFs regulating neurogenesis (such as *ASCL1, FOXM1* and *SOX2*; Figure 21A). Later in life (12 years and older), the interactions with *SOX2* increased whereas overlap of target genes with *ASCL1* and *FOXM1* was lost (Figure 21B). In contrast, *FOXO1* gained shared target genes with *PAX6* and *SOX2* in the adult group (Figure 21B) after being isolated in fetal age.


A
FOXO1
C9orf6
ASCL1
SOX2
PAX6
FOXM1

B

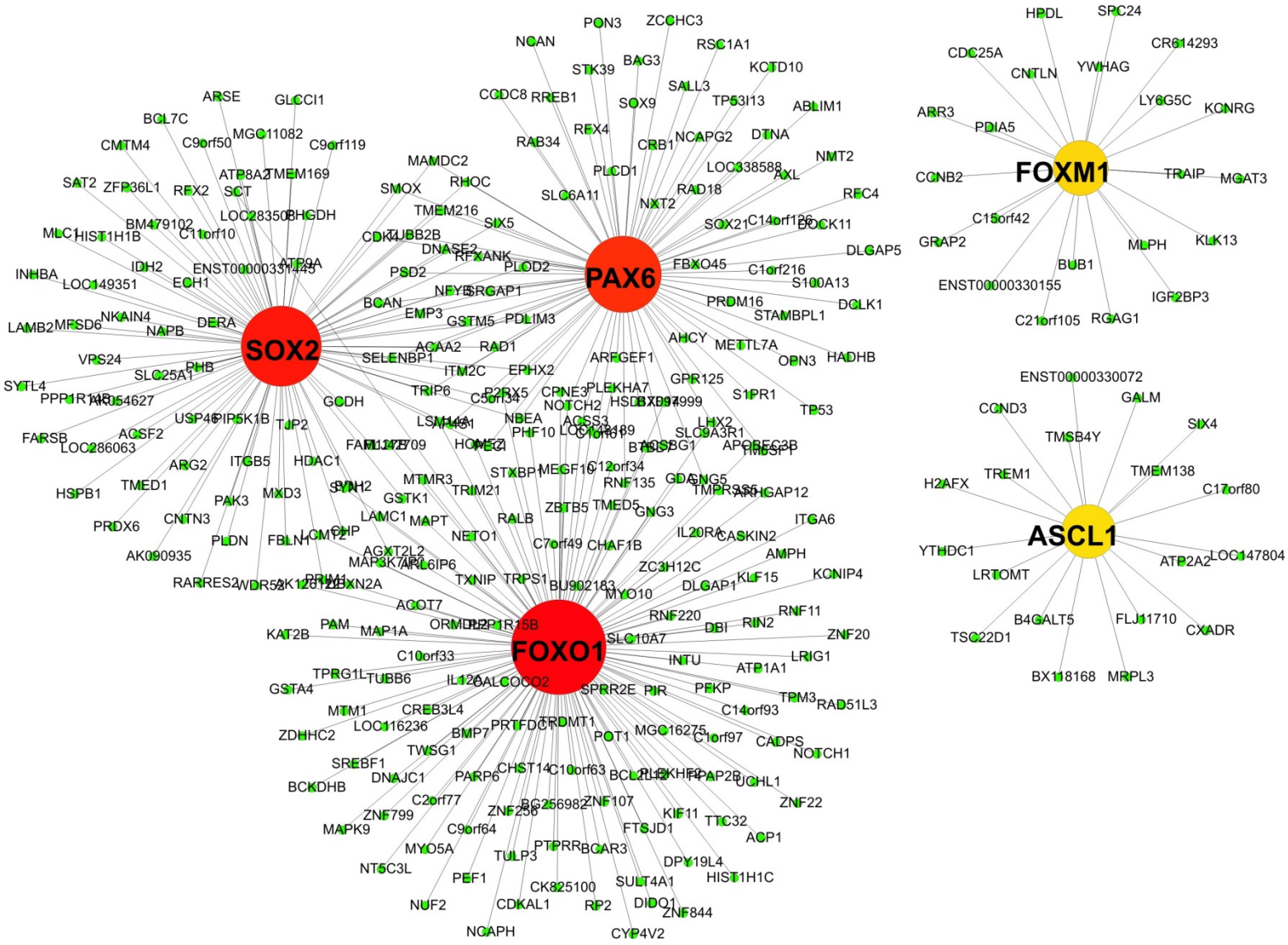

SOX2
PAX6
FOXO1
FOXM1
ASCL1
ARSE
GLCCI1
BCL7C
MGC11082
CMTM4
C9orf50
C9orf119
ATP8A2
TMEM169
SAT2
ZFP36L1
RFX2
SCT
LOC283508
BHGDH
BM479102
C11orf10
MLC1
HIST1H1B
IDH2
ENST00000331445
ATP9A
INHBA
LOC149351
ECH1
LAMB2
MFSD6
NKAIN4
NAPB
DERA
VPS24
SLC25A1
PHB
SYTL4
PPP1R14B
AK054627
USP48
PIP5K1B
TJP2
FARSB
LOC286063
ACSF2
ARG2
ITGB5
HDAC1
HSPB1
TMED1
PAK3
MXD3
PRDX6
CNTN3
PLDN
AK090935
FBLN1
RARRES2
WDR52
PRIM1
ZXDA
ZXDB
ZXDC
ZXN2A
KAT2B
PAM
MAP1A
ACOT7
ORMDL2
PPP1R15B
C10orf33
TPRG1L
TUBB6
GSTA4
MTM1
CREB3L4
ZDHHC2
LOC116236
BMP7
PRTFDC1
CRDMT1
TWSG1
SREBF1
DNAJC1
BCKDHB
PARP6
CHST14
C10orf63
C2orf77
BG256982
ZNF107
ZNF799
ZNF256
MAPK9
C9orf64
KIF11
TTC32
FTSJD1
ACP1
MYO5A
TULP3
PTPRR
BCAR3
NT5C3L
DPY19L4
PEF1
CK825100
SULT4A1
HIST1H1C
NUF2
CDKAL1
RP2
DIDO1
ZNF844
NCAPH
CYP4V2
PON3
ZCCHC3
NCAN
RSC1A1
STK39
BAG3
KCTD10
CCDC8
RREB1
SOX9
SALL3
TP53I13
ABLIM1
REX4
RAB34
CRB1
NCAPG2
DTNA
NMT2
PLCD1
LOC338588
AXL
SMOX
RHOC
RAD18
SLC6A11
NXT2
RFC4
MAMDC2
TMEM216
SIX5
SOX21
C14orf126
DOCK11
CDK4
TUBB2B
DNASE2
RFXANK
PLOD2
DLGAP5
PSD2
FBXO45
C1orf216
NFYB
BRCAP1
S100A13
BCAN
EMP3
GSTM5
PDLIM3
PRDM16
DCLK1
STAMBPL1
ACAA2
RAD1
AHCY
METTL7A
SELENBP1
ARFGEF1
OPN3
HADHB
ITM2C
EPHX2
GPR125
TRIP6
P2RX5
CRNE3
PLEKHA7
S1PR1
C5orf34
NOTCH2
HSD3B7
BX537999
TP53
GCDH
NBEA
AOSS3
LHX2
PHF10
SLC9A3R1
APOBEC3B
HMGSPT
STXBP1
MEGF10
C12orf34
RNF135
GDAP1
GNG5
MTMR3
TRIM21
ARHGAP12
GSTK1
ZBTB5
TMED5
GNG3
ITGA6
MAPT
RALB
LAMC1
IL20RA
CASKIN2
NETO1
C7orf49
CHAF1B
AMPH
AGXT2L2
MAP3K7AP2
ARL6IP6
ZC3H12C
TXNIP
TRPS1
BU902183
KLF15
KCNIP4
MYO10
DLGAP1
RNF220
RNF11
DBI
RIN2
SLC10A7
ZNF20
INTU
LRIG1
ATP1A1
GALCOCO2
SPRR2E
PIR
PFKP
TPM3
RAD51L3
C14orf93
C1orf97
CADPS
NOTCH1
POT1
MGC16275
UCHL1
ZNF22
FAP2B
BCL2L14
PLEKHF2
C10orf63
CHST14
PARP6
DNAJC1
BCKDHB
C2orf77
BG256982
ZNF107
ZNF256
ZNF799
MAPK9
C9orf64
KIF11
TTC32
FTSJD1
ACP1
MYO5A
TULP3
PTPRR
BCAR3
NT5C3L
DPY19L4
PEF1
CK825100
SULT4A1
HIST1H1C
NUF2
CDKAL1
RP2
DIDO1
ZNF844
NCAPH
CYP4V2
HPDL
SRC24
CR614293
CDC25A
CNTLN
YWHAG
LY6G5C
KCNRG
ARR3
PDIA5
CCNB2
TRAIP
MGAT3
C15orf42
GRAP2
MLPH
KLK13
BUB1
ENST00000330155
IGF2BP3
C21orf105
RGAG1
ENST00000330072
GALM
CCND3
SIX4
TMSB4Y
H2AFX
TREM1
TMEM138
C17orf80
YTHDC1
ATP2A2
LOC147804
LRTOMT
B4GALT5
FLJ11710
CXADR
TSC22D1
BX118168
MRPL3

Figure 21. Predicted target genes of five TFs identified as "neurogenesis TFs" for cell cycle modules during fetal life (A) compared to adult life (B). Nodes' color reflects outdegree, with large and red nodes having the largest number of direct edges in contrast to small and green nodes which have the smallest possible number (one edge per node). Yellow nodes are the intermediates between red and green. For simplicity, only direct predicted edges between the five TFs and their target genes are displayed.

In line with the observation that the same TFs, e.g. *SOX2* and *EMX2*, changed the majority of their target genes from fetal to adult life, the change in predicted target genes shared between two TFs indicates that neurogenesis TFs play different roles across the lifespan. It has been shown that neurogenesis TFs fulfil various functions linked to neurons' life cycle. For example, as described before, *FOXO1* has been associated with lifelong neurogenesis, neuronal polarity, and neuronal death [217]. Obviously, it would be expected that gene regulation in the ageing brain is more strongly influenced by neuronal death than neurogenesis when compared to fetal brains. The observed moderate-to-strong preservation of fetal cell cycle modules in adult brain, as well as replication of neurogenesis TFs, e.g. *SOX2*, could indicate that, although the coordination of cell cycle modules is no longer sufficiently strong enough to be detected *de novo*, the cell cycle modules retain aspects of their embryonic regulatory patterns during ageing. Potentially, these patterns could contribute to the ability of brain tissue lying outside of neurogenic niches, specifically the prefrontal cortex, to react to injury or to neurodegeneration.

4.4 Discussion

Neurogenesis in the human adult brain is still poorly understood. Whereas the adult CNS was originally believed to consist of non-renewable elements, the discovery of neurogenic niches, namely forebrain subventricular zone (SVZ) and dentate gyrus subgranular zone (SGZ) of the hippocampus, proved that neurogenesis is a lifelong process [207, 225]. Recently, an increasing number of brain areas have been associated with adult neurogenesis [194], including cortex [195, 196], amygdala [197-199], hypothalamus [200, 201], striatum [202, 203], and substantia nigra [204, 205]. In turn, neurogenesis can be induced outside of known neurogenic zones, using transplantation of niche-astrocyte-produced signals [206].

However, compared to fetal development, which is specialized for cell proliferation [194], the adult brain consists predominantly of mature neurons and quiescent cells and supports neurogenesis only at low levels or under nonphysiological conditions [194, 226]. It is still unknown whether the low level of neurogenesis reflects a limitation of either endogenous neural precursors' potential or microenvironmental signals required for neuronal differentiation and survival [227]. Maintenance of these microenvironments that support proliferation and differentiation of neural stem cells across the lifespan, or adult neurogenic niches, is potentially regulated by TFs, including the *FOXO* family [217]. It has been suggested that within the neurogenic niches, SVZ and SGZ, molecular morphogens and signals from embryogenesis are maintained and are key to adult neurogenesis [225]. Preservation of microenvironments that have the potential to support neurogenesis could be a necessary prerequisite for cells' ability to respond to mitogenic signals and to subsequently reenter cell cycle following brain injury or neurodegeneration. Here, I showed

that transcriptional modules enriched for cell cycle genes with strong cohesiveness during fetal age are preserved during adult age, although their coordination of expression is no longer sufficiently strong to detect *de novo* cell cycle modules. This would be expected as the fetal brains are 'poised' for development, in contrast to adult brains which support only low levels of cell proliferation yet or maintain the ability to respond to aversive stimuli, such as injury, with neurogenesis [194, 226]. Some TFs regulated a significant subset of the cell cycle modules' genes during fetal but also during adult life (e.g. *SOX2, SOX9, EMX2*), although they changed most of their predicted target genes within the two cell cycle modules. Different known neurogenesis TFs were found to be cell cycle-module specific only during adult life (e.g. *FOXO1* [217]). The preservation of cell cycle module-specific TFs as well as regulation by known neurogenesis TFs indicate that regulatory TF-gene interactions are partially maintained during ageing in the prefrontal cortex. However, as the prefrontal cortex has no known neurogenic area harbouring comparably strong *de novo* neurogenesis during adulthood, the observed transcriptional coordination between neurogenesis genes, and genes working in the same transcriptional subnetwork, might indicate the maintenance of a microenvironment that facilitates plasticity or neurogenesis in response to occasional brain injury rather than spontaneous, continuous neurogenesis. The shift in TF-target genes from fetal to adult age (e.g. *SOX2* and *EMX2*) further indicates that the identified neurogenesis TFs could play pleiotropic functional roles, in line with previous studies (e.g. [217]). However, the transition of TF-target genes from fetal to older age would need (*i*) replication in an independent data set with sufficient sample size covering the same age range, and (*ii*) experimental validation based for example on ChIP-Seq data for the given TFs.

The present study was limited by the availability of suitable ChIP-Seq data. Ideally, the computational predictions of TF-targets would be validated using ChIP-Seq data from the same cell type, namely adult neuronal cells, potentially from consented tissue donations during surgery. The publicly available data was from a different cell type, and adapted from a different organism (see **Chapter 4.5.1**). For this reason, the non-significant overlap of ChIP-Seq peaks with basal regulatory domains of predicted *FOXO1*-target genes may result from differences in the underlying cell type and model organism (mouse) compared to human. In addition, adult life is characterized by more subtle gene expression changes than the strong dynamics observed during embryogenesis and before or after birth. Therefore, TF-target predictions during adult life rely on a relatively low signal-to-noise ratio and could be associated with a higher error rate compared to fetal life. Consequently, independent data used for validation – in this case ChIP-Seq data – would need to be well represented by the sample from where transcriptional data was collected to capture the signal of the true positives.

The two modules linked to "cell cycle" contained a large number of genes. Potentially, the high number of genes in the two modules – which were derived from fetal age – reflects "poising" of the fetal cortex for cell proliferation. Clearly, cortical growth is one of the central biological functions during fetal age. For this reason, a large number of genes might be involved in neuronal growth, or are coordinated with genes directly involved in cell cycle. However, some of these genes might switch to other functions later in life which would decrease the signal-to-noise ratio in the adult data. In addition, although many of the TFs identified as "neurogenesis TFs" are known to play a role in neurogenesis (Supplementary Table S4.5), it would be possible that a TF identified as adult

"neurogenesis TF" regulates a subset of genes in the cell cycle-modules which switch to alternative functions during brain ageing. Clearly, the present data does not allow inference of causal effects of the predicted target genes on cell proliferation. The question whether these genes maintain an ability to revert back to their earlier functions for neurogenesis, for example in response to brain injury, would need to be tested experimentally, such as by assessing TF-target genes' expression levels in a model organism before and after applying an insult to brain tissue.

Heterogeneous processes linked to neuro- and gliogenesis, cell cycle and cell proliferation [207] are potentially linked to a microenvironment that is supportive of neurogenesis. These microenvironments do not necessarily promote spontaneous, continuous neurogenesis, but could provide potential to react to aversive stimuli with neurogenic processes. Identifying key factors in regulating the maintenance of these microenvironments that support neurogenesis could be crucial for the development of approaches that induce neuronal proliferation after brain injury such as strokes or transplantation. Experimental approaches could aim to test whether injury of brain tissue can revert these regulatory patterns to their earlier functions, for example as would be reflected by neurogenesis TFs that are maintained over the adult lifespan regaining control of their target genes from earlier age groups.

The present study aimed to investigate whether modules linked to cell cycle are preserved outside of known neurogenic niches. Future work could aim to compare the present results from prefrontal cortex to transcriptional networks and their TF-target interactions within zones known to have neurogenic potential during ageing, such as regions of the

hippocampus. Ideally, module preservation between transcriptional modules inside and outside of neurogenic niches could further be compared to gene expression data from non-neuronal tissue of the same subjects. In this way, it would be possible to differentiate between general markers, or regulators, for cell proliferation and those that are specific to brain tissues and to neuro- or gliogenesis. In turn, depending on accuracy of dissection and the subsequent availability of tissue samples restricted to specific cell types, the analysis could potentially aim to investigate more narrowly defined aspects of the very heterogeneous processes linked to neuro- and gliogenesis in the brain.

4.5 Methods

Details of (*i*) microarray gene expression data set used to build the transcriptional networks (see **Chapter 2.1**), (*ii*) tools for testing for enrichment (GAT (see **Chapter 2.3.1**), DAVID, MGI phenotypes; see **Chapter 2.4**), and (*iii*) network algorithms (WGCNA and ARACNE; see **Chapter 2.5**), are described in detail in the indicated chapters. Master regulator analysis was performed as described in **Chapter 3.5.7**.

4.5.1 ChIP-Seq data

I used publicly available ChIP-Seq data to validate predicted TF targets (for details on ChIP-Seq data, see **Chapter 2.2**). If available I annotated peaks from cell lines provided in the ENCODE data [137]; alternatively I searched for ChIP-Seq data for predicted TFs in the literature. All peaks were annotated using HOMER (v4.3 [189]; see **Chapter 2.3.2**).

- For *FOXM1*, I used two replicates of broad ChIP-Seq peaks from GM12878 (B-lymphocyte, lymphoblastoid; 34792 peaks in one replicate and 20154 peaks in the second replicate) in the ENCODE data.
- For *FOXO1*, I used peaks from 2 replicates (n=4013 and n=3021 peaks) from CD8+ cells (GSM141666 and GSM141667). Genome coordinates were converted from mm9 (*Mus musculus*) to hg19 (*Homo sapiens*) using the online liftOver tool.

4.5.2 Functional enrichment: GO-terms and MGI phenotypes

I tested for enrichment of functional terms using DAVID and MGI phenotypes as described in **Chapter 2.4**. Here, I tested separately for functional enrichment in both cell cycle modules. More precisely, the gene sets comprising the two modules were each tested against a background of all brain-expressed genes that were initially used to build the transcriptional network.

4.5.3 Enrichment of genomic regions

I tested for the enrichment of ChIP-Seq peaks within the basal regulatory domains of TF-predicted target genes. Basal regulatory domains include 5kb upstream and 1kb downstream of a gene's transcription starts site (TSS) [121]. Definition of basal regulatory domains are described in detail in **Chapter 2.4.5**.

4.5.4 TF-binding motif enrichment

The MEME Suite (version 4.10.0 [221]) was used to test for the enrichment of *FOXO1*-binding motifs within the sequences of basal regulatory domains (5kb upstream and 1kb downstream of transcription starts site (TSS) [121]) of *FOXO1*-predicted target genes compared to a background of sequences of basal regulatory domains of all brain-expressed genes initially used to build the transcriptional networks. Fasta files, containing sequence information, for basal regulatory domains of the predicted TF-target genes and the whole set of brain-expressed genes were generated using the FastaFromBed tool from the Bedtools Suite (v2.11.2). The consensus sequence of a binding motif for a transcription factor can be derived by (*i*) experimentally generating a library of short DNA sequences

directly bound by a transcription factor and subsequently (*ii*) performing a multiple sequence alignment of the identified DNA sequences. The conserved DNA sequence at each position in the motif can then be written using IUPAC (International Union of Pure and Applied Chemistry) codes. This notation uses the characters G, C, A, and T to refer to the four nucleotides commonly found in DNA. Ambiguities can be expressed using additional letters, for examples IUPAC code "N" indicates G, C, A, or T [228]. The IUPAC codes for two *FOXO1*-motifs in vertebrates (NRWAAACAAN and GNNTTGTTTACNTT) were downloaded from the TRANSFAC data base [222], and transformed into MEME motif files using the MEME Suite tools. No *FOXO1*-motifs specific to humans were available. A tool from the MEME suite, AME (Analysis of Motif Enrichment) [223], was used to assess p-values for enrichment of motifs within the *FOXO1*-predicted target genes compared to a background of all brain-expressed genes. The analysis using AME was run separately for both *FOXO1*-motifs. Briefly, AME takes as input a set of DNA-binding motifs and scores a set of DNA sequences, whereby each position in the sequence is treated as a potential starting point for a binding event. Here, I used default settings, in which AME performs a Fisher exact test for the binding events of each motif versus the number of binding events in the background sequence set [223].

4.6 Appendix

Table S4.1. Enrichment of GO-terms in the first module linked to "cell cycle", referred to as "Interphases module".* Categories highlighted in yellow are non-redundant GO-terms for the given modules, summarized by REViGO [117] (see Chapter 2.4.1).

Term	Term Name	gene count	fold	p-value (Bonferroni corrected)
GO:0007049	cell cycle	151	1.5	2.95E-04
GO:0000278	mitotic cell cycle	80	1.7	9.27E-03
GO:0022403	cell cycle phase	86	1.6	1.14E-02
GO:0022402	cell cycle process	111	1.5	1.86E-02
GO:0051301	cell division	63	1.7	4.20E-02
GO:0030529	ribonucleoprotein complex	101	1.5	3.53E-02

*2445 genes in the module had annotated DAVID IDs compared to a background of 16475 brain-expressed genes used to build the transcriptional network with annotated DAVID IDs.

Table S4.2. Enrichment of GO-terms in the second module linked to "cell cycle", referred to as the "G2/M module".* Categories highlighted in yellow are non-redundant GO-terms for the given modules, summarized by REViGO [117] (see Chapter 2.4.1).

Term	Term Name	gene count	fold	p-value (Bonferroni corrected)
GO:0022403	cell cycle phase	72	3.8	1.3E-19
GO:0007049	cell cycle	101	2.8	2.2E-18
GO:0000279	M phase	61	4.1	3.3E-18
GO:0022402	cell cycle process	82	3.1	5.9E-17
GO:0000278	mitotic cell cycle	62	3.6	2.8E-15
GO:0007067	mitosis	46	4.7	3.6E-15
GO:0000280	nuclear division	46	4.7	3.6E-15
GO:0000087	M phase of mitotic cell cycle	46	4.6	7.1E-15
GO:0048285	organelle fission	46	4.5	2.7E-14
GO:0044427	chromosomal part	62	3.4	7.8E-15
GO:0005694	chromosome	68	3.1	1.2E-14
GO:0051301	cell division	51	3.9	2.8E-13
GO:0006260	DNA replication	33	3.8	2.2E-07
GO:0000793	condensed chromosome	26	4.5	1.6E-07
GO:0043228	non-membrane-bounded organelle	181	1.5	6.2E-07
GO:0043232	intracellular non-membrane-bounded organelle	181	1.5	6.2E-07
GO:0065004	protein-DNA complex assembly	20	5.5	4.1E-06
GO:0000785	chromatin	32	3.3	2.4E-06
GO:0006334	nucleosome assembly	18	5.3	5.3E-05
GO:0031497	chromatin assembly	18	5.1	8.8E-05

GO:0006323	DNA packaging	21	4.4	9.1E-05
GO:0006259	DNA metabolic process	53	2.2	1.3E-04
GO:0007017	microtubule-based process	34	2.9	1.3E-04
GO:0032993	protein-DNA complex	18	5.0	2.6E-05
GO:0051276	chromosome organization	50	2.3	1.9E-04
GO:0034728	nucleosome organization	18	4.7	3.6E-04
GO:0000786	nucleosome	15	5.7	7.1E-05
GO:0000777	condensed chromosome kinetochore	15	5.6	9.5E-05
GO:0006333	chromatin assembly or disassembly	21	3.9	7.0E-04
GO:0000775	chromosome, centromeric region	22	3.7	1.2E-04
GO:0005819	spindle	23	3.5	1.9E-04
GO:0007346	regulation of mitotic cell cycle	24	3.4	1.4E-03
GO:0000776	kinetochore	17	4.5	2.5E-04
GO:0000779	condensed chromosome, centromeric region	15	5.0	3.5E-04
GO:0006261	DNA-dependent DNA replication	14	5.4	2.4E-03
GO:0015630	microtubule cytoskeleton	52	2.0	6.1E-04
GO:0051726	regulation of cell cycle	38	2.4	4.0E-03
GO:0007051	spindle organization	11	6.2	1.4E-02
GO:0007059	chromosome segregation	15	4.1	3.0E-02
GO:0010564	regulation of cell cycle process	18	3.5	3.0E-02
GO:0007052	mitotic spindle organization	7	10.8	4.1E-02
GO:0000228	nuclear chromosome	21	2.8	2.6E-02

*868 genes in the module had annotated DAVID IDs compared to a background of 16475 brain-expressed genes used to build the transcriptional network with annotated DAVID IDs

Table S4.3. Transcription Factors ("TF") whose predicted target genes ("genes in regulon") show significant overlap with the Interphases module ("overlap with Interphases module") based on significant Fisher's exact test p-value ("FET p-value"; α=0.05 corrected for 1556 TFs in the transcriptional network generated by ARACNE).

Interphases module

Fetal age				**Adult age (>12 years)**			
TF	**genes in regulon**	**overlap with Interphases module**	**FET p-value**	**TF**	**genes in regulon**	**overlap with Interphases module**	**FET p-value**
SOX21	198	139	6.3E-72	*EMX2*	295	95	3.4E-15
HMGB2	234	141	5.7E-60	*SOX9*	215	69	2.4E-11
EMX2	348	174	3.1E-57	*TIPARP*	261	75	1.1E-09
KLF15	292	156	2.2E-56	*HES6*	193	60	1.9E-09
NR2E1	233	131	7.1E-51	*FOXO1*	356	91	1.3E-08
PAX6	296	150	2.3E-50	*HOMEZ*	125	42	4.3E-08
DZIP1L	262	138	5.6E-49	*TRPS1*	154	48	7.0E-08
ASCL1	120	88	2.0E-48	*HIST1H1C*	154	48	7.0E-08
USP39	154	98	5.2E-45	*ZIC1*	168	51	7.2E-08
RFX2	129	88	3.1E-44	*SOX2*	291	76	7.8E-08
NCOA1	345	154	8.2E-43	*TSC22D4*	482	112	8.4E-08
MGMT	219	114	5.8E-40	*ELAVL2*	216	60	1.8E-07
TCF19	241	118	6.9E-38	*DNAJC1*	218	60	2.6E-07
QKI	116	76	1.8E-36	*ZFHX4*	169	49	6.0E-07
PRDM2	151	88	3.4E-36	*ZNF107*	103	34	1.3E-06
ZNF22	107	71	1.0E-34	*ONECUT2*	131	40	1.5E-06
DEK	127	78	1.8E-34	*FOXN2*	235	61	1.8E-06
NR1D2	214	105	6.1E-34	*KLF15*	236	61	2.1E-06
ELF1	167	90	2.5E-33	*SOX21*	134	40	2.8E-06

RFXANK	159	87	6.2E-33	*GLI3*	116	36	3.3E-06
NOC3L	177	90	9.0E-31	*RFX2*	198	52	7.1E-06
BHLHE22	118	70	1.0E-29	*ZNF146*	115	34	1.9E-05
EMX1	336	130	1.0E-28				
DNAJC1	196	91	2.7E-27				
ZFHX4	163	80	3.6E-26				
MBD2	115	65	5.1E-26				
PRDM16	125	68	8.0E-26				
SOX9	99	59	2.1E-25				
HES1	143	72	1.5E-24				
ZNF217	189	85	2.1E-24				
SUZ12	136	69	7.9E-24				
ZNF143	73	48	1.3E-23				
CSRNP3	184	82	2.9E-23				
RFX5	262	102	4.9E-23				
TRPS1	81	50	1.0E-22				
ZGPAT	231	93	2.5E-22				
ZNF227	71	46	2.7E-22				
MBD4	142	68	1.1E-21				
MXD3	224	90	1.4E-21				
HOMEZ	107	56	2.0E-20				
PLAG1	224	88	2.3E-20				
RFX4	86	49	3.7E-20				
CREM	143	66	5.4E-20				
SALL1	97	52	1.1E-19				
HIVEP2	112	56	3.3E-19				
GLI3	205	81	5.0E-19				
ZFP30	48	33	1.4E-17				
ZNF367	266	92	6.1E-17				
ZNF573	145	62	7.5E-17				
ZNF706	264	91	1.1E-16				
ESRRG	199	75	2.2E-16				
ID4	85	44	4.0E-16				
DNMT3A	79	42	5.3E-16				

EIF3K	121	54	7.9E-16
MYBL2	156	63	1.2E-15
TIAL1	87	44	1.2E-15
ZNF628	164	64	5.1E-15
SIX5	185	69	6.5E-15
HBP1	92	44	1.7E-14
ZNF691	60	34	2.6E-14
ZNF174	77	39	4.8E-14
ZC3HAV1	52	31	5.2E-14
CREB3L4	113	49	6.7E-14
FOXJ1	81	40	6.8E-14
ZNF256	59	33	1.0E-13
KLF6	126	52	1.3E-13
ZNF337	115	49	1.5E-13
NFYB	34	24	1.6E-13
AKNA	97	44	1.8E-13
BCL6	237	78	2.8E-13
ZFP36L1	77	38	2.9E-13
TUT1	107	46	5.7E-13
SATB1	103	44	2.3E-12
KLF9	115	47	2.9E-12
NFATC4	127	50	3.2E-12
DRAP1	125	49	6.4E-12
ETV6	98	42	6.5E-12
MYBL1	47	27	7.7E-12
TGIF1	152	55	1.4E-11
ZFP36L2	112	45	1.7E-11
ZBTB25	58	30	2.0E-11
RAD51	243	75	2.8E-11
ZNF429	76	35	3.1E-11
TAL2	159	56	3.2E-11

FBXO41	195	64	4.4E-11
SNAI2	74	34	6.4E-11
E2F6	147	52	1.3E-10
U2AF1	303	86	1.3E-10
TSC22D1	110	43	1.4E-10
CHD7	76	34	1.6E-10
MBNL2	99	40	1.8E-10
ELAVL2	66	31	2.2E-10
ARNT	66	31	2.2E-10
TSHZ3	104	41	2.7E-10
ZEB1	93	38	3.5E-10
NFIL3	50	26	3.7E-10
ZSCAN5A	72	32	6.7E-10
SLC2A4RG	91	37	7.0E-10
AHR	52	26	1.1E-09
ARID5B	110	41	2.0E-09
MYT1L	79	33	2.6E-09
ARNTL2	145	49	2.7E-09
ZNF219	154	51	2.8E-09
PROX1	104	39	4.0E-09
HLX	55	26	5.3E-09
REPIN1	217	64	5.9E-09
BDP1	70	30	6.5E-09
SCRT2	158	51	7.5E-09
TERF2	228	66	7.8E-09
THRB	209	62	7.9E-09
ZMAT1	49	24	8.4E-09
FOXN2	182	56	9.9E-09
HES6	79	32	1.1E-08
THRA	138	46	1.3E-08
TCF4	57	26	1.4E-08

MAFF	108	39	1.4E-08
SSRP1	60	26	5.0E-08
CUX2	122	41	6.1E-08
GPR155	136	44	7.3E-08
HEY2	89	33	8.4E-08
ZNF99	217	61	9.1E-08
PDCD6	128	42	9.2E-08
POLE3	174	52	9.5E-08
TRIM23	200	57	1.4E-07
CAMTA1	31	17	1.5E-07
SOX2	198	56	2.4E-07
INSM1	129	41	3.4E-07
ZNF626	65	26	3.5E-07
ZNF534	40	19	5.5E-07
ARID3B	96	33	6.5E-07
NFE2L2	165	48	7.0E-07
TCF7L1	92	32	7.0E-07
NR3C2	190	53	8.1E-07
ZNF107	52	22	8.7E-07
PRDM8	60	24	9.7E-07
FOXP1	102	34	9.9E-07
EXOC2	85	30	1.1E-06
GTF2F2	107	35	1.1E-06
ZFPM2	77	28	1.2E-06
HOPX	28	15	1.3E-06
ZNF454	32	16	1.9E-06
RPA2	79	28	2.2E-06
ZBTB41	177	49	2.6E-06
AKAP8	63	24	2.7E-06
THAP10	275	68	3.1E-06
PLXNC1	178	49	3.1E-06

HSF1	89	30	3.3E-06
LARP6	116	36	3.3E-06
ZNF764	99	32	4.3E-06
ZNF426	41	18	4.5E-06
SIM1	104	33	4.8E-06
SMARCA5	49	20	5.3E-06
MEF2C	78	27	5.6E-06
RNF166	143	41	6.5E-06
NFIB	208	54	6.8E-06
ZIC3	64	23	1.4E-05
TP63	86	28	1.4E-05
DACH1	173	46	1.7E-05
RC3H2	61	22	1.9E-05
SETBP1	101	31	2.0E-05
MST1R	57	21	2.1E-05
SP110	301	70	2.1E-05
ZFYVE20	274	65	2.1E-05
YEATS2	79	26	2.3E-05
SREBF1	107	32	2.6E-05
NFE2L3	103	31	3.0E-05
ZNF714	42	17	3.1E-05

Table S4.4. Transcription Factors ("TF") whose predicted target genes ("genes in regulon") show significant overlap with the G2/M module ("overlap with G2/M module") based on significant Fisher's exact test p-value ("FET p-value"; α=0.05 corrected for 1556 TFs in the transcriptional network generated by ARACNE).

G2/M module							
Fetal age				**Adult age (>12 years)**			
TF	**genes in regulon**	**overlap with Interphases module**	**FET p-value**	**TF**	**genes in regulon**	**overlap with Interphases module**	**FET p-value**
HMGA2	275	146	3.9E-115	*FOXO1*	356	38	1.6E-05
TGIF2	216	115	1.7E-90	*ZNF114*	76	14	2.8E-05
E2F2	240	119	7.3E-89				
KNTC1	261	123	1.9E-88				
RAD51	243	112	4.2E-79				
TP53	132	85	9.4E-77				
FOXM1	193	90	4.6E-64				
MXD3	224	82	2.7E-48				
FOXN4	238	79	6.8E-43				
PSMD12	134	60	1.3E-41				
ZNF20	248	76	1.6E-38				
NR3C2	190	66	2.3E-37				
ADAMTS17	293	81	2.5E-37				
TFAP2C	230	69	2.2E-34				
TFAP2E	191	62	3.4E-33				
CBX2	133	51	1.6E-31				
EOMES	127	48	2.1E-29				
NHLH1	156	52	1.2E-28				
DEPDC7	127	47	2.5E-28				
EZH2	345	76	6.0E-28				
ZNF385B	176	53	9.6E-27				

GLI3	205	57	1.1E-26
TFDP2	157	49	1.4E-25
TIGD4	157	49	1.4E-25
DEPDC1	134	45	4.1E-25
ZBTB47	189	53	4.4E-25
BACH1	123	42	7.6E-24
ZNF367	266	61	1.3E-23
E2F7	110	37	6.6E-21
GLI1	82	32	1.6E-20
RCOR2	140	41	1.8E-20
ZNF491	83	32	2.5E-20
MYT1	180	46	4.0E-20
MKRN3	96	33	4.0E-19
ZBTB16	77	29	3.2E-18
WDHD1	229	49	7.4E-18
SOX5	130	35	2.2E-16
TSHZ2	186	42	2.2E-16
HIST1H1D	106	31	8.8E-16
ZNF197	250	48	1.5E-15
ZNF780A	156	37	2.6E-15
ARID2	104	30	3.9E-15
E2F5	95	28	1.8E-14
EMX2	348	55	7.5E-14
E2F8	131	32	8.2E-14
GABPB1	309	51	1.1E-13
THRA	138	32	3.9E-13
NR5A1	174	36	5.3E-13
HESX1	193	38	6.1E-13
TCF19	241	43	6.4E-13
ATF1	119	29	1.3E-12
ZNF425	75	23	1.5E-12

SEMA4A	156	32	1.3E-11
ZNF572	72	21	4.1E-11
ZNF514	67	19	6.0E-10
KLF16	152	29	7.3E-10
ARNTL2	145	28	1.1E-09
ZNF443	102	23	1.4E-09
MYB	177	31	1.7E-09
FOXN2	182	31	3.5E-09
ZNF516	52	16	3.6E-09
SORBS2	144	27	4.2E-09
ZNF700	195	32	5.0E-09
HMGB3	147	27	6.6E-09
THAP9	37	13	1.8E-08
NFYA	166	28	2.5E-08
ZNF93	325	42	3.4E-08
ZNF493	68	17	3.8E-08
DEPDC4	33	12	4.0E-08
LARP6	116	22	9.1E-08
ZNF155	73	17	1.2E-07
ZNF175	205	30	2.1E-07
DACH1	173	27	2.3E-07
NR4A2	184	28	2.3E-07
DNMT3A	79	17	4.0E-07
ZNF397	136	23	4.1E-07
RBM20	91	18	6.9E-07
ZNF385A	110	20	7.0E-07
WHSC1	111	20	8.2E-07
NRL	111	20	8.2E-07
TIA1	132	22	9.4E-07
SOX3	93	18	9.7E-07
ZBTB40	43	12	1.1E-06

CUL4A	43	12	1.1E-06
AKNA	97	18	1.8E-06
GLI2	131	21	3.1E-06
STAT6	215	28	5.5E-06
SCRT2	158	23	5.9E-06
ZNF492	105	18	6.0E-06
POU3F1	243	30	7.6E-06
ZSCAN16	78	15	8.4E-06
DDIT3	82	15	1.6E-05
OTX1	124	19	1.8E-05
FGD1	183	24	2.2E-05
BATF3	85	15	2.5E-05
ZNF530	76	14	2.8E-05

Table S4.5. Putative functions of 23 TFs predicted to regulate Interphases – and G2/M module genes during adulthood, based on literature information.

TF	Keywords Function	Reference
EMX2	CNS development, cell proliferation	[229]
SOX9	Neuron-glia switch	[230]
TIPARP	Skeletal development, protein biosynthesis	[231]
HES6	Neuronal differentiation	[232]
FOXO1	Lifelong neurogenesis, neuronal polarity, and neuronal death	[217]
HOMEZ	Expressed in the brain	[233]
TRPS1	Skeletal and hair development	[234]
HIST1H1C	Apoptosis	[235]
ZIC1	Neurogenesis	[236]
SOX2	Neurogenesis	[237]
TSC22D4	Pituitary organogenesis	[238]
ELAVL2	Neuronal degeneration	[239]
DNAJC1	Candidate tumour suppressor gene	[240]
ZFHX4	Neural and muscle differentiation	[241]
ZNF107	Cell proliferation and survival	[242]
ONECUT2	Liver differentiation and metabolism	[243]
FOXN2	Leukaemia	[244]
KLF15	Clock-dependent oscillator, astrocyte development	[245, 246]
SOX21	Neuronal differentiation	[247]
GLI3	Limb development	[248]
RFX2	Neurosphere formation	[249]
ZNF146	Carcinogenesis	[250]
ZNF114	Oligodendroglioma progression-ass. candidate gene	[251]

Chapter 5 - Brain-expressed 3'UTR extensions strengthen miRNA cross-talk between ion channel/transporter encoding mRNAs

5.1 Abstract

Why protein-coding genes express transcripts with longer 3' untranslated regions (3'UTRs) in the brain rather than in other tissues remains poorly understood. Given the established role of 3'UTRs in post-transcriptional regulation of transcript abundance and their recently highlighted contributions to miRNA-mediated cross-talk between mRNAs, I hypothesized that 3'UTR lengthening enhances coordinated expression between functionally-related genes in the brain.

To test this hypothesis, I annotated 3'UTRs of human brain-expressed genes and found that transcripts encoding ion channels or transporters are specifically enriched among those genes expressing their longest 3'UTR extension in this tissue. These 3'UTR extensions have high density of response elements predicted for those miRNAs that are specifically expressed in the human frontal cortex. Importantly, these miRNA response elements are more frequently shared among ion channel/transporter-encoding mRNAs than expected by chance. This indicates that miRNA-mediated cross-talk accounts, at least in part, for the observed coordinated expression of ion channel/transporter genes in the adult human brain.

I conclude that extension of these genes' 3'UTRs enhances the miRNA-mediated cross-talk among their transcripts which post-transcriptionally regulates their mRNAs' relative levels.

The following chapter is a modification of our paper published in Frontiers in Genetics (2014) [252]. All analysis done by co-authors is explicitly indicated in **Chapter 5.5**. I wrote the first draft of the paper and created the figures. The final draft was completed with Ana Marques and Chris Ponting.

5.2 Introduction

Post-transcriptional regulation of transcript abundance can be achieved by altering RNA stability or translation [253]. Untranslated regions (UTRs) located at 5' and 3' ends of mRNAs contain the majority of post-transcriptional regulatory elements. Of these, the 3'UTRs contain the highest density of motifs contributing to mRNA stability, translation and localization [254, 255]. A preeminent class of such motifs are miRNA recognition elements (MREs) that bind to near complementary miRNAs (~22 nucleotide noncoding RNAs), which in turn induce either degradation or translational inhibition of their cognate mRNAs targets [256]. Conserved miRNA sites in 3'UTRs experience stronger negative selection compared to other functional elements [257]. The significance of miRNAs in regulating transcript abundance is exemplified by the prediction of their MREs in over 60% of human genes [9] and their demonstrated roles in many aspects of neuronal cell physiology, including terminal neuronal differentiation [258], polarization of neurons [259], axonogenesis [260], neuroplasticity and stress response [261].

Spatially and temporally regulated expression of alternative 3'UTRs, resulting either from alternative cleavage and polyadenylation or alternative splicing, are a common feature of animal genes [262, 263]. Alternative 3'UTR usage is often preserved across evolution as indicated by, for example, similar levels of between-species sequence conservation at proximal and distal poly(A) sites [264]. Differences in 3'UTR length are in many cases correlated with transcript abundance [265]. The 3'UTRs of transcripts that are expressed in the brain tend to be longer than those expressed in other tissues, a property that is widely conserved across evolution [265-269]. For example transcripts whose proteins have roles in development, morphogenesis and signal transduction express longer 3'UTRs compared to

transcripts for proteins involved in metabolism or RNA processing [270] and 3'UTRs of genes encoding ribosomal genes have been reported to be ~6 fold shorter than 3'UTRs of neurogenesis genes [256].

Here I sought to address why such transcripts might require longer 3'UTRs using an assumption that their greater lengths reflect a greater degree of posttranscriptional regulation by miRNAs. Long 3'UTRs of neuronally expressed isoforms are enriched in conserved binding sites (MREs) for well-known neural miRNAs [269] that are thought to contribute to the fine-tuning of their encoded products' abundance. For example, the predominant *MeCP2* isoform in human fetal brain contains an unusually long 3'UTR [271, 272] which harbors functional binding sites for miR-483-5p, a fetal brain-enriched miRNA that post-transcriptionally represses *MeCP2* in this tissue [273]. Interestingly, miR-483-5p also regulates transcript levels of other *MeCP2* protein interacting partners [273]. This suggests that differential 3'UTR usage may fine-tune gene-gene interactions via miRNA-mediated modulation of their transcripts' abundance [273].

The ability of mRNAs to compete for miRNAs depends both on their relative abundance and their MRE repertoire [274]. The latter is determined, largely, by the sequence and length of their 3'UTRs [275] which control, in a miRNA-dependent manner, their relative abundance [276]. I hypothesized that genes with brain-specific 3'UTR extensions might engage in the miRNA-mediated cross-talk with other functionally related transcripts. Given that there is a known excess of long alternative 3'UTRs in brain-expressed isoforms [265-269], I expected such transcripts to contribute to brain-specific biological processes.

A key feature of brain cells, particularly neurons, is their ability to respond to electrochemical changes in their extracellular environment [277, 278]. Crucial to this

process are genes encoding ion channels and transporters. These proteins form, often gated, multi-subunit pores in the membrane which permit efficient and specific transport of extracellular ions [279]. Ion channels and transporters are present in most cell types [279] but their importance in the central nervous system is illustrated by the many channelopathies that affect these tissues [280]. Importantly, the transcript abundance of ion channel encoding genes is influenced by environmental cues, including the electrical activity present within the cell [281]. Given that ion channels require the coordinated assembly of several protein subunits to convey their function, transcript abundance may be tightly regulated to ensure proper channel assembly and stoichiometry. Such coordination can be achieved via miRNA-mediated post-transcriptional regulation. For example, miR-103 simultaneously regulates the abundance of at least three transcripts encoding subunits of the $Ca_v1.2$ L-type calcium channel, and modulates its activity [282]. Variants in this channel play a role in chronic neuropathic pain [282] as well as in Timothy's [283] and Brugada Syndrome [284].

Here I investigated the biological relevance of 3'UTR lengthening among transcripts expressed in the human adult brain. Genes that express their longest alternative 3'UTR in this organ were found to be substantially enriched for those encoding ion channels and transporters. I provide evidence that extended 3'UTR transcription in the brain provides an additional miRNA-mediated layer of posttranscriptional regulation of these genes' transcript levels.

5.3 Results

5.3.1 Annotation and characterization of human genes 3' untranslated regions

Long poly-A^+ selected single- and paired-end RNA sequencing data from 16 normal human tissues was used to annotate transcripts *de novo* using Cufflinks (version 1.3.0 [114], Figure 26A). For simplicity, I only considered genes whose complete set of protein-coding transcripts contain the same constitutive stop codon. Exonic nucleotides downstream of these stop codons form part of these protein-coding transcripts 3' untranslated region (3'UTR). For each gene, I defined its constitutive 3'UTR as all 3'UTR nucleotides expressed across the complete set of transcripts. The remaining 3'UTR nucleotides were defined as contributing to alternative 3'UTRs. For each gene I also defined its longest alternative 3'UTR as its alternative 3'UTR with the largest number of nucleotides observed in any tissue (Figure 26B).

I next sought to compare the density of functional elements in alternative and constitutive 3'UTRs using the derived allele frequency (DAF) test [257]. A fraction of single nucleotide polymorphisms (SNPs) with low allele frequencies that is higher than found for SNPs in neutrally evolving sequence suggests that negative selection has acted to prevent the accumulation of loss-of-function mutations, an evolutionary signature of functionality [257]. I compared the ratio, *r*, of rare (DAF<10%) to intermediate (10%<DAF<90%) frequency SNPs [285, 286] found in constitutive or longest alternative 3'UTRs against the ratio calculated for neutrally evolving sequences. As neutral proxy I used transposable element-derived sequences (Ancestral Repeats or ARs [287]) that inserted prior to the common ancestor of human and mouse. To ensure that ARs shared the same phylogenetic

history as the 3'UTRs, I selected ARs in the regions in the vicinity of 3'UTRs (20, 50 or 100 kilobases (kb)) [285]. 3'UTRs without ARs in their vicinity were excluded (n=4324 genes with longest alternative 3'UTRs contained ARs within 20kb vicinity). Both longest alternative (r =15304/6334=2.42; two tailed Fisher's exact test (FET) $p\text{-}value<7\text{x}10^{-3}$) and constitutive ($r$ =12900/5144=2.51; FET $p\text{-}value<2\text{x}10^{-5}$) 3'UTRs exhibit a moderate excess of SNPs at low frequency (DAF<10%) relative to ARs within 20kb (r =18143/7928=2.29), consistent with some nucleotides in these regions being preferentially preserved by natural selection. Constitutive 3'UTRs were no different to longest alternative 3'UTRs in this test (FET p-value=0.10), suggesting that they evolved under similar selective constraints. Similar results were obtained when ARs within 50kb and 100kb either up- or downstream of 3'UTRs were considered. I concluded that 3'UTR extensions are subject to selective constraint in humans, consistent with these regions harboring functionally conserved motifs.

Next, I evaluated tissue differences in 3'UTR usage. Alternative 3'UTRs expressed in the human adult brain are significantly longer than those found in other tissues (two-tailed Wilcoxon test, $p\text{-}value < 5\text{x}10^{-8}$), consistent with previous reports [268, 269]. The median length of brain-expressed alternative 3'UTRs (272 nucleotides) is 8% higher than for ovary (251 nucleotides, Figure 22A), the tissue with the next longest median alternative 3'UTRs, and 24% higher than the median length of 3'UTRs of transcripts from all other tissues (219 nucleotides, Figure 22A). The observed extension of brain-expressed alternative 3'UTRs could reflect a brain-specific usage of distal poly-adenylation signals or splice sites or be a consequence of technical bias, such as differences in RNA library preparation or sequencing coverage. To distinguish between these two explanations, I compared the ratio

between the expression levels of alternative and their corresponding constitutive 3'UTRs, across all tissues tested. Relative expression level (*E*) was defined as the ratio between the numbers of reads per nucleotide in alternative and constitutive 3'UTRs (see **Chapter 5.5.3.3**). I found that isoforms expressed in the brain have the highest relative expression of alternative compared to constitutive 3'UTRs (Figure 22B-C). More specifically, the median relative expression level of all genes with annotated alternative 3'UTRs was significantly higher in the brain (*E*=0.35) than the median relative expression of the other 15 tissues investigated (*E*=0.28; Figure 22B; Wilcoxon test, *p-value*<2.2×10^{-16}). This is consistent with brain-expressed transcripts having the longest alternative 3'UTRs owing to biological rather than technical reasons.

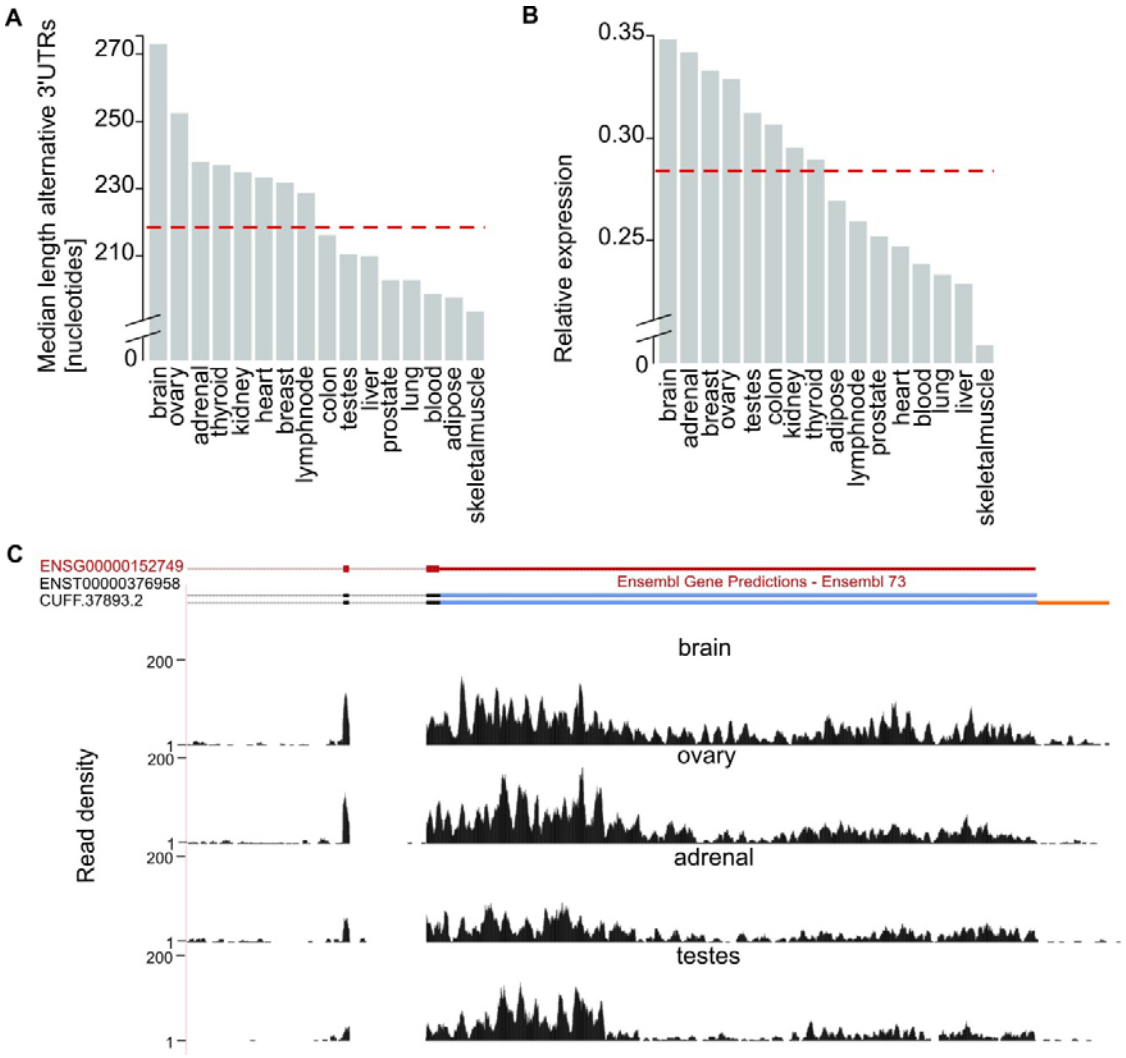


Figure 22. Brain-expressed isoforms have the longest 3'UTR extensions. A) Median length (in nucleotides, y-axis) of alternative 3'UTRs across 16 human tissues. The median across all non-neuronal tissues (219 nucleotides) is depicted as a horizontal dashed red line. B) Median fold difference in expression (calculated as the number of reads per nucleotide of expressed sequence) of alternative relative to constitutive 3'UTR (y-axis) across 16 tissues. The median across all non-neuronal tissues (median relative expression=0.28) is depicted as a horizontal dashed red line. C) G protein-

coupled receptor 180 (*GPR180*; ENSG00000152749; chr13:95,277,556-95,288,768, red) expresses its longest 3'UTR in the brain. Alongside support for the expression of the only transcript annotated by ENSEMBL for this gene (ENST00000376958) our approach allowed the annotation of an isoform with an extended 3'UTR in the brain (CUFF.37893.2). Read density (black, y-axis) across the last coding exons (black), the constitutive (blue) and extended (orange) portions of *GPR180* 3'UTR in brain, ovary, adrenal and testes.

5.3.2 Isoforms with long alternative 3'UTRs are often co-expressed and functionally related

Next I tested whether genes expressing isoforms with the longest alternative 3'UTRs in the brain are functionally related, contributing to the same biological process or being part of the same pathway. To do so I compared the frequency of gene ontology (GO) annotations (described in **Chapter 2.4**) for the 544 genes that express their longest alternative 3'UTRs only in the brain against GO-terms for all brain-expressed genes with annotated 3'UTRs.

After correcting for multiple testing (Bonferroni correction, α=0.05), I found that genes with the longest alternative 3'UTR isoform expressed in the brain were significantly enriched in 28 GO-terms that were related to either ion channel or transporter function, as summarized using REViGO ([117], Figure 23A). Almost half of the genes expressing their longest 3'UTR in the brain contributed to the observed enrichment (*n*=226 genes; Supplementary Table S5.2). The protein-coding genes that contribute to this enrichment are, for simplicity, hereafter termed *ion channel/transporter genes*. None of the terms that were enriched for genes with longest alternative 3'UTRs in brain was replicated in any of

the other 15 tissues tested (Supplementary Table S5.3). *Ion channel/transporter genes* tended to express especially long alternative 3'UTRs, showing a median length of n=1322 nucleotides compared to n=798 in the remaining n=322 genes with longest 3'UTRs only in brain (two-tailed Wilcoxon test, *p-value* < 0.05).

Relative to non-related genes, the expression of genes functioning in the same pathway or sharing biological roles is expected to be more strongly co-regulated. Indeed, I found that *ion channel/transporter genes* (n=192 genes associated with probes in the arrays used by [288]) showed significantly increased pairwise correlation in expression across the prefrontal cortex of postmortem brains from 148 healthy adult subjects [288] compared to randomly selected gene sets of brain expressed genes of the same sample size (median Spearman's correlation=0.0155; *empirical p-value*<0.001, Figure 23B). Similar results were found when I compared *ion channel/transporter genes* to random sets from genes expressing their longest 3'UTRs in the brain (n=1261 probes; *empirical p-value*<0.001, Figure 23B) or only in the brain (brain-specific; n=449 probes; *empirical p-value*<0.001, Figure 23B). The expression of genes contributing to the same GO-terms are more likely to be correlated with each other than random sets of brain-expressed genes. For that reason, I compared *ion channel/transporter genes* to random sets of brain-expressed genes annotated with the GO-term "ion channel activity" and found significantly increased correlation for *ion channel/transporter genes* (n=269 probes; *empirical p-value*<0.05, Figure 23B). The increased correlation in expression of *ion channel/transporter genes* indicates that extended alternative 3'UTRs facilitate coordination of these genes in the brain.

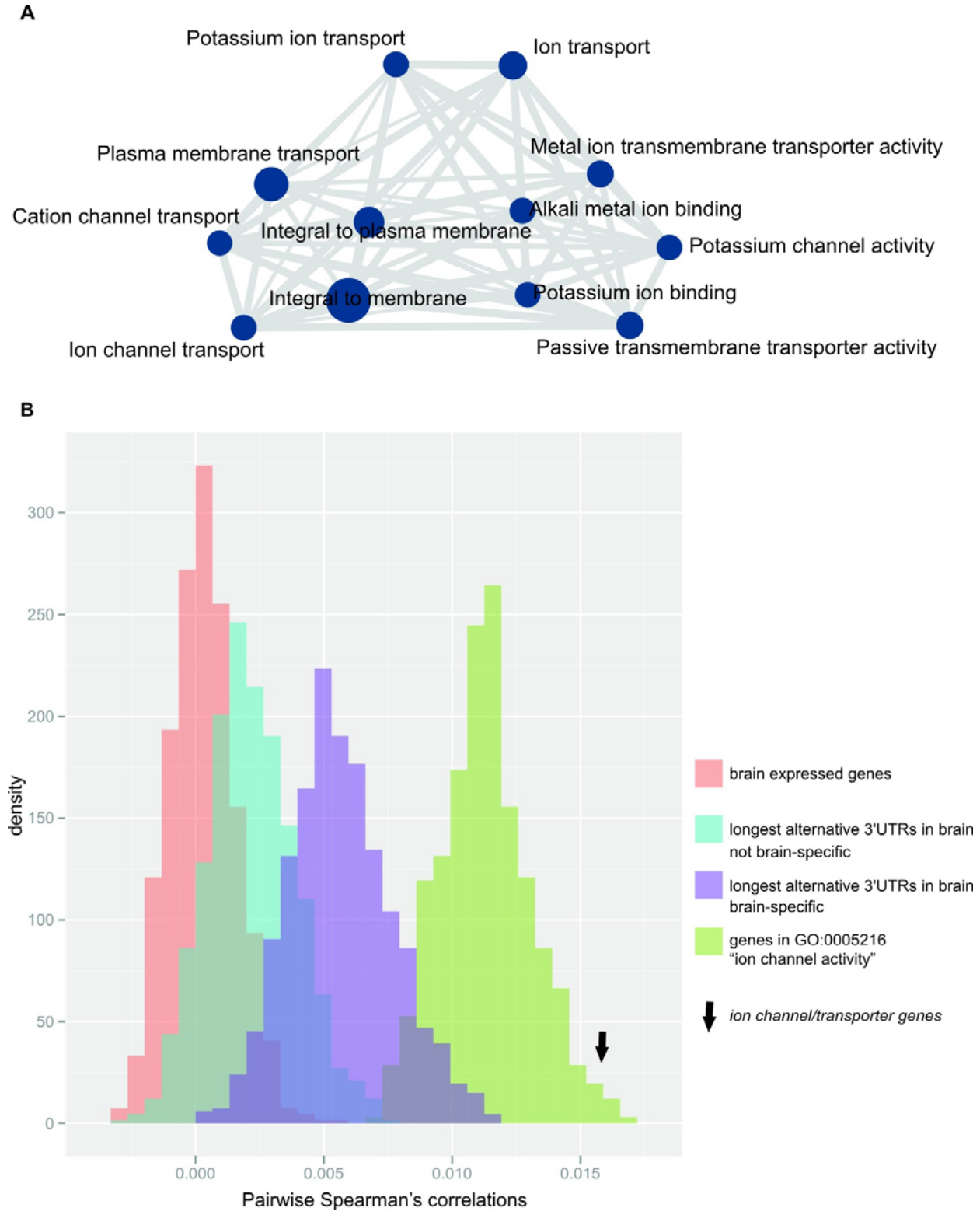

A
Potassium ion transport
Ion transport
Plasma membrane transport
Metal ion transmembrane transporter activity
Cation channel transport
Integral to plasma membrane
Alkali metal ion binding
Potassium channel activity
Integral to membrane
Potassium ion binding
Ion channel transport
Passive transmembrane transporter activity
B
density
300
250
200
150
100
50
0
0.000
0.005
0.010
0.015
Pairwise Spearman's correlations
brain expressed genes
longest alternative 3'UTRs in brain not brain-specific
longest alternative 3'UTRs in brain brain-specific
genes in GO:0005216 "ion channel activity"
ion channel/transporter genes

Figure 23. Ion channel/transporter genes are over-represented within genes expressing their longest 3'UTR in the brain A) Enrichment map [289] of significantly enriched GO-terms after GO-terms were summarized using REViGO [117]. Node size (blue circles) reflects the number of contributing genes, and edge size (grey lines) corresponds to the number of genes that are shared between nodes. B) Distribution of the median pairwise Spearman's correlations for 1000 bootstrap samples of random gene sets (each *n*=192, matched to *ion channel transporter genes*) from the populations: (i) brain-expressed genes (*n*=16812 probes; red), (ii) genes with longest alternative 3'UTRs expressed in brain (*n*=1261 probes; blue), (iii) genes with longest 3'UTRs expressed only in brain (*n*=449 probes; purple) or (iv) genes contributing to GO-term "ion channel activity" (*n*=269 probes; green). Arrow indicates the median pairwise Spearman's correlations for the *ion channel/transporter genes*.

5.3.3 Long alternative 3'UTRs strengthen miRNA mediated cross-talk among *ion channel/transporter genes* in the brain

The finding that the expression levels of *ion channel/transporter genes* show increased correlation in the human adult brain, prompted me to investigate the mechanism underlying these genes' coordinated expression. I hypothesized that alternative 3'UTRs might contain miRNA recognition elements (MREs) for miRNAs that are shared between *ion channel/transporter genes* thereby enhancing the posttranscriptional coordination of their relative abundance in the adult prefrontal cortex, by means of miRNA mediated cross-talk between transcripts.

I compared the percentages of predicted MREs per gene in constitutive or longest alternative 3'UTRs of *ion channel/transporter genes* for those miRNA families that are known to be expressed in the human adult frontal cortex (FC) [290]. I defined sets of miRNA families whose miRNAs are transcribed (denoted 'expressed') or specifically expressed in the FC at levels that are at least 10-, 5- or 2-times (denoted FC10, FC5 and FC2, respectively) higher than their median level across all non-neuronal tissues ([290]; Supplementary Table S1). For each gene, MRE percentage was defined as the ratio of predicted MREs for miRNA families expressed in or specific to the frontal cortex divided by the total number of predicted MREs. Only *ion channel/transporter genes* with at least one predicted MRE in alternative and constitutive 3'UTRs were considered (n=209 genes).

Longest alternative 3'UTRs for *ion channel/transporter genes* expressed in the brain showed a significantly higher percentage of MREs for miRNA families that are specifically expressed in the human frontal cortex relative to constitutive 3'UTRs (Figure 24A). This observation for *ion channel/transporter genes* is consistent with a recent report of a general enrichment of neuronal specific miRNAs across 3'UTR extensions in the human brain [269]. The difference in MRE percentage for alternative or constitutive 3'UTRs widened for increased tissue specificity of miRNA families in the human frontal cortex (Figure 24A). For the FC10 set, the median percentage of MREs predicted in brain alternative 3'UTRs was twice as high as for constitutive 3'UTRs. This large increase in the percentage of MREs for FC-specifically expressed miRNAs was not found in genes expressing their longest 3'UTRs only in brain after removing *ion channel/transporter genes* (n=300 genes, two-sample KS-test *p-value*=0.45).

I hypothesized that competition for shared miRNAs contributes to the observed increased coordination in expression between *ion channel/transporter genes*. To test this hypothesis, the proportion, *s*, of all MREs that were predicted to bind frontal cortex (FC)-specific miRNAs and that were present within the constitutive and alternative 3'UTRs of a pair of *ion channel/transporter genes* expressing their longest 3' UTR in the brain and shared between the two genes 3'UTRs was calculated; I will refer to this as the 'percentage of shared MREs' (Figure 24B). The median of *s* for these *ion channel/transporter genes* was then compared to values from randomly selected brain-expressed genes. *Ion channel/transporter genes* were found to share a significantly (*empirical p-value*<0.001) higher percentage of predicted MREs with each other (median *s*=47% for FC10 miRNA families) than other brain-expressed genes (Figure 24C). The median *s*-value between gene-pairs was highest for miRNA families that were specifically expressed in the frontal cortex (FC10 set). Values decreased to 40% (*p-value*<0.001) for FC5 miRNA families, and to 33% for FC2 miRNA families (*p-value*<0.001) as would be expected if the most frontal cortex-specific miRNAs contributed the most to the observed coordination of *ion channel/transporter gene* transcript abundance.

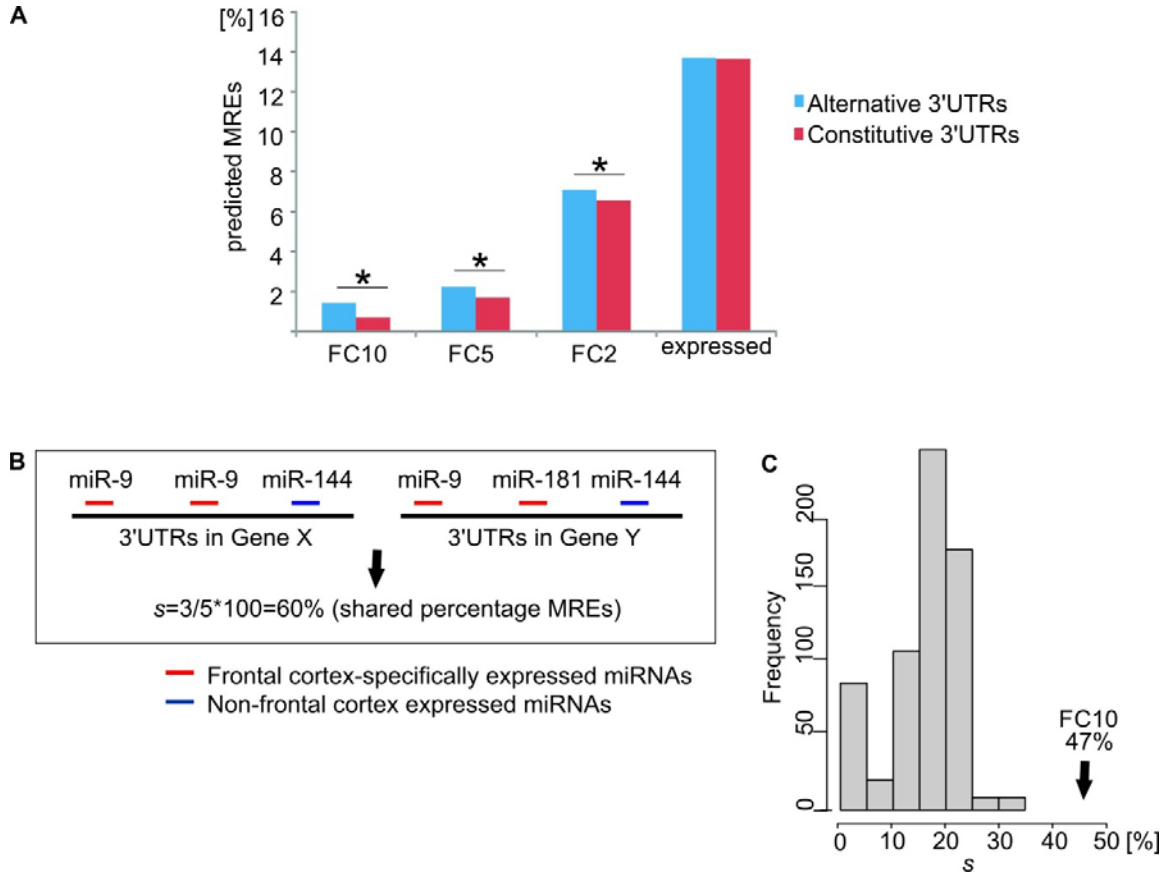


Figure 24. Frontal cortex-specifically expressed miRNAs targeting 3'UTR extensions strengthen the cross-talk between *ion channel/transporter genes* A) Differences in the percentage of predicted MREs for miRNA families expressed in human adult frontal cortex (y-axis for different levels of relative expression: FC10, FC5, FC2 and expressed) between constitutive (red) and alternative (blue) 3'UTRs of *ion channel/transporter genes*. * indicates a significant difference ($p < 0.05$). B) Calculation of the percentage (%) of shared MREs, *s*, between two *ion channel/transporter* genes. Gene X and Gene Y have two and one predicted MREs for a shared frontal cortex-specific miRNA (miR-9; red), respectively. Gene X and Y have predicted MREs for

miRNAs that are either not expressed specifically in the frontal cortex (Gene X, miR-144) or not shared (Gene Y, miR-181). Of the 5 shared predicted MREs in both genes 3 are for frontal cortex-specific miRNAs (miR-9; red), resulting in s=3/5x100 =60% shared frontal cortex-specific MREs C) Percentage shared MREs for FC10 in *ion channel/transporter genes* (median percentage shared MREs *s*=47%, black arrow). Histogram represents the null distribution of the same metric for 1000 random brain-expressed gene sets excluding *ion channel/transporter genes* (each *n*=226).

GABRA4 (Gamma-aminobutyric acid receptor, alpha-4) is an appropriate illustration of the ion channel/transporter network described above. *GABRA4* belongs to the receptor gene family GABA-A of ion channels and interacts with GABA, the main inhibitory neurotransmitter in the brain [291]. The expression level of *GABRA4* has been linked to learning and long-term potentiation [292] and its expression levels are increased following seizures [293]. I identified *ion channel/transporter genes* that show significantly correlated expression (α=0.05; Bonferroni corrected) in the adult human brain with *GABRA4* (Figure 25). The genes highly co-expressed with *GABRA4* included those encoding potassium channels which are involved in nervous excitability [294] and activity-dependent regulation of neuronal signaling (e.g. *KCNB1* [295]), as well as G-protein coupled receptors (e.g. *GPR158*) which are crucial for many physiological and disease pathways [296]. Significantly co-expressed genes with *GABRA4* share a large number of MREs for frontal cortex specifically expressed miRNAs (Figure 25) consistent with miRNA-mediated crosstalk between *ion channel/transporter genes* in the *GABRA4*-subnetwork potentially contributing to their coordinated expression.

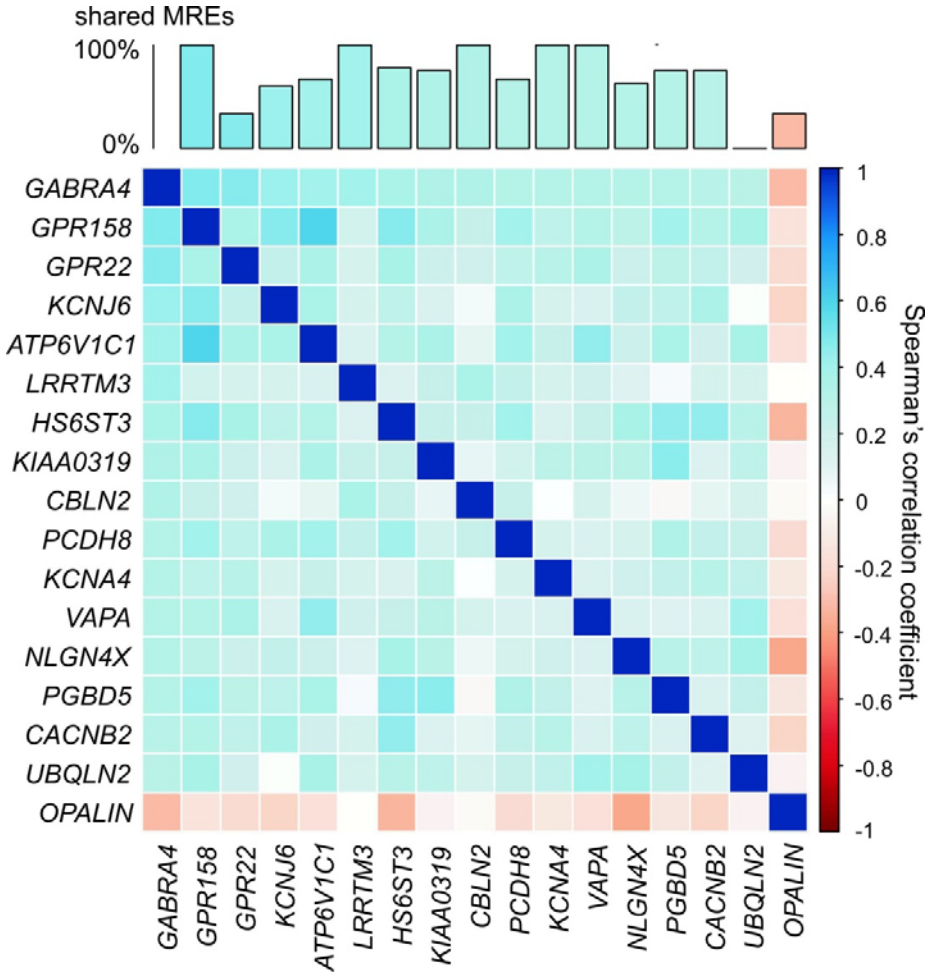


Figure 25. ***GABRA4*-subnetwork. Example of gene in ion channel/transporter network. Correlation matrix of *ion channel/transporter genes* significantly correlated with *GABRA4* (α=0.05, Bonferroni corrected, pairwise Spearman's correlation). Bar chart shows percentage of MREs *GABRA4* shares with every other gene in *GABRA4*-subnetwork.**

5.4 Discussion

The biological relevance of the widespread and conserved tendency of protein-coding genes to express isoforms with longer alternative 3' untranslated regions (3' UTRs) in brain [268, 269] has been poorly understood. Given the role of these regions in the post-transcriptional modulation of relative transcript abundance by means of miRNA mediated cross-talk [27], I hypothesized that long alternative 3'UTRs could enhance the miRNA-regulated abundance of functionally related genes. To test this hypothesis, I used mRNA sequencing data to annotate 3'UTRs in adult human tissues. Consistent with previous analysis in humans [268, 269] and other animals [265-267] I find that brain-expressed isoforms have longer alternative 3'UTRs which are enriched for brain-expressed miRNA binding sites [269]. Two major novel findings emerged from the analysis of genes expressing their longest alternative 3'UTRs in the brain. Firstly, that ion channel/transporter genes are enriched within this gene set; secondly that these genes exhibit increased transcriptional coordination mediated by competition for shared frontal cortex-specific miRNAs.

Transport across the cellular membrane is primarily mediated by ion channels and other transporter genes and this function is crucial to ensure appropriate responses to changing extra- and intracellular environments. In the brain, transport across the cellular membrane is crucial for cellular excitability. Controlled responses to both excitatory and inhibitory signals that underlie communication between cells in the brain require the dynamic regulation of the abundance and repertoire of ion channel and other transporter proteins [297]. This is achieved at several levels including transcriptional and post-translational control. For example, in neurons some ion channel/transporter gene encoding transcripts

are enriched at dendrites [298]. Translation of transcripts localized at dendrites is evolutionarily conserved in multiple species [298] and is believed to allow efficient spatial and temporal resolution of extracellular cues [299, 300].

Translation of localized mRNAs is tightly regulated post-transcriptionally by miRNAs, with several brain-expressed miRNAs present within axonal compartments [301]. These miRNAs are capable of actively regulating local mRNA translation [302, 303]. Importantly, miRNA expression and repertoires are highly sensitive to microenvironmental changes, with several miRNAs exhibiting increased transcription, for example upon long term potentiation [304]. The observations that both ion channel/transporter genes and miRNAs respond to changes in environment and have specific and overlapping subcellular localizations suggest that their interactions mediate responses to milieu changes. Transmembrane transport that is mediated by ion channels/transporters, in response to changes in cellular microenvironment results in changes in miRNA abundance that in turn will translate in post-transcriptional repression of their targets, including transporter protein-encoding mRNAs. This regulatory feedback loop is likely to provide dynamic yet robust changes in ion channel/transporter abundance [305].

The findings that a subset of ion channel/transporter genes tend to express longer alternative 3'UTRs in the brain and that the sequence of these extensions is enriched for frontal cortex-specific miRNA response elements is consistent with this proposal. Furthermore I show that predicted MREs located at alternative 3'UTRs can enhance the miRNA mediated crosstalk between these transcripts, likely contributing to the observed coordination of ion-channel/transporter genes in the brain.

I propose that the alternative 3'UTRs of ion channel/transporter mRNAs in the brain post-transcriptionally enhance the robustness of dynamic chances of their product abundance in response to differences in cellular microenvironment ensuring coordinated assembly of functional transmembrane channels.

5.5 Methods

5.5.1 RNA sequencing data

I obtained single- and paired-end long poly-adenylated RNA sequencing (RNA-Seq) reads derived from 16 human adult tissues of one individual each (adipose, adrenal, blood, brain, breast, colon, heart, kidney, liver, lung, lymph-node, ovary, prostate, skeletal-muscle, testes, and thyroid) from the Illumina's Human BodyMap 2.0 project, downloaded from http://www.ensembl.info/blog/2011/05/24/human-bodymap-2-0-data-from-illumina/.

The expression profiles of human miRNAs were obtained from [290] and used to annotate miRNA families expressed in human adult frontal cortex. A miRNA was considered to be expressed, if it had ≥ 1 reads supporting its expression in this tissue (resulting in 76 expressed miRNA families [290]). A miRNA family was defined as frontal cortex-specifically expressed if its expression was supported by a number of reads which was 2-, 5- or 10-times higher than this miRNA's median across all non-neuronal tissues (Supplementary Table S5.1) [290].

5.5.2 Annotation of constitutive and alternative 3' untranslated regions

Single- and paired-reads for each of the 16 tissues were mapped to the human genome (hg19) using TopHat (version 1.3.0 [306]) (mapping sequencing reads using TopHat was done by Ana Marques). I assembled transcripts *de novo* with Cufflinks (version 1.0.3, [114]) using ENSEMBL (build 71, [307]) protein-coding gene transcripts as reference. I considered transcripts that overlapped exonic bases of ENSEMBL protein-coding genes to be part of that gene locus.

For each tissue, I excluded protein-coding genes that were not expressed (have an FPKM - Fragments Per Kilobase of transcript per Million mapped reads- equal to zero). I included genes annotated in ENSEMBL and newly assembled transcripts. Next I considered only protein-coding genes with stop codon genomic locations that were common to all its annotated transcripts in ENSEMBL; I refer to these as constitutive stop codons (*n*=8533 genes contained constitutive stop codons (Figure 26A) out of *n*=19985 protein-coding genes with annotated stop codons in ENSEMBL build 71). For each gene and in each tissue, I annotated a consensus transcript, containing all transcribed nucleotides and defined its 3' untranslated region (3'UTR) as the transcribed region downstream of its constitutive stop codon (Figure 26B). The nucleotide regions common to all transcripts from a gene were annotated as that gene's constitutive 3' UTRs and the remaining 3' UTR nucleotides were considered to be alternative 3'UTRs. For each gene, I annotated the alternative 3' UTRs with the maximal length in nucleotides as its longest alternative 3' UTRs (Figure 26B).

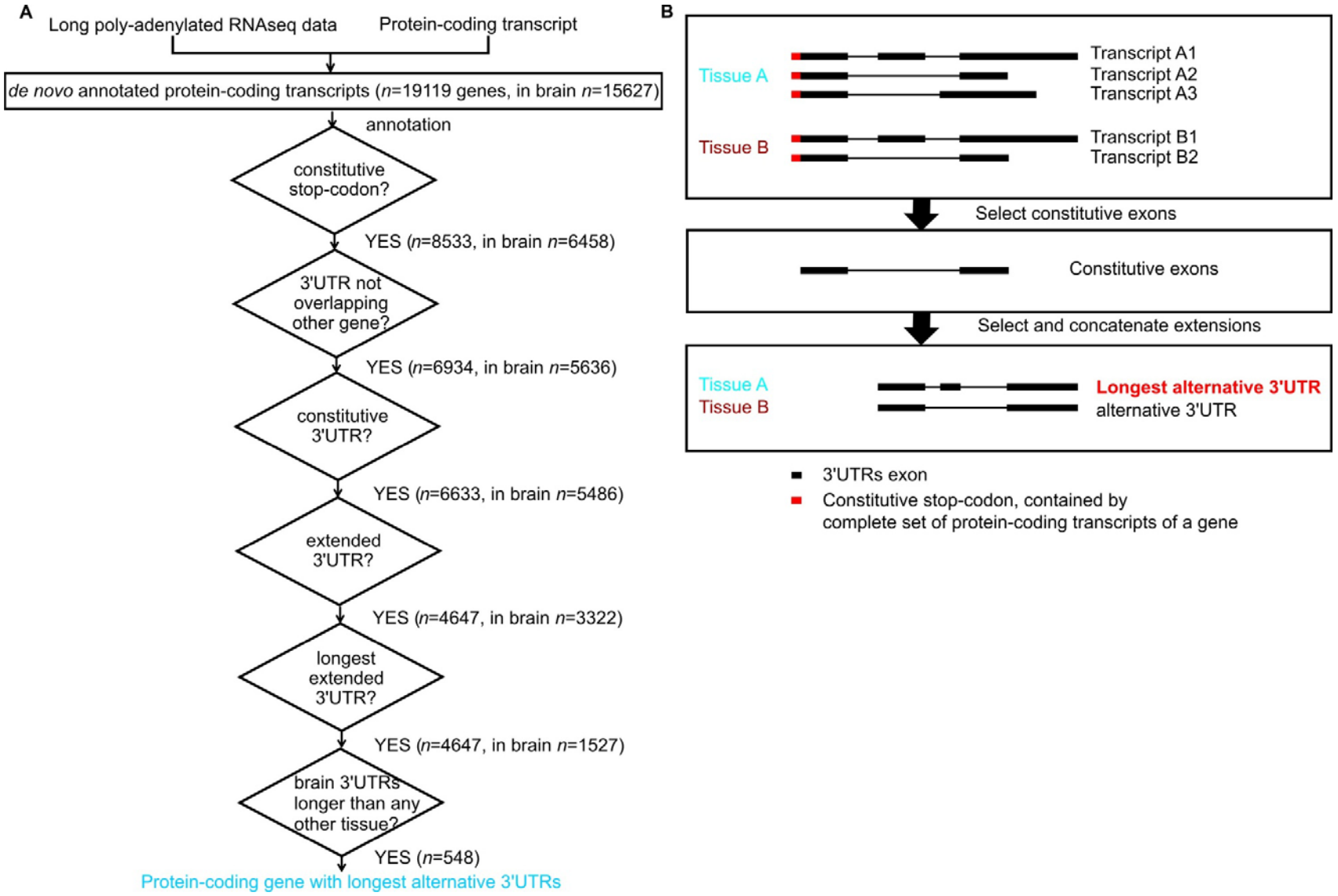

A
Long poly-adenylated RNAseq data
Protein-coding transcript
de novo annotated protein-coding transcripts (n=19119 genes, in brain n=15627)
annotation
constitutive stop-codon?
YES (n=8533, in brain n=6458)
3'UTR not overlapping other gene?
YES (n=6934, in brain n=5636)
constitutive 3'UTR?
YES (n=6633, in brain n=5486)
extended 3'UTR?
YES (n=4647, in brain n=3322)
longest extended 3'UTR?
YES (n=4647, in brain n=1527)
brain 3'UTRs longer than any other tissue?
YES (n=548)
Protein-coding gene with longest alternative 3'UTRs
B
Tissue A
Transcript A1
Transcript A2
Transcript A3
Tissue B
Transcript B1
Transcript B2
Select constitutive exons
Constitutive exons
Select and concatenate extensions
Tissue A
Tissue B
Longest alternative 3'UTR
alternative 3'UTR
3'UTRs exon
Constitutive stop-codon, contained by complete set of protein-coding transcripts of a gene

Figure 26. **Annotation of constitutive and alternative 3'UTRs. A) Flowchart for annotation of protein-coding genes with longest alternative 3'UTRs; *n* indicates number of genes. B) Example of a gene with constitutive stop codon (red), expressed only in tissue A (blue) and tissue B (brown). Tissue A expresses three transcripts, and tissue B expresses 2 transcripts (top panel). The constitutive 3'UTRs consists in the nucleotides expressed in all transcript of a gene across all tissues where the gene is expressed (middle panel). The extended alternative 3'UTRs is the consensus transcript for a gene in a given tissue after constitutive parts have been removed. Tissue A expresses the longest alternative 3'UTRs for this gene, as it contains more nucleotides than the alternative 3'UTRs in tissue B (bottom panel).**

5.5.3 3'UTR expression

I used CoverageBed (Bedtools version 2.17.0; [308]) to determine the number of long mRNASeq reads derived from each adult human tissue mapping to constitutive and alternative 3' UTRs. For each tissue, I estimated constitutive and alternative 3'UTR expression (*E*) as the number of reads mapping to that region divided by its total length. To identify differences in 3'UTRs usage between tissues, I estimated the ratio between the expression in tissue *i* of the longest alternative (*L*) and constitutive (*C*) 3'UTRs ($E_i=E_{Li}/E_{Ci}$).

5.5.4 Lineage specific constraint

I used the derived frequencies of 1000 Genomes single nucleotide polymorphisms (SNP; [285, 286]) to test for differences in selective constraint within the human lineage between ancestral repeats (AR), constitutive 3'UTRs and the longest alternative 3'UTRs

across all tissues, as described in [257, 309]. I considered SNPs with frequency f<10% to be rare and intermediate if 10%<f<90%. I considered only ARs within the vicinity of 3'UTRs (20kb, 50kb and 100kb downstream of the alternative 3'UTRs or upstream of the constitutive stop codon).

5.5.5 Functional enrichment

For each tissue, I tested whether genes with the longest alternative 3'UTR expressed in only that tissue were functionally enriched using the functional classification tool Database for Annotation, Visualization, and Integrated Discovery (DAVID; [310]) using default parameters (count = 2 and ease = 0.1). Genes expressed in only one tissue were discarded, for example in brain, out of n=548 genes expressing their longest alternative 3'UTRs in brain only 4 were excluded. For each tissue, I compared the gene ontology (GO) annotations of genes with a longest alternative 3'UTR expressed in that tissue to a background containing all genes with annotated 3'UTRs expressed in that tissue. I considered GO-terms that were significantly enriched at α=0.05, after Bonferroni multiple testing correction. Online DAVID tool and GO-terms are described in detail in **Chapter 2.4**.

5.5.6 Expression correlation

I tested whether the genes (n=226) contributing to functional enrichment of ion channel and transporter encoding genes (referred to as '*ion channel/transporter genes'*) showed increased correlation in expression. In order to match the developmental stage of the subject from which the RNA used to annotate 3'UTRs in the brain was obtained (a 77 year old female) I restricted the analyses of gene expression to adulthood (20 to 78 years; n=148) using the microarray data from 'braincloud' [288] (see **Chapter 2.1**), preprocessed and quantified as described by the authors [288]. 'Braincloud' contains

gene expression data from the prefrontal cortex (PFC) of human postmortem tissue. For each gene, I used only one probe, namely the probe that showed the strongest correlation with age during the lifespan [288] (number of probes=16812; the expression matrix containing the probes correlated with age was provided by Andrew Jaffe) to avoid bias arising from using multiple probes for one gene. *N*=192 probes were available for *ion channel/transporter genes*. I compared the median of all pairwise Spearman's correlations among the *ion channel/transporter genes* to the same measure from *n*=1000 random bootstrap samples of same set size (*n*=192 genes, matched to *ion channel/transporter genes*) from (*i*) brain-expressed genes (*n*=16812 probes available in 'braincloud'), (*ii*) genes with longest 3'UTRs expressed in brain (*n*=1261 probes) or (*iii*) genes with longest 3'UTRs expressed only in brain (*n*=449 probes). I repeated the analysis for 1000 random samples of 192 genes annotated with the GO-term "Ion channel activity" (GO:0005216), which was the most significant p-value in enrichment analysis (*n*=269 probes).

5.5.7 Prediction of miRNA response elements

MiRNA response elements (MREs) in all annotated constitutive and longest alternative 3'UTRs, were predicted using TargetScan (version 5.0; [115]; prediction of MREs using TargetScan was performed by Ana Marques). For all constitutive and longest alternative 3'UTRs annotated in brain, I estimated the percentage of MREs expressed in adult frontal cortex (FC) [290] by dividing the number of MREs for miRNAs expressed in adult FC by the total number of MREs predicted for the region of interest. The difference between MRE percentages in longest alternative 3' UTRs compared to constitutive 3'UTRs in the brain was assessed using a Kolmogorov-Smirnov test (KS-test). I repeated the analysis considering only families containing miRNAs whose expression was 10-times, 5-times, or 2-times higher than the median expression across

all non-neuronal tissues [290]; hereafter referred to as FC10, FC5 and FC2 sets. I included only *ion channel/transporter genes* which had at least 1 MRE predicted in alternative and constitutive 3'UTRs (*n*=209). The same analysis was done for the remaining *n*=300 genes expressing the longest alternative 3'UTRs only in brain after removing *ion channel/transporter genes*.

Next, the percentage of shared MREs for FC10, FC5 and FC2 ([290], Supplementary Table S1) predicted in constitutive and alternative 3'UTRs for each pair of *ion channel/transporter genes* (*n*=226) was compared (percentage of shared MREs was calculated by Ana Marques). Precisely, the observed median value was compared to what would be expected based on the median percentage of shared MREs between 1000 random sets of *n*=226 genes expressed in the brain, excluding *ion channel/transporter genes*.

5.6 Appendix

Table S5.1. MiRNA families expressed in human frontal cortex and expressed specifically in human frontal cortex [290].

miRNA families expressed in human frontal cortex	miRNA families 2-times higher expressed in human frontal cortex than in not neuronal tissue	miRNA families 5-times higher expressed in human frontal cortex than in not neuronal tissue	miRNA families 10-times higher expressed in human frontal cortex than in not neuronal tissue
let-7/98	let-7/98	let-7/98	miR-124/506
miR-101	miR-103/107	miR-124/506	miR-181
miR-103/107	miR-124/506	miR-125/351	miR-99ab/100
miR-105.h	miR-125/351	miR-126/126-3p	miR-125/351
miR-124/506	miR-126/126-3p	miR-129/129-5p	miR-126/126-3p
miR-125/351	miR-127	miR-139-5p	miR-9
miR-126/126-3p	miR-129/129-5p	miR-143	miR-143
miR-127	miR-132/212	miR-181	miR-129/129-5p
miR-129-3p	miR-137	miR-27ab	
miR-129/129-5p	miR-138	miR-29abc	
miR-130/301	miR-139-3p	miR-451	
miR-132/212	miR-139-5p	miR-9	
miR-136	miR-143	miR-99ab/100	
miR-137	miR-145		
miR-138	miR-149		
miR-139-3p	miR-151		
miR-139-5p	miR-154		
miR-140/140-5p/876-3p	miR-181		
miR-142-5p	miR-197		
miR-143	miR-22		
miR-145	miR-221/222		
miR-146	miR-23ab		
miR-148/152	miR-24		
miR-149	miR-26ab/1297		
miR-15/16/195/424/497	miR-27ab		
miR-151	miR-29abc		
miR-153	miR-30a/30a-5p/30b/30b-5p/30cde/384-5p		
miR-154	miR-324-5p		
miR-17-5p/20/93.mr/106/519.d	miR-338/338-3p		
miR-181	miR-342/342-3p		

miR-185/882
miR-186
miR-191
miR-194
miR-197
miR-199/199-3p
miR-21/590-5p
miR-214/761
miR-219-2-3p
miR-22
miR-221/222
miR-23ab
miR-24
miR-25/32/92/92ab/363/367
miR-26ab/1297
miR-27ab
miR-28/28-5p/708
miR-29abc
miR-30a/30a-5p/30b/30b-5p/30cde/384-5p
miR-323/323-3p
miR-324-3p.h
miR-324-5p
miR-330-3p
miR-338/338-3p
miR-342/342-3p
miR-361/361-5p
miR-365
miR-370
miR-374/374ab
miR-377
miR-379
miR-411
miR-425/489
miR-431
miR-451
miR-487/487b
miR-518a-3p/518bcf/518d-3p
miR-539
miR-574/574-3p
miR-628-3p
miR-652
miR-7/7ab
miR-744
miR-885-5p
miR-365
miR-379
miR-411
miR-425/489
miR-431
miR-451
miR-574/574-3p
miR-628-3p
miR-7/7ab
miR-744
miR-885-5p
miR-9
miR-99ab/100

miR-9
miR-99ab/100

Table S5.2. Genes expressing their longest 3'UTRs in the brain and contributing to observed GO-term enrichment.

ENSEMBL GENE-ID	Name
ENSG00000168830	5-hydroxytryptamine (serotonin) receptor 1E
ENSG00000179097	5-hydroxytryptamine (serotonin) receptor 1F
ENSG00000102468	5-hydroxytryptamine (serotonin) receptor 2A
ENSG00000157219	5-hydroxytryptamine (serotonin) receptor 5A
ENSG00000158748	5-hydroxytryptamine (serotonin) receptor 6
ENSG00000077522	actinin, alpha 2
ENSG00000121989	activin A receptor, type IIA
ENSG00000143199	adenylate cyclase 10 (soluble)
ENSG00000196581	adherens junctions associated protein 1
ENSG00000108684	amiloride-sensitive cation channel 1, neuronal
ENSG00000155097	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1
ENSG00000103222	ATP-binding cassette, sub-family C (CFTR/MRP), member 1
ENSG00000138075	ATP-binding cassette, sub-family G (WHITE), member 5
ENSG00000142867	B-cell CLL/lymphoma 10; hypothetical LOC646626
ENSG00000176788	brain abundant, membrane attached signal protein 1
ENSG00000154162	cadherin 12, type 2 (N-cadherin 2)
ENSG00000143126	cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, Drosophila)
ENSG00000165995	calcium channel, voltage-dependent, beta 2 subunit
ENSG00000166862	calcium channel, voltage-dependent, gamma subunit 2
ENSG00000142408	calcium channel, voltage-dependent, gamma subunit 8
ENSG00000143119	CD53 molecule
ENSG00000153563	CD8a molecule
ENSG00000149798	CDC42 effector protein (Rho GTPase binding) 2
ENSG00000175161	cell adhesion molecule 2
ENSG00000141668	cerebellin 2 precursor
ENSG00000054803	cerebellin 4 precursor
ENSG00000129048	chemokine (C-C motif) receptor-like 1
ENSG00000101204	cholinergic receptor, nicotinic, alpha 4
ENSG00000204006	chromosome 1 open reading frame 185
ENSG00000185742	chromosome 11 open reading frame 87
ENSG00000139971	chromosome 14 open reading frame 37
ENSG00000107020	chromosome 9 open reading frame 46
ENSG00000122705	clathrin, light chain (Lca)

ENSG00000049656	CLPTM1-like
ENSG00000155052	contactin associated protein-like 5
ENSG00000143786	cornichon homolog 3 (Drosophila)
ENSG00000163626	COX18 cytochrome c oxidase assembly homolog (S. cerevisiae)
ENSG00000179979	cysteine-rich PAK1 inhibitor
ENSG00000152207	cysteinyl leukotriene receptor 2
ENSG00000186684	cytochrome P450, family 27, subfamily C, polypeptide 1
ENSG00000171451	dermatan sulfate epimerase-like
ENSG00000134780	diacylglycerol lipase, alpha
ENSG00000107984	dickkopf homolog 1 (Xenopus laevis)
ENSG00000116544	discs, large (Drosophila) homolog-associated protein 3
ENSG00000080166	dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)
ENSG00000184845	dopamine receptor D1
ENSG00000169676	dopamine receptor D5
ENSG00000171587	Down syndrome cell adhesion molecule
ENSG00000108947	ephrin-B3
ENSG00000204442	family with sequence similarity 155, member A
ENSG00000183807	family with sequence similarity 162, member B
ENSG00000196666	family with sequence similarity 180, member B
ENSG00000086570	FAT tumor suppressor homolog 2 (Drosophila)
ENSG00000126500	fibronectin leucine rich transmembrane protein 1
ENSG00000185070	fibronectin leucine rich transmembrane protein 2
ENSG00000126266	free fatty acid receptor 1
ENSG00000172461	fucosyltransferase 9 (alpha (1,3) fucosyltransferase)
ENSG00000132975	G protein-coupled receptor 12
ENSG00000178015	G protein-coupled receptor 150
ENSG00000158292	G protein-coupled receptor 153
ENSG00000151025	G protein-coupled receptor 158
ENSG00000188888	G protein-coupled receptor 179
ENSG00000152749	G protein-coupled receptor 180
ENSG00000172209	G protein-coupled receptor 22
ENSG00000154478	G protein-coupled receptor 26
ENSG00000170775	G protein-coupled receptor 37 (endothelin receptor type B-like)
ENSG00000146360	G protein-coupled receptor 6
ENSG00000112218	G protein-coupled receptor 63
ENSG00000175229	galactose-3-O-sulfotransferase 3
ENSG00000109158	gamma-aminobutyric acid (GABA) A receptor, alpha 4
ENSG00000145864	gamma-aminobutyric acid (GABA) A receptor, beta 2
ENSG00000215045	glutamate receptor, ionotropic, delta 2 (Grid2) interacting protein
ENSG00000150086	glutamate receptor, ionotropic, N-methyl D-aspartate 2B
ENSG00000130055	glycerophosphodiester phosphodiesterase domain containing 2
ENSG00000188828	glycine receptor, alpha 4

ENSG00000185245	glycoprotein Ib (platelet), alpha polypeptide
ENSG00000179399	glypican 5
ENSG00000173905	golgi integral membrane protein 4
ENSG00000120063	guanine nucleotide binding protein (G protein), alpha 13
ENSG00000167414	guanine nucleotide binding protein (G protein), gamma 8
ENSG00000173706	HEG homolog 1 (zebrafish)
ENSG00000182601	heparan sulfate (glucosamine) 3-O-sulfotransferase 4
ENSG00000136720	heparan sulfate 6-O-sulfotransferase 1
ENSG00000185352	heparan sulfate 6-O-sulfotransferase 3
ENSG00000164588	hyperpolarization activated cyclic nucleotide-gated potassium channel 1
ENSG00000205664	hypothetical protein LOC441528; similar to hCG1981372; similar to Serine/threonine-protein kinase PRKX (Protein kinase PKX1)
ENSG00000134049	immediate early response 3 interacting protein 1
ENSG00000169306	interleukin 1 receptor accessory protein-like 1
ENSG00000137261	KIAA0319
ENSG00000134962	klotho beta
ENSG00000142235	lemur tyrosine kinase 3
ENSG00000128011	leucine rich repeat and fibronectin type III domain containing 1
ENSG00000156564	leucine rich repeat and fibronectin type III domain containing 2
ENSG00000174482	leucine rich repeat and Ig domain containing 2
ENSG00000170382	leucine rich repeat neuronal 2
ENSG00000198739	leucine rich repeat transmembrane neuronal 3
ENSG00000168924	leucine zipper-EF-hand containing transmembrane protein 1
ENSG00000204033	leucine-rich repeat, immunoglobulin-like and transmembrane domains 2
ENSG00000182508	lipoma HMGIC fusion partner-like 1
ENSG00000156959	lipoma HMGIC fusion partner-like 4
ENSG00000197753	lipoma HMGIC fusion partner-like 5
ENSG00000157978	low density lipoprotein receptor adaptor protein 1
ENSG00000176454	lysophosphatidylcholine acyltransferase 4
ENSG00000109270	MAPK scaffold protein 1
ENSG00000130368	MAS1 oncogene
ENSG00000152034	melanin-concentrating hormone receptor 2
ENSG00000189332	mitochondrial carrier triple repeat 1 pseudogene
ENSG00000147065	moesin
ENSG00000074771	NADPH oxidase 3
ENSG00000138653	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 4
ENSG00000185737	neuregulin 3
ENSG00000146938	neuroligin 4, X-linked
ENSG00000132911	neuromedin U receptor 2
ENSG00000172260	neuronal growth regulator 1
ENSG00000148400	Notch homolog 1, translocation-associated (Drosophila)
ENSG00000153989	nuclear undecaprenyl pyrophosphate synthase 1 pseudogene;

	nuclear undecaprenyl pyrophosphate synthase 1 homolog (S. cerevisiae)
ENSG00000113569	nucleoporin 155kDa
ENSG00000165204	olfactory receptor, family 1, subfamily K, member 1
ENSG00000168158	olfactory receptor, family 2, subfamily C, member 1
ENSG00000181074	olfactory receptor, family 52, subfamily N, member 4
ENSG00000197430	oligodendrocytic myelin paranodal and inner loop protein
ENSG00000116329	opioid receptor, delta 1
ENSG00000082556	opioid receptor, kappa 1
ENSG00000099864	paralemmin
ENSG00000205808	phosphatidic acid phosphatase type 2 domain containing 2
ENSG00000155629	phosphoinositide-3-kinase adaptor protein 1
ENSG00000127472	phospholipase A2, group V
ENSG00000177614	piggyBac transposable element derived 5
ENSG00000076356	plexin A2
ENSG00000118762	polycystic kidney disease 2 (autosomal dominant)
ENSG00000183775	potassium channel tetramerisation domain containing 16
ENSG00000184261	potassium channel, subfamily K, member 12
ENSG00000152315	potassium channel, subfamily K, member 13
ENSG00000164794	potassium channel, subfamily V, member 1
ENSG00000143603	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3
ENSG00000123700	potassium inwardly-rectifying channel, subfamily J, member 2
ENSG00000120457	potassium inwardly-rectifying channel, subfamily J, member 5
ENSG00000157542	potassium inwardly-rectifying channel, subfamily J, member 6
ENSG00000124134	potassium voltage-gated channel, delayed-rectifier, subfamily S, member 1
ENSG00000156486	potassium voltage-gated channel, delayed-rectifier, subfamily S, member 2
ENSG00000184156	potassium voltage-gated channel, KQT-like subfamily, member 3
ENSG00000158445	potassium voltage-gated channel, Shab-related subfamily, member 1
ENSG00000182674	potassium voltage-gated channel, Shab-related subfamily, member 2
ENSG00000170049	potassium voltage-gated channel, shaker-related subfamily, beta member 3
ENSG00000182255	potassium voltage-gated channel, shaker-related subfamily, member 4
ENSG00000130037	potassium voltage-gated channel, shaker-related subfamily, member 5
ENSG00000143473	potassium voltage-gated channel, subfamily H (eag-related), member 1
ENSG00000101856	progesterone receptor membrane component 1
ENSG00000119973	prolactin releasing hormone receptor
ENSG00000197870	proline-rich protein BstNI subfamily 3

ENSG00000111218	protein arginine methyltransferase 8
ENSG00000138650	protocadherin 10
ENSG00000136099	protocadherin 8
ENSG00000112852	protocadherin beta 2
ENSG00000113209	protocadherin beta 5
ENSG00000113211	protocadherin beta 6
ENSG00000179331	RAB39, member RAS oncogene family
ENSG00000152932	RAB3C, member RAS oncogene family
ENSG00000117280	RAB7, member RAS oncogene family-like 1
ENSG00000181467	RAP2B, member of RAS oncogene family
ENSG00000100302	RASD family, member 2
ENSG00000143622	Ras-like without CAAX 1
ENSG00000137275	receptor (TNFRSF)-interacting serine-threonine kinase 1
ENSG00000186479	regulator of G-protein signaling 7 binding protein
ENSG00000125744	reticulon 2
ENSG00000185924	reticulon 4 receptor-like 1
ENSG00000180245	retinal pigment epithelium-derived rhodopsin homolog
ENSG00000132694	Rho guanine nucleotide exchange factor (GEF) 11
ENSG00000102683	sarcoglycan, gamma (35kDa dystrophin-associated glycoprotein)
ENSG00000158352	shroom family member 4
ENSG00000164647	six transmembrane epithelial antigen of the prostate 1
ENSG00000178235	SLIT and NTRK-like family, member 1
ENSG00000018280	solute carrier family 11 (proton-coupled divalent metal ion transporters), member 1
ENSG00000164707	solute carrier family 13 (sodium/sulfate symporters), member 4
ENSG00000197901	solute carrier family 22 (organic anion transporter), member 6
ENSG00000092096	solute carrier family 22, member 17
ENSG00000155886	solute carrier family 24 (sodium/potassium/calcium exchanger), member 2
ENSG00000151729	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 4
ENSG00000155287	solute carrier family 25, member 28
ENSG00000205060	solute carrier family 35, member B4
ENSG00000183780	solute carrier family 35, member F3
ENSG00000100191	solute carrier family 5 (low affinity glucose cotransporter), member 4
ENSG00000198689	solute carrier family 9 (sodium/hydrogen exchanger), member 6
ENSG00000101187	solute carrier organic anion transporter family, member 4A1
ENSG00000132671	somatostatin receptor 4
ENSG00000184005	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 3
ENSG00000101638	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 5
ENSG00000130052	StAR-related lipid transfer (START) domain containing 8

ENSG00000185518	synaptic vesicle glycoprotein 2B; hypothetical protein LOC100128403
ENSG00000110975	synaptotagmin X
ENSG00000162512	syndecan 3
ENSG00000124145	syndecan 4
ENSG00000075073	tachykinin receptor 2
ENSG00000006128	tachykinin, precursor 1
ENSG00000127364	taste receptor, type 2, member 4
ENSG00000221937	taste receptor, type 2, member 40
ENSG00000127366	taste receptor, type 2, member 5
ENSG00000005108	thrombospondin, type I, domain containing 7A
ENSG00000072315	transient receptor potential cation channel, subfamily C, member 5
ENSG00000111199	transient receptor potential cation channel, subfamily V, member 4
ENSG00000167664	transmembrane and immunoglobulin domain containing 2
ENSG00000181284	transmembrane protein 102
ENSG00000139364	transmembrane protein 132B; hypothetical LOC121296
ENSG00000151952	transmembrane protein 132D
ENSG00000181291	transmembrane protein 132E
ENSG00000181264	transmembrane protein 136
ENSG00000179292	transmembrane protein 151A
ENSG00000152128	transmembrane protein 163
ENSG00000157600	transmembrane protein 164
ENSG00000106565	transmembrane protein 176B
ENSG00000151353	transmembrane protein 18
ENSG00000163444	transmembrane protein 183A; transmembrane protein 183B
ENSG00000206069	transmembrane protein 211
ENSG00000187049	transmembrane protein 216
ENSG00000187783	transmembrane protein 72
ENSG00000164841	transmembrane protein 74
ENSG00000120337	tumor necrosis factor (ligand) superfamily, member 18
ENSG00000125657	tumor necrosis factor (ligand) superfamily, member 9
ENSG00000141540	tweety homolog 2 (Drosophila)
ENSG00000188021	ubiquilin 2
ENSG00000172318	UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 1
ENSG00000101558	VAMP (vesicle-associated membrane protein)-associated protein A, 33kDa
ENSG00000160293	vav 2 guanine nucleotide exchange factor
ENSG00000178201	vomeronasal 1 receptor 1
ENSG00000187135	V-set and transmembrane domain containing 2B
ENSG00000177054	zinc finger, DHHC-type containing 13

Table S5.3. Functional enrichment for GO-terms across all 16 tissues tested. Only GO-terms significant at Bonferroni-corrected p-values (α=0.05) are shown. 'Count' refers to number of genes contributing to enrichment of a GO-term. 'Fold' refers to fold enrichment.

Tissue	Term		Count	Fold	p-values
Brain	ion channel activity	GO:0005216	33	2.8	2.7E-05
	cation channel activity	GO:0005261	28	3.2	2.7E-05
	metal ion transmembrane transporter activity	GO:0046873	30	3.0	2.9E-05
	gated channel activity	GO:0022836	30	2.9	7.0E-05
	substrate specific channel activity	GO:0022838	33	2.7	7.5E-05
	channel activity	GO:0015267	34	2.6	1.0E-04
	passive transmembrane transporter activity	GO:0022803	34	2.6	1.0E-04
	intrinsic to membrane	GO:0031224	197	1.3	1.4E-04
	integral to membrane	GO:0016021	191	1.3	1.4E-04
	metal ion transport	GO:0030001	35	2.4	1.6E-03
	potassium channel activity	GO:0005267	19	3.7	5.5E-04
	voltage-gated cation channel activity	GO:0022843	19	3.6	7.5E-04
	plasma membrane	GO:0005886	143	1.4	4.5E-04
	ion transport	GO:0006811	44	2.1	3.4E-03
	cation transport	GO:0006812	36	2.2	8.1E-03
	potassium ion binding	GO:0030955	17	3.6	4.1E-03
	voltage-gated channel activity	GO:0022832	21	3.0	4.3E-03
	voltage-gated ion channel activity	GO:0005244	21	3.0	4.3E-03
	voltage-gated potassium channel activity	GO:0005249	15	3.9	4.8E-03
	potassium ion transport	GO:0006813	19	3.2	2.0E-02
	intrinsic to plasma membrane	GO:0031226	65	1.7	4.7E-03
	ion channel complex	GO:0034702	21	2.9	5.5E-03
	integral to plasma membrane	GO:0005887	64	1.7	5.5E-03
	alkali metal ion binding	GO:0031420	21	2.8	1.3E-02
	plasma membrane part	GO:0044459	98	1.4	1.2E-02
	cation channel complex	GO:0034703	16	3.2	2.2E-02
	voltage-gated potassium channel complex	GO:0008076	13	3.6	3.5E-02
	potassium channel complex	GO:0034705	13	3.6	3.5E-02
Kidney	insoluble fraction	GO:0005626	22	2.4	3.4E-02
Lung	cullin-RING ubiquitin ligase complex	GO:0031461	4	39.7	1.2E-02
Skeletal muscle	muscle system process	GO:0003012	6	14.3	1.3E-02

Chapter 6 – Identification of candidate long noncoding RNAs acting as competitive regulatory elements in a transcriptional network linked to hormone activity

6.1 Abstract

The adult human brain's transcriptome shows a vast diversity in RNA types, their relative expression among different brain regions, and cell types [27, 67]. An increasing number of studies provide evidence for the biological role of noncoding RNAs (ncRNAs), for example for fine-tuning the expression of coding transcripts and the modulation of cross-talk among them [27]. However, the functional role of the majority of ncRNAs remains unknown. In the present study, I sought to identify a brain region-specific transcriptional co-expression network and to detect candidate long noncoding RNAs (lncRNAs) that act as regulatory RNAs for these region-specific transcriptional networks. I hypothesized that the association of the lncRNAs with the brain region-specific network would be reflected by an increased co-expression between the lncRNAs and the protein-coding genes in the network, and, as an independent line of evidence, an increased number of microRNAs (miRNAs) whose binding is shared between the lncRNAs and the network's mRNAs. The analysis of transcriptional networks from the human adult brains yielded one module that was specifically expressed in the hypothalamus and was functionally linked to hormone activity. Different sets of lncRNAs whose expression levels were significantly correlated with the "hormone module's" representative expression pattern showed an increased number of shared miRNAs with the mRNAs of the protein-coding genes in the "hormone module". More specifically, the lncRNAs that were more strongly correlated with the

expression of the “hormone module” genes tended to also share a higher median percentage of response elements for the same miRNAs. Thus, the present study identified a set of candidate lncRNAs that could act in a hypothalamus-specific transcriptional network that regulates fine-tuning of expression among genes that modulate hormone activity.

6.2 Introduction

Brain regions show characteristic gene expression patterns that are highly conserved across individuals [3, 73, 77]. This regional identity partly depends on differences in cell type composition, such as varying relative abundances of neurons, oligodendrocytes, astrocytes and microglia [67, 73, 78].

Transcript abundance is fine-tuned by transcriptional and post-transcriptional regulatory elements, including different types of noncoding RNAs [27]. Here, I considered two types of noncoding RNAs which have been suggested to interact during post-transcriptional regulation. The first type were miRNAs (~22 nucleotide long microRNA) which bind to near complementary sequences on their target RNAs and subsequently repress transcript abundance or translation [9]. The second type were lncRNAs (long noncoding RNAs, arbitrarily defined as ≥200 nucleotides [6, 311, 312]) that display moderate evolutionary conservation compared to neutral sequences, indicating some degree of biological functionality [313]. In this study, I considered intergenic or antisense lncRNAs, with the latter defined as transcripts that intersect exons of protein-coding loci on the opposite strand or that have published evidence for antisense regulation of a coding gene [311].

One potential mechanism for the regulation of transcript abundance involves 'competitive endogenous RNAs' (ceRNAs) [27]. These ceRNAs – for example mRNAs or lncRNAs - share miRNA response elements (MREs) for the same miRNAs, thereby competing for these miRNAs for binding and regulating each other's transcript abundance [27] (Figure 27).

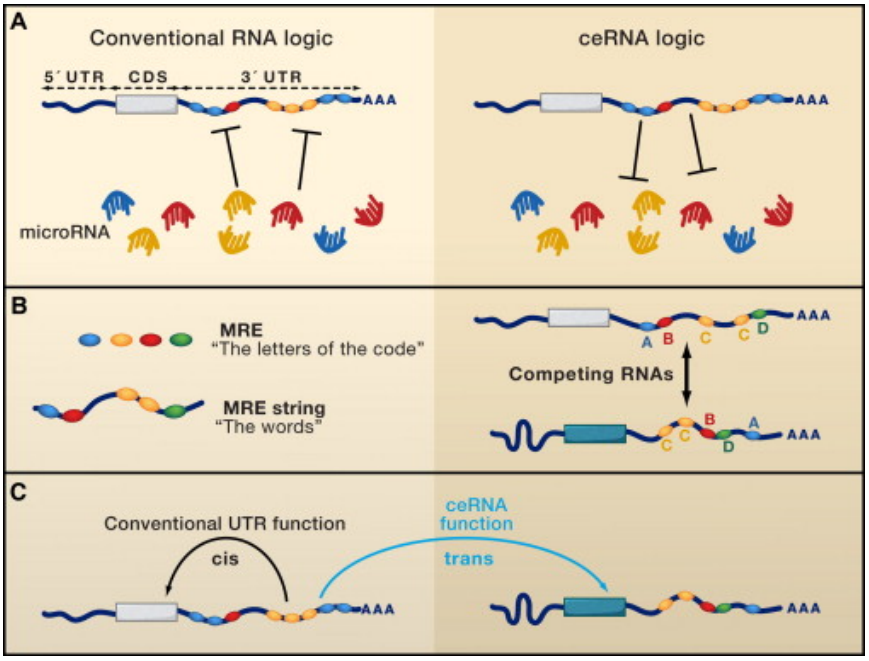


Figure 27. The potential mechanism of ceRNA activity. A) mRNAs could regulate each other's transcript abundance by a reciprocal relationship between mRNAs and miRNAs. B) This reciprocal relationship between mRNAs and miRNAs could be 'communicated' by miRNA response elements (MREs) for miRNAs shared between two target genes. Thus, the higher the number of shared interacting miRNAs, the stronger the cross-talk between the transcripts could be. C) The 3'UTRs of mRNAs contain a high density of MREs. MiRNAs can bind to these MREs and act as cis regulatory elements to regulate the stability and the abundance of the 3'UTRs' own transcript. Furthermore, ceRNAs activity could allow regulation of transcript levels in trans by altering the availability of miRNAs Figure reproduced from [27] with permission from Elsevier, www.cell.com.

Both lncRNAs and miRNAs show high regional specificity in their expression patterns, and these depend on brain region as well as cell type [314, 315]. Importantly, noncoding (nc) RNAs are trafficked to distinct subcellular locations, which has been interpreted as evidence for biological functionality as opposed to cell type specific transcriptional noise, for example due to differential open chromatin conformation between cell types [314].

The regional specificity of ncRNAs leads to the question of whether these ncRNAs engage in region-specific gene regulatory networks (GRNs), for example by miRNA-mediated ceRNA cross-talk between 3'UTRs of protein-coding transcripts and lncRNAs. In the present study, I aimed to identify region-specific GRNs of protein-coding genes and their regulatory ncRNAs. More specifically, I first used a data-driven approach to generate transcriptional networks and then aimed to detect lncRNAs whose association with a region-specific transcriptional subnetwork was supported by (*i*) the correlation of the lncRNAs' expression levels with the representative expression pattern of the brain region-specific transcriptional subnetworks, (*ii*) the increased expression of the lncRNA in the respective brain region compared to other regions, and (*iii*) an increase in shared MREs between the lncRNA and the 3'UTRs of the transcripts assigned to the brain region-specific transcriptional subnetworks, indicated by a high number of predicted MREs for miRNAs targeting both genes.

6.3 Results

6.3.1 Transcriptional network

The coexpression transcriptional networks were built applying WGCNA [122] to RNA-Seq gene expression data from 20195 protein-coding genes with annotated ENSEMBL identifiers, expressed in 225 brain tissue samples from the Genotype-Tissue Expression project (GTEx [113]), covering 13 brain regions (Figure 28) in 35 human adult subjects (20-70 years, 7.0<RIN<9.8). The coordination of gene expression across time and different brain regions allows detection of gene modules that each share a representative expression pattern, quantified by their first principal component, which in the following is referred to as the "module eigengene" (ME) [122]. I detected 42 gene modules in the GTEx data set, with 18733 genes being assigned to one of these modules. Each of the 18733 genes was expressed in at least one sample (FPKM>0.1). All of the modules showed significant enrichment for GO-terms (Table 2), compared to a background of all genes that were assigned to a module. To test whether the module's representative expression pattern showed a link to external traits, I correlated the ME of each module with the following seven variables: Age, sex, Hardy index (death classification, see **Chapter 6.6.1**), brain region, subject identifier, sequencing library depth, and RIN value (Pearson's correlation, threshold for significance $\alpha<0.05$). The correlation with the subject identifier was used to control for effects caused by inclusion or exclusion of single subjects.

6.3.2 Modules specific to brain regions

One module's ME only showed significant correlation with brain region and no other external variable after multiple testing correction for the 42 modules that were tested (Table 2 and Figure 28A, Pearson's correlation=0.3, $\alpha<0.05$). Two modules' MEs were

significantly correlated with brain region and no other external variables at uncorrected p-values (Table 2 and Figure 28B (Pearson's correlation=0.2) and Figure 28C (Pearson's correlation=0.1), $\alpha<0.05$). These three modules showed significant enrichment for GO-terms (Table 2) linked to (*i*) regionalization and pattern formation (Figure 28A; genes shown in Supplementary Table S6.1), (*ii*) hormone activity (Figure 28B; genes shown in Supplementary Table S6.2), and (*iii*) feeding behaviour with a trend for hormone activity and sleeping behaviour (Figure 28C; genes shown in Supplementary Table S6.3). The module enriched for regionalization was most highly expressed in the spinal cord (Figure 28A) and the modules linked to hormone activity and sleeping/feeding behaviour in the hypothalamus (Figue 28B and Figure 28C) which is in line with the known role of the hypothalamus in hormone regulation [316].

Table 2. Significant GO-term enrichment in the three modules associated with brain regions. Each module was assigned a name, indicated in the header lines. A yellow background highlights the non-redundant terms, summarized using REViGO [117]. P-values are Benjamini-Hochberg corrected (α<0.05); 'fold' refers to fold enrichment of each GO-term.

Regionalization module (42 genes with DAVID identifiers out of 45 genes)			
category	**term**	**fold**	**p-value**
GO:0048706	embryonic skeletal system development	66.4	1.30E-21
GO:0009952	anterior/posterior pattern formation	42	1.20E-21
GO:0003002	regionalization	31.3	6.30E-21
GO:0048562	embryonic organ morphogenesis	40.8	2.80E-20
GO:0007389	pattern specification process	24.2	2.40E-20
GO:0048704	embryonic skeletal system morphogenesis	78.4	3.90E-20
GO:0043565	sequence-specific DNA binding	13.3	7.00E-20
GO:0048568	embryonic organ development	31.7	1.00E-18
GO:0048705	skeletal system morphogenesis	42.8	5.20E-18
GO:0001501	skeletal system development	19.1	1.60E-17
GO:0048598	embryonic morphogenesis	18.9	2.50E-16
GO:0003700	transcription factor activity	8.3	1.50E-15
GO:0030528	transcription regulator activity	6	6.70E-15
GO:0043009	chordate embryonic development	16.6	2.50E-14
GO:0009792	embryonic development ending in birth or egg	16.5	2.30E-14
GO:0006355	regulation of transcription, DNA-dependent	5.3	3.80E-14
GO:0051252	regulation of RNA metabolic process	5.2	6.20E-14
GO:0045449	regulation of transcription	3.7	5.50E-11
GO:0003677	DNA binding	4	8.20E-11
GO:0006350	transcription	4.1	3.50E-10
GO:0051216	cartilage development	25.9	8.10E-05
GO:0021602	cranial nerve morphogenesis	116.1	1.20E-04
GO:0030878	thyroid gland development	106.4	1.50E-04
GO:0021545	cranial nerve development	67.2	6.10E-04
GO:0021563	glossopharyngeal nerve development	319.3	6.40E-04
GO:0021615	glossopharyngeal nerve morphogenesis	319.3	6.40E-04
GO:0030902	hindbrain development	26.6	7.20E-04
GO:0048732	gland development	14.3	1.10E-03
GO:0021675	nerve development	51.1	1.20E-03
GO:0006357	regulation of transcription from RNA polymerase II promoter	4.5	4.70E-03

GO:0045935	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	4.7	8.50E-03
GO:0051173	positive regulation of nitrogen compound metabolic process	4.5	1.00E-02
GO:0031328	positive regulation of cellular biosynthetic process	4.3	1.40E-02
GO:0009891	positive regulation of biosynthetic process	4.2	1.50E-02
GO:0035270	endocrine system development	18.5	1.90E-02
GO:0045941	positive regulation of transcription	4.6	2.00E-02
GO:0010628	positive regulation of gene expression	4.5	2.40E-02
GO:0010557	positive regulation of macromolecule biosynthetic process	4	4.40E-02
GO:0002009	morphogenesis of an epithelium	12.8	4.90E-02

Hormone module (100 genes with DAVID identifiers out of 112 genes)

category	term	fold	p-value
GO:0042277	peptide binding	9.7	3.50E-05
GO:0005184	neuropeptide hormone activity	44.3	2.10E-05
GO:0007631	feeding behavior	18.3	2.30E-04
GO:0008528	peptide receptor activity, G-protein coupled	13.6	2.00E-05
GO:0001653	peptide receptor activity	13.6	2.00E-05
GO:0007186	G-protein coupled receptor protein signaling pathway	3.3	1.00E-03
GO:0009725	response to hormone stimulus	5.7	7.20E-04
GO:0007610	behavior	4.8	1.40E-03
GO:0009719	response to endogenous stimulus	5.2	1.20E-03
GO:0030594	neurotransmitter receptor activity	12.9	8.90E-04
GO:0048545	response to steroid hormone stimulus	7.6	3.50E-03
GO:0042165	neurotransmitter binding	11.9	1.10E-03
GO:0005179	hormone activity	11.6	1.10E-03
GO:0043025	cell soma	8.6	4.30E-03
GO:0030141	secretory granule	8.1	3.00E-03
GO:0051047	positive regulation of secretion	10.2	8.30E-03
GO:0006972	hyperosmotic response	48.5	8.20E-03
GO:0007267	cell-cell signaling	3.8	7.40E-03
GO:0008188	neuropeptide receptor activity	21.8	2.50E-03
GO:0042923	neuropeptide binding	21.2	2.40E-03
GO:0007166	cell surface receptor linked signal transduction	2.2	2.30E-02
GO:0051046	regulation of secretion	6.3	2.40E-02
GO:0010817	regulation of hormone levels	7.4	2.80E-02
GO:0045777	positive regulation of blood pressure	27.4	3.00E-02
GO:0008217	regulation of blood pressure	9.6	2.70E-02
GO:0010033	response to organic substance	3.1	2.70E-02
GO:0017046	peptide hormone binding	25.1	1.30E-02
GO:0042538	hyperosmotic salinity response	78.8	3.60E-02
GO:0044057	regulation of system process	4.7	3.50E-02
GO:0007618	mating	23.4	3.40E-02

GO:0006970	response to osmotic stress	20.3	4.90E-02
GO:0043565	sequence-specific DNA binding	3.2	4.30E-02

Feeding module (29 genes with DAVID identifiers out of 33 genes)

category	term	fold	p-value
GO:0007610	behavior	8.8	2.50E-02
GO:0007631	feeding behavior	33.7	3.10E-02
GO:0007218	neuropeptide signaling pathway	25.2	4.80E-02

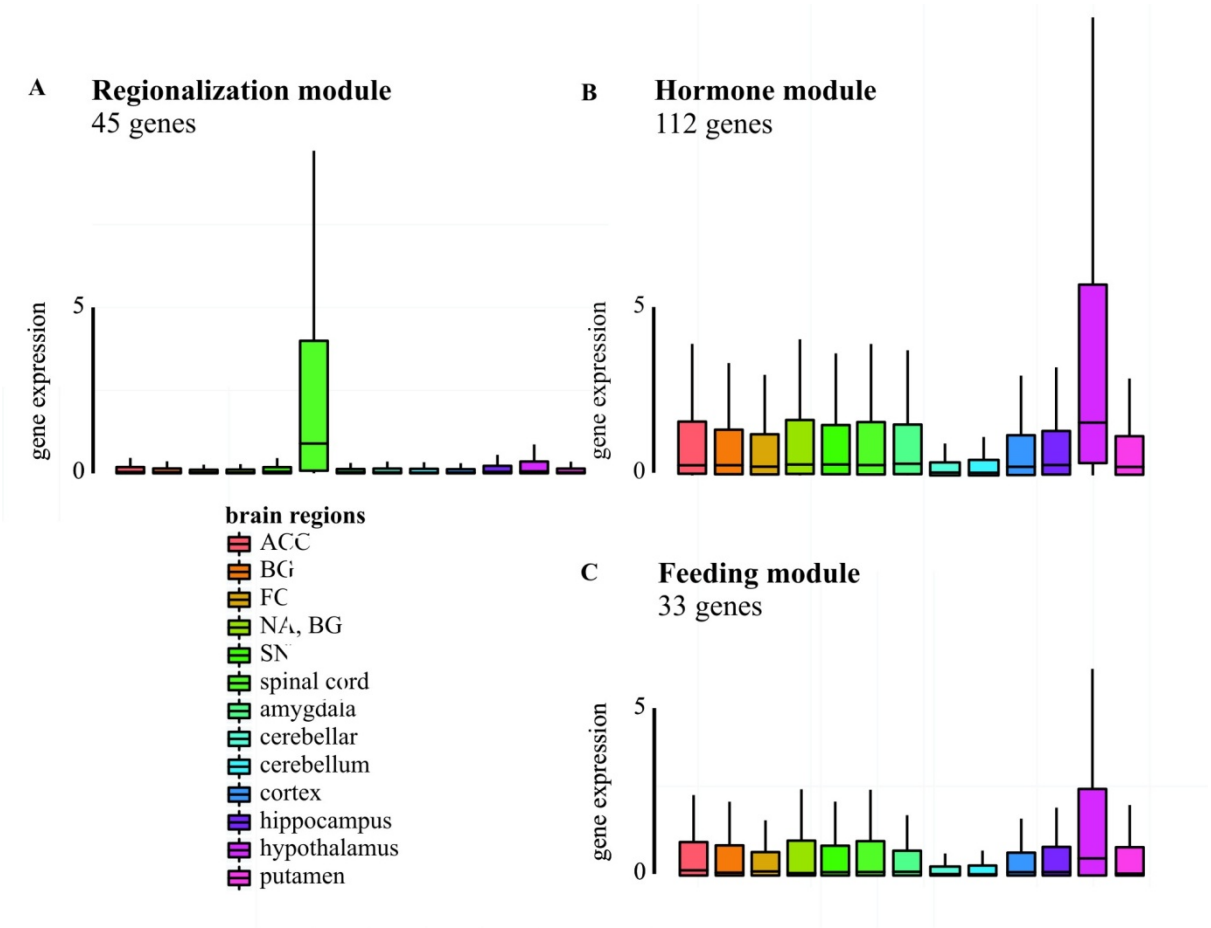


Figure 28. Expression of all genes in the three modules correlated with brain region (α<0.05, uncorrected) and no other external variable, separately for all available brain regions. The legend indicates brain regions with the abbreviations: ACC (anterior cingulate cortex), BG (basal ganglia), FC (frontal cortex), NA, BG (nucleus accumbens, basal ganglia), cerebellar (cerebellar hemisphere), and SN (substantia nigra). Gene expression is given in FPKM, normalized across all libraries (see Chapter 6.6.1).

6.3.3 LncRNAs correlated with modules' eigengenes

Next I hypothesized that a network of regulatory lncRNAs plays a role in fine-tuning the expression of the protein-coding genes in the modules. The annotation of lncRNAs and brain-expressed miRNAs is described in detail in **Chapter 6.6.3** and **Chapter 6.6.4**. To identify lncRNAs that regulate the modules' gene expression, I selected lncRNAs based on (*i*) the correlation of the lncRNAs' expression levels with the module eigengene, (*ii*) the increased expression of the lncRNA in the respective brain region, and (*iii*) an increase in shared miRNAs predicted to bind the lncRNA and the mRNAs in the region-specific modules (described in further detail below; Figure 29).

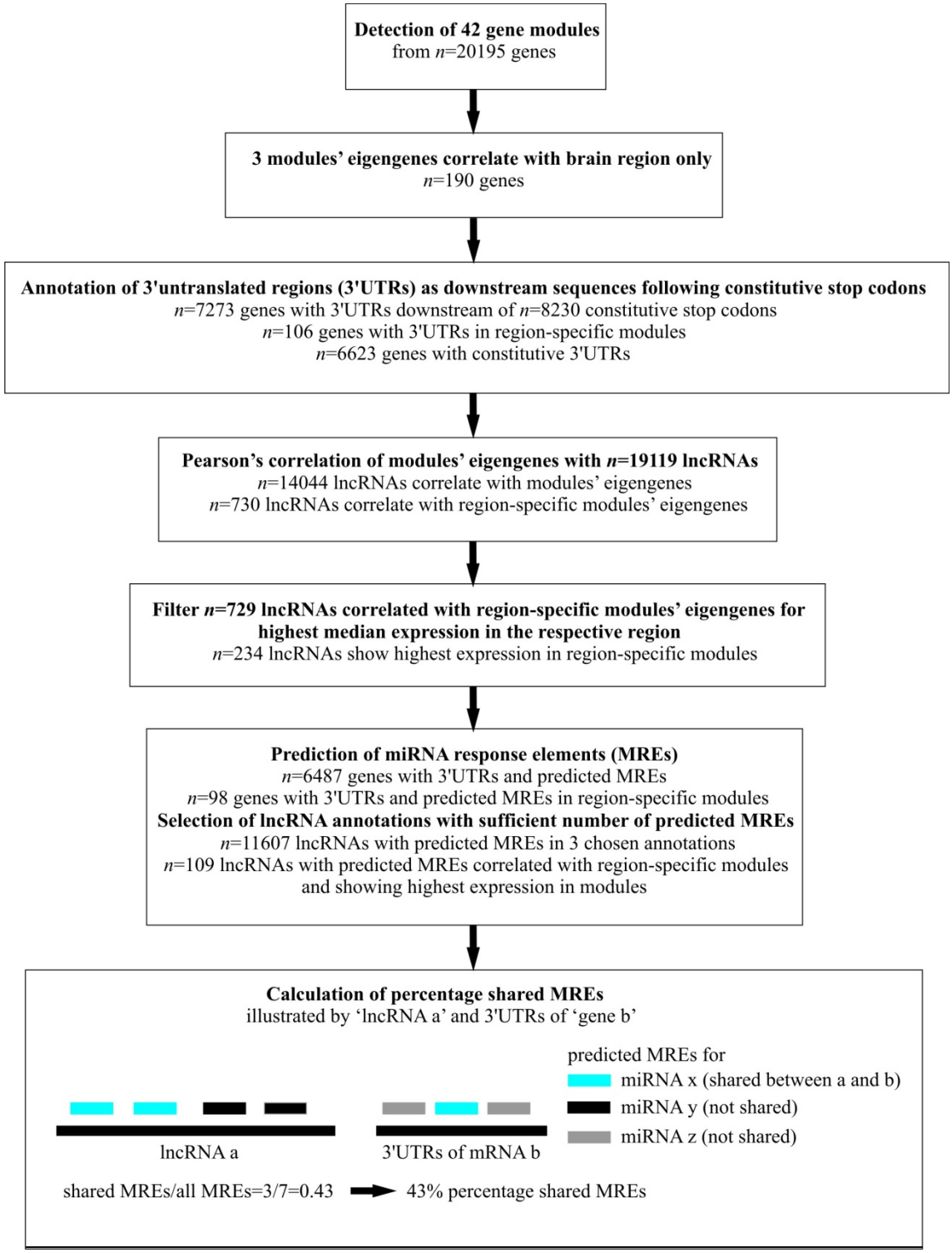


Figure 29. Flowchart for detection of gene modules, annotation of 3'UTRs, and association of lncRNAs with modules' eigengenes (MEs) using correlation and highest median expression values (see Chapter 6.6). The bottom panel shows an

illustration of how the 'percentage shared MREs' for 'lncRNA a' and the 3'UTRs of the mRNA of 'gene b' were calculated. In this example, a total of 7 MREs were predicted to bind to 3 different miRNAs (referred to as x, y, and z). Only 'miRNA x' targets both 'lncRNA a' and the 3'UTRs of the mRNA of 'gene b'. The percentage shared MREs therefore is calculated as the ratio of 3 predicted MREs for the shared miRNA x, divided by all 7 predicted MREs for both genes.

The correlation between the lncRNAs' expression levels and the modules' MEs was calculated using Pearson's correlation ($\alpha<0.05$, corrected for all n=19119 lncRNAs tested). To exclude correlations confounded by single outliers, I used only the lncRNAs with the highest expression in spinal cord for the regionalization module, or in the hypothalamus for the hormone and feeding modules, respectively. A lncRNA was considered to have its 'highest expression' in a specific region if it showed either the highest median value across individuals in a particular tissue sample type or the highest 3rd quantile (the value ranked higher than 75% of all other values) of its expression in this region (Figure 30).

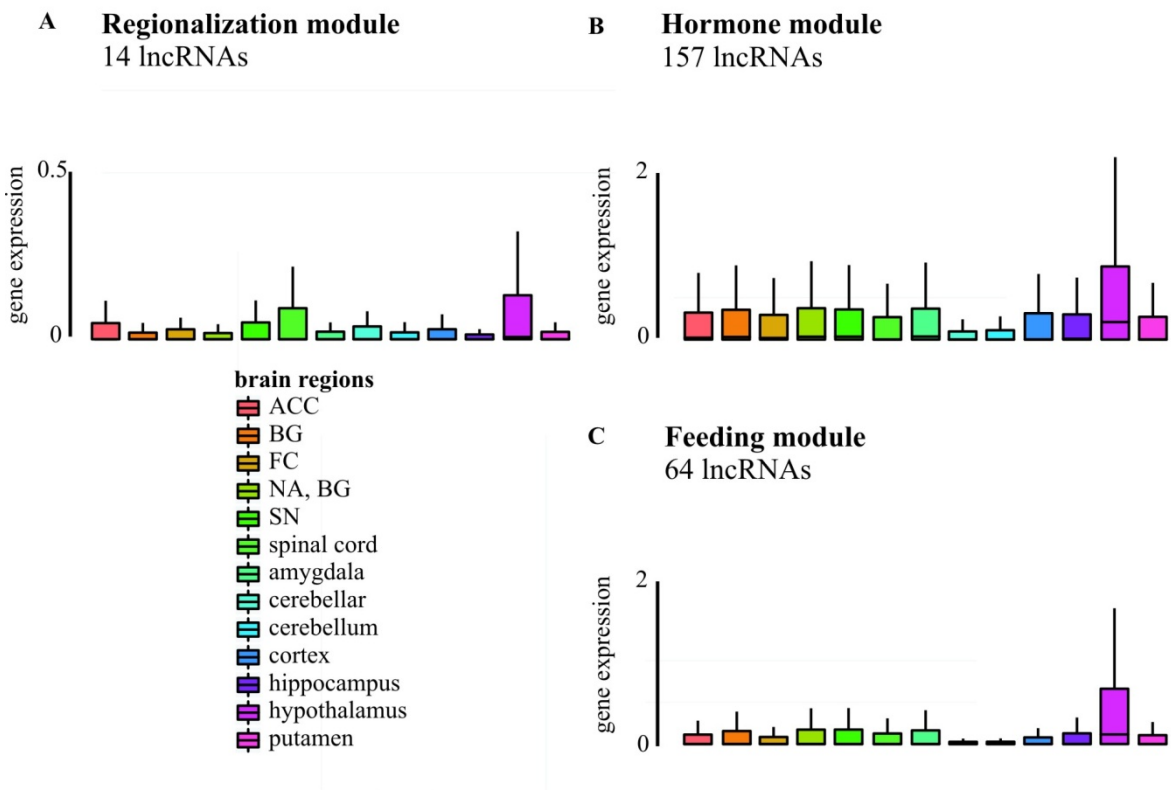


Figure 30. Expression of all lncRNAs whose expression correlated with the module eigengenes (MEs) of the three modules that were correlated with brain regions ($\alpha<0.05$, corrected for all lncRNAs tested) and no other external variable, separately for all available brain regions. The legend indicates brain regions with the abbreviations: ACC (anterior cingulate cortex), BG (basal ganglia), FC (frontal cortex), NA, BG (nucleus accumbens, basal ganglia), cerebellar (cerebellar hemisphere), and SN (substantia nigra). Gene expression is given in FPKM, normalized across all libraries (see Chapter 6.6.1).

Next, I asked whether the association between the genes in the brain region-specific modules and the selected lncRNAs is further supported by shared miRNA-mediated regulatory elements, which could fine-tune the co-expression of these genes. The 3' untranslated regions (3'UTRs) of mRNAs show the highest density of binding motifs for miRNAs [256, 317]. The diversity of extended 3'UTR isoforms leads to the gain and loss of regulatory elements and therefore influences transcript stability, localization and translational efficiency [318] (see **Chapter 5**). Tissues express 3'UTRs with varying lengths with the longest extended isoforms localized in the central nervous system (CNS) [318, 319] (see **Chapter 5**). The diversity of 3'UTR isoforms among human tissues leads to the question of whether 3'UTR lengths are uniformly distributed across brain regions or are region-specific. The length distribution of 3'UTRs in brain regions determines the availability of binding elements for miRNAs within these regions. For this reason, I asked whether the lengths of 3'UTRs differed across brain regions and whether the observed genome-wide length distribution of 3'UTRs was recapitulated in the 3'UTRs of the mRNAs in one of the brain region-specific modules, namely the hormone module which included the highest number of mRNAs with predicted MREs (described in detail below).

To do so, I compared the lengths of alternatively spliced 3'UTR isoforms across 13 CNS tissues available in the GTEx RNA-Seq data set (Figure 31). Only two tissues differed significantly from all other tissues that were considered (using 2-sample Kolmogorov-Smirnov (KS) tests; $\alpha<0.05$, corrected for all tests performed). More specifically, the cerebellum showed the longest 3'UTR isoforms and the anterior cingulate cortex (ACC) contained the shortest extensions (Figure 31). The length distribution of alternative 3'UTRs of the mRNAs of the genes assigned to the hormone module is shown as an example in Figure 31.

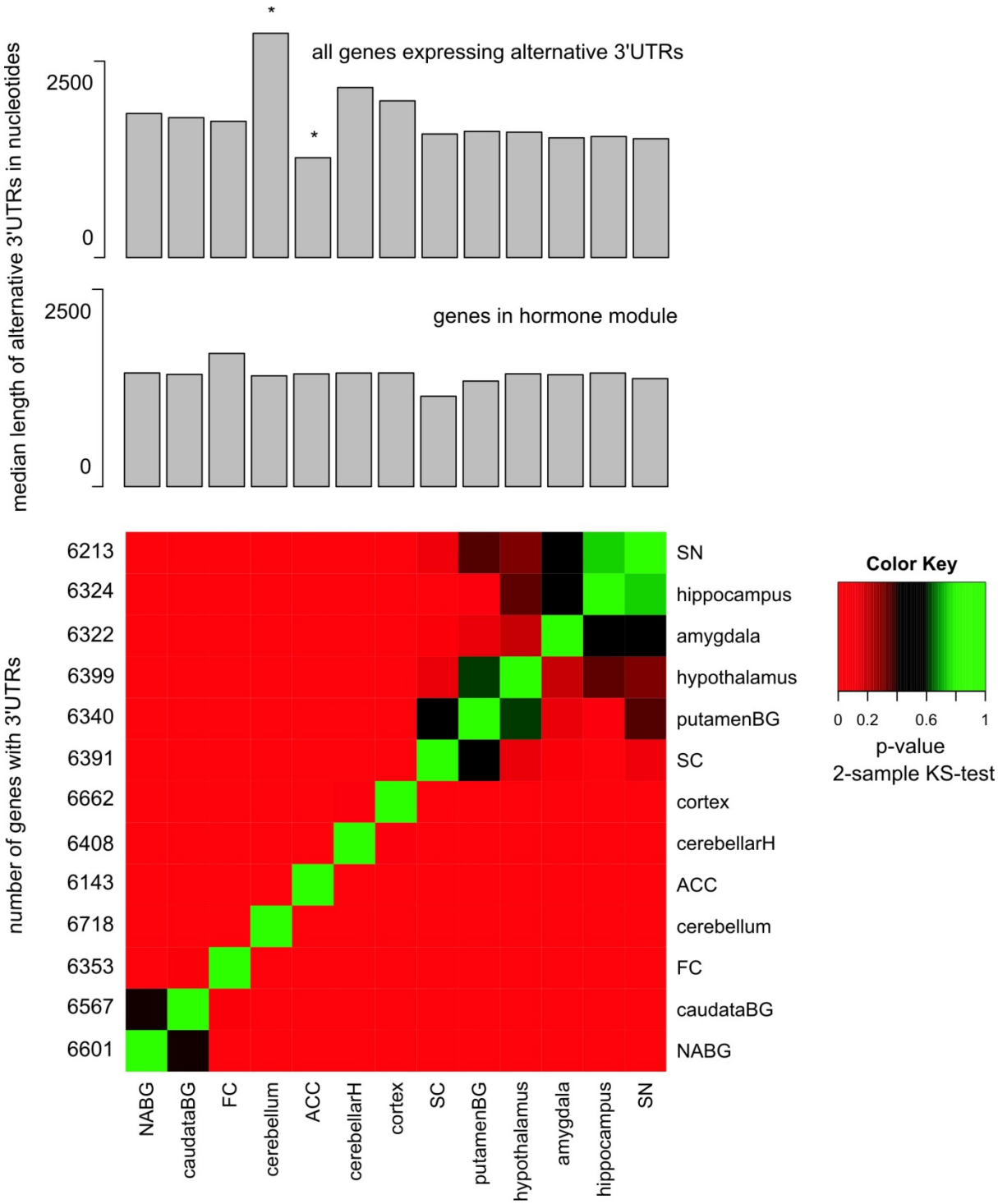


Figure 31. Genome-wide length distribution of alternative 3'UTRs shows significantly longer extensions in the cerebellum and significantly shorter extensions in the ACC, but no significant differences among the alternative

3'UTRs of the mRNAs in the hormone module. The barplots show median lengths of expressed alternative 3'UTRs in number of nucleotides merged across all samples for a brain region (indicated on the x-axis of heatmap). The barplot on the top includes all genes with alternative 3'UTRs, and the barplot below includes only the genes in the hormone module with alternative 3'UTRs. Asterisks indicate the only two regions (cerebellum and ACC) which differed significantly from all other brain regions using 2-sample Kolmogorov-Smirnov (KS) tests ($\alpha<0.05$, corrected for all tests performed). The heatmap shows all p-values for the 2-sample KS tests between all brain regions. The 3'UTRs were annotated as described in Chapter 6.6.2 and illustrated in Figure 29. The x-axis indicates brain regions with the following abbreviations: ACC (anterior cingulate cortex), BG (basal ganglia), cerebellarH (cerebellar hemisphere), FC (frontal cortex), NABG (nucleus accumbens, basal ganglia), and SN (substantia nigra).

6.3.4 Shared MREs among the hormone module genes' transcripts and the associated lncRNAs

Next, I hypothesized that the lncRNAs that significantly correlated with the MEs' of brain region-specific modules, and that were highly expressed in the same region as the given module (referred to as "associated lncRNAs", see **Chapter 6.6.6**), would share a higher proportion of predicted miRNA response elements (MREs) for miRNAs with the transcripts in the module than would randomly selected lncRNAs. To test for evidence of miRNA modulated cross-talk between the genes in the brain region-specific modules and the "associated" lncRNAs, I predicted MREs only for 80 brain-expressed miRNAs (see Supplementary Table S6.4 and **Chapter 6.6.7**) for all lncRNAs and the 3'UTRs of the mRNAs from the region-specific modules. The 3'UTRs of 6487 genes contained MREs for the 80 miRNAs (Supplementary Table S6.4; for details on 3'UTRs annotation and MREs prediction see **Chapter 6.6.2** and **Chapter 6.6.7**).

The hormone module included 58 genes with at least one MRE predicted in their 3'UTRs. These 58 genes retained the strong enrichment for GO-terms linked to hormone activity (Supplementary Table S6.5). The median number of predicted MREs (n=15) for the 58 hormone module genes was not significantly different compared to all 6487 genes which had at least one MRE predicted (n=22, Wilcoxon rank sum test, p-*value*=0.1). Furthermore, the number of shared MREs per gene-pair normalised by the total number of MREs for the gene-pair (see bottom panel in Figure 29) accounted for differences in the 3'UTR lengths and the MRE densities.

The regionalization and feeding modules showed substantially lower numbers of genes with at least one predicted MRE, namely 23 genes in the case of the regionalization module, and 17 genes in the case of the feeding module. For this reason, the following analysis focused on the hormone module only.

The percentage of shared predicted MREs was calculated for all gene-pairs, including the 58 protein-coding genes in the hormone module with at least one predicted MRE in their 3'UTRs and the associated antisense or intergenic lncRNAs (lincRNAs). This was achieved by dividing the number of predicted MREs for miRNAs targeting both genes/lncRNAs for a given pair by the total number of predicted MREs for both genes/lncRNAs (Figure 29, for an example see Figure 35). In the following, this measure is referred to as the "percentage of shared MREs". The median percentage of shared MREs between all possible gene-pairs among the 3'UTRs of the genes in the hormone module and the associated lncRNAs was compared to the same measure from 1000 random sample-size matched sets of lncRNAs, from which the associated lncRNAs were excluded. The analysis was performed separately for all lncRNA annotations that were considered, whereby no lncRNA was part of more than one annotation, and none of the lncRNAs overlapped with other lncRNAs (see **Chapter 6.6.4** for details on the lncRNA annotations).

The lincRNA sets referred to as "Scripture" [312], and "Hangauer" [6] were not further investigated as they contained only 3 and 4 associated lincRNAs respectively after all filtering steps were applied.

6.3.4.1 Antisense lncRNAs

All 33 associated antisense lncRNAs [311, 320] contained at least one predicted MRE. Overall, the median number of predicted MREs per lncRNA (n=51) was not significantly different from the whole set of antisense lncRNAs in which each had at least one predicted MRE (n= 34, Wilcoxon rank sum test, *p-value*=0.1).

The 33 associated antisense lncRNAs showed a trend towards a higher median percentage of shared MREs with the genes in the hormone module than expected by

random sampling of an equal number of lncRNAs (19%, *empirical p-value*=0.06, Figure 32). This result was highly specific to the 58 coding genes in the hormone module. If the hormone module genes and the associated lncRNAs were instead compared to randomly sampled coding genes, the median percentage of shared MREs was not higher than expected by chance. The same held true when only the subsets of coding-genes in the module, or associated lncRNAs were considered separately. The specificity of the results against different background randomisations indicates that the observed increase in shared MREs is due to the genes in the hormone module sharing a larger number of MREs with the associated lncRNAs in particular compared to other lncRNAs, but that the complete set of hormone module genes and associated lncRNAs did not share more MREs across all gene-pairs than expected by chance.

The observed difference between sampling lncRNAs only or protein-coding genes could be influenced by the availability of brain-region specific miRNAs. Clearly, a high number of brain region-specific miRNAs that target region-specific genes would lead to a higher number of shared miRNAs between these genes relative to other regions. To test the relative density of MREs for all 80 miRNA families considered in this analysis, I compared the number of each miRNA's predicted MREs per gene, normalized by the total number of MREs per gene, across all genes in the hormone module and the associated lncRNAs to the same measure from 1000 sample-size matched random samples of coding genes and lncRNAs. Only one miRNA, namely miRNA-488 which is highly expressed in the hypothalamus [321], showed a higher density of MREs in the 3'UTRs of the hormone module mRNAs and the associated antisense lncRNAs (MRE density=0.035, *empirical p-value*=0.03, uncorrected for multiple testing). In comparison, the set of 58 hormone module genes showed a significantly higher MREs density (α=0.05, uncorrected for multiple testing) for 23 miRNAs, including miR-488

(MRE density =0.041, *empirical p-value*=0.009), but no miRNA had a significantly higher density in the 33 associated antisense lncRNAs. The observed enrichment for only one miRNA is in line with only few published studies that reported high or even specific expression of single miRNAs in the hypothalamus [322, 323]. Thus, the observed specificity of an increased number of shared MREs to the hormone module genes and the associated lncRNAs compared to random lncRNAs as described above could be influenced by a relatively higher number of miRNAs targeting primarily other brain regions and not the hypothalamus.

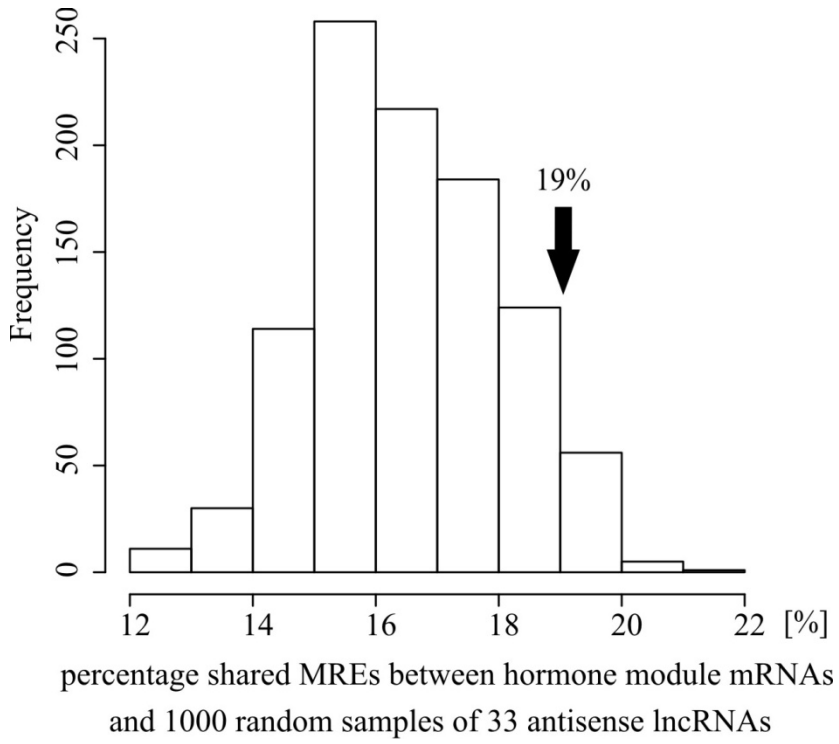


Figure 32. Median percentage shared MREs between the 33 antisense lncRNAs associated with the hormone module and the 3'UTRs of 58 genes in the hormone module, which had at least one predicted MRE, compared to an empirical null distribution. The median percentage shared MREs for the genes in the hormone module and the associated lncRNAs is indicated by a black arrow. The histogram shows the same measure for 1000 random samples of 33 lncRNAs that were not associated with the hormone module.

6.3.4.2 LincRNAs

The median percentage shared MREs was calculated for two further sets of lincRNAs (described in detail in **Chapter 6.6.4**), namely 54 lincRNAs from the GENCODE set with at least one predicted MRE referred to as "lincRNAs" [311], and 23 lincRNAs from the Cabili et al. [312] data set with at least one predicted MRE referred to as "Cufflinks" [312], in the same way as described above. Briefly, the median percentage of shared MREs between all possible gene-pairs among the hormone module genes and the associated lncRNAs was compared to the same measure considering randomly sampled lncRNAs. The "lincRNAs" set showed a trend for an increased percentage of predicted MREs for shared MREs (22%, *empirical p-value*=0.06). The only miRNAs that showed increased density in the "lincRNA" set combined with the hormone module genes were miR-488 (MRE density=0.031, *empirical p-value*=0.03, uncorrected for multiple testing), and miR-338-5p (MRE density =0.029, *empirical p-value*=0.046, uncorrected for multiple testing).

Similar results were observed for the "Cufflinks" set (percentage shared MREs=17%, *empirical p-value*=0.079). The "Cufflinks" set and the hormone module genes showed nominally significant enrichment for 11 miRNAs (miR-101/101ab, miR-149, miR-194, miR-197, miR-338-5p, miR-370, miR-377, miR-432, miR-488, miR-9/9ab, and miR-300-5p/4709-3p).

6.3.5 Correlation between the antisense lncRNA's module membership and the median percentage of MREs it shares with the hormone module genes

I next asked whether the module membership (MM) of the lncRNAs, defined as the correlation of a lncRNA's expression with the ME, was correlated with the median percentage of MREs that are shared with the hormone module genes. To test whether the MM could predict the median percentage of shared MREs, I calculated the Spearman's correlation between a lncRNA's MM and the median percentage of shared MREs between the lncRNA and the genes in the module.

6.3.5.1 Antisense lncRNAs

There was a significant positive correlation between the MM and the median percentage of shared MREs between the 33 associated antisense lncRNAs and the genes in the hormone module (Rho=0.37, *p-value*=0.03, median percentage shared MREs across 33 lncRNAs=24%, median MM across 33 lncRNAs Rho=0.40). This correlation indicates that lncRNAs with a high MM tend to share an increased number of MREs with the hormone module genes. The same measure was calculated for 1000 randomly sampled sets of 33 antisense lncRNAs to test whether the observed correlation was specific to the hormone module-associated lncRNAs (Figure 33). The 33 antisense lncRNAs that were associated with the hormone module did indeed show a significantly stronger correlation between the MM and the median percentage MREs than random sets of antisense lncRNAs (*empirical p-value*=0.03).

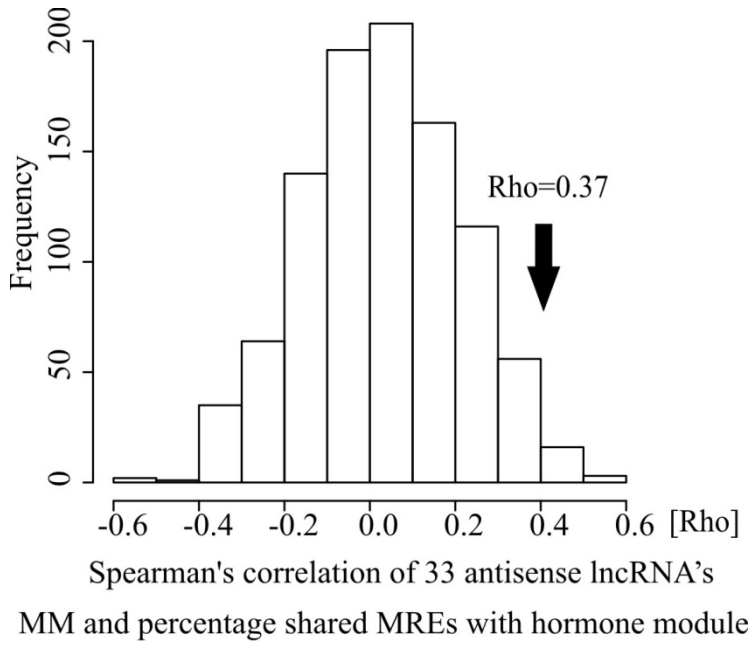


Figure 33. **Spearman's correlation between the module membership (MM; defined as the correlation of a gene's expression and the module eigengene (ME)) of the 33 associated antisense lncRNAs and the median percentage shared MREs between the lncRNA and the 3'UTRs of the mRNAs in the hormone module compared to an empirical null distribution. The correlation between a lncRNA's MM and the median percentage shared MREs between the lncRNA and the hormone module genes is indicated by a black arrow. The histogram shows the same measure for 1000 random samples of 33 lncRNAs that were not associated with the hormone module.**

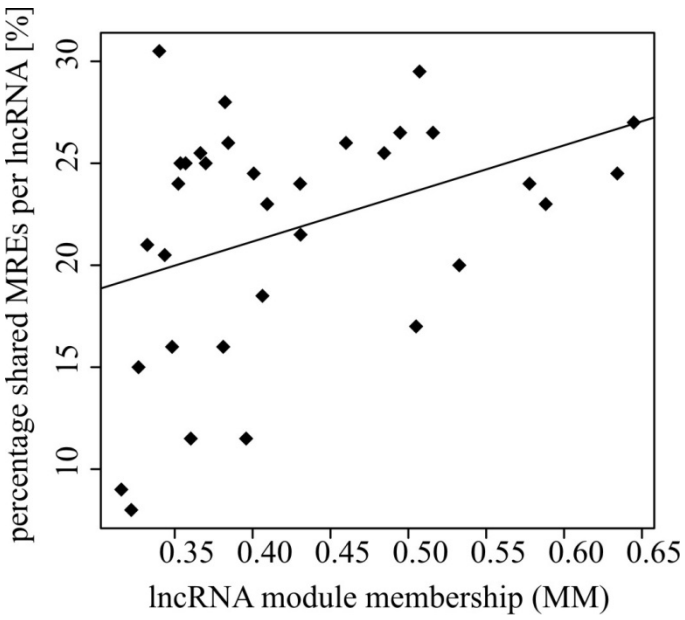


Figure 34. **Scatterplot for the 33 hormone module-associated antisense lncRNAs showing each lncRNA's module membership (MM) versus the median percentage of shared MREs between the lncRNA and all genes in the hormone module. The black demarcates the fit of the linear regression model (Spearman's correlation=0.37).**

The 33 associated antisense lncRNAs are summarized in Table 3. The description indicates whether each lncRNA was (*i*) intersecting, (*ii*) the nearest neighbour or (*iii*) within 100kb vicinity (upstream and downstream) of a protein-coding gene that is part of the hormone module.

The percentage shared MREs between the genes in the hormone module and the associated antisense lncRNAs ranged from 0% to 100%, with a median of 19% across all gene-pairs. 100% indicates that the two genes share all their predicted MREs, which was the case for one single gene-pair, namely ENSG00000213904 (LIPE; Hormone-

sensitive lipase; antisense RNA) and Neuroglobin (*NGB*), whereby the two genes contained predicted MREs for almost all of the considered miRNA families (*n*=79). Figure 35 shows an example of the shared predicted MREs and hypothalamus-specific expression for a gene in the hormone module (*PGR*, ENSG00000082175, progesterone receptor; receptors for sex steroids, including progesterone, are present in neuron and glial cells where they mediate steroid effects [324]) and the associated antisense lncRNA ENSG00000225764 (*LEPREL1* antisense RNA 1).

Table 3 33 antisense lncRNAs associated with the hormone module. 'Median percentage shared MREs' and module membership (MM) are explained in detail in the text.

chr	**start**	**stop**	**strand**	**ENSEMBL** identifier	**MM**	**median percentage shared MREs**	**HGNC**	**description of antisense lncRNAs and the protein-coding genes they overlap**
chr3	189838752	189862635	+	ENSG00000225764	0.34	30.5	LEPREL1-AS1	LEPRECAN-LIKE 1; *LEPREL1*; PROLYL 3-HYDROXYLASE 2; *P3H2*; required for proper collagen biosynthesis, folding, and assembly
chr4	106461346	106473512	-	ENSG00000250522	0.38	26.5	AC004066.3	
chr5	985229	997432	-	ENSG00000215246	0.49	24	RP11-43F13.3	
chr5	10479482	10482421	+	ENSG00000246016	0.32	6	RP11-1C1.7	
chr5	58037407	58112926	-	ENSG00000248475	0.37	22.5	RP11-479O16.1	
chr5	146559766	146614422	-	ENSG00000250343	0.35	23	CTC-255N20.1	
chr5	50668569	50679166	-	ENSG00000259663	0.51	27.5	CTD-2314G24.2	antisense lncRNA intersects with the protein-coding gene *ISL1* that is part of the hormone module

chr6	28864306	28865099	+	ENSG00000224157	0.35	17.5	HCG14	HLA complex group 14 (non-protein coding)
chr7	100606877	100611410	-	ENSG00000225946	0.37	21	ID retired in ENSEMBL 76	
chr7	121943711	121950131	+	ENSG00000230316	0.46	27.5	FEZF1-AS1	FOREBRAIN EMBRYONIC ZINC FINGER; *FEZ*; ZINC FINGER PROTEIN 312B; *ZNF312B*
chr7	102670550	102671801	+	ENSG00000236226	0.50	15	RP11-645N11.3	
chr8	124213411	124214983	-	ENSG00000204949	0.38	16	FAM83A antisense RNA 1	*HCCC11*
chr9	136519707	136522435	-	ENSG00000225756	0.43	21	DBH-AS1	DOPAMINE BETA-MONOOXYGENASE; catalyzes the oxidative hydroxylation of dopamine to norepinephrine
chr10	118429613	118432250	+	ENSG00000232767	0.35	23	RP11-498B4.5	antisense lncRNA intersects with the protein-coding gene *C10orf82* that is part of the hormone module

chr11	22305275	22359791	-	ENSG00000254768	0.52	26	CTD-2140G10.2	antisense lncRNA intersects with the protein-coding gene *SLC17A6* that is part of the hormone module
chr11	8056992	8061265	-	ENSG00000254921	0.43	22.5	RP11-236J17.6	
chr11	22282754	22283802	-	ENSG00000255372	0.36	10	CTD-3064C13.1	lncRNA is located within 75840 nucleotides distance to the protein-coding gene *SLC17A6* that is pat of the hormone module
chr11	6642681	6643553	+	ENSG00000255390	0.34	19	RP11-732A19.5	
chr12	113909807	113918286	+	ENSG00000257935	0.63	24	LHX5-AS1	LIM HOMEOBOX PROTEIN 5; *LHX5*; Lhx5 mRNA expression in the developing central nervous system of mouse embryos and concluded that the expression pattern suggested a regulatory function in forebrain formation.
chr13	31456069	31507983	-	ENSG00000224743	0.38	24	TEX26-AS1	

chr14	36988482	36992221	+	ENSG00000253563	0.53	20.5	NKX2-1-AS1	NK2 HOMEOBOX 1; *NKX2-1*; THYROID TRANSCRIPTION FACTOR 1; *TITF1*; *TTF1*; The *NKX2-1* gene encodes a transcription factor that is expressed during early development of thyroid, lung, and forebrain regions, particularly the basal ganglia and hypothalamus
chr14	60981836	61021634	+	ENSG00000258670	0.64	27	RP11-1042B17.3	antisense lncRNA intersects with the protein-coding gene *C14orf39* that is part of the hormone module
chr15	91475815	91478360	-	ENSG00000224441	0.41	18	AC068831.3	
chr15	27728180	27787137	-	ENSG00000259168	0.58	23.5	RP11-100M12.3	
chr17	42980528	42981914	-	ENSG00000224911	0.40	10	AC015936.3	
chr18	21574176	21595303	-	ENSG00000264745	0.48	25	RP11-403A21.2	
chr19	42901279	43156171	+	ENSG00000213904	0.33	20	LIPE-AS1	LIPASE, HORMONE-

								SENSITIVE; LIPE; *HSL*
chr19	42656720	42661675	+	ENSG00000259436	0.59	24	CTC-378H22.2	
chr20	31148894	31161690	+	ENSG00000233293	0.41	22.5	RP11-410N8.3	
chr22	43608679	43609667	+	ENSG00000236272	0.32	9	Z82214.3	
chrX	56755691	56844813	+	ENSG00000204272	0.40	22.5	LINC01420	long intergenic non-protein coding RNA 1420
chrX	119055223	119056803	-	ENSG00000227285	0.33	16	RP3-327A19.5	
chrX	130115563	130192120	-	ENSG00000228659	0.36	23.5	LINC01201	lncRNA is located within 95 nucleotides distance to the protein-coding gene *ARHGAP36* that is pat of the hormone module

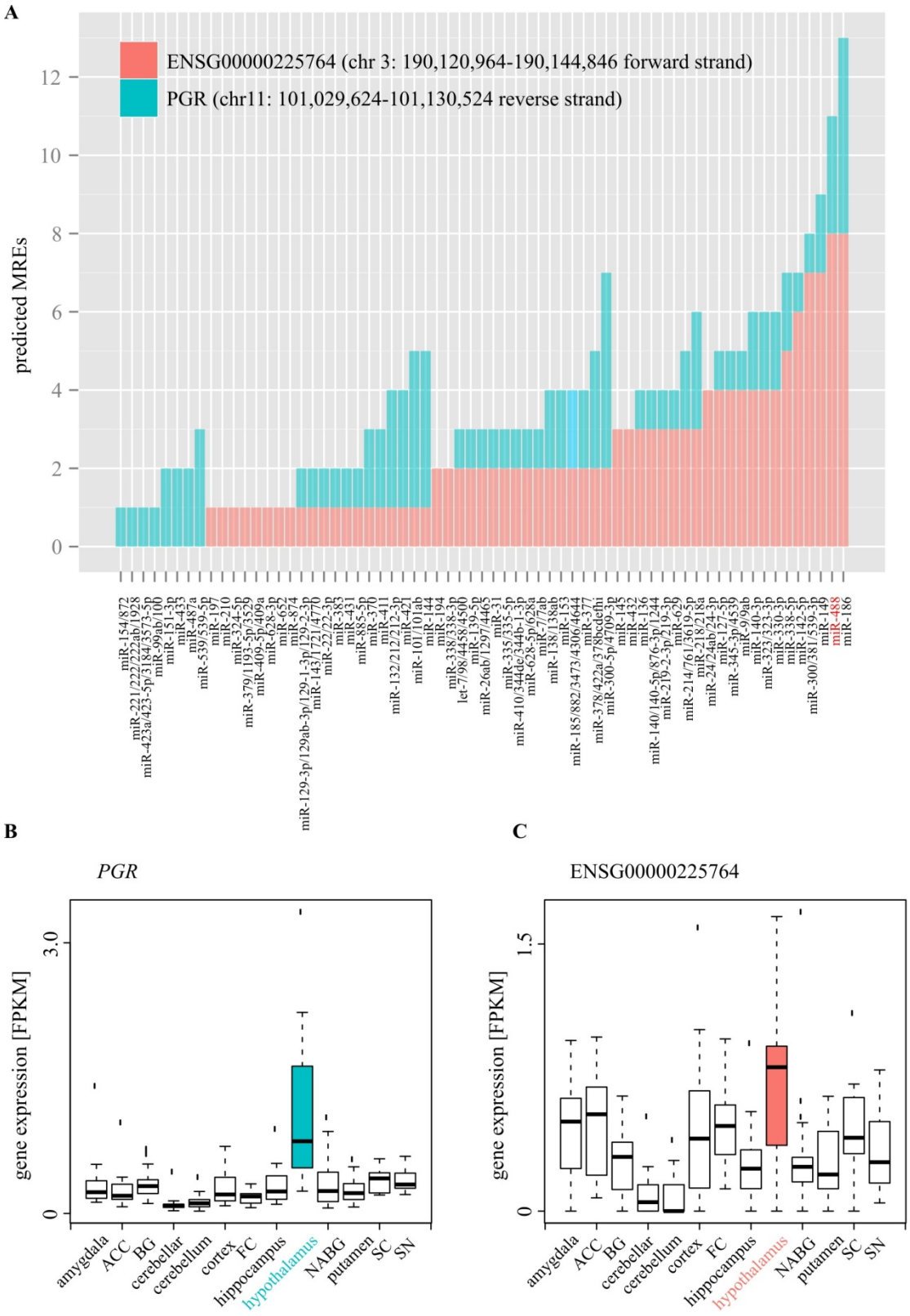

A
ENSG00000225764 (chr 3: 190,120,964-190,144,846 forward strand)
PGR (chr11: 101,029,624-101,130,524 reverse strand)
predicted MREs
0
2
4
6
8
10
12
miR-154/872
miR-221/222/222ab/1928
miR-423a/423-5p/3184/3573-5p
miR-99ab/100
miR-151-3p
miR-433
miR-487a
miR-539/539-5p
miR-197
miR-210
miR-324-5p
miR-379/1193-5p/3529
miR-409-5p/409a
miR-628-3p
miR-652
miR-874
miR-129-3p/129ab-3p/129-1-3p/129-2-3p
miR-143/1721/4770
miR-22/22-3p
miR-383
miR-431
miR-885-5p
miR-370
miR-411
miR-132/212/212-3p
miR-421
miR-101/101ab
miR-144
miR-194
miR-338/338-3p
let-7i/98/4458/4500
miR-139-5p
miR-26ab/1297/4465
miR-31
miR-335/335-5p
miR-410/344de/344b-1-3p
miR-628-5p/628a
miR-7/7ab
miR-138/138ab
miR-153
miR-185/882/3473/4306/4644
miR-377
miR-378/422a/378bcdefhi
miR-300-5p/4709-3p
miR-145
miR-432
miR-136
miR-140/140-5p/876-3p/1244
miR-219-2-3p/219-3p
miR-629
miR-214/761/3619-5p
miR-218/218a
miR-24/24ab/24-3p
miR-127-5p
miR-345-3p/4539
miR-9/9ab
miR-140-3p
miR-323/323-3p
miR-330-3p
miR-338-5p
miR-142-5p
miR-300/381/539-3p
miR-149
miR-488
miR-186
B
PGR
gene expression [FPKM]
0
3.0
amygdala
ACC
BG
cerebellar
cerebellum
cortex
FC
hippocampus
hypothalamus
NABG
putamen
SC
SN
C
ENSG00000225764
gene expression [FPKM]
0
1.5
amygdala
ACC
BG
cerebellar
cerebellum
cortex
FC
hippocampus
hypothalamus
NABG
putamen
SC
SN

Figure 35. A) MREs shared between the 3'UTRs of the mRNAs of *PGR* (protein-coding gene assigned to the hormone module) and the associated antisense lncRNA ENSG00000225764. The barplot shows the number of predicted MREs for each miRNA that targets one or both of these genes (the names of the miRNAs are shown on the x-axis). Both genes combined contained 238 predicted MREs, including 203 MREs (85%) for miRNAs that are predicted to target both genes. MiRNA-488 that showed increased density of MREs for the hormone module and the associated lncRNAs in three independent annotations is highlighted in red. B) Boxplot shows expression of *PGR* across the 13 brain regions. C) Boxplot shows expression of ENSG00000225764 across the 13 brain regions. The x-axes of both boxplots indicate brain regions with the following abbreviations: ACC (anterior cingulate cortex), BG (basal ganglia), cerebellarH (cerebellar hemisphere), FC (frontal cortex), NABG (nucleus accumbens, basal ganglia), and SN (substantia nigra).

6.3.5.2 LincRNAs

The correlation between the MM and the median percentage shared MREs was calculated for two further sets of lincRNAs, namely 54 intergenic lncRNAs (lincRNAs) with at least one predicted MRE referred to as "lincRNAs" [311], and 22 lincRNAs with at least one predicted MRE referred to as "Cufflinks" [312], in the same way as described above.

The "Cufflinks" set showed a non-significant positive correlation between the MM and an increased number of predicted shared MREs (Rho=0.28, *p-value*=0.2, median percentage shared MREs across 23 lincRNA=24%, median MM across 23 lincRNA Rho=0.38).

No significant correlation was observed between the MM and the median percentage of predicted MREs for shared MREs for the "lincRNA" set (Rho=-0.06, *p-value*=0.64, median percentage shared MREs across 54 lincRNA=25, median MM across 54 lincRNA Rho=0.39).

6.4 Discussion

The human brain's transcriptome includes many diverse noncoding RNAs [27, 325, 326] whose biological role, in the main, remains poorly understood. It has been suggested that networks of RNAs could have relevance for fine-tuning the coordination of gene expression [27]. For example, long noncoding RNAs (lncRNAs) and microRNAs (miRNAs) could form a network of "competitive endogenous RNAs" (ceRNAs) where genes regulate each other's transcript abundance by competing for the same miRNAs [27]. In this study, I first detected transcriptional co-expression networks – referred to as modules - that are specific to distinct brain regions, and next identified lncRNAs whose expression levels are associated with these modules. This association was quantified by (*i*) module membership of the lncRNA, defined as a significant correlation of a lncRNA's expression with the representative gene expression pattern of the module (the so called "module eigengene"), (*ii*) increased expression of the lncRNA in the same brain region as the module, and (*iii*) evidence for a higher number of miRNA response elements (MREs) that are shared between the lncRNA and the genes in the region-specific module than expected by chance.

This analysis detected one hypothalamus-specific gene module that was strongly enriched for functional GO-terms linked to hormone activity. This "hormone module" was associated with both a set of antisense lncRNAs and, independently, showed a trend for an association with two sets of intergenic lncRNAs. The association was reflected by the significant correlation of expression between the lncRNAs and the protein-coding genes in the hormone module, in addition to the elevated expression of the lncRNA in the hypothalamus. On the miRNA level, the genes in the hormone module were found to share a higher number of MREs with the associated lncRNAs compared

to other lncRNAs. This increase in shared MREs between the hormone module genes and the associated lncRNAs was further reflected by a significantly positive correlation between a lncRNA's module membership and the median percentage of MREs it shared with the set of hormone module genes. This analysis thus shows how the combination of co-expression and evidence for competition for the same miRNAs can be used to identify candidate lncRNAs for competitive endogenous RNA networks. Future work could aim to experimentally validate the role the suggested lncRNAs play within the hormone module, for example by testing the hypothesis that the depletion of an associated lncRNA would lead to decreased expression of the hormone module, modulated by the increased availability of miRNAs targeting many transcripts from this gene set.

The identification of candidate lncRNAs acting as competitive regulatory elements for the hormone module suffers from different limitations. Generally, the computational prediction of MREs shows a high rate of false positives due to challenges such as the short length of binding elements' sequences which increases the number of false positives [327]. However, although the TargetScan tool [115] shows an estimated false positive rate of 22-31% [327], previous studies found a relatively high degree of complementarity between computationally predicted and experimentally determined binding sites [327]. Furthermore, the present analysis partially accounted for the 'noise' in the MRE predictions using relative measures such as the ratio of MREs for shared miRNAs divided by the total predicted MREs between a gene-pair, thereby providing an internal standard. Clearly, this approach cannot fully account for extreme length biases which allow either very few predicted MREs or in turn predicted MREs for all included miRNA families. However, as described in **Chapter 6.3.5**, especially very high numbers of shared percentage MREs were only observed for single gene-pairs.

Another general limitation is the comparably low expression of lncRNAs [311]. Low expression of RNA elements can lead to less reliable detection and quantification, and thus introduce further 'noise' [328]. Here, I addressed this issue by selecting lncRNAs based on the correlation of the expression across all available samples and, additionally, based on increased expression in the hypothalamus. The application of two 'filtering steps' was expected to reduce the number of false positives.

Evidence for shared MREs between the hormone module and the associated lncRNAs was highly specific only to the comparison of the hormone module genes to different sets of lncRNAs. Comparison of the percentage shared MREs between the hormone module genes to other protein-coding genes revealed that the hormone module did not share more MREs than other sets of protein-coding genes. Potentially, this observation could be influenced by a larger number of miRNAs that are specific to different brain regions. Clearly, the enrichment of MREs for region-specific miRNAs in the 3'UTRs of region-specifically expressed mRNAs determines how many of these genes could share MREs. Consequently, a higher number of miRNAs whose expression is specific to regions outside of the hypothalamus could lead to the observed average number of percentage shared MREs among the genes in the hormone module. In line with this hypothesis, only one miRNA showed nominally significant enrichment in the hormone module mRNAs and the associated lncRNAs across all independent annotations, namely miR-488.

In summary, the present study used a combination of quantifying co-expression between different RNA types and the number of shared MREs between these RNA elements to suggest candidate lncRNAs for the regulation of genes linked to hormonal activity. Future work could aim to quantify the effect of the suggested key regulators by experimental network perturbation. In this way, RNA regulatory elements with

relevance for medical conditions linked to hypothalamic function, such as obesity [323, 329], could be identified.

6.6 Methods

6.6.1 RNA-Sequencing data

RNA-Sequencing data were obtained from the Genotype-Tissue Expression project (GTEx [113]). I used gene expression data of 20195 protein-coding genes with annotated ENSEMBL identifiers, that were expressed in at least one of 225 brain tissue samples, covering 13 brain regions (Figure 28) in 35 human adult subjects (20-70 years). Gene expression was quantified as FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and quantile normalization was applied across all libraries that were used. Quantile normalization transforms the data to equalize the data distributions across different samples, to allow meaningful comparisons among the samples [330] (see **Chapter 3.5.2**). The quantification of gene expression was provided by Wilfried Haerty. The GTEx data provide information about different external traits, namely age, sex, the brain region from which the sample was taken, RIN values (see **Chapter 1.2**, all included samples had RIN values of 7.0<RIN<9.8), RNA read depth of libraries, subject identifier (described below), and Hardy Index. The Hardy Index (variable accession phv00169092.v2.p1 [113]) classifies the manner of death on a 4-point scale. It ranges from 1) violent and fast death (for example by accident with terminal phase < 10 min), over 2) fast death of natural causes (terminal phase < 1 hour), and 3) intermediate death which comes unexpectedly after illness (terminal phase between 1 and 24 hours), to 4) slow death after long illness (terminal phase longer than 1 day). The subject identifier was included to test for the impact of single subjects on the observed results, due to numbers of samples from different brain regions not being equally distributed across all subjects.

6.6.2 Annotation of 3' untranslated regions

RNA-Seq reads from 13 human tissues derived from the central nervous system were mapped to the human genome (built hg19) using TopHat ([43] version 2.0.12). Transcripts were assembled using Cufflinks ([43] version 2.2.1) with ENSEMBL (build 73) as a reference, allowing the assembly of novel transcripts (see **Chapter 2.3.3**). The transcript assembly was undertaken by Wilfried Haerty.

The 3' untranslated regions (3'UTRs) of all genes were annotated as all nucleotides downstream of stop codons constitutively expressed across all transcripts using gene annotations from ENSEMBL (build 73; n=8230 genes with ENSEMBL gene identifiers contained constitutive stop codons, out of n=22208 protein coding genes which contained at least one stop codon in the ENSEMBL build; see Figure 29).

I considered only expressed genes in the GTEx RNA-Seq data, defined as FPKM≥0.1 in at least one sample, which contained a constitutive stop codon (n=7273 genes with constitutive stop codon and downstream 3'UTRs).

For all 13 tissues, I annotated constitutively and alternatively expressed 3'UTRs. Constitutive 3'UTRs were defined as all nucleotides which were expressed in all transcripts of a gene across all samples. To annotate alternative 3'UTRs for each tissue (*i*) the 3'UTRs of all transcripts expressed in a given tissue were concatenated per gene, and (*ii*) the portion of the 3'UTRs which was constitutively expressed across all considered tissues was discarded. In the following analysis, I used 3'UTRs from 223 samples with 7.0<RIN<9.8, after excluding 2 samples which expressed a very low number of 3'UTRs compared to the other samples.

6.6.3 Expression of miRNAs

MiRNA families were considered as brain-expressed if they contained at least one miRNA whose expression was supported by at least one copy in cerebellum, frontal cortex, midbrain or hippocampus [331], resulting in 121 brain-expressed miRNA families. 79 of these miRNA families showed one or more predicted MRE in the considered gene sets (see **Chapter 6.6.7** and Supplementary Table S6.4).

6.6.4 Annotation of long noncoding RNAs

Loci of long noncoding RNAs (lncRNAs) and long intergenic noncoding RNAs (lincRNAs) were annotated based on five independent publicly available data sets from human tissues (Figure 36, described in detail below). Expression was quantified per locus, including all transcripts overlapping the given locus, using FPKMs (quantile normalized across all libraries). Annotations of lncRNAs and quantification of their expression were provided by Wilfried Haerty. In total, 19119 multi-exonic lncRNAs were expressed in at least one sample with RIN>7. If multiple transcripts from different annotations overlapped the same locus, they were merged, so that no lncRNA was tested twice in different annotation sets.

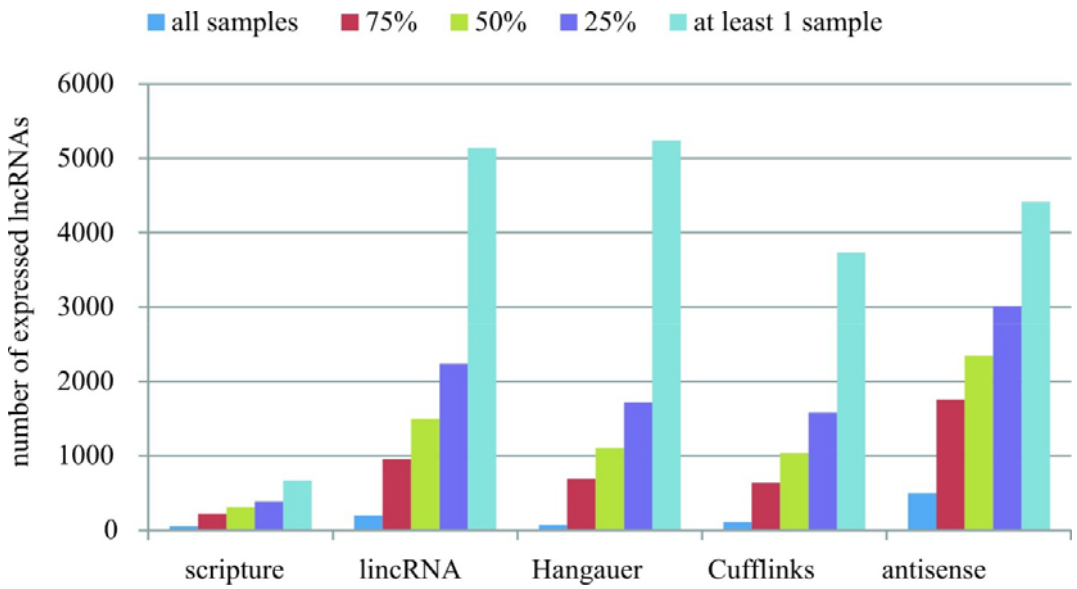


Figure 36. **Number of multi-exonic lncRNAs expressed in all samples, or 75%, 50%, or 25% of them, or in at least one sample, separately for five lncRNA annotation sets.**

1) LncRNAs annotated by Derrien et al. [311]

LncRNA annotations from Derrien et al. contributed to the sets indicated as "lincRNA" and "antisense". The authors presented comprehensive lncRNA annotations based on data produced by the GENCODE consortium [332]. Here, I considered only lncRNAs not overlapping genes on the sense strand, namely the group "lincRNA", defined as transcripts at least >200 base pairs long located at intergenic noncoding RNA loci, and "antisense" [311, 320], which intersect protein-coding genes on the antisense strand or published evidence suggests antisense regulation of a coding gene. Boundaries of protein-coding genes were defined as 5kb upstream of the start codon, and 30kb downstream of the stop codon [311].

2) LincRNAs annotated by Cabili et al. [312]

The identification of lincRNAs by Cabili et al. combined RNA-Seq data with available annotations, as described in [312]. Briefly, compilation of the lincRNA catalog was completed in four steps: (*i*) transcript assembly using Cufflinks [114] and Scripture [333], (*ii*) compilation of previously annotated transcripts, (*iii*) integration of RNA-Seq with annotation resources to identify isoforms for each transcription locus, and (*iv*) identification of lincRNAs based on the following criteria: those that were reliably expressed, large, multi-exonic, noncoding, and intergenic [312].

3) LincRNAs annotated by Hangauer et al. [6]

Hangauer et al. used multiple publicly available data sets containing human tissue samples to annotate lincRNAs as described in [6]. The authors considered transcripts with length >200 nucleotides, excluding transcripts which were (*i*) overlapping extended protein coding gene structures, (*ii*) located within 1kb of another gene on the same strand, or overlapped another gene on the opposite strand, (*iii*) not expressed at FPKM>1 in at least one sample [6].

6.6.5 Construction of transcriptional networks using WGCNA

Transcriptional networks were built using step-by-step network construction and module detection as implemented in WGCNA (Weighted correlation network analysis; see **Chapter 2.5.1**) [122]. Briefly, in the first step a pairwise Pearson's correlation matrix of all genes across all samples is calculated. Here, an unsigned network which does not preserve the sign of the correlation was generated. This correlation matrix is then converted into an adjacency matrix. The soft thresholding power was chosen to reach a scale-free topology fit index indicating approximate scale-free topology, as suggested by the authors [122]. Next, the adjacency matrix is transformed to

Topological Overlap Matrix (TOM), reducing spurious associations, and dissimilarity is calculated. The dissimilarity matrix is then used to produce a hierarchical tree. The minimum cluster size was arbitrarily set to 30 genes. The standard method for branch cutting, namely Dynamic Tree Cut, was applied [122]. This function returns the gene modules, whereby each module gets assigned a colour. To reduce redundant information, the modules with highly similar expression profiles were merged. The merging procedure was applied using a height cut of 0.25 (corresponding to a correlation of 0.75). Merging is based on calculation of the module eigengene (ME) which provides a summary expression profile. These MEs can be correlated to external quantitative traits to analyse the dependency between the given trait and the module's representative expression. Here, I used age, Hardy Index (see **Chapter 6.6.1**), sex, brain region, subject identifier, RNA read depth of the libraries, and RIN values.

6.6.6 Correlation of lncRNAs with the ME

The correlation of a lncRNA's expression with the module's ME was calculated using Pearson's correlation. This measure is referred to as "module membership" (MM) [122]. Pearson's correlation was used when the two variables that were correlated included either the variable indicating brain region or a gene's expression which was specific to a brain region (or generally categorical data). To remove lncRNAs whose expression correlated with the ME because of single outliers, I applied an additional filtering step for selecting lncRNAs. More specifically, I included only lncRNAs that showed either the highest median expression or the highest 3rd quantile of expression in the same brain region as the protein-coding genes assigned to the respective module in the further analysis. LncRNAs that passed all filters are referred to as lncRNAs that are associated with the given module.

6.6.7 Prediction of miRNA response elements

The prediction of miRNA response elements (MREs) was undertaken using TargetScan (version 6.0 [115]). I predicted MREs in all annotated lincRNAs, and the 3'UTRs of the mRNAs in the transcriptional network. TargetScan predicts miRNA targets based on miRNA names and their 7 nucleotide long seed region sequence [115] (see **Chapter 2.3.4**) within sequences of interest. I included only miRNA families that have been identified as human brain-expressed [331] and that contained seed regions for human miRNAs as defined by the TargetScan tool [115]. The sequences of the genomic regions of interest (lncRNAs and 3' untranslated regions) were derived using the "fastaFromBed" tool from the Bedtools Suite (v2.11.22) using UCSC genome build hg19 as a reference.

6.6.8 Density of miRNA response elements

I tested for enrichment of MREs for specific miRNAs by calculating the MREs density for all 80 miRNAs. MREs density of a miRNA was defined as the median ratio of the predicted MREs for the miRNA in a gene divided by the total number of predicted MREs in the same gene across all considered genes. More specifically, I compared the median MREs density of each of the 80 miRNAs within a gene set of interest to the same measure across n=1000 sample-size matched random samples.

6.6.9 Calculation of percentage shared MREs

The 'percentage shared MREs' was calculated per gene pair (Figure 29). More specifically, I first selected the miRNAs that were predicted to target transcripts from both genes in a given gene-pair. The sum of the number of MREs for all of the shared predicted MREs was divided by the total number of predicted MREs for all miRNAs

which were predicted to target at least one of the genes in the gene-pair (bottom panel in Figure 29, for an example see Figure 35A). This ratio was converted to a percentage. For each gene set of interest (for example all hormone module genes and a set of associated lncRNAs), I compared the median of the percentage shared MREs for this gene set to the same measure from n=1000 sample size-matched random samples from a suitable null distribution for the gene set, restricted to the lncRNAs that were expressed (FPKM≥0.1) in at least one sample. The empirical p-value was calculated as the number of random samples showing a higher percentage shared MREs than the gene set of interest, divided by the total number of observations.

6.7 Appendix

Supplementary Table S6.1. Genes in the "Regionalization module".

ENSG00000054938	*CHRDL2*	chordin-like 2
ENSG00000165606	*DRGX*	dorsal root ganglia homeobox
ENSG00000221818	*EBF2*	early B-cell factor 2
ENSG00000108001	*EBF3*	early B-cell factor 3
ENSG00000198759	*Egfl6*	EGF-like-domain, multiple 6
ENSG00000070193	*fgf10*	fibroblast growth factor 10
ENSG00000164900	*GBX1*	gastrulation brain homeobox 1
ENSG00000145888	*GLRA1*	glycine receptor, alpha 1
ENSG00000105996	*hoxa2*	homeobox A2
ENSG00000105997	*HOXA3*	homeobox A3
ENSG00000197576	*HOXA4*	homeobox A4
ENSG00000106004	*HOXA5*	homeobox A5
ENSG00000122592	*HOXA7*	homeobox A7
ENSG00000173917	*HOXB2*	homeobox B2
ENSG00000120093	*HOXB3*	homeobox B3
ENSG00000182742	*HOXB4*	homeobox B4
ENSG00000120075	*hoxb5*	homeobox B5
ENSG00000108511	*HOXB6*	homeobox B6
ENSG00000120087	*HOXB7*	homeobox B7
ENSG00000120068	*HOXB8*	homeobox B8
ENSG00000170689	*HOXB9*	homeobox B9
ENSG00000198353	*HOXC4*	homeobox C4
ENSG00000172789	*HOXC5*	homeobox C5
ENSG00000128652	*hoxd3*	homeobox D3
ENSG00000170166	*HOXD4*	homeobox D4
ENSG00000175879	*HOXD8*	homeobox D8

ENSG00000138136	*Lbx1*	ladybird homeobox 1
ENSG00000136944	*lmx1b*	LIM homeobox transcription factor 1, beta
ENSG00000181541	*Mab21l2*	mab-21-like 2 (C. elegans)
ENSG00000184350	*MRGPRE*	MAS-related GPR, member E
ENSG00000160505	*NLRP4*	NLR family, pyrin domain containing 4
ENSG00000163273	*Nppc*	natriuretic peptide precursor C
ENSG00000172774	*OR1S1*	olfactory receptor, family 1, subfamily S, member 1
ENSG00000075891	*PAX2*	paired box 2
ENSG00000165462	*PHOX2A*	paired-like homeobox 2a
ENSG00000109132	*Phox2b*	paired-like homeobox 2b
ENSG00000159182	*PRAC*	prostate cancer susceptibility candidate
ENSG00000071677	*PRLH*	prolactin releasing hormone
ENSG00000143125	*PROK1*	prokineticin 1
ENSG00000141293	*skap1*	src kinase associated phosphoprotein 1
ENSG00000103546	*SLC6A2*	solute carrier family 6 (neurotransmitter transporter, noradrenalin), member 2
ENSG00000165970	*SLC6A5*	solute carrier family 6 (neurotransmitter transporter, glycine), member 5

Supplementary Table S6.2. Genes in the "Hormone module".

ENSG00000141433	*ADCYAP1*	adenylate cyclase activating polypeptide 1 (pituitary)
ENSG00000144891	*Agtr1*	angiotensin II receptor, type 1
ENSG00000160224	*AIRE*	autoimmune regulator
ENSG00000139211	*AMIGO2*	adhesion molecule with Ig-like domain 2
ENSG00000164512	*ANKRD55*	ankyrin repeat domain 55
ENSG00000147256	*ARHGAP36*	hypothetical protein FLJ30058
ENSG00000101200	*AVP*	arginine vasopressin
ENSG00000102239	*BRS3*	bombesin-like receptor 3
ENSG00000165863	*C10orf82*	chromosome 10 open reading frame 82
ENSG00000179008	*C14orf39*	chromosome 14 open reading frame 39
ENSG00000205856	*C22orf42*	chromosome 22 open reading frame 42
ENSG00000187944	*C2orf66*	chromosome 2 open reading frame 66
ENSG00000172568	*C5orf40*	chromosome 5 open reading frame 40
ENSG00000004948	*CALCR*	calcitonin receptor
ENSG00000164326	*CARTPT*	CART prepropeptide
ENSG00000164045	*CDC25A*	cell division cycle 25 homolog A (S. pombe)
ENSG00000140488	*Celf6*	bruno-like 6, RNA binding protein (Drosophila)
ENSG00000136698	*cfc1*	cripto, FRL-1, cryptic family 1
ENSG00000152093	*CFC1B*	cripto, FRL-1, cryptic family 1B
ENSG00000154645	*CHODL*	chondrolectin
ENSG00000125931	*CITED1*	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 1
ENSG00000144191	*CNGA3*	cyclic nucleotide gated channel alpha 3
ENSG00000128510	*CPA4*	carboxypeptidase A4
ENSG00000147571	*crh*	corticotropin releasing hormone
ENSG00000147231	*CXorf57*	chromosome X open reading frame 57
ENSG00000019186	*Cyp24a1*	cytochrome P450, family 24, subfamily A, polypeptide 1
ENSG00000163331	*DAPL1*	death associated protein-like 1

		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000205916	*DAZ1*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000188120	*DAZ1*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000187191	*DAZ1*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000205944	*DAZ1*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000205916	*DAZ2*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000188120	*DAZ2*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000187191	*DAZ2*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000205944	*DAZ2*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000205916	*DAZ3*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000188120	*DAZ3*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000187191	*DAZ3*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000205944	*DAZ3*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000205916	*DAZ4*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000188120	*DAZ4*	deleted in azoospermia 2; deleted in azoospermia 4
		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000187191	*DAZ4*	deleted in azoospermia 2; deleted in azoospermia 4

		deleted in azoospermia 1; deleted in azoospermia 3;
ENSG00000205944	*DAZ4*	deleted in azoospermia 2; deleted in azoospermia 4
ENSG00000198889	*DCAF12L1*	WD repeat domain 40B
ENSG00000198354	*DCAF12L2*	WD repeat domain 40C
ENSG00000100867	*Dhrs2*	dehydrogenase/reductase (SDR family) member 2
ENSG00000162595	*DIRAS3*	DIRAS family, GTP-binding RAS-like 3
ENSG00000185559	*DLK1*	delta-like 1 homolog (Drosophila)
ENSG00000144355	*dlx1*	distal-less homeobox 1
ENSG00000171551	*ECEL1*	endothelin converting enzyme-like 1
ENSG00000164220	*f2rl2*	coagulation factor II (thrombin) receptor-like 2
ENSG00000182183	*FAM159A*	family with sequence similarity 159, member A
ENSG00000164451	*Fam26d*	family with sequence similarity 26, member D
ENSG00000125355	*fam70a*	family with sequence similarity 70, member A
ENSG00000102531	*FNDC3A*	fibronectin type III domain containing 3A
ENSG00000204612	*FOXB2*	forkhead box B2
ENSG00000152254	*G6PC2*	glucose-6-phosphatase, catalytic, 2
ENSG00000147402	*GABRQ*	gamma-aminobutyric acid (GABA) receptor, theta
ENSG00000101958	*Glra2*	glycine receptor, alpha 2
ENSG00000167414	*GNG8*	guanine nucleotide binding protein (G protein), gamma 8
ENSG00000188816	*HMX2*	H6 family homeobox 2
ENSG00000188620	*HMX3*	H6 family homeobox 3
ENSG00000147255	*IGSF1*	immunoglobulin superfamily, member 1
ENSG00000133124	*IRS4*	insulin receptor substrate 4
ENSG00000016082	*Isl1*	ISL LIM homeobox 1
ENSG00000109265	*KIAA1211*	KIAA1211
ENSG00000116014	*Kiss1r*	KISS1 receptor
ENSG00000150361	*Klhl1*	kelch-like 1 (Drosophila)
ENSG00000166603	*MC4R*	melanocortin 4 receptor
ENSG00000169446	*MMGT1*	membrane magnesium transporter 1
ENSG00000182636	*NDN*	necdin homolog (mouse)

ENSG00000165553	*ngb*	neuroglobin
ENSG00000187258	*NPSR1*	neuropeptide S receptor 1
ENSG00000185149	*NPY2R*	neuropeptide Y receptor Y2
ENSG00000144852	*nr1i2*	nuclear receptor subfamily 1, group I, member 2
ENSG00000185554	*NXF2*	nuclear RNA export factor 2; nuclear RNA export factor 2B
ENSG00000185945	*NXF2*	nuclear RNA export factor 2; nuclear RNA export factor 2B
ENSG00000185554	*NXF2B*	nuclear RNA export factor 2; nuclear RNA export factor 2B
ENSG00000185945	*NXF2B*	nuclear RNA export factor 2; nuclear RNA export factor 2B
ENSG00000109205	*odam*	odontogenic, ameloblast asssociated
ENSG00000205922	*Onecut3*	one cut homeobox 3
ENSG00000175398	*OR10P1*	olfactory receptor, family 10, subfamily P, member 1
ENSG00000171540	*otp*	orthopedia homeobox
ENSG00000101405	*OXT*	oxytocin, prepropeptide
ENSG00000145730	*pam*	peptidylglycine alpha-amidating monooxygenase
ENSG00000124253	*PCK1*	phosphoenolpyruvate carboxykinase 1 (soluble)
ENSG00000175426	*PCSK1*	proprotein convertase subtilisin/kexin type 1
ENSG00000198300	*PEG3*	paternally expressed 3; PEG3 antisense RNA (non-protein coding); zinc finger, imprinted 2
ENSG00000198300	*PEG3AS*	paternally expressed 3; PEG3 antisense RNA (non-protein coding); zinc finger, imprinted 2
ENSG00000082175	*PGR*	progesterone receptor
ENSG00000112238	*prdm13*	PR domain containing 13
ENSG00000119973	*PRLHR*	prolactin releasing hormone receptor
ENSG00000050628	*PTGER3*	prostaglandin E receptor 3 (subtype EP3)
ENSG00000122420	*PTGFR*	prostaglandin F receptor (FP)
ENSG00000198570	*Rd3*	retinal degeneration 3
ENSG00000157828	*RPS4Y2*	ribosomal protein S4, Y-linked 2
ENSG00000171951	*scg2*	secretogranin II (chromogranin C)
ENSG00000136546	*SCN7A*	sodium channel, voltage-gated, type VII, alpha
ENSG00000169432	*SCN9A*	sodium channel, voltage-gated, type IX, alpha subunit

ENSG00000184860	*sdr42e1*	short chain dehydrogenase/reductase family 42E, member 1
ENSG00000184302	*SIX6*	SIX homeobox 6
ENSG00000091664	*SLC17A6*	solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 6
ENSG00000165349	*SLC7A3*	solute carrier family 7 (cationic amino acid transporter, y+ system), member 3
ENSG00000157005	*sst*	somatostatin
ENSG00000126549	*STATH*	statherin
ENSG00000169836	*TACR3*	tachykinin receptor 3
ENSG00000159648	*TEPP*	testis, prostate and placenta expressed
ENSG00000180176	*TH*	tyrosine hydroxylase
ENSG00000144229	*THSD7B*	thrombospondin, type I, domain containing 7B
ENSG00000104953	*TLE6*	transducin-like enhancer of split 6 (E(sp1) homolog, Drosophila)
ENSG00000186889	*TMEM17*	transmembrane protein 17
ENSG00000170893	*trh*	thyrotropin-releasing hormone
ENSG00000069018	*TRPC7*	transient receptor potential cation channel, subfamily C, member 7
ENSG00000085831	*TTC39A*	tetratricopeptide repeat domain 39A
ENSG00000171724	*VAT1L*	vesicle amine transport protein 1 homolog (T. californica)-like
ENSG00000128564	*VGF*	VGF nerve growth factor inducible
ENSG00000110002	*vwa5a*	von Willebrand factor A domain containing 5A
ENSG00000162552	*WNT4*	wingless-type MMTV integration site family, member 4
ENSG00000174460	*ZCCHC12*	zinc finger, CCHC domain containing 12
ENSG00000198300	*ZIM2*	paternally expressed 3; PEG3 antisense RNA (non-protein coding); zinc finger, imprinted 2

Supplementary Table S6.3. Genes in the "Feeding module".

ENSG00000188909	*BSX*	brain-specific homeobox
ENSG00000185480	*C12orf48*	chromosome 12 open reading frame 48
ENSG00000172137	*CALB2*	calbindin 2
ENSG00000110448	*CD5*	CD5 molecule
ENSG00000105205	*clc*	Charcot-Leyden crystal protein
ENSG00000145708	*CRHBP*	corticotropin releasing hormone binding protein
ENSG00000069482	*GAL*	galanin prepropeptide
ENSG00000118702	*GHRH*	growth hormone releasing hormone
ENSG00000147257	*GPC3*	glypican 3
ENSG00000137252	*HCRTR2*	hypocretin (orexin) receptor 2
ENSG00000140287	*HDC*	histidine decarboxylase
ENSG00000123496	*IL13RA2*	interleukin 13 receptor, alpha 2
ENSG00000101230	*ism1*	isthmin 1 homolog (zebrafish)
ENSG00000162761	*LMX1A*	LIM homeobox transcription factor 1, alpha
ENSG00000187123	*lypd6*	LY6/PLAUR domain containing 6
ENSG00000180815	*MAP3K15*	mitogen-activated protein kinase kinase kinase 15
ENSG00000140563	*MCTP2*	multiple C2 domains, transmembrane 2
ENSG00000167346	*MMP26*	matrix metallopeptidase 26
ENSG00000105954	*NPVF*	neuropeptide VF precursor
ENSG00000205330	*OR6C1*	olfactory receptor, family 6, subfamily C, member 1
ENSG00000164093	*pitx2*	paired-like homeodomain 2
ENSG00000183395	*PMCH*	pro-melanin-concentrating hormone
ENSG00000115828	*QPCT*	glutaminyl-peptide cyclotransferase
ENSG00000186867	*qrfpr*	pyroglutamylated RFamide peptide receptor
ENSG00000101074	*R3HDML*	R3H domain containing-like
ENSG00000165646	*SLC18A2*	solute carrier family 18 (vesicular monoamine), member 2
ENSG00000134207	*SYT6*	synaptotagmin VI
ENSG00000160472	*Tmem190*	transmembrane protein 190

ENSG00000108828	*vat1*	vesicle amine transport protein 1 homolog (T. californica)

Supplementary Table S6.4. 80 MiRNA families with predicted MREs in 3'UTRs and lncRNAs.

let-7/98/4458/4500
miR-101/101ab
miR-127-5p
miR-127/127-3p
miR-129-3p/129ab-3p/129-1-3p/129-2-3p
miR-132/212/212-3p
miR-134/3118
miR-136
miR-137/137ab
miR-138/138ab
miR-139-3p
miR-139-5p
miR-140-3p
miR-140/140-5p/876-3p/1244
miR-142-3p
miR-142-5p
miR-143/1721/4770
miR-144
miR-145
miR-149
miR-151-3p
miR-151-5p/151b
miR-153
miR-154/872
miR-184
miR-185/882/3473/4306/4644
miR-186

miR-191

miR-194

miR-197

miR-21/590-5p

miR-210

miR-214/761/3619-5p

miR-218/218a

miR-219-2-3p/219-3p

miR-22/22-3p

miR-221/222/222ab/1928

miR-24/24ab/24-3p

miR-26ab/1297/4465

miR-300-5p/4709-3p

miR-300/381/539-3p

miR-31

miR-323-5p/1421qns

miR-323/323-3p

miR-324-5p

miR-330-3p

miR-335/335-5p

miR-338-5p

miR-338/338-3p

miR-339-3p

miR-345-3p/4539

miR-345/345-5p

miR-370

miR-377

miR-378/422a/378bcdefhi

miR-379/1193-5p/3529

miR-383

miR-409-5p/409a

miR-410/344de/344b-1-3p

miR-411

miR-421

miR-423a/423-5p/3184/3573-5p

miR-431

miR-432

miR-433

miR-451

miR-487a

miR-488

miR-539/539-5p

miR-628-3p

miR-628-5p/628a

miR-629

miR-652

miR-7/7ab

miR-744/1716

miR-874

miR-885-5p

miR-9/9ab

miR-95

miR-99ab/100

Supplementary Table S6.5. Significant GO-term enrichment in the 58 genes from the "hormone module" (Table 2) that had at least one MRE predicted. Yellow highlights the non-redundant terms, summarized using REViGO [117]. P-values are Benjamini-Hochberg corrected (α<0.05), 'fold' refers to the fold enrichment of each GO-term.

Hormone module (58 genes with at least one MRE predicted)			
category	**term**	**fold**	**p-value**
GO:0005184	neuropeptide hormone activity	77.7	1.40E-06
GO:0007631	feeding behavior	28.2	9.90E-05
GO:0005179	hormone activity	20.5	6.80E-05
GO:0001653	peptide receptor activity	18.6	7.80E-05
GO:0008528	peptide receptor activity, G-protein coupled	18.6	7.80E-05
GO:0042277	peptide binding	12.4	9.70E-05
GO:0007186	G-protein coupled receptor protein signaling pathway	4.2	1.60E-03
GO:0006972	hyperosmotic response	85.4	2.70E-03
GO:0008217	regulation of blood pressure	16.8	4.30E-03
GO:0007610	behavior	6	4.10E-03
GO:0007267	cell-cell signaling	5.2	3.50E-03
GO:0009725	response to hormone stimulus	7	3.20E-03
GO:0043025	cell soma	11.8	1.20E-02
GO:0030141	secretory granule	11.2	7.60E-03
GO:0006970	response to osmotic stress	35.8	1.20E-02
GO:0042538	hyperosmotic salinity response	138.8	1.20E-02
GO:0030594	neurotransmitter receptor activity	16.2	6.80E-03
GO:0008188	neuropeptide receptor activity	30.6	6.90E-03
GO:0042923	neuropeptide binding	29.8	6.40E-03
GO:0042165	neurotransmitter binding	14.9	5.80E-03
GO:0003078	regulation of natriuresis	104.1	2.00E-02
GO:0044062	regulation of excretion	92.6	2.40E-02
GO:0008015	blood circulation	9.1	2.20E-02
GO:0048545	response to steroid hormone stimulus	9	2.30E-02
GO:0007166	cell surface receptor linked signal transduction	2.6	2.30E-02
GO:0007565	female pregnancy	13	2.20E-02
GO:0044057	regulation of system process	6.4	2.30E-02
GO:0010033	response to organic substance	3.9	2.20E-02
GO:0009651	response to salt stress	64.1	2.90E-02
GO:0051954	positive regulation of amine transport	59.5	3.20E-02
GO:0007620	copulation	55.5	3.50E-02
GO:0051240	positive regulation of multicellular organismal process	7	4.00E-02

Supplementary Table S6.6. 51 genes in hormone module annotated in the DAVID data base and including at least one predicted MRE as described in the main text.

ENSG00000141433	*ADCYAP1*	adenylate cyclase activating polypeptide 1 (pituitary)
ENSG00000144891	*AGTR1*	angiotensin II receptor, type 1
ENSG00000139211	*AMIGO2*	adhesion molecule with Ig-like domain 2
ENSG00000101200	*AVP*	arginine vasopressin
ENSG00000102239	*BRS3*	bombesin-like receptor 3
ENSG00000187944	*C2orf66*	chromosome 2 open reading frame 66
ENSG00000004948	*CALCR*	calcitonin receptor
ENSG00000164326	*CARTPT*	CART prepropeptide
ENSG00000136698	*CFC1*	cripto, FRL-1, cryptic family 1
ENSG00000152093	*CFC1B*	cripto, FRL-1, cryptic family 1B
ENSG00000144191	*CNGA3*	cyclic nucleotide gated channel alpha 3
ENSG00000147571	*CRH*	corticotropin releasing hormone
ENSG00000147231	*CXorf57*	chromosome X open reading frame 57
ENSG00000198354	*DCAF12L2*	WD repeat domain 40C
ENSG00000162595	*DIRAS3*	DIRAS family, GTP-binding RAS-like 3
ENSG00000171551	*ECEL1*	endothelin converting enzyme-like 1
ENSG00000164220	*F2RL2*	coagulation factor II (thrombin) receptor-like 2
ENSG00000182183	*FAM159A*	family with sequence similarity 159, member A
ENSG00000102531	*FNDC3A*	fibronectin type III domain containing 3A
ENSG00000204612	*FOXB2*	forkhead box B2
ENSG00000147402	*GABRQ*	gamma-aminobutyric acid (GABA) receptor, theta
ENSG00000167414	*GNG8*	guanine nucleotide binding protein (G protein), gamma 8

ENSG00000188816	*HMX2*	H6 family homeobox 2
ENSG00000188620	*HMX3*	H6 family homeobox 3
ENSG00000133124	*IRS4*	insulin receptor substrate 4
ENSG00000016082	*ISL1*	ISL LIM homeobox 1
ENSG00000150361	*KLHL1*	kelch-like 1 (Drosophila)
ENSG00000166603	*MC4R*	melanocortin 4 receptor
ENSG00000182636	*NDN*	necdin homolog (mouse)
ENSG00000165553	*NGB*	neuroglobin
ENSG00000185149	*NPY2R*	neuropeptide Y receptor Y2
ENSG00000205922	*ONECUT3*	one cut homeobox 3
ENSG00000171540	*OTP*	orthopedia homeobox
ENSG00000101405	*OXT*	oxytocin, prepropeptide
ENSG00000082175	*PGR*	progesterone receptor
ENSG00000112238	*PRDM13*	PR domain containing 13
ENSG00000119973	*PRLHR*	prolactin releasing hormone receptor
ENSG00000198570	*RD3*	retinal degeneration 3
ENSG00000184302	*SIX6*	SIX homeobox 6
ENSG00000091664	*SLC17A6*	solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 6
ENSG00000165349	*SLC7A3*	solute carrier family 7 (cationic amino acid transporter, y+ system), member 3
ENSG00000157005	*SST*	somatostatin
ENSG00000126549	*STATH*	statherin
ENSG00000169836	*TACR3*	tachykinin receptor 3
ENSG00000159648	*TEPP*	testis, prostate and placenta expressed

ENSG00000186889	*TMEM17*	transmembrane protein 17
ENSG00000170893	*TRH*	thyrotropin-releasing hormone
ENSG00000069018	*TRPC7*	transient receptor potential cation channel, subfamily C, member 7
ENSG00000171724	*VAT1L*	vesicle amine transport protein 1 homolog (T. californica)-like
ENSG00000128564	*VGF*	VGF nerve growth factor inducible
ENSG00000174460	*ZCCHC12*	zinc finger, CCHC domain containing 12

Chapter 7 – Global networks dynamics in the human brain's transcriptome across the human adult lifespan

7.1 Abstract

The human brain's transcriptome shows characteristic changes during distinct developmental stages, across all genes or restricted to transcriptional subnetworks [2, 3]. These changes differ in their orders of magnitude, with the greatest changes appearing around the time of birth compared to more subtle changes later in life [2, 3]. Prompted by these age-dependent dynamics in the brain's transcriptome, I hypothesized that transcriptional networks derived from microarray gene expression data in human post-mortem brains from fetal to later life would show changes in their global network characteristics and in comparison to networks covering the whole lifespan. I tested this hypothesis by quantifying a selection of global network metrics separately for transcriptional networks from defined age bins and in comparison to sample size-matched transcriptional networks covering the whole available age range. Global network metrics for defined age groups showed differences compared to random permutations and also to the same network metrics based on the whole lifespan. More specifically global network metrics indicated highest cohesiveness in networks across the whole age range, in line with strong expression changes previously observed before and after birth. However, global network metrics did not differ among the age groups. The same trend was observed after quantifying global network measures for gene subnetworks as opposed to all expressed genes genome-wide. Thus, the present data indicate that the development of the human brain's transcriptome shows age-dependent

turning points during the lifespan but retains a similar global network structure across different age groups.

7.2 Introduction

The human brain's transcriptome changes during ageing, whereby distinct developmental stages show characteristic gene expression dynamics [2, 3]. Strongest expression changes occur directly before and after birth, with sets of age-related genes showing reversed patterns of expression fold change during later life, in subjects aged 50 years and older [2, 3]. Age-related gene expression changes have been described on a global level [2], as well as on a local level for genes clustered by their gene expression across time [3].

Global transcriptional network dynamics, defined as network metrics quantifying expression changes for gene sets, either genome-wide or based on gene clusters, as opposed to dynamics of single genes, provide a way to quantify differences between transcriptional networks across time or different conditions ([334]; for example [335]). Characterization of global network statistics across different conditions [335] could provide information about genes' or gene subnetworks' context [61, 68, 101]. For example, a gene mutation could potentially have varying downstream effects during different developmental stages, influenced by the transcriptional context of this gene's products, such as segregation into subnetworks or transcriptional network stability.

A variety of metrics allows characterization of networks on global and local levels [334, 336]. A simple way to construct a transcriptional network is to define genes as nodes and the correlation between all gene-pairs across time as network edges, using correlation coefficients as edge weights [334, 335]. Global network dynamics may then be quantified for example based on mean edge weights across the network, reflecting network cohesiveness, or modularity reflecting network segregation [335]. Further global measures include path length, a way to quantify connectivity in a network by

identifying the shortest distance between any given pair of nodes. The lower the average path length in a network, the higher the connectivity between nodes, potentially influencing the impact changes in one edge have on another one [334]. Network stability can be quantified using measures of entropy, assessing the randomness in a network [337], and network resilience to insults, e.g. using the assortativity coefficient [336]. Network metrics are described in detail in **Chapter 7.5.3**.

Based on observed age-dependence of gene expression changes across the lifespan as described above [2, 3], I tested the hypothesis that global network characteristics, such as network cohesiveness, segregation or stability, differ between distinct age groups and in comparison to transcriptional networks covering the complete available age range. Differences in global metrics for transcriptional networks derived from age groups could provide information about the context of genetic interactions and accordingly the potential downstream effects of genetic mutations or environmental insults to the transcriptional networks. Global network metrics could thus be useful for identifying critical developmental periods for environmental or genetic risk factors and their interactions.

7.3 Results

7.3.1 Transcriptional networks covering the lifespan show stronger coordination of expression than within age groups

Whole genome microarray gene expression data from *n*=248 human post-mortem brains aged between -0.5 to 78 years [2] were used to build co-expression networks (described in detail in **Chapter 2.1**). Network nodes are genes and network edges their pairwise correlation across a given age range. Each network contained correlations between all gene-pairs for 18846 expressed genes in 38 subjects, either within an age bin (Figure) or across the whole age range (*n*=10 random samples of 38 subjects each across the lifespan; boxplots in Figure 37). Correlation between gene-pairs was assessed as partial Pearson's correlation, accounting for age and two potential confounding variables, namely RIN value (see **Chapter 1.2**) and the pH of the post-mortem brains. I used an unbiased approach and analyzed the full network without introducing thresholds for correlations (except when necessary for specific network metrics, see **Chapter 7.5.3**).

First I sought to compare the temporal coordination of gene expression within age bins to the coordination across the whole lifespan, as reflected by the dynamics of global network metrics (each metric is described in detail in **Chapter 7.5.3**). Figure 37 shows distributions of seven global network metrics (red dots), compared to the same metric for random samples drawn from the lifespan (boxplots for the same metrics from *n*=10 random samples). All global metrics were compared to temporal and nodal null models. More precisely, I compared the global network metrics for each age bin and for the whole age range to the same metric after permuting (*i*) all matrix elements $a_{i,j}$ in the expression matrix *A=[i-j]* (darkblue dots in Figure 37), (*ii*) matrix elements within a given column *j* (permutation across genes, preserving gene expression values per

subject, lightblue dots in Figure 37), and (*iii*) matrix elements within a given row *i* (permutation across subjects, preserving gene expression values per gene, purple dots in Figure 37). Permutations were performed for each age bin separately and across the whole lifespan. Permutation (*i*) destroys all network structure and therefore leads to the most 'liberal' comparison of the null model to true measures. Here, results for permutation (*i*) and (*ii*) yielded highly similar global network measures (reflected by almost exactly overlapping dots in Figure 37). Results of the null models for a random sample of subjects (*n*=38) across the lifespan were essentially identical in a second random sample (not displayed). All global network metrics for different age bins and across the whole age range were markedly different from the three null distributions (*i-iii*), with the exception that of network resilience (Figure 37F), as discussed below in further detail.

Mean partial Pearson's correlation (Figure 37A) was similar among the age bins compared to the whole age range, with the exception of teenage age (network 2, Figure 37H) which showed increased mean correlation. However, the mean across all positive and negative correlations in the complete network was very close to zero. Clearly, correlations close to zero could be based on all correlations being close to zero or else similar numbers of strong correlations being highly negative or highly positive. For this reason, I next compared the absolute mean partial Pearson's correlations within the age bins to the whole age range (Figrue 37B). Absolute correlations ranged from 0.18 to 0.28 (Figure 37B). Given that correlations were assessed as partial correlations (which are expected to show lower values than correlations which do not account for confounders), this indicates that the network included highly correlated gene-pairs as opposed to all correlations centring around zero. Mean absolute correlation (Figure 37B), reflecting network cohesiveness, as well as modularity (Figure 37C), reflecting

network segregation, were each significantly lower within age bins compared to networks from the whole available age range. The increased coordination of networks covering the whole age range suggests that genes tend to show highly correlated timecourses over the whole lifespan, but less so on a shorter time scale. This is in line with previous results that showed that greatest changes in gene expression occur shortly before and after birth, and waves of gene expression changes during earliest years are mirrored during later life, in subjects aged 50 years and older [2, 3]. All transcriptional networks from all age bins showed lower entropy than the networks across the whole age range, using two different approaches to assess network entropy (see **Chapter 7.5.3.3**; Figure 37D and Figure 37E). This indicates that entropy in the transcriptional networks does not show a distinct trend towards lower or higher entropy from fetal to old age. Network resilience, reflecting the resilience of network structure in response to insults, such as removal of nodes, showed a tendency to increase from lower to higher resilience over the lifespan (Figure 37F). Only the youngest and the oldest transcriptional network did not overlap the distribution of network resilience across the 10 random samples across the whole age range. More precisely, fetal brains showed lower resilience, and the oldest group (network 5, containing subjects older than 50 years, as shown in Figure 37H) showed higher resilience. This trend was replicated after permutation within rows, which preserves gene expression data within a given gene. Overlap with this temporal null model indicates that the observed resilience was driven by genes which showed comparably small changes across time – and therefore show a similar pattern before and after permuting subjects - and not by coordinated gene expression changes. Path length (Figure 37G) showed the opposite pattern as absolute mean correlation, as would be expected in a network which defines edges as correlation between nodes (see **Chapter 7.5.3.5**).

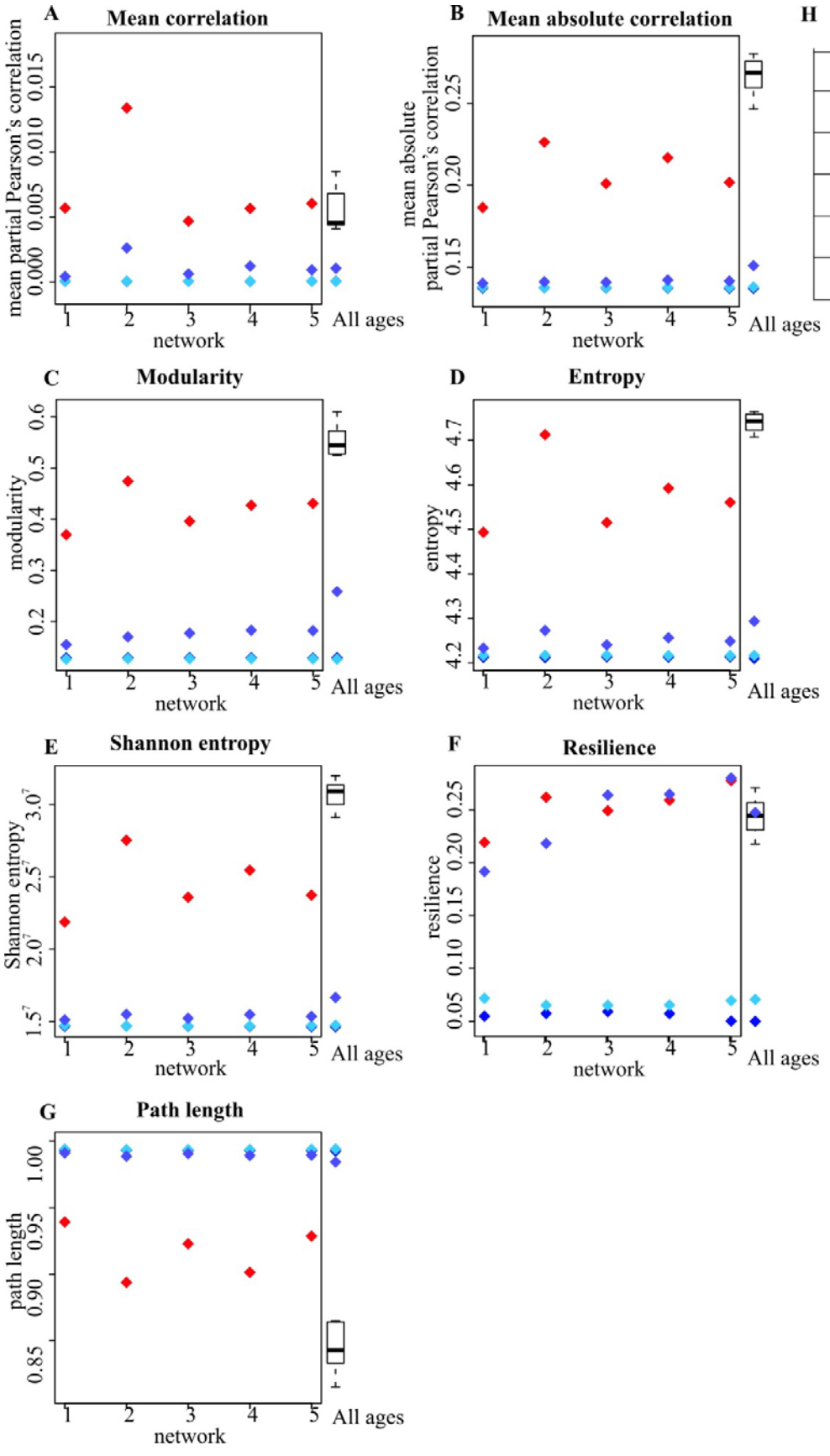

A
Mean correlation
mean partial Pearson's correlation
0.000
0.005
0.010
0.015
1
2
3
4
5
network
All ages
B
Mean absolute correlation
mean absolute partial Pearson's correlation
0.15
0.20
0.25
1
2
3
4
5
network
All ages
H
Network n=38
Age [years]
1
-0.5-0
2
12-20
3
20-35
4
35-50
5
50+
Boxplots:Lifespan
C
Modularity
modularity
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
network
All ages
D
Entropy
entropy
4.2
4.3
4.4
4.5
4.6
4.7
1
2
3
4
5
network
All ages
E
Shannon entropy
Shannon entropy
1.5^7
2.0^7
2.5^7
3.0^7
1
2
3
4
5
network
All ages
F
Resilience
resilience
0.05
0.10
0.15
0.20
0.25
1
2
3
4
5
network
All ages
G
Path length
path length
0.85
0.90
0.95
1.00
1
2
3
4
5
network
All ages

Figure 37. A-G) Seven global network metrics (y-axis) for all age groups (networks 1-5 on x-axis, each network was built from data from *n*=38 subjects). Age ranges for networks 1-5 are shown in H). Each red dot shows the mean value of the given global network metric in the respective age group. Blue and purple dots indicate the same global metrics derived after random permutations (*i-iii*) of gene expression data as described in the main text. Boxplots on the right: Each plot shows the distribution of the same global network metrics for a network of *n*=38 subjects that were randomly sampled across the full age range (*n*=10 sets of random samples). Details of network metrics are given in Chapter 7.5.3.

Next I tested the impact of each individual subject on the global network metrics by re-calculating each metric 38 times, whereby each of the 38 subjects in each network was removed once (leave-one-out (LOO) null model; path length was not re-calculated due to high computational requirements of this measure). The LOO null model generates a null distribution for variance in the data explained by single subjects as opposed to variance among the age bins (see **Chapter 7.5.4**). This procedure yielded global metrics for 38 networks (each derived from 38-1=37 subjects), showing the variance in metrics explained by exclusion of one subject. Figure 38 illustrates that the variance of global network metrics between age groups was mostly explained by the impact of single subjects. For example, mean correlation (Figure 38A) in the teenage group (12-20 years) was markedly higher than in all other groups, except after removal of one single subject. (Due to the decreased sample size, global network measures for LOO null model are not directly comparable to measures for *n*=38 subjects and are shown separately in Figure 38).

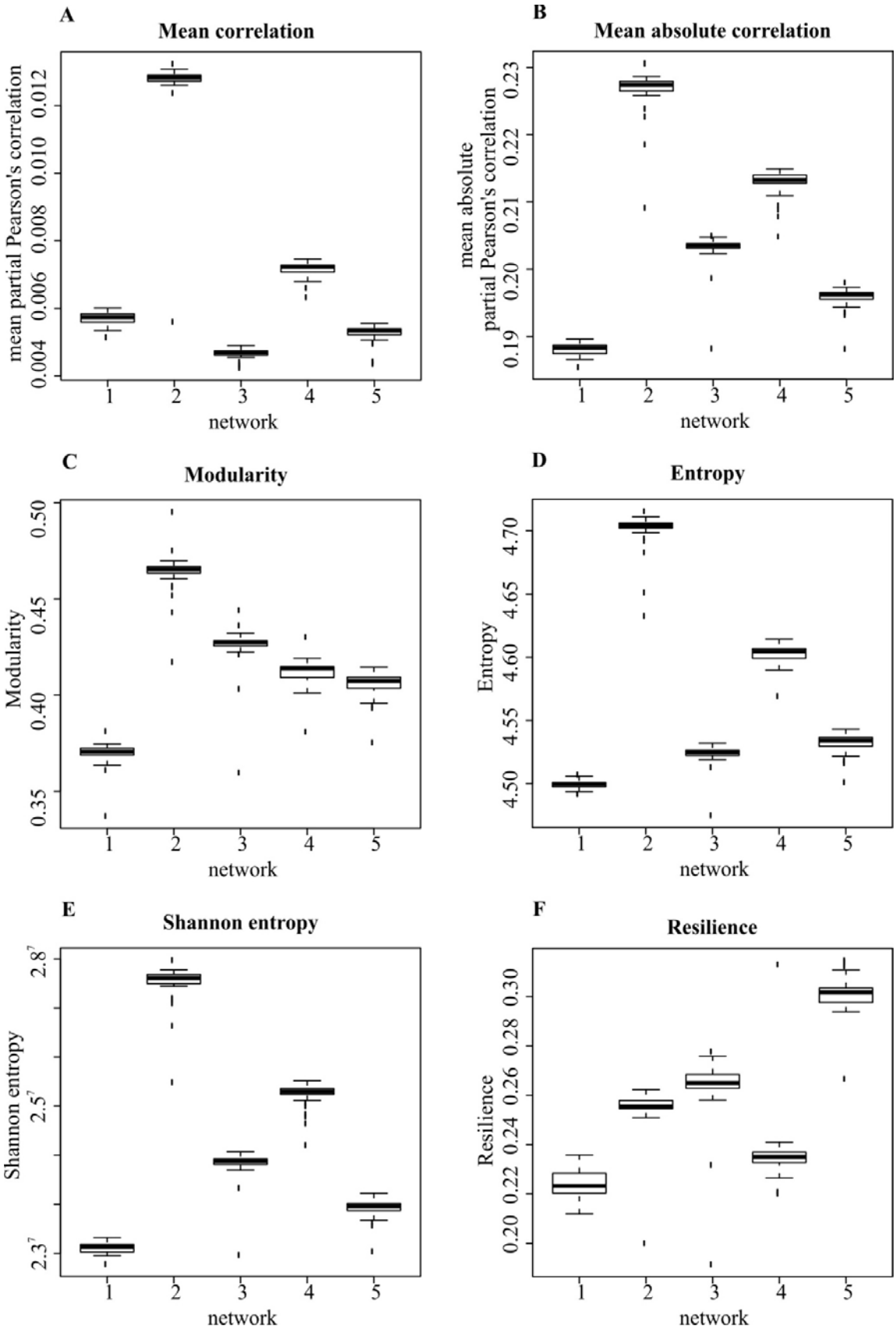

A
Mean correlation
mean partial Pearson's correlation
0.004
0.006
0.008
0.010
0.012
1
2
3
4
5
network
B
Mean absolute correlation
mean absolute partial Pearson's correlation
0.19
0.20
0.21
0.22
0.23
1
2
3
4
5
network
C
Modularity
Modularity
0.35
0.40
0.45
0.50
1
2
3
4
5
network
D
Entropy
Entropy
4.50
4.55
4.60
4.65
4.70
1
2
3
4
5
network
E
Shannon entropy
Shannon entropy
2.3^7
2.5^7
2.8^7
1
2
3
4
5
network
F
Resilience
Resilience
0.20
0.22
0.24
0.26
0.28
0.30
1
2
3
4
5
network

Figure 38. Leave-one-out null model for global network metrics. Each boxplot shows global network metrics for 38 samples of 37 subjects each, after each subject had been removed from the sample once (38-1=37). Age ranges of networks 1-5 are displayed in Figure 37H.

In summary, the comparison of global network measures across the lifespan showed that network structure is broadly similar across all distinct age groups. Any observed variance in the global network measures for the different age groups was highly sensitive to the inclusion or exclusion of single subjects (Figure 38), which indicates that global network metrics did not show stable differences across age groups. However, stronger network cohesiveness (reflected by mean absolute correlation) and modularity were observed robustly for sample size-matched networks covering the full age range compared to transcriptional networks restricted to age groups. Gene sets are known to show strong expression changes before and after birth, mirrored by reversed changes during later age [2, 3], which potentially explain the higher coordination of gene expression across the lifespan than within particular age groups.

7.3.2 Transcriptional networks from age groups but not covering the full lifespan are enriched for gene functional annotations

The observed differences in global network metrics for networks covering the full age range compared to networks within age bins led to the question whether the stronger coordination in networks covering the whole lifespan, reflected by higher mean absolute correlation and modularity (Figure 37B and Figure 37C), derived from the whole network or alternatively from distinct subnetworks. To test whether global network metrics were observed homogeneously for the whole network or differed among subnetworks, I used a Louvain-like greedy community detection method to identify

communities (or in other words ‘modules’) in transcriptional networks from each age bin and across the full age range, using an arbitrary minimum threshold of 30 genes for a module. The youngest three age groups (fetal, 12-20 years and 20-35 years) each segregated into three modules (Figure 39A). The older age groups yielded about twice as many modules compared to the younger ones (7 in case of 35-50 years and 5 for subjects aged 50 years and older). In contrast, networks covering the whole age range segregated into only two large modules (illustrated by two examples, indicated as “R1” and “R2”, Figure 39A). Overall, mean correlation within the modules – reflecting network cohesiveness – showed stronger variance in the older transcriptional networks, which obviously is influenced by the larger number of modules in this age group (mean absolute correlation in Figure 39B and mean correlation in Figure 39C). Mean correlation within the modules did not depend on module size (not displayed). Modules in fetal networks showed the lowest correlation, in contrast to networks across the entire lifespan which had the highest correlation. Across the different age bins, older networks tended to show higher correlations in their modules (Figure 39B and Figure 39C), following the same trend as mean correlations across the whole network.

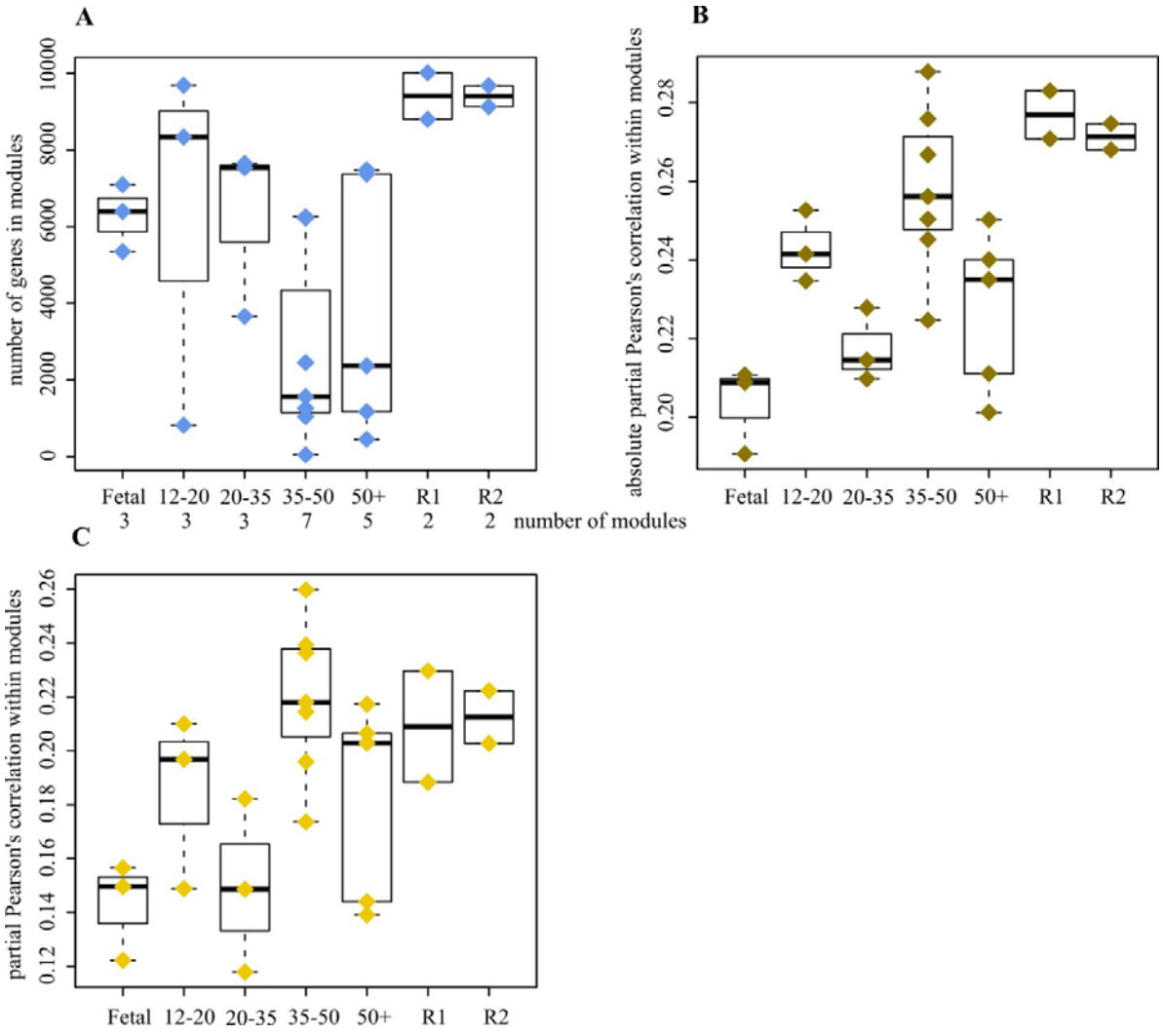


Figure 39. Transcriptional network modules within five age bins and across the whole age range (referred to as "R1" and "R2"). A) Size of network modules, in terms of numbers of genes per module. Numbers of detected modules per age bin and across age ranges [years] are listed below the x-axis. B) Partial Pearson's correlations within network modules in age bins or across the whole age range. C) Absolute partial Pearson's correlations within network modules in age bins or across the whole age range.

Next I asked the question whether transcriptional network modules in different age bins were enriched for molecular and cellular functional annotations. Enrichment for functional terms can be used as a possible approach to validate whether biological information is preserved in transcriptional modules, as would be reflected by genes annotated with the same functional term being assigned to the same module [3, 68, 101]. More specifically, I used GO-terms (described in **Chapter 2.4**) to test for functional enrichment within modules against a background of all genes that were originally used to build the transcriptional networks (Table 4). Despite having the lowest mean correlation, the modules derived from fetal transcriptional networks showed the strongest enrichments in all three modules, in terms of the number of enriched GO-terms, fold enrichment and significance. All detected modules still covered a relatively wide range of functional terms, as indicated in Table 4. Potentially, this diversity within the modules might be decreased by adjusting parameters of the Louvain-like greedy community detection method until it assembles smaller modules. However, such tuning of parameters during community detection based on characteristics of the detected modules would disallow objective comparison of global networks across the age bins, which was the aim of the present study. Modules derived from transcriptional networks across the full age range showed the opposite pattern with no significant enrichment in both highly correlated modules found in two examples (indicated “R1” and “R2” in Figure 39). The lack of enrichment for the modules across the lifespan could reflect that the majority of genes show strong expression changes around the age of birth, as was shown previously [2, 3], resulting in two highly correlated modules, that are not specific for specific molecular and cellular pathways.

Table 4. GO-terms that were enriched (α=0.05, Bonferroni corrected) in modules detected within age bins. Categories highlighted in yellow are non-redundant GO-terms for the given modules, summarized by REViGO [117] (see Chapter 2.4.1). "fold" indicates fold enrichment of this GO-term. Modules which did not show significant enrichment after p-value correction are not displayed. The numbers of detected modules (including > 30 genes) are indicated in brackets.

Fetal

Module 1

category	term	fold	p-value
GO:0003677	DNA binding	1.14	5.2E-04
GO:0003700	transcription factor activity	1.20	2.4E-02
GO:0004672	protein kinase activity	1.25	1.8E-02
GO:0004674	protein serine/threonine kinase activity	1.36	2.6E-04
GO:0004984	olfactory receptor activity	1.54	1.6E-05
GO:0005886	plasma membrane	1.14	1.7E-08
GO:0006350	transcription	1.16	6.6E-05
GO:0006355	regulation of transcription, DNA-dependent	1.20	1.6E-06
GO:0006468	protein amino acid phosphorylation	1.27	3.5E-03
GO:0007186	G-protein coupled receptor protein signaling pathway	1.21	2.7E-02
GO:0007606	sensory perception of chemical stimulus	1.54	5.3E-07
GO:0007608	sensory perception of smell	1.53	1.8E-05
GO:0016021	integral to membrane	1.07	2.7E-03
GO:0016563	transcription activator activity	1.30	2.5E-02
GO:0030528	transcription regulator activity	1.15	1.4E-02
GO:0031224	intrinsic to membrane	1.07	4.6E-03
GO:0045111	intermediate filament cytoskeleton	1.56	2.3E-02
GO:0045449	regulation of transcription	1.15	3.1E-05
GO:0051252	regulation of RNA metabolic process	1.20	8.3E-07

Module 2

category	term	fold	p-value
GO:0005576	extracellular region	1.31	4.3E-16
GO:0001775	cell activation	1.46	2.8E-02
GO:0002526	acute inflammatory response	1.90	1.2E-02
GO:0002684	positive regulation of immune system process	1.56	5.5E-03
GO:0004857	enzyme inhibitor activity	1.50	9.2E-03
GO:0005216	ion channel activity	1.37	2.3E-02
GO:0005509	calcium ion binding	1.31	1.6E-05

GO:0005615	extracellular space	1.36	7.4E-06
GO:0005773	vacuole	1.42	2.7E-02
GO:0005886	plasma membrane	1.17	2.9E-12
GO:0005887	integral to plasma membrane	1.31	2.0E-09
GO:0006811	ion transport	1.31	7.3E-04
GO:0006873	cellular ion homeostasis	1.47	4.8E-04
GO:0006875	cellular metal ion homeostasis	1.63	2.4E-03
GO:0006952	defense response	1.56	7.2E-11
GO:0006954	inflammatory response	1.63	1.4E-06
GO:0006955	immune response	1.59	1.2E-13
GO:0007155	cell adhesion	1.29	2.0E-02
GO:0007166	cell surface receptor linked signal transduction	1.17	2.8E-02
GO:0007267	cell-cell signaling	1.34	4.5E-03
GO:0009611	response to wounding	1.49	5.3E-07
GO:0010033	response to organic substance	1.28	1.2E-02
GO:0015267	channel activity	1.37	1.4E-02
GO:0016021	integral to membrane	1.12	3.9E-09
GO:0016023	cytoplasmic membrane-bounded vesicle	1.26	4.4E-02
GO:0019226	transmission of nerve impulse	1.48	5.5E-04
GO:0019725	cellular homeostasis	1.35	3.0E-02
GO:0019842	vitamin binding	1.65	4.9E-02
GO:0022610	biological adhesion	1.29	1.6E-02
GO:0022803	passive transmembrane transporter activity	1.37	1.4E-02
GO:0022838	substrate specific channel activity	1.38	1.5E-02
GO:0030003	cellular cation homeostasis	1.52	9.4E-03
GO:0031012	extracellular matrix	1.37	1.3E-02
GO:0031224	intrinsic to membrane	1.12	7.3E-11
GO:0031226	intrinsic to plasma membrane	1.30	2.4E-09
GO:0031406	carboxylic acid binding	1.64	1.8E-02
GO:0031410	cytoplasmic vesicle	1.24	4.9E-02
GO:0031988	membrane-bounded vesicle	1.25	4.6E-02
GO:0042592	homeostatic process	1.30	3.1E-03
GO:0044057	regulation of system process	1.48	2.6E-03
GO:0044421	extracellular region part	1.36	1.8E-09
GO:0044459	plasma membrane part	1.23	5.0E-11
GO:0045087	innate immune response	1.97	1.2E-05
GO:0048584	positive regulation of response to stimulus	1.52	2.2E-02
GO:0048878	chemical homeostasis	1.49	2.1E-07
GO:0050801	ion homeostasis	1.49	2.8E-05
GO:0055065	metal ion homeostasis	1.64	1.1E-03
GO:0055080	cation homeostasis	1.54	6.1E-04
GO:0055082	cellular chemical homeostasis	1.48	1.5E-04

Module 3

category	term	fold	p-value
GO:0031974	membrane-enclosed lumen	1.37	4.6E-27
GO:0000087	M phase of mitotic cell cycle	1.76	2.8E-09
GO:0000226	microtubule cytoskeleton organization	1.60	1.9E-02
GO:0000278	mitotic cell cycle	1.60	1.1E-10
GO:0000279	M phase	1.59	2.2E-08
GO:0000280	nuclear division	1.77	1.7E-09
GO:0000313	organellar ribosome	2.08	2.0E-03
GO:0000375	RNA splicing, via transesterification reactions	1.80	1.4E-07
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	1.80	1.4E-07
GO:0000387	spliceosomal snRNP biogenesis	2.52	7.2E-03
GO:0000398	nuclear mRNA splicing, via spliceosome	1.80	1.4E-07
GO:0000775	chromosome, centromeric region	2.04	5.6E-10
GO:0000776	kinetochore	2.20	6.1E-08
GO:0000777	condensed chromosome kinetochore	2.38	2.3E-07
GO:0000779	condensed chromosome, centromeric region	2.27	8.8E-07
GO:0000793	condensed chromosome	1.92	2.4E-07
GO:0003677	DNA binding	1.17	2.3E-06
GO:0003723	RNA binding	1.34	1.7E-06
GO:0003735	structural constituent of ribosome	1.61	6.7E-04
GO:0003954	NADH dehydrogenase activity	2.31	1.2E-03
GO:0005643	nuclear pore	1.71	3.7E-02
GO:0005654	nucleoplasm	1.41	2.3E-14
GO:0005681	spliceosome	1.89	2.3E-08
GO:0005694	chromosome	1.51	1.4E-09
GO:0005730	nucleolus	1.32	1.6E-05
GO:0005739	mitochondrion	1.40	1.1E-16
GO:0005740	mitochondrial envelope	1.34	4.5E-03
GO:0005743	mitochondrial inner membrane	1.48	7.8E-05
GO:0005746	mitochondrial respiratory chain	1.95	1.4E-02
GO:0005747	mitochondrial respiratory chain complex I	2.35	6.1E-04
GO:0005759	mitochondrial matrix	1.61	1.5E-06
GO:0005761	mitochondrial ribosome	2.08	2.0E-03
GO:0005815	microtubule organizing center	1.44	6.7E-03
GO:0005819	spindle	1.61	6.6E-03
GO:0005840	ribosome	1.59	2.8E-05
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	2.18	4.8E-02
GO:0006259	DNA metabolic process	1.40	7.8E-06
GO:0006260	DNA replication	1.70	3.3E-06

GO:0006261	DNA-dependent DNA replication	1.98	3.3E-02
GO:0006281	DNA repair	1.45	1.8E-03
GO:0006323	DNA packaging	1.84	2.7E-04
GO:0006334	nucleosome assembly	1.92	4.5E-03
GO:0006350	transcription	1.14	1.6E-02
GO:0006396	RNA processing	1.56	1.6E-15
GO:0006397	mRNA processing	1.70	1.6E-13
GO:0006403	RNA localization	1.77	2.4E-03
GO:0006412	translation	1.49	9.6E-06
GO:0006974	response to DNA damage stimulus	1.36	1.4E-02
GO:0007049	cell cycle	1.41	2.6E-11
GO:0007067	mitosis	1.77	1.7E-09
GO:0008137	NADH dehydrogenase (ubiquinone) activity	2.31	1.2E-03
GO:0008380	RNA splicing	1.70	9.8E-12
GO:0015630	microtubule cytoskeleton	1.28	1.1E-02
GO:0015931	nucleobase, nucleoside, nucleotide and nucleic acid transport	1.67	2.1E-02
GO:0015935	small ribosomal subunit	1.87	2.0E-02
GO:0016071	mRNA metabolic process	1.62	7.5E-12
GO:0016604	nuclear body	1.50	1.7E-02
GO:0016651	oxidoreductase activity, acting on NADH or NADPH	1.79	2.1E-02
GO:0016655	oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor	2.31	1.3E-04
GO:0019866	organelle inner membrane	1.40	1.7E-03
GO:0022402	cell cycle process	1.46	6.4E-10
GO:0022403	cell cycle phase	1.53	4.5E-09
GO:0022613	ribonucleoprotein complex biogenesis	1.68	8.1E-06
GO:0022618	ribonucleoprotein complex assembly	1.87	1.5E-02
GO:0022900	electron transport chain	1.71	2.1E-02
GO:0030529	ribonucleoprotein complex	1.60	1.0E-16
GO:0030532	small nuclear ribonucleoprotein complex	2.61	5.4E-04
GO:0030964	NADH dehydrogenase complex	2.35	6.1E-04
GO:0031497	chromatin assembly	1.90	4.4E-03
GO:0031966	mitochondrial membrane	1.33	1.3E-02
GO:0031967	organelle envelope	1.34	6.3E-06
GO:0031975	envelope	1.34	7.7E-06
GO:0031980	mitochondrial lumen	1.61	1.5E-06
GO:0031981	nuclear lumen	1.36	5.1E-19
GO:0033279	ribosomal subunit	1.64	4.0E-03
GO:0034621	cellular macromolecular complex subunit organization	1.59	9.9E-10

GO:0034622	cellular macromolecular complex assembly	1.55	3.8E-07
GO:0034728	nucleosome organization	1.86	4.0E-03
GO:0043228	non-membrane-bounded organelle	1.21	6.7E-12
GO:0043232	intracellular non-membrane-bounded organelle	1.21	6.7E-12
GO:0043233	organelle lumen	1.36	6.0E-26
GO:0043933	macromolecular complex subunit organization	1.30	1.4E-04
GO:0044427	chromosomal part	1.58	1.4E-10
GO:0044429	mitochondrial part	1.42	5.7E-09
GO:0044451	nucleoplasm part	1.32	2.4E-04
GO:0044455	mitochondrial membrane part	1.80	5.2E-05
GO:0045271	respiratory chain complex I	2.35	6.1E-04
GO:0048285	organelle fission	1.73	6.6E-09
GO:0050136	NADH dehydrogenase (quinone) activity	2.31	1.2E-03
GO:0050657	nucleic acid transport	1.83	4.9E-04
GO:0050658	RNA transport	1.83	4.9E-04
GO:0051028	mRNA transport	1.86	1.1E-03
GO:0051236	establishment of RNA localization	1.83	4.9E-04
GO:0051276	chromosome organization	1.31	2.6E-02
GO:0051301	cell division	1.60	1.4E-07
GO:0065003	macromolecular complex assembly	1.28	4.6E-03
GO:0065004	protein-DNA complex assembly	1.94	6.3E-04
GO:0070013	intracellular organelle lumen	1.37	3.2E-26
GO:0070469	respiratory chain	2.03	4.6E-04

12-20 years (3 modules)

Module 1

category	term	fold	p-value
GO:0004175	endopeptidase activity	1.25	2.7E-02
GO:0004252	serine-type endopeptidase activity	1.44	6.9E-03
GO:0004857	enzyme inhibitor activity	1.30	1.4E-02
GO:0004984	olfactory receptor activity	1.35	4.1E-04
GO:0005576	extracellular region	1.22	7.2E-20
GO:0005578	proteinaceous extracellular matrix	1.29	4.4E-04
GO:0005615	extracellular space	1.23	2.6E-06
GO:0005886	plasma membrane	1.08	3.3E-05
GO:0005887	integral to plasma membrane	1.14	7.2E-04
GO:0006952	defense response	1.27	9.6E-06
GO:0006955	immune response	1.35	1.4E-13
GO:0007166	cell surface receptor linked signal transduction	1.16	7.6E-08
GO:0007186	G-protein coupled receptor protein signaling pathway	1.23	1.3E-07
GO:0007398	ectoderm development	1.38	4.0E-03

GO:0007600	sensory perception	1.24	4.4E-05
GO:0007606	sensory perception of chemical stimulus	1.32	4.4E-04
GO:0007608	sensory perception of smell	1.31	5.0E-03
GO:0008544	epidermis development	1.37	2.3E-02
GO:0009611	response to wounding	1.21	3.1E-02
GO:0030855	epithelial cell differentiation	1.53	3.0E-05
GO:0031012	extracellular matrix	1.25	3.3E-03
GO:0031226	intrinsic to plasma membrane	1.14	3.5E-04
GO:0044421	extracellular region part	1.24	8.5E-11
GO:0044459	plasma membrane part	1.09	7.7E-03
GO:0050877	neurological system process	1.15	7.4E-03
GO:0050890	cognition	1.21	2.6E-04

Module 2

category	term	fold	p-value
GO:0000775	chromosome, centromeric region	1.48	1.8E-02
GO:0003677	DNA binding	1.12	1.1E-04
GO:0004221	ubiquitin thiolesterase activity	1.69	2.8E-03
GO:0005654	nucleoplasm	1.24	9.2E-07
GO:0005694	chromosome	1.27	1.4E-03
GO:0005730	nucleolus	1.20	8.6E-03
GO:0005739	mitochondrion	1.15	7.4E-03
GO:0005794	Golgi apparatus	1.20	2.8E-04
GO:0006350	transcription	1.14	8.8E-06
GO:0006396	RNA processing	1.24	5.5E-03
GO:0006511	ubiquitin-dependent protein catabolic process	1.47	1.5E-06
GO:0008104	protein localization	1.18	5.0E-03
GO:0008270	zinc ion binding	1.16	1.7E-09
GO:0009057	macromolecule catabolic process	1.27	4.8E-07
GO:0015031	protein transport	1.21	8.8E-04
GO:0016790	thiolester hydrolase activity	1.56	1.3E-02
GO:0016879	ligase activity, forming carbon-nitrogen bonds	1.35	9.7E-03
GO:0019941	modification-dependent protein catabolic process	1.39	5.5E-12
GO:0030163	protein catabolic process	1.34	1.3E-09
GO:0031974	membrane-enclosed lumen	1.18	2.7E-09
GO:0031981	nuclear lumen	1.20	1.4E-08
GO:0043167	ion binding	1.07	1.8E-03
GO:0043169	cation binding	1.07	3.0E-03
GO:0043228	non-membrane-bounded organelle	1.11	6.7E-04
GO:0043232	intracellular non-membrane-bounded organelle	1.11	6.7E-04
GO:0043233	organelle lumen	1.18	4.9E-09

GO:0043632	modification-dependent macromolecule catabolic process	1.39	5.5E-12
GO:0044257	cellular protein catabolic process	1.35	6.9E-10
GO:0044265	cellular macromolecule catabolic process	1.30	6.9E-09
GO:0044427	chromosomal part	1.34	5.6E-05
GO:0044451	nucleoplasm part	1.21	2.6E-02
GO:0045184	establishment of protein localization	1.21	7.1E-04
GO:0045449	regulation of transcription	1.10	3.8E-03
GO:0046872	metal ion binding	1.07	2.7E-03
GO:0046914	transition metal ion binding	1.12	1.3E-06
GO:0051603	proteolysis involved in cellular protein catabolic process	1.35	9.1E-10
GO:0070013	intracellular organelle lumen	1.19	1.6E-09

20-35 years (3 modules)

Module 1

category	term	fold	p-value
GO:0001653	peptide receptor activity	1.62	1.0E-02
GO:0003700	transcription factor activity	1.24	2.5E-05
GO:0004984	olfactory receptor activity	1.74	8.4E-15
GO:0005576	extracellular region	1.20	3.5E-09
GO:0005886	plasma membrane	1.17	1.6E-17
GO:0005887	integral to plasma membrane	1.24	4.5E-08
GO:0006955	immune response	1.34	2.0E-06
GO:0007166	cell surface receptor linked signal transduction	1.26	1.2E-12
GO:0007186	G-protein coupled receptor protein signaling pathway	1.40	8.0E-16
GO:0007600	sensory perception	1.45	9.2E-14
GO:0007606	sensory perception of chemical stimulus	1.69	2.1E-15
GO:0007608	sensory perception of smell	1.70	8.4E-14
GO:0008528	peptide receptor activity, G-protein coupled	1.62	1.0E-02
GO:0016021	integral to membrane	1.12	8.9E-14
GO:0031224	intrinsic to membrane	1.12	8.9E-14
GO:0031226	intrinsic to plasma membrane	1.24	8.8E-09
GO:0043565	sequence-specific DNA binding	1.34	1.3E-06
GO:0044421	extracellular region part	1.18	7.7E-03
GO:0044459	plasma membrane part	1.17	3.1E-08
GO:0050877	neurological system process	1.28	1.8E-08
GO:0050890	cognition	1.39	1.6E-11

Module 2

category	term	fold	p-value
GO:0000166	nucleotide binding	1.13	3.3E-04

GO:0000313	organellar ribosome	2.08	1.7E-06
GO:0000375	RNA splicing, via transesterification reactions	1.57	8.6E-05
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	1.57	8.6E-05
GO:0000398	nuclear mRNA splicing, via spliceosome	1.57	8.6E-05
GO:0000502	proteasome complex	1.73	1.6E-02
GO:0000775	chromosome, centromeric region	1.48	4.4E-02
GO:0003723	RNA binding	1.44	2.2E-18
GO:0003735	structural constituent of ribosome	1.56	6.5E-05
GO:0003743	translation initiation factor activity	1.87	2.5E-04
GO:0005654	nucleoplasm	1.17	1.5E-02
GO:0005681	spliceosome	1.62	1.5E-05
GO:0005730	nucleolus	1.29	5.1E-07
GO:0005739	mitochondrion	1.34	1.6E-17
GO:0005740	mitochondrial envelope	1.36	2.8E-06
GO:0005743	mitochondrial inner membrane	1.37	5.0E-04
GO:0005746	mitochondrial respiratory chain	1.86	1.3E-03
GO:0005759	mitochondrial matrix	1.54	5.2E-08
GO:0005761	mitochondrial ribosome	2.08	1.7E-06
GO:0005794	Golgi apparatus	1.17	1.9E-02
GO:0005829	cytosol	1.22	2.5E-08
GO:0005840	ribosome	1.59	1.1E-08
GO:0006091	generation of precursor metabolites and energy	1.34	1.3E-02
GO:0006119	oxidative phosphorylation	1.68	5.8E-03
GO:0006364	rRNA processing	1.60	4.2E-02
GO:0006396	RNA processing	1.48	2.7E-16
GO:0006397	mRNA processing	1.49	9.5E-09
GO:0006412	translation	1.62	2.6E-16
GO:0006413	translational initiation	2.01	1.1E-03
GO:0006457	protein folding	1.59	1.4E-05
GO:0006511	ubiquitin-dependent protein catabolic process	1.50	3.8E-06
GO:0006605	protein targeting	1.41	4.1E-03
GO:0006886	intracellular protein transport	1.47	3.5E-10
GO:0007005	mitochondrion organization	1.52	6.1E-03
GO:0008104	protein localization	1.31	5.2E-11
GO:0008135	translation factor activity, nucleic acid binding	1.73	2.3E-05
GO:0008380	RNA splicing	1.52	7.9E-09
GO:0009057	macromolecule catabolic process	1.29	4.4E-07
GO:0010498	proteasomal protein catabolic process	1.74	2.2E-05
GO:0012505	endomembrane system	1.19	1.1E-02
GO:0015031	protein transport	1.38	8.6E-15
GO:0015980	energy derivation by oxidation of organic compounds	1.50	4.1E-02

GO:0016071	mRNA metabolic process	1.43	2.5E-07
GO:0016072	rRNA metabolic process	1.61	1.8E-02
GO:0019866	organelle inner membrane	1.35	7.8E-04
GO:0019941	modification-dependent protein catabolic process	1.36	2.2E-08
GO:0022613	ribonucleoprotein complex biogenesis	1.70	2.6E-10
GO:0022618	ribonucleoprotein complex assembly	1.73	1.3E-02
GO:0022900	electron transport chain	1.60	2.0E-02
GO:0022904	respiratory electron transport chain	1.81	3.8E-02
GO:0030163	protein catabolic process	1.32	8.7E-07
GO:0030529	ribonucleoprotein complex	1.53	1.6E-19
GO:0031090	organelle membrane	1.22	1.8E-06
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	1.89	1.2E-04
GO:0031396	regulation of protein ubiquitination	1.65	1.6E-03
GO:0031397	negative regulation of protein ubiquitination	1.77	1.5E-03
GO:0031398	positive regulation of protein ubiquitination	1.66	1.1E-02
GO:0031966	mitochondrial membrane	1.34	5.5E-05
GO:0031967	organelle envelope	1.33	1.3E-08
GO:0031974	membrane-enclosed lumen	1.24	4.8E-17
GO:0031975	envelope	1.33	1.7E-08
GO:0031980	mitochondrial lumen	1.54	5.2E-08
GO:0031981	nuclear lumen	1.21	3.2E-09
GO:0033365	protein localization in organelle	1.46	3.6E-02
GO:0034470	ncRNA processing	1.46	3.3E-03
GO:0034613	cellular protein localization	1.46	1.6E-10
GO:0034621	cellular macromolecular complex subunit organization	1.42	1.5E-06
GO:0034622	cellular macromolecular complex assembly	1.40	9.8E-05
GO:0034660	ncRNA metabolic process	1.48	4.2E-05
GO:0042254	ribosome biogenesis	1.70	5.0E-06
GO:0043161	proteasomal ubiquitin-dependent protein catabolic process	1.74	2.2E-05
GO:0043228	non-membrane-bounded organelle	1.16	8.6E-10
GO:0043232	intracellular non-membrane-bounded organelle	1.16	8.6E-10
GO:0043233	organelle lumen	1.24	7.8E-16
GO:0043632	modification-dependent macromolecule catabolic process	1.36	2.2E-08
GO:0043933	macromolecular complex subunit organization	1.26	8.0E-05
GO:0044257	cellular protein catabolic process	1.33	4.1E-07
GO:0044265	cellular macromolecule catabolic process	1.32	8.6E-09
GO:0044429	mitochondrial part	1.44	3.2E-15
GO:0044455	mitochondrial membrane part	1.77	1.2E-07

GO:0045184	establishment of protein localization	1.38	1.2E-14
GO:0045333	cellular respiration	1.72	1.0E-03
GO:0046907	intracellular transport	1.38	3.7E-12
GO:0048193	Golgi vesicle transport	1.49	3.0E-02
GO:0051082	unfolded protein binding	1.71	1.9E-05
GO:0051340	regulation of ligase activity	1.82	2.5E-05
GO:0051351	positive regulation of ligase activity	1.87	3.5E-05
GO:0051352	negative regulation of ligase activity	1.79	3.0E-03
GO:0051436	negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle	1.85	5.6E-04
GO:0051437	positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle	1.83	7.3E-04
GO:0051438	regulation of ubiquitin-protein ligase activity	1.80	1.5E-04
GO:0051439	regulation of ubiquitin-protein ligase activity during mitotic cell cycle	1.85	1.2E-04
GO:0051443	positive regulation of ubiquitin-protein ligase activity	1.84	2.2E-04
GO:0051444	negative regulation of ubiquitin-protein ligase activity	1.79	3.0E-03
GO:0051603	proteolysis involved in cellular protein catabolic process	1.33	3.1E-07
GO:0065003	macromolecular complex assembly	1.25	1.1E-03
GO:0070013	intracellular organelle lumen	1.24	6.0E-16
GO:0070469	respiratory chain	1.84	3.6E-04
GO:0070727	cellular macromolecule localization	1.46	1.1E-10

Module 3

category	term	fold	p-value
GO:0008270	zinc ion binding	1.22	6.0E-04
GO:0045449	regulation of transcription	1.20	2.5E-03
GO:0044421	extracellular region part	1.36	1.5E-03
GO:0046914	transition metal ion binding	1.16	2.6E-02

30-50 years (7 modules)

Module 1

category	term	fold	p-value
GO:0000313	organellar ribosome	1.97	2.3E-02
GO:0003723	RNA binding	1.28	1.4E-03
GO:0003743	translation initiation factor activity	1.97	4.7E-03
GO:0005739	mitochondrion	1.39	1.8E-15
GO:0005740	mitochondrial envelope	1.40	8.8E-05
GO:0005743	mitochondrial inner membrane	1.41	3.7E-03
GO:0005759	mitochondrial matrix	1.52	2.2E-04
GO:0005761	mitochondrial ribosome	1.97	2.3E-02

GO:0006091	generation of precursor metabolites and energy	1.44	3.5E-03
GO:0006396	RNA processing	1.34	4.6E-04
GO:0006397	mRNA processing	1.39	1.9E-02
GO:0006412	translation	1.45	4.5E-04
GO:0006413	translational initiation	2.15	1.4E-02
GO:0006457	protein folding	1.58	7.5E-03
GO:0006511	ubiquitin-dependent protein catabolic process	1.52	8.3E-04
GO:0007005	mitochondrion organization	1.60	2.1E-02
GO:0009057	macromolecule catabolic process	1.24	2.4E-02
GO:0019866	organelle inner membrane	1.38	9.1E-03
GO:0019941	modification-dependent protein catabolic process	1.36	8.4E-05
GO:0030163	protein catabolic process	1.30	4.5E-03
GO:0030529	ribonucleoprotein complex	1.35	6.3E-05
GO:0031398	positive regulation of protein ubiquitination	1.77	3.5E-02
GO:0031966	mitochondrial membrane	1.37	1.5E-03
GO:0031967	organelle envelope	1.27	6.7E-03
GO:0031974	membrane-enclosed lumen	1.14	4.4E-03
GO:0031975	envelope	1.26	7.7E-03
GO:0031980	mitochondrial lumen	1.52	2.2E-04
GO:0043233	organelle lumen	1.14	1.6E-02
GO:0043632	modification-dependent macromolecule catabolic process	1.36	8.4E-05
GO:0044257	cellular protein catabolic process	1.31	2.0E-03
GO:0044265	cellular macromolecule catabolic process	1.27	2.8E-03
GO:0044429	mitochondrial part	1.46	1.2E-10
GO:0044455	mitochondrial membrane part	1.64	1.2E-02
GO:0046907	intracellular transport	1.26	2.2E-02
GO:0051082	unfolded protein binding	1.83	1.1E-04
GO:0051351	positive regulation of ligase activity	1.89	1.0E-02
GO:0051437	positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle	1.89	2.2E-02
GO:0051439	regulation of ubiquitin-protein ligase activity during mitotic cell cycle	1.85	3.5E-02
GO:0051443	positive regulation of ubiquitin-protein ligase activity	1.93	6.1E-03
GO:0051603	proteolysis involved in cellular protein catabolic process	1.32	1.2E-03
GO:0070013	intracellular organelle lumen	1.15	2.1E-03

Module 2

category	term	fold	p-value
GO:0005576	extracellular region	1.23	2.3E-08

GO:0003700	transcription factor activity	1.19	6.9E-04
GO:0005578	proteinaceous extracellular matrix	1.23	2.0E-03
GO:0006350	transcription	1.24	1.4E-03
GO:0006357	regulation of transcription from RNA polymerase II promoter	1.29	5.9E-03
GO:0030528	transcription regulator activity	1.14	2.0E-02
GO:0043565	sequence-specific DNA binding	1.25	2.5E-02
GO:0044421	extracellular region part	1.38	1.2E-02

Module 3

category	term	fold	p-value
GO:0001568	blood vessel development	2.01	5.7E-04
GO:0001570	vasculogenesis	3.62	1.4E-02
GO:0001944	vasculature development	1.97	1.3E-03
GO:0002684	positive regulation of immune system process	2.02	1.9E-03
GO:0005615	extracellular space	1.49	4.1E-03
GO:0006952	defense response	1.85	1.1E-07
GO:0006954	inflammatory response	1.85	4.5E-03
GO:0006955	immune response	1.73	4.3E-06
GO:0009611	response to wounding	1.79	6.9E-06
GO:0010033	response to organic substance	1.52	2.3E-03
GO:0010941	regulation of cell death	1.42	4.0E-02
GO:0042127	regulation of cell proliferation	1.44	3.1E-02
GO:0042981	regulation of apoptosis	1.42	4.5E-02
GO:0043067	regulation of programmed cell death	1.43	3.2E-02
GO:0044421	extracellular region part	1.42	1.9E-03
GO:0048514	blood vessel morphogenesis	2.06	1.7E-03
GO:0050778	positive regulation of immune response	2.23	2.8E-02

Module 5

category	term	fold	p-value
GO:0005216	ion channel activity	2.10	1.3E-04
GO:0005244	voltage-gated ion channel activity	2.74	1.3E-05
GO:0005249	voltage-gated potassium channel activity	2.76	4.1E-02
GO:0005261	cation channel activity	2.34	1.2E-04
GO:0005267	potassium channel activity	2.58	2.2E-02
GO:0005509	calcium ion binding	1.51	3.2E-02
GO:0005886	plasma membrane	1.25	2.3E-04
GO:0015267	channel activity	2.00	5.3E-04
GO:0016021	integral to membrane	1.17	5.9E-03
GO:0022803	passive transmembrane transporter activity	2.00	5.3E-04
GO:0022832	voltage-gated channel activity	2.74	1.3E-05
GO:0022836	gated channel activity	2.40	3.3E-06
GO:0022838	substrate specific channel activity	2.04	3.2E-04

GO:0022843	voltage-gated cation channel activity	2.85	2.1E-04
GO:0030001	metal ion transport	1.87	7.8E-03
GO:0031224	intrinsic to membrane	1.16	6.5E-03
GO:0031420	alkali metal ion binding	2.18	2.4E-02
GO:0034702	ion channel complex	2.30	2.8E-03
GO:0034703	cation channel complex	2.82	3.6E-04
GO:0045202	synapse	1.90	4.6E-03
GO:0046873	metal ion transmembrane transporter activity	2.04	3.2E-03
GO:0050877	neurological system process	1.60	9.4E-04

50 years and older (5 modules)

Module 2

category	term	fold	p-value
GO:0003700	transcription factor activity	1.19	1.0E-02
GO:0004984	olfactory receptor activity	1.54	5.0E-07
GO:0005576	extracellular region	1.19	8.1E-08
GO:0005578	proteinaceous extracellular matrix	1.33	1.1E-02
GO:0005886	plasma membrane	1.09	4.5E-04
GO:0005887	integral to plasma membrane	1.14	4.8E-02
GO:0007166	cell surface receptor linked signal transduction	1.17	9.2E-05
GO:0007186	G-protein coupled receptor protein signaling pathway	1.25	9.0E-05
GO:0007600	sensory perception	1.30	4.6E-05
GO:0007606	sensory perception of chemical stimulus	1.48	3.6E-06
GO:0007608	sensory perception of smell	1.52	2.1E-06
GO:0016021	integral to membrane	1.06	1.6E-02
GO:0030528	transcription regulator activity	1.13	4.8E-02
GO:0031224	intrinsic to membrane	1.06	5.3E-03
GO:0031226	intrinsic to plasma membrane	1.15	2.0E-02
GO:0043565	sequence-specific DNA binding	1.26	2.9E-03
GO:0044421	extracellular region part	1.17	2.8E-02
GO:0050890	cognition	1.26	5.5E-04

Module 3

category	term	fold	p-value
GO:0000166	nucleotide binding	1.15	1.8E-05
GO:0003723	RNA binding	1.34	2.6E-08
GO:0003743	translation initiation factor activity	1.79	2.1E-02
GO:0005739	mitochondrion	1.25	6.7E-08
GO:0005759	mitochondrial matrix	1.48	4.5E-05
GO:0005829	cytosol	1.15	1.2E-02
GO:0006259	DNA metabolic process	1.27	3.0E-02
GO:0006364	rRNA processing	1.66	2.5E-02

GO:0006396	RNA processing	1.36	4.6E-07
GO:0006397	mRNA processing	1.35	1.0E-02
GO:0006412	translation	1.44	9.5E-06
GO:0006511	ubiquitin-dependent protein catabolic process	1.50	3.2E-05
GO:0006886	intracellular protein transport	1.37	2.3E-04
GO:0008104	protein localization	1.23	2.0E-04
GO:0008135	translation factor activity, nucleic acid binding	1.58	5.0E-02
GO:0008380	RNA splicing	1.39	4.5E-03
GO:0009057	macromolecule catabolic process	1.26	2.0E-04
GO:0010498	proteasomal protein catabolic process	1.62	2.7E-02
GO:0015031	protein transport	1.29	1.9E-06
GO:0016071	mRNA metabolic process	1.31	2.3E-02
GO:0017076	purine nucleotide binding	1.14	3.0E-03
GO:0019941	modification-dependent protein catabolic process	1.33	9.5E-06
GO:0022613	ribonucleoprotein complex biogenesis	1.53	6.3E-04
GO:0030163	protein catabolic process	1.28	7.6E-04
GO:0030529	ribonucleoprotein complex	1.34	1.9E-06
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	1.85	3.4E-03
GO:0031398	positive regulation of protein ubiquitination	1.72	7.7E-03
GO:0031974	membrane-enclosed lumen	1.18	4.8E-07
GO:0031980	mitochondrial lumen	1.48	4.5E-05
GO:0031981	nuclear lumen	1.16	1.0E-03
GO:0032553	ribonucleotide binding	1.14	1.4E-03
GO:0032555	purine ribonucleotide binding	1.14	1.4E-03
GO:0034613	cellular protein localization	1.34	7.8E-04
GO:0034660	ncRNA metabolic process	1.46	8.0E-04
GO:0042254	ribosome biogenesis	1.60	7.0E-03
GO:0043161	proteasomal ubiquitin-dependent protein catabolic process	1.62	2.7E-02
GO:0043233	organelle lumen	1.17	9.8E-07
GO:0043632	modification-dependent macromolecule catabolic process	1.33	9.5E-06
GO:0044257	cellular protein catabolic process	1.29	5.5E-04
GO:0044265	cellular macromolecule catabolic process	1.30	6.1E-06
GO:0044429	mitochondrial part	1.31	7.9E-06
GO:0045184	establishment of protein localization	1.29	1.2E-06
GO:0046907	intracellular transport	1.30	6.2E-06
GO:0051340	regulation of ligase activity	1.75	3.9E-03
GO:0051351	positive regulation of ligase activity	1.89	1.8E-04
GO:0051352	negative regulation of ligase activity	1.75	4.6E-02
GO:0051436	negative regulation of ubiquitin-protein	1.81	1.3E-02

	ligase activity during mitotic cell cycle		
GO:0051437	positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle	1.84	3.9E-03
GO:0051438	regulation of ubiquitin-protein ligase activity	1.72	2.0E-02
GO:0051439	regulation of ubiquitin-protein ligase activity during mitotic cell cycle	1.83	2.4E-03
GO:0051443	positive regulation of ubiquitin-protein ligase activity	1.86	1.2E-03
GO:0051444	negative regulation of ubiquitin-protein ligase activity	1.75	4.6E-02
GO:0051603	proteolysis involved in cellular protein catabolic process	1.29	4.7E-04
GO:0070013	intracellular organelle lumen	1.18	2.3E-07
GO:0070727	cellular macromolecule localization	1.33	9.5E-04

Module 4

category	term	fold	p-value
GO:0006955	immune response	1.81	3.3E-02

Module 5

category	term	fold	p-value
GO:0009611	response to wounding	1.69	8.7E-04
GO:0022612	gland morphogenesis	6.02	3.4E-03
GO:0005615	extracellular space	1.57	4.9E-04
GO:0044421	extracellular region part	1.42	3.9E-03
GO:0006954	inflammatory response	1.78	3.3E-02

7.4 Discussion

The temporal dynamics of the human brain's transcriptome show characteristic patterns at distinct developmental stages with strongest changes occurring directly before and after birth [2, 3]. Based on these trends observed from large numbers of genes, I tested the hypothesis that transcriptional networks from distinct age bins or across the whole age range show differences in global network metrics.

Comparison of (*i*) transcriptional networks derived from five age bins, (*ii*) transcriptional networks from the whole lifespan, and (*iii*) temporal and nodal null distributions of transcriptional networks consistently showed differences between these three groups. More precisely, mean absolute correlation, modularity, entropy and path lengths all showed the most extreme values for networks covering the whole age range, with less extreme values for transcriptional networks based on the age bins, and the least extreme observed values for the null distributions. The only exception was network resilience which overlapped with the temporal null distribution. Potentially, the overlap with the null distribution could be caused by observed resilience values deriving predominantly from correlations of genes with similar timecourses across the given age range, which consequentially show the highest similarity to the temporal null distribution. The observed increased coordination of transcriptional networks for different age groups (for example higher mean absolute correlation and modularity compared to random permutations) indicates that transcriptional networks preserve information about coordination between genes across different age stages and after correcting for confounding variables. Networks covering the full age range showed the most extreme values, potentially due to high numbers of genes undergoing strong expression changes just before and after birth, mirrored by reversed changes later in life

[2, 3]. However, the leave-one-out null model showed that differences between the transcriptional networks from different age groups were mostly due to variance explained by inclusion or exclusion of single subjects. For example, 37 out of 38 transcriptional networks containing 37 subjects each (after removing one subject, 38-1=37 subjects) showed increased mean correlation in the group aged 12-20 years (Figure 37A). In contrast, removal of one specific subject resulted in the mean correlation dropping to the same level as all other four age groups, indicating that one subject's data can be sufficient to significantly change global network metrics for networks of this size.

Next I aimed to test whether the observed changes in global network measures were specific to gene modules within the networks or were homogeneously distributed across all genes in the network. To this end, I used a Louvain-like greedy community detection method to derive gene modules separately for each of the five age bins and for the whole available age range. Overall, global network metrics for only genes in these modules followed similar patterns to global metrics calculated for the whole network. Testing for functional enrichment, using GO-terms, further indicated that genes assigned to the modules were still relatively diverse, reflected by multiple 'themes' of functional terms being assigned to the same modules. Two replicates of networks built on 38 randomly chosen subjects from the whole age range each yielded two large modules that were not significantly enriched for any functional terms. This clearly indicates that both modules covered a wide range of genes. Potentially, this diversity in the modules covering the full age range is influenced by the way the algorithm assigns genes to modules (described in detail in **Chapter 7.5.3.2**). Briefly, the algorithm applies forced module assignment which does not allow genes not to be assigned to any module. In consequence, modules contain genes which can show comparably weak

similarity to the other genes in the modules, because they show even less similarity to the remaining modules.

The present study has several limitations, in part due to the data set used and in part due to general drawbacks of the methods I applied. In general, differences in global network metrics can be difficult to interpret. For example, differences in modularity between two or more transcriptional networks could derive from (*i*) differences in mean correlation within the modules, (*ii*) differences in partitions, for example the number of modules, or (*iii*) differences in the amount of inter-modular connections. Furthermore, validation of global network metrics in transcriptional networks is problematic. By definition, these metrics can not be tested directly by disrupting selected candidate genes. Potentially, 'marker genes' could be identified whose expression dynamics are highly representative of global network changes. However, any validation based on marker genes would obviously only be an imperfect proxy for the observed differences in global metrics, and thus would not necessarily experimentally validate the quantified differences. The most important limitation of the data set used to build the networks in the present study was the available number of subjects within each age bin. The smallest number of post-mortem brains was available for the fetal sample. Clearly, it is not possible to determine a minimum number of data points in a timeseries that are required to assess partial correlations in a reliable way. However, because global network measures are susceptible to the inclusion or exclusion of single subjects, it is likely that network stability could have been increased by using an increased number of subjects within age groups, particularly if these age groups did not encompass developmental stages which are accompanied by strong fold changes, such as the time around birth [2, 3]. In addition, studies using human post-mortem brains always face challenges concerning a lack of control over the conditions under which samples were acquired, because study

design obviously cannot influence medical decisions before death. These problems are associated with multiple confounding variables, such as a tendency for older people to die of diseases in hospitals, whereas younger brains are more often collected after accidents or homicides. Accordingly, further quantitative variables characterizing the post-mortem brains could have been included when calculating correlations as additional confounders, in order to control for their effects on the observed variance. Examples of additional phenotypes are measures reflecting lifestyle, such as nutrition, exercise or experience with legal or illegal drugs during earlier life.

7.5 Methods

7.5.1 Data

Microarray gene expression data from braincloud [2] are described in **Chapter 2.1**.

7.5.2 Network construction

Transcriptional networks were defined as correlation matrices in which each gene represented a network node and each correlation between a gene-pair a network edge [335]. Correlations between gene pairs were quantified as partial Pearson's correlations, correcting for age (*range(age[years])*=-0.5 to 78), pH (*range(pH)*=1 to 19) and RIN value (*range(RIN*)=7 to 10). Each network contained *n*=38 subjects, restricted to each of five age categories; fetal (between -0.5 and -0.38 years), 12 to 20 years, 20 to 35 years, 35 to 50 years and 50 years and older. The choice of *n*=38 subjects was determined by the maximum number of samples with RIN>7 available in the fetal group. All older age bins were randomly downsampled to match this set. Transcriptional networks derived from gene expression changes within an age group were compared to transcriptional networks based on sample size-matched random samples covering the whole age range, containing at least one subject from each age group.

All computations, including network construction and calculation of network metrics, were performed using matlab, matlab functions and the publicly available brain connectivity toolbox (BCT) [336] which provides matlab code for a wide range of network metrics.

7.5.3 Global network metrics

7.5.3.1 Network connectivity

The mean correlation of a transcriptional network – in this case represented by a correlation matrix – can be interpreted as a measure of network connectivity or cohesiveness [335]. I calculated the mean correlation across each of the transcriptional networks, as well as mean absolute correlation, resetting all correlations to positive values. Mean correlation reflects (*i*) the strength of correlations and (*ii*) the number of positive correlations compared to negative ones. Thus, mean correlations close to zero can either derive from all correlations being close to zero or alternatively from similar numbers of high positive and negative correlations. In turn mean absolute correlation provides an overall measure of connective strength in the network.

7.5.3.2 Network segregation and community detection

Communities within the transcriptional network were detected and network segregation, defined as modularity (Q), was quantified using a Louvain-like greedy community detection method. The algorithm determines optimal partitions of the network into communities, with stronger connectivity within each of the communities than across all the communities [335]. More precisely, the algorithm detects communities by identifying nonoverlapping groups of nodes by maximizing the number of within-group edges and subsequentially minimizing the number of edges between groups [336]. The Louvain-like greedy community detection method derives communities iteratively in two phases: First, it moves each node at a time until quality is not further improved. Next, the identified communities are aggregated to build a new network in which each node is assigned to a community [338]. Forced assignment as applied here does not permit nodes, in this case genes, to not belong to any module. The global network

metric “modularity (*Q*)” quantifies to what degree the network could be subdivided into groups [336]. As suggested previously [335], modularity (*Q*) is reported as the average across multiple runs for each network (*n*=100 repetitions). For consistency, network communities are referred to as ‘modules’ throughout the text.

7.5.3.3 Entropy

Statistical randomness can be quantified using network entropy. I calculated entropy in two alternative ways:

1) Entropy [337] defined as

$$-\sum(p \cdot \log_2(p))$$

where p contains all histogram counts for quantitative values in the input matrix [337].

2) Shannon entropy [339]

Entropy (*E*) is defined as a cost function fulfilling

$$E(s) = \sum_i E(\mathrm{s}_i)$$

and $E(0) = 0$

with *s* being the signal. Indices *i* can be interpreted as “position” or “frequency” of the function for *s* [339]. The Shannon entropy, without normalization, is then defined as

$$E1(s_i) = s_i^2 \log(s_i^2)$$

$$E1(s) = -\sum_i s_i^2 \log(\mathrm{s}_i^2)$$

with convention $\log(0) = 0$ [339].

7.5.3.4 Network resilience

Network resilience reflects the network's resilience to insults, such as the removal of a node. One indirect measure – as opposed to direct measures quantifying changes in the network before and after an insult – is the assortativity coefficient [340]. The assortativity coefficient is the correlation coefficient between degrees of nodes on both ends of network edges. Degree of a node refers to the number of edges assigned to the given node. Positive assortativity coefficients reflect nodes with similar degrees linking to one another [340], which can be interpreted as mutually interconnected high-degree hubs, forming a comparably resilient core. In contrast, negative assortativity coefficients have distributed and more vulnerable high-degree nodes [336]. Here, I used the assortativity coefficient for a binary, unweighted and undirected network. Network edges were either 0 or 1, with 1 indicating a significant partial Pearson's correlation between two genes at $\alpha=0.05$, uncorrected for multiple tests.

7.5.3.5 Path length

Network paths are defined as any sequence of nodes where every consecutive pair of nodes is connected by an edge [334, 336]. The length of a path is the number of edges it includes, not the number of nodes. Generally, nodes can be visited more than once by the same path [334]. Here, I calculated the characteristic path length of a network, defined as the average shortest path length across the shortest paths between all pairs of nodes. The characteristic path length can be interpreted as the potential for functional integration between parts of the network, whereby short path lengths in a network lead to higher integration [336].

7.5.4 Null models

I compared the global network metrics for transcriptional networks within the age bins and the lifespan, as described above, to the same metrics for different null models.

(i) Random permutation

The fully randomized null model was derived by randomly permuting all matrix elements $a_{i,j}$ in the gene expression matrix *A[i-j]* with *i* genes (in matrix rows) from *j* samples (in matrix columns). Clearly, this procedure destroys all network structure and consequentially leads to the most 'liberal' p-values.

(ii) Nodal null model

The nodal null model was calculated by permuting expression levels of all genes within one sample from one subject. More precisely, in the expression matrix containing genes as rows and timepoints as columns, matrix elements were permuted across all rows while preserving the entries per column.

(iii) Temporal null model

The temporal null model consisted of the expression matrix after random permutation of gene expression values across samples. Thus, in the expression matrix containing genes as rows and samples as columns, matrix elements were permuted across all columns while preserving entries per row.

(iv) Leave-one-out (LOO) null model

The leave-one-out (LOO) null model recalculated all global network measures after leaving each subject out exactly once. Therefore, for each age-bin that initially contained 38 subjects, measures were re-calculated for 38 transcriptional networks each

containing 37=38-1 subjects, wherein a different subject was removed each time. The LOO null model tests for the impact of single subjects on the given network metric. The null hypothesis is that single subjects have only a minor influence on the global network metric per transcriptional network, compared to the difference between the global network metrics across for example different age categories. Alternatively, the differences in global network metrics across the age categories could be smaller than the variance within an age category deriving from inclusion or exclusion of one subject. The latter case indicates that there is no significant difference in the given global network measure between the age groups.

7.5.5 Functional enrichment

GO-term enrichment for all genes within a module was tested against a background of all brain-expressed probes with annotated DAVID IDs (n=16475 out of 18846 probes) that were used originally to derive the transcriptional network, using the functional classification tool Database for Annotation, Visualization, and Integrated Discovery (DAVID [120], described in detail in **Chapter 2.4**) with default parameters (count = 2 and ease = 0.1).

Chapter 8 – Conclusions and Outlook

In this thesis I have described analyses of transcriptional networks based on gene expression data from human post-mortem brains. I used publicly available microarray and RNA-Sequencing data to derive gene expression networks across time and space, covering all developmental stages in the human brain from fetal to old age, and across different adult tissues. Transcriptional modules were detected based on various criteria including (*i*) their relation to external variables, such as age, or (*ii*) genomic regions, such as brain-specifically expressed extended alternative 3' untranslated regions. I found that the transcriptional modules recapitulate cell types, for example microglia, indicating that coordination of gene expression preserves information about tissue composition and its development across the lifespan. Furthermore, transcriptional networks initially identified based on genomic features were enriched for functional categories, such as GO-terms, and showed increased co-expression during adult age, providing further evidence for biologically relevant information being preserved in the transcriptional networks. I identified key regulatory elements for the brain's transcriptional network, including transcription factors and noncoding RNA elements such as miRNAs. The computational predictions were validated using independent publicly available experimental data sets, including ChIP-Sequencing data for transcription factors suggested as key regulators for distinct transcriptional modules. In summary, the work presented in my thesis shows that transcriptional networks preserve information about underlying functionally related gene groups, cell types and therefore tissue composition and its dynamics during ageing. Importantly, transcriptional networks can therefore be used to identify crucial genetic elements that fine-tune their co-expression in specific circumstances – depending on localization in the body or the developmental stage. Here, the key regulators were detected directly from human post-

mortem brains without the necessity to use a proxy, such as model organisms or cell lines. In this way, the transcriptional networks could be studied in their true cellular context, which is especially important for genes encoding proteins involved in cell-cell communication, including surface receptors or ion channels. Furthermore, although post-mortem tissue studies may suffer from some general limitations, such as confounding effects of death and circumstances of dying, in addition to a lack of control over the previous living conditions, these studies allow questions to be addressed that directly interrogate genetic interactions at distinct developmental stages. For example, as described in **Chapter 3**, post-mortem brains were used to analyse what role transcription factors, known to regulate early microglia development, play in microglia homeostasis of the human adult brain. Ideally, these questions would be addressed using consented tissue donations from living donors after surgical procedures. However, firstly even in this case the procedure itself and the given medical condition could be confounding variables, and secondly, no sufficiently large data set derived from brain surgery currently exists.

The studies presented in this thesis suggest future directions for investigating genetic networks in the brain, their regulatory elements and their interdependence with the cellular context.

The gene sets linked to distinct biological functions, such as microglia, the "immune system" of the brain [141, 142], and their key regulators identified here, provide candidate genes for experimental follow-up. Currently, methods for perturbing whole gene subnetworks as opposed to isolated genes are not well-established; however, the disruption of specific regulatory elements, and limited numbers of direct interaction partners in model systems using CRISPR/Cas9 technology [341] for example, could be used to (*i*) validate the causal effect of identified regulators, and to (*ii*) elucidate the

underlying molecular mechanisms of interacting genes and their downstream effects on the transcriptional networks. An in-depth analysis of interacting factors could also allow translation of candidate regulatory elements into clinical applications. One example is the recently developed method Combinatorial Genetics En Masse (CombiGEM), that allows the generation of high-order barcoded combinations of genetic elements and subsequently their characterization using next-generation sequencing [342]. CombiGEM has been successfully applied to identify genetic combinations that facilitate killing of highly antibiotic-resistant bacteria [342].

The rapid development of sequencing technologies provides new opportunities for studying an organism's transcriptome, ranging from increasing sequencing depth and speed at lower costs for massively parallel sequencing [343, 344] to single-cell sequencing [345, 346]. Potentially, single-cell sequencing will allow the study of specific cell- and tissue types with unprecedented accuracy, including the quantification of heterogeneity within a given cell type [347] across development, and could even help to identify new cell types and subclasses. Importantly, single-cell sequencing may allow possibilities for analysing the details of cellular context, which is a key question for understanding developmental changes and for various applications in synthetic biology [348]. More specifically, previous studies that constructed genetic circuits that were designed to be functional in living cells suffered from simplified cell parameters, due to lack of more comprehensive models of cellular composition [348], including the relative abundance of cell types and their interactions, such as the modulation of glial gene expression by the presence of neurons [97]. The age-dependence for microglia surface receptor expression observed here and for single genes in previous studies [164, 165], highlights the role cell-cell communication plays for biological systems. To investigate these interacting systems, the cells could be categorized based on their

cellular microenvironment and sequencing could be applied to subsequently quantify the transcriptional impact of these microenvironmental factors.

Recent advances in optogenetics, which allow direct manipulation of gene expression by the integration of genes whose expression can be switched on or off by the application of light, combined with sequencing, could lead to an inference of causality in genetic circuits or networks [349, 350]. Optogenetics directly manipulates genetic targets and their downstream interaction partners in model systems, whereas reverse engineering of transcriptional networks based on gene expression relies on some similarity measure, for example correlation or mutual information. The combination of RNA-Seq, potentially down to single cell resolution, with optogenetics could be used to validate the causal effects of genes and regulatory elements in a defined microenvironment. Interestingly, both technologies can be applied in living animals [351], leading to possible applications for studying development and even manipulating behaviour.

Improved sequencing technologies further allow the detailed analysis of noise in a system [90, 352, 353]. Especially in the brain this would be of great interest, because neuronal activation is transmitted between cells based on activation thresholds, depending on the differential potentials across the membrane [354]. Thus, cells exist in states which either allow or prevent activation, by reacting to action potentials, based on previous activation [354]. This kind of regulation of responsiveness could be studied in further detail based on the ‘resting state’ of gene expression oscillations (or their perturbation) using transcriptome sequencing of single cells in specific contexts. Future studies could ask whether temporal gene expression patterns play a role in cell responsiveness, how ‘resting state’ gene expression is maintained or passed on between cells, and whether cells show a memory effect for previous activation. Furthermore,

noise and resting state oscillations of gene expression could also play a role in stochastic resonance. In this phenomenon, a signal is amplified by adding white noise of various frequencies, whereby only the noise frequencies corresponding to the original signal will resonate, and thereby 'boost' the original signal [355].

Summing up, the studies presented in this thesis have provided evidence for regulatory interactions of protein-coding and noncoding genes being preserved in transcriptional networks, as well as the specificity of cell types and the importance of cellular context. Thus, future studies, applying sequencing tools with higher resolution down to single cells, and methods which allow the direct manipulation of genetic circuits, could aim to validate causal effects and quantify the interactions between genes and their microenvironment at different developmental stages.

Table of Figures

Figure 1. The genetic code is transcribed into pre-mRNA, consisting in 3'end and 5'end untranslated regions (UTRs), the protein coding regions consisting of expressed regions (exons, colored cylinders) and intragenic regions (introns, black lines). In the mature mRNA, introns are spliced out, and exclusion of exonic regions can lead to alternative isoforms (demarcated by orange and purple arrows). (In addition to the exclusion of exonic regions, alternative isoforms can be generated by the use of alternative exon boundaries, which is not displayed here). The terminal untranslated regions can regulate translation of mRNA into proteins. Here, only one isoform (orange arrows) includes the alternatively spliced region (red), leading to alternative protein products; Figure adapted from [17] (Copyright Romero 2006) 11

Figure 2. Comparison of alignments of reads to genome followed by transcript assembly (left) or transcript assembly followed by alignment of transcripts to genome (right). Reference [42] suggests that align-then-assemble (left) approaches are more sensitive because they can detect less abundant isoforms in contrast to assemble-then-align (right) methods. Figure reproduced from [42] with permission from the Nature Publishing Group, www.nature.com. 15

Figure 3. Transcription factors (TFs) target specific DNA sequences, called transcription factor binding sites (TFBS), which can be located proximal or distal to a transcription start site (TSS). Multiple TFs may form functional *cis*-regulatory modules (CRMs). The transcription initiation complex is further stabilized by interactions between bound TFs and cofactors. Figure reproduced from [47] with permission from the Nature Publishing Group, www.nature.com 17

Figure 4. Illustration of gene expression patterns observed across the human lifespan using four individual genes. The gene expression level is shown on the y-axis and age on the x-axis. The four genes illustrate strong expression changes before and around the age of birth, followed by more stable gene expression during adult age. Figure reproduced from [2] with permission from the Nature Publishing Group, www.nature.com.....22

Figure 5. Gene expression patterns in neocortex reflect cortical organization. The first three principal components (a-c) were derived from 1000 genes with the highest variance across 57 cortical regions. All three components differentiate between distinct cortical divisions: (a) primary sensory and motor areas, (b) subdivisions in frontal, temporal and occipital lobe, and (c) caudal portion of the frontal lobe. Transcriptional profiles are mapped back onto the three dimensional brain topography (d-e). Gyri are labelled in a MRI representation of the brain (d). Multi-dimensional scaling allows mapping of gene expression profiles back into the brain space, with distance reflecting the similarity in gene expression. Gene expression clusters reflect lobe and gyral

positions as indicated by the arrows (e). Figure reproduced from [73] with permission from the Nature Publishing Group, www.nature.com 25

Figure 6. The DNA fragments enriched by ChIP are sequenced from 5'ends and aligned to plus and minus strands of the genome. The resulting two peaks, one on each strand, flank the binding region of the protein of interest. The distribution of the 'true' peak is approximated by extending reads on each strand in the appropriate direction. Figure reproduced from [103] with permission from the Nature Publishing Group, www.nature.com 36

Figure 7. Flowchart for RNA-Sequencing (RNA-Seq) reads' alignment and assembly using Cufflinks [114]. A) Cufflinks' input consists of aligned cDNA fragment sequences. The software TopHat performs the required spliced alignments. B-C) Cufflinks defines paired-end RNA-Seq reads as single alignments and assembles the overlapping fragment alignments separately. D-E) The next steps estimate the abundance of the assembled transcripts. B) First, 'incompatible' fragments are identified which originate from differentially spliced mRNA isoforms. The 'compatible' fragments, with alignments overlapping in the genome, on the other hand are connected in an 'overlap graph'. In the example shown here, yellow, blue, and red fragments must have originated from different isoforms, whereas all remaining fragments could belong to any of the three shown paths. C) The isoforms are assembled from the overlap graph. Each path (three in the example), corresponds to fragments that could be merged into the same isoform. Cufflinks works based on Dilworth's Theorem, which states that the number of incompatible reads equals the minimum number of transcripts needed to 'explain' the fragments. More precisely, Dilworth's Theorem is used to produce a minimal set of paths that cover the fragments in the overlap graph, by identifying the maximum sets of reads that could not come from the same isoform. D) The transcript abundance is calculated after matching the fragments to their 'original' isoforms (illustrated here by colours). For example, the violet fragment could have come from the blue or the red isoforms, whereas the grey fragments could have come from any of the three isoforms. The probability of observing a fragment is modelled with a linear function based on the abundance of the transcript from which it could have originated. Cufflinks incorporates the distribution of the fragment lengths for the assignment of fragments to the isoforms. This is important as the fragment assignment often implies different lengths, although only the ends of fragments are sequenced and therefore the lengths maybe unknown. This is illustrated here by the violet fragment, which could be longer if it came from the red instead of the blue isoform. E) The pie chart shows the isoform abundances that best explain the observed fragments, based on the numerical maximization of a likelihood function for all possible sets of the relative abundances of the red, yellow, and blue isoforms (γ_1,γ_2,γ_3) [114]. Figure reproduced from [114] with permission from the Nature Publishing Group, www.nature.com 40

Figure 8. Venn diagram showing the overlap of genes in the three microarray gene expression data sets and in the microglia module. A) Overlap of genes from the three microarray gene expression data sets braincloud, Medical Research Council (MRC/UK)

and Harvard Brain Tissue Resource Center (HBTRC). B) Demographics and coverage of the three data sets. C) Venn diagram showing the microglia modules and their gene overlaps..... 64

Figure 9. Microglia modules' ME's [gene relative expression: $\log_2$(sample/reference)] negative correlation with age. A) Pearson's correlation of all *n*=162 genes in the microglia module with age, derived from braincloud data ranked by increasing correlations. The red dashed line indicates the Pearson's correlation for the microglia module's ME with age (Pearson's correlation=-0.35). B) showing a negative slope of a linear model fit with the microglia module's ME's expression as the dependent variable and age [years] as the independent variable. C and D display the same metrics for the network derived from MRC/UK data (*n*=141 genes in the microglia module); E and F correspondingly display the same metrics for the network derived from HBTRC data (*n*=326 genes in the microglia module)..... 65

Figure 10. Scatterplot of Pearson's correlation coefficient (y-axis) for the correlation between module eigengene (ME) and age [years] at constant RIN values (x-axis)..... 67

Figure 11. Microglia surface receptors' and microglia M1 and M1/M2b phenotype markers' expression correlation with age for MRC/UK data. The regression model fit and confidence interval for normalized gene expression versus age is for A) microglia module's ME (shown in grey), B) surface receptors for neuron-microglia crosstalk (blue), C) Toll-like receptors (red), and D) M1 and M1/M2b activational phenotype markers (green). The x-axis shows age and the y-axis shows normalized gene expression of the microglia module's ME. Legends in B-D indicate all respective marker genes that are expressed in the MRC/UK data set and therefore were included in the regression. Pearson's correlation coefficient of the mean expression of the expressed marker genes (B-D) or module ME versus age (A) is shown in the bottom-left corner of each plot..... 69

Figure 12. Microglia surface receptors' and microglia M1 and M1/M2b phenotype markers' expression correlation with age for (A) braincloud data and (B) HBTRC brain bank data. The regression model fit and confidence interval for average gene expression versus age are for the microglia module's ME (grey), surface receptors for neuron-microglia crosstalk (blue), Toll-like receptors (red) or M1 and M1/M2b activational phenotype markers (green). The x-axis shows age and the y-axis shows normalized gene expression of the microglia module's ME..... 70

Figure 13. Overlap of master regulator's (*RUNX1*, *PU.1* and *IRF8*) predicted target genes in the microglia module (A) compared to genes from whole data set (MRC/UK) (B). A) Predicted regulons (in other word target genes) of *RUNX1, PU.1,* and *IRF8* in the MRC/UK gene expression data for predicted regulons which are part of the microglia module. Size and colour of nodes reflect indegree of nodes. More specifically, small and green nodes have a small number of incoming edges, and, by contrast, large

and red nodes have a high number of incoming edges. Yellow nodes are intermediate. Here, the possible range of indegree goes from zero (a master regulator which is not regulated by another master regulator and therefore has no incoming edges) to two (an incoming edge from both other master regulators.) For simplicity, only edges that connect at least one of the three master regulators (*RUNX1*, *PU.1* and *IRF8*) with another gene are displayed, excluding edges between target genes. B) Same measures including all predicted target genes. It is of note that *n*=14 genes that are regulated by all three master regulators are present in both A and B..... 75

Figure 14. Protein-protein interactions are recapitulated in a master regulator (MR) network. A protein-protein network, derived from direct and indirect interactions among proteins encoded by MRs' genes. Colours are DAPPLE p-values (see legend), calculated from 1000 within-degree node-label permutations. The network was constructed on the basis of physical protein-protein interactions that have been reported in the literature [169]. 77

Figure 15. ChIP-Seq peaks for SPI1 (PU.1), RUNX1, IRF8 and TAL1 in genes in the combined microglia module. A) Fold enrichment for SPI1 (PU.1) ChIP-Seq peaks in all annotated basal regulatory domains of all genes in the combined microglia module (*n*=406 assigned to the microglia module in all three data sets combined). Stars demarcate significant (*p-value*<0.05) enrichment. Legend indicates the cell lines that were used for each master regulator. B) Illustration of master regulator bindings sites overlapping the basal regulatory domain of *AIF1* taken from UCSC genome browser. 80

Figure 16. Scatterplot of module eigengenes (ME, the first principal component) of each of the two modules enriched for GO-terms linked to "cell cycle". The y-axis indicates relative gene expression and the x-axis indicates fetal age. Age of the fetal brains is given in years, relative to date of birth. The total number of genes in each module is indicated in the plot, more precisely *n*=2701 genes in one module (A) and *n*=966 genes in the other module (B)..... 113

Figure 17. Enrichment of marker genes for cell cycle stages in two modules enriched for GO-terms linked to "cell cycle". Hereafter, the first module is referred to as the "Interphases module" (left) and the other module is referred to as the "G2/M module" (right). Cell cycle stages referred to using colours are indicated in the legend, including the total numbers of marker genes available for each stage [217]. 115

Figure 18. Barplot of Pearson's correlation (Pearson's correlation coefficient [-1...1] on the y-axis) of each cell cycle stage marker gene (x-axis) expressed in Interphases or G2/M module with age. Each bar is the Pearson's correlation of a given expressed marker gene with age, over the age bin indicated on the y-axis. Cell cycle stages referred to using colours are indicated in the legend, including total numbers of available marker genes [217]..... 117

Figure 19. Barplot of Pearson's correlation (Pearson's correlation coefficient [-1...1] on the y-axis) with age of each cell cycle stage marker gene (x-axis) expressed in the complete microarray gene expression data set used to build the transcriptional network. Each bar is the Pearson's correlation of a given expressed marker gene with age over the age bin indicated on the y-axis. Cell cycle stages referred to using colours are indicated in the legend [217]. 118

Figure 20. Venn diagrams indicating the overlap of the Interphases module's genes with predicted *SOX2* target genes during fetal or adult life (A), and G2/M module's genes with *FOXM1*-predicted target genes during fetal life compared to *FOXO1*-predicted target genes during adult life (B). 125

Figure 21. Predicted target genes of five TFs identified as "neurogenesis TFs" for cell cycle modules during fetal life (A) compared to adult life (B). Nodes' color reflects outdegree, with large and red nodes having the largest number of direct edges in contrast to small and green nodes which have the smallest possible number (one edge per node). Yellow nodes are the intermediates between red and green. For simplicity, only direct predicted edges between the five TFs and their target genes are displayed. 130

Figure 22. Brain-expressed isoforms have the longest 3'UTR extensions. A) Median length (in nucleotides, y-axis) of alternative 3'UTRs across 16 human tissues. The median across all non-neuronal tissues (219 nucleotides) is depicted as a horizontal dashed red line. B) Median fold difference in expression (calculated as the number of reads per nucleotide of expressed sequence) of alternative relative to constitutive 3'UTR (y-axis) across 16 tissues. The median across all non-neuronal tissues (median relative expression=0.28) is depicted as a horizontal dashed red line. C) G protein-coupled receptor 180 (*GPR180*; ENSG00000152749; chr13:95,277,556-95,288,768, red) expresses its longest 3'UTR in the brain. Alongside support for the expression of the only transcript annotated by ENSEMBL for this gene (ENST00000376958) our approach allowed the annotation of an isoform with an extended 3'UTR in the brain (CUFF.37893.2). Read density (black, y-axis) across the last coding exons (black), the constitutive (blue) and extended (orange) portions of *GPR180* 3'UTR in brain, ovary, adrenal and testes. 161

Figure 23. Ion channel/transporter genes are over-represented within genes expressing their longest 3'UTR in the brain A) Enrichment map [287] of significantly enriched GO-terms after GO-terms were summarized using REViGO [116]. Node size (blue circles) reflects the number of contributing genes, and edge size (grey lines) corresponds to the number of genes that are shared between nodes. B) Distribution of the median pairwise Spearman's correlations for 1000 bootstrap samples of random gene sets (each n=192, matched to *ion channel transporter genes*) from the populations: (i) brain-expressed genes (n=16812 probes; red), (ii) genes with longest alternative 3'UTRs expressed in brain (n=1261 probes; blue), (iii) genes with longest 3'UTRs expressed only in brain (n=449 probes; purple) or (iv) genes contributing to GO-term

"ion channel activity" (*n*=192 probes; green). Arrow indicates the median pairwise Spearman's correlations for the *ion channel/transporter genes*..... 165

Figure 24. Frontal cortex-specifically expressed miRNAs targeting 3'UTR extensions strengthen the cross-talk between *ion channel/transporter genes* A) Differences in the percentage of predicted MREs for miRNA families expressed in human adult frontal cortex (y-axis for different levels of relative expression: FC10, FC5, FC2 and expressed) between constitutive (red) and alternative (blue) 3'UTRs of *ion channel/transporter genes*. * indicates a significant difference ($p < 0.05$). B) Calculation of the percentage (%) of shared MREs, *s*, between two *ion channel/transporter* genes. Gene X and Gene Y have two and one predicted MREs for a shared frontal cortex-specific miRNA (miR-9; red), respectively. Gene X and Y have predicted MREs for miRNAs that are either not expressed specifically in the frontal cortex (Gene X, miR-144) or not shared (Gene Y, miR-181). Of the 5 shared predicted MREs in both genes 3 are for frontal cortex-specific miRNAs (miR-9; red), resulting in s=3/5x100 =60% shared frontal cortex-specific MREs C) Percentage shared MREs for FC10 in *ion channel/transporter genes* (median percentage shared MREs *s*=47%, black arrow). Histogram represents the null distribution of the same metric for 1000 random brain-expressed gene sets excluding *ion channel/transporter genes* (each *n*=226)..... 168

Figure 25.*GABRA4*-subnetwork. Example of gene in ion channel/transporter network. Correlation matrix of *ion channel/transporter genes* significantly correlated with *GABRA4* (α=0.05, Bonferroni corrected, pairwise Spearman's correlation). Bar chart shows percentage of MREs *GABRA4* shares with every other gene in *GABRA4*-subnetwork. 170

Figure 26. Annotation of constitutive and alternative 3'UTRs. A) Flowchart for annotation of protein-coding genes with longest alternative 3'UTRs; *n* indicates number of genes. B) Example of a gene with constitutive stop codon (red), expressed only in tissue A (blue) and tissue B (brown). Tissue A expresses three transcripts, and tissue B expresses 2 transcripts (top panel). The constitutive 3'UTRs consists in the nucleotides expressed in all transcript of a gene across all tissues where the gene is expressed (middle panel). The extended alternative 3'UTRs is the consensus transcript for a gene in a given tissue after constitutive parts have been removed. Tissue A expresses the longest alternative 3'UTRs for this gene, as it contains more nucleotides than the alternative 3'UTRs in tissue B (bottom panel)..... 177

Figure 27. The potential mechanism of ceRNA activity. A) mRNAs could regulate each other's transcript abundance by a reciprocal relationship between mRNAs and miRNAs. B) This reciprocal relationship between mRNAs and miRNAs could be 'communicated' by miRNA response elements (MREs) for miRNAs shared between two target genes. Thus, the higher the number of shared interacting miRNAs, the stronger the cross-talk between the transcripts could be. C) The 3'UTRs of mRNAs contain a high density of MREs. MiRNAs can bind to these MREs and act as cis regulatory elements to regulate the stability and the abundance of the 3'UTRs' own transcript. Furthermore, ceRNAs

activity could allow regulation of transcript levels in trans by altering the availability of miRNAs. Figure reproduced from [27] with permission from Elsevier, www.cell.com. 193

Figure 28. Expression of all genes in the three modules correlated with brain region ($\alpha<0.05$, uncorrected) and no other external variable, separately for all available brain regions. The legend indicates brain regions with the abbreviations: ACC (anterior cingulate cortex), BG (basal ganglia), FC (frontal cortex), NA, BG (nucleus accumbens, basal ganglia), cerebellar (cerebellar hemisphere), and SN (substantia nigra). Gene expression is given in FPKM, normalized across all libraries (see Chapter 6.6.1). 200

Figure 29. Flowchart for detection of gene modules, annotation of 3'UTRs, and association of lncRNAs with modules' eigengenes (MEs) using correlation and highest median expression values (see Chapter 6.6). The bottom box shows an illustration of how the 'percentage shared MREs' for 'lncRNA a' and the 3'UTRs of the mRNA of 'gene b' were calculated. In this example, a total of 7 MREs were predicted bind to 3 different miRNAs (referred to as x, y, and z). Only 'miRNA x' targets both 'lncRNA a' and the 3'UTRs of mRNA of 'gene b'. The percentage shared MREs therefore is calculated as the ratio of 3 predicted MREs for the shared miRNA x, divided by all 7 predicted MREs for both genes. 202

Figure 30. Expression of all lncRNAs whose expression correlated with the module eigengenes (MEs) of the three modules that were correlated with brain regions ($\alpha<0.05$, corrected for all lncRNAs tested) and no other external variable, separately for all available brain regions. The legend indicates brain regions with the abbreviations: ACC (anterior cingulate cortex), BG (basal ganglia), FC (frontal cortex), NA, BG (nucleus accumbens, basal ganglia), cerebellar (cerebellar hemisphere), and SN (substantia nigra). Gene expression is given in FPKM, normalized across all libraries (see Chapter 6.6.1). 204

Figure 31. Genome-wide length distribution of alternative 3'UTRs shows significantly longer extensions in the cerebellum and significantly shorter extensions in the ACC, but no significant differences among the alternative 3'UTRs of the mRNAs in the hormone module. The barplots show median lengths of expressed alternative 3'UTRs in number of nucleotides merged across all samples for a brain region (indicated on the x-axis of heatmap). The barplot on the top includes all genes with alternative 3'UTRs, and the barplot below includes only the genes in the hormone module with alternative 3'UTRs. Asterisks indicate the only two regions (cerebellum and ACC) which differed significantly from all other brain regions using 2-sample Kolmogorov-Smirnov (KS) tests ($\alpha<0.05$, corrected for all tests performed). The heatmap shows all p-values for the 2-sample KS tests between all brain regions. The 3'UTRs were annotated as described in Chapter 6.6.2 and illustrated in Figure 28. The x-axis indicates brain regions with the following abbreviations: ACC (anterior cingulate cortex), BG (basal ganglia), cerebellarH (cerebellar hemisphere), FC (frontal cortex), NABG (nucleus accumbens, basal ganglia), and SN (substantia nigra). 206

Figure 32. Median percentage shared MREs between the 33 antisense lncRNAs associated with the hormone module and the 3'UTRs of 58 genes in the hormone module, which had at least one predicted MRE, compared to an empirical null distribution. The median percentage shared MREs for the genes in the hormone module and the associated lncRNAs is indicated by a black arrow. The histogram shows the same measure for 1000 random samples of 33 lncRNAs that were not associated with the hormone module. 212

Figure 33. Spearman's correlation between the module membership (MM; defined as the correlation of a gene's expression and the module eigengene (ME)) of the 33 associated antisense lncRNAs and the median percentage shared MREs between the lncRNA and the 3'UTRs of the mRNAs in the hormone module compared to an empirical null distribution. The correlation between a lncRNA's MM and the median percentage shared MREs between the lncRNA and the hormone module genes is indicated by a black arrow. The histogram shows the same measure for 1000 random samples of 33 lncRNAs that were not associated with the hormone module..... 215

Figure 34. Scatterplot for the 33 hormone module-associated antisense lncRNAs showing each lncRNA's module membership (MM) versus the median percentage of shared MREs between the lncRNA and all genes in the hormone module. The black demarcates the fit of the linear regression model (Spearman's correlation=0.37)..... 216

Figure 35. A) MREs shared between the 3'UTRs of the mRNAs of *PGR* (protein-coding gene assigned to the hormone module) and the associated antisense lncRNA ENSG00000225764. The barplot shows the number of predicted MREs for each miRNA that targets one or both of these genes (the names of the miRNAs are shown on the x-axis). Both genes combined contained 238 predicted MREs, including 203 MREs (85%) for miRNAs that are predicted to target both genes. MiRNA-488 that showed increased density of MREs for the hormone module and the associated lncRNAs in three independent annotations is highlighted in red. B) Boxplot shows expression of *PGR* across the 13 brain regions. C) Boxplot shows expression of ENSG00000225764 across the 13 brain regions. The x-axes of both boxplots indicate brain regions with the following abbreviations: ACC (anterior cingulate cortex), BG (basal ganglia), cerebellarH (cerebellar hemisphere), FC (frontal cortex), NABG (nucleus accumbens, basal ganglia), and SN (substantia nigra). 224

Figure 36. Number of multi-exonic lncRNAs expressed in all samples, or 75%, 50%, or 25% of them, or in at least one sample, separately for five lncRNA annotation sets. . 233
Figure 36. A-G) Seven global network metrics (y-axis) for all age groups (networks 1-5 on x-axis, each network was built from data from *n*=38 subjects). Age ranges for networks 1-5 are shown in H). Each red dot shows the mean value of the given global network metric in the respective age group. Blue and purple dots indicate the same global metrics derived after random permutations (*i*-*iii*) of gene expression data as

described in the main text. Boxplots on the right: Each plot shows the distribution of the same global network metrics for a network of *n*=38 subjects that were randomly sampled across the full age range (*n*=10 sets of random samples). Details of network metrics are given in Chapter 7.5.3. 262

Figure 37. A-G) Seven global network metrics (y-axis) for all age groups (networks 1-5 on x-axis, each network was built from data from *n*=38 subjects). Age ranges for networks 1-5 are shown in H). Each red dot shows the mean value of the given global network metric in the respective age group. Blue and purple dots indicate the same global metrics derived after random permutations (*i-iii*) of gene expression data as described in the main text. Boxplots on the right: Each plot shows the distribution of the same global network metrics for a network of *n*=38 subjects that were randomly sampled across the full age range (*n*=10 sets of random samples). Details of network metrics are given in Chapter 7.5.3.

Figure 38. Leave-one-out null model for global network metrics. Each boxplot shows global network metrics for 38 samples of 37 subjects each, after each subject had been removed from the sample once (38-1=37). Age ranges of networks 1-5 are displayed in Figure 37H. 264

Figure 39. Transcriptional network modules within five age bins and across the whole age range (referred to as “R1” and “R2”). A) Size of network modules, in terms of numbers of genes per module. Numbers of detected modules per age bin and across age ranges [years] are listed below the x-axis. B) Partial Pearson’s correlations within network modules in age bins or across the whole age range. C) Absolute partial Pearson’s correlations within network modules in age bins or across the whole age range. 266

Table of Tables

Table 1. Functional enrichment for GO-terms for genes contributing to the age-dependent module in all three data sets combined (*n*=426 genes), calculated using a hypergeometric test in GREAT (described in Chapter 2.4.5 [120]). P-values were corrected for multiple testing by applying false discovery rate (FDR Q-Val, see Chapter 2.3.1). 'Fold' indicates fold enrichment. 60

Table 2. Significant GO-term enrichment in the three modules associated with brain regions. Each module was assigned a name, indicated in the header lines. A yellow background highlights the non-redundant terms, summarized using REViGO [116]. P-values are Benjamini-Hochberg corrected ($\alpha<0.05$); 'fold' refers to fold enrichment of each GO-term. 197

Table 3 33 antisense lncRNAs associated with the hormone module. 'Median percentage shared MREs' and module membership (MM) are explained in detail in the text. 218

Table 4. GO-terms that were enriched ($\alpha=0.05$, Bonferroni corrected) in modules detected within age bins. Categories highlighted in yellow are non-redundant GO-terms for the given modules, summarized by REViGO [116] (see Chapter 2.4.1). "fold" indicates fold enrichment of this GO-term. Modules which did not show significant enrichment after p-value correction are not displayed. The numbers of detected modules (including > 30 genes) are indicated in brackets. 268

References

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F *et al*: **Landscape of transcription in human cells**. *Nature* 2012, **489**(7414):101-108.
2. Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, Colantuoni EA, Elkahloun AG, Herman MM, Weinberger DR *et al*: **Temporal dynamics and genetic control of transcription in the human prefrontal cortex**. *Nature* 2011, **478**(7370):519-523.
3. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G *et al*: **Spatio-temporal transcriptome of the human brain**. *Nature* 2011, **478**(7370):483-489.
4. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing**. *Science* 2008, **320**(5881):1344-1349.
5. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S *et al*: **Global identification of human transcribed sequences with genome tiling arrays**. *Science* 2004, **306**(5705):2242-2246.
6. Hangauer MJ, Vaughn IW, McManus MT: **Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs**. *Plos Genet* 2013, **9**(6).
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
8. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R *et al*: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.
9. Bartel DP: **MicroRNAs: Target Recognition and Regulatory Functions**. *Cell* 2009, **136**(2):215-233.
10. Friedman RC, Farh KK, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome research* 2009, **19**(1):92-105.
11. Siomi MC, Sato K, Pezic D, Aravin AA: **PIWI-interacting small RNAs: the vanguard of genome defence**. *Nature reviews Molecular cell biology* 2011, **12**(4):246-258.
12. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ: **Developmentally regulated piRNA clusters implicate MILI in transposon control**. *Science* 2007, **316**(5825):744-747.
13. Ponting CP, Oliver PL, Reik W: **Evolution and functions of long noncoding RNAs**. *Cell* 2009, **136**(4):629-641.
14. Mercer TR, Dinger ME, Mattick JS: **Long non-coding RNAs: insights into functions**. *Nature reviews Genetics* 2009, **10**(3):155-159.
15. Lajoie MJ, Rovner AJ, Goodman DB, Aerni HR, Haimovich AD, Kuznetsov G, Mercer JA, Wang HH, Carr PA, Mosberg JA *et al*: **Genomically recoded organisms expand biological functions**. *Science* 2013, **342**(6156):357-360.
16. Neumann H, Wang K, Davis L, Garcia-Alai M, Chin JW: **Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome**. *Nature* 2010, **464**(7287):441-444.
17. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z *et al*: **Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(22):8390-8395.
18. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.

19. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D *et al*: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome**. *Science* 2008, **321**(5891):956-960.
20. Maier T, Guell M, Serrano L: **Correlation of mRNA and protein in complex biological samples**. *FEBS letters* 2009, **583**(24):3966-3973.
21. Vogel C, Marcotte EM: **Insights into the regulation of protein abundance from proteomic and transcriptomic analyses**. *Nature reviews Genetics* 2012, **13**(4):227-232.
22. Wu L, Belasco JG: **Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs**. *Molecular cell* 2008, **29**(1):1-7.
23. Eulalio A, Huntzinger E, Izaurralde E: **Getting to the root of miRNA-mediated gene silencing**. *Cell* 2008, **132**(1):9-14.
24. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA *et al*: **MicroRNA expression profiles classify human cancers**. *Nature* 2005, **435**(7043):834-838.
25. Bazzini AA, Lee MT, Giraldez AJ: **Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish**. *Science* 2012, **336**(6078):233-237.
26. Djuranovic S, Nahvi A, Green R: **miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay**. *Science* 2012, **336**(6078):237-240.
27. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP: **A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?** *Cell* 2011, **146**(3):353-358.
28. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.
29. Schena M: **Genome analysis with gene expression microarrays**. *BioEssays : news and reviews in molecular, cellular and developmental biology* 1996, **18**(5):427-431.
30. Shalon D, Smith SJ, Brown PO: **A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization**. *Genome research* 1996, **6**(7):639-645.
31. Tang T, Francois N, Glatigny A, Agier N, Mucchielli MH, Aggerbeck L, Delacroix H: **Expression ratio evaluation in two-colour microarray experiments is significantly improved by correcting image misalignment**. *Bioinformatics* 2007, **23**(20):2686-2691.
32. Clark TA, Sugnet CW, Ares M, Jr.: **Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays**. *Science* 2002, **296**(5569):907-910.
33. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews Genetics* 2009, **10**(1):57-63.
34. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(12):5463-5467.
35. Harbers M, Carninci P: **Tag-based approaches for transcriptome research and genome annotation**. *Nature methods* 2005, **2**(7):495-502.
36. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T *et al*: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(26):15776-15781.
37. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature methods* 2008, **5**(7):621-628.

38. Katz Y, Wang ET, Airoldi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation**. *Nature methods* 2010, **7**(12):1009-U1101.
39. McGettigan PA: **Transcriptomics in the RNA-seq era**. *Curr Opin Chem Biol* 2013, **17**(1):4-11.
40. Chappell L: **Finding a needle in a haystack. Microbial metatranscriptomes**. *Nature reviews Microbiology* 2012, **10**(7):446.
41. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T *et al*: **Comparative analysis of RNA sequencing methods for degraded or low-input samples**. *Nature methods* 2013, **10**(7):623-629.
42. Haas BJ, Zody MC: **Advancing RNA-Seq analysis**. *Nature biotechnology* 2010, **28**(5):421-423.
43. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nature protocols* 2012, **7**(3):562-578.
44. Steijger T, Abril JF, Engstrom PG, Kokocinski F, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J, Bertone P *et al*: **Assessment of transcript reconstruction methods for RNA-seq**. *Nature methods* 2013, **10**(12):1177-1184.
45. van der Brug M, Nalls MA, Cookson MR: **Deep sequencing of coding and non-coding RNA in the CNS**. *Brain research* 2010, **1338**:146-154.
46. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T: **The RIN: an RNA integrity number for assigning integrity values to RNA measurements**. *BMC molecular biology* 2006, **7**:3.
47. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nature reviews Genetics* 2004, **5**(4):276-287.
48. Carter SL, Brechbuhler CM, Griffin M, Bond AT: **Gene co-expression network topology provides a framework for molecular characterization of cellular state**. *Bioinformatics* 2004, **20**(14):2242-2250.
49. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis**. *Statistical applications in genetics and molecular biology* 2005, **4**:Article17.
50. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context**. *BMC bioinformatics* 2006, **7 Suppl 1**:S7.
51. Dunlop MJ, Cox RS, 3rd, Levine JH, Murray RM, Elowitz MB: **Regulatory activity revealed by dynamic correlations in gene expression noise**. *Nature genetics* 2008, **40**(12):1493-1498.
52. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes**. *Nature* 2004, **431**(7006):308-312.
53. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of Escherichia coli**. *Nature genetics* 2002, **31**(1):64-68.
54. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks**. *Science* 2002, **298**(5594):824-827.
55. Shinozaki K, Yamaguchi-Shinozaki K, Seki M: **Regulatory network of gene expression in the drought and cold stress responses**. *Current opinion in plant biology* 2003, **6**(5):410-417.
56. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG *et al*: **Core transcriptional regulatory circuitry in human embryonic stem cells**. *Cell* 2005, **122**(6):947-956.
57. Teichmann SA, Babu MM: **Gene regulatory network growth by duplication**. *Nature genetics* 2004, **36**(5):492-496.

58. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G: **Transposable elements have rewired the core regulatory network of human embryonic stem cells**. *Nature genetics* 2010, **42**(7):631-634.
59. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH *et al*: **Evolution of the mammalian transcription factor binding repertoire via transposable elements**. *Genome research* 2008, **18**(11):1752-1762.
60. Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK: **Effect of natural genetic variation on enhancer selection and function**. *Nature* 2013, **503**(7477):487-492.
61. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H *et al*: **The transcriptional network for mesenchymal transformation of brain tumours**. *Nature* 2010, **463**(7279):318-325.
62. Fletcher MN, Castro MA, Wang X, de Santiago I, O'Reilly M, Chin SF, Rueda OM, Caldas C, Ponder BA, Markowetz F *et al*: **Master regulators of FGFR2 signalling and breast cancer risk**. *Nature communications* 2013, **4**:2464.
63. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C *et al*: **A genomic regulatory network for development**. *Science* 2002, **295**(5560):1669-1678.
64. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al*: **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science* 2002, **298**(5594):799-804.
65. Bullmore E, Sporns O: **Complex brain networks: graph theoretical analysis of structural and functional systems**. *Nature reviews Neuroscience* 2009, **10**(3):186-198.
66. Hickman SE, Kingery ND, Ohsumi TK, Borowsky ML, Wang LC, Means TK, El Khoury J: **The microglial sensome revealed by direct RNA sequencing**. *Nature neuroscience* 2013, **16**(12):1896-1905.
67. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH: **Functional organization of the transcriptome in human brain**. *Nature neuroscience* 2008, **11**(11):1271-1282.
68. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R *et al*: **Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease**. *Cell* 2013, **153**(3):707-720.
69. Berchtold NC, Cribbs DH, Coleman PD, Rogers J, Head E, Kim R, Beach T, Miller C, Troncoso J, Trojanowski JQ *et al*: **Gene expression changes in the course of normal brain aging are sexually dimorphic**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(40):15605-15610.
70. Mazin P, Xiong J, Liu X, Yan Z, Zhang X, Li M, He L, Somel M, Yuan Y, Phoebe Chen YP *et al*: **Widespread splicing changes in human brain development and aging**. *Molecular systems biology* 2013, **9**:633.
71. Pletikos M, Sousa AM, Sedmak G, Meyer KA, Zhu Y, Cheng F, Li M, Kawasawa YI, Sestan N: **Temporal Specification and Bilaterality of Human Neocortical Topographic Gene Expression**. *Neuron* 2013.
72. Kumar A, Gibbs JR, Beilina A, Dillman A, Kumaran R, Trabzuni D, Ryten M, Walker R, Smith C, Traynor BJ *et al*: **Age-associated changes in gene expression in human brain and isolated neurons**. *Neurobiology of aging* 2013, **34**(4):1199-1209.
73. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL *et al*: **An anatomically comprehensive atlas of the adult human brain transcriptome**. *Nature* 2012, **489**(7416):391-399.
74. Loerch PM, Lu T, Dakin KA, Vann JM, Isaacs A, Geula C, Wang J, Pan Y, Gabuzda DH, Li C *et al*: **Evolution of the aging brain transcriptome and synaptic regulation**. *PloS one* 2008, **3**(10):e3329.
75. Sterner KN, Weckle A, Chugani HT, Tarca AL, Sherwood CC, Hof PR, Kuzawa CW, Boddy AM, Abbas A, Raaum RL *et al*: **Dynamic gene expression in the human cerebral cortex distinguishes children from adults**. *PloS one* 2012, **7**(5):e37714.

76. Patani R, Lewis PA, Trabzuni D, Puddifoot CA, Wyllie DJ, Walker R, Smith C, Hardingham GE, Weale M, Hardy J *et al*: **Investigating the utility of human embryonic stem cell-derived neurons to model ageing and neurodegenerative disease using whole-genome gene expression and splicing analysis**. *Journal of neurochemistry* 2012, **122**(4):738-751.
77. Belgard TG, Marques AC, Oliver PL, Abaan HO, Sirey TM, Hoerder-Suabedissen A, Garcia-Moreno F, Molnar Z, Margulies EH, Ponting CP: **A transcriptomic atlas of mouse neocortical layers**. *Neuron* 2011, **71**(4):605-616.
78. Tan PP, French L, Pavlidis P: **Neuron-Enriched Gene Expression Patterns are Regionally Anti-Correlated with Oligodendrocyte-Enriched Patterns in the Adult Mouse and Human Brain**. *Frontiers in neuroscience* 2013, **7**:5.
79. Ponting CP, Oliver PL: **Brain, know thy transcriptome, know thyself**. *Neuron* 2012, **75**(4):543-545.
80. Xu XX, Nehorai A, Dougherty J: **Cell Type Specific Analysis of Human Transcriptome Data**. *2012 Ieee International Workshop on Genomic Signal Processing and Statistics (Gensips)* 2012:99-100.
81. Bernard A, Lubbers LS, Tanis KQ, Luo R, Podtelezhnikov AA, Finney EM, McWhorter MM, Serikawa K, Lemon T, Morgan R *et al*: **Transcriptional architecture of the primate neocortex**. *Neuron* 2012, **73**(6):1083-1099.
82. Konopka G, Friedrich T, Davis-Turak J, Winden K, Oldham MC, Gao F, Chen L, Wang GZ, Luo R, Preuss TM *et al*: **Human-specific transcriptional networks in the brain**. *Neuron* 2012, **75**(4):601-617.
83. Somel M, Liu X, Khaitovich P: **Human brain evolution: transcripts, metabolites and their regulators**. *Nature reviews Neuroscience* 2013, **14**(2):112-127.
84. Buckner RL, Krienen FM: **The evolution of distributed association networks in the human brain**. *Trends in cognitive sciences* 2013, **17**(12):648-665.
85. Geschwind DH, Rakic P: **Cortical evolution: judge the brain by its cover**. *Neuron* 2013, **80**(3):633-647.
86. Johnson MB, Kawasawa YI, Mason CE, Krsnik Z, Coppola G, Bogdanovic D, Geschwind DH, Mane SM, State MW, Sestan N: **Functional and evolutionary insights into human brain development through global transcriptome analysis**. *Neuron* 2009, **62**(4):494-509.
87. Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, Yuan Y, He L, Oleksiak A, Zhang Y *et al*: **Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques**. *Genome research* 2012, **22**(4):611-622.
88. Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S: **A neutral model of transcriptome evolution**. *PLoS biology* 2004, **2**(5):E132.
89. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D *et al*: **Population genomics of human gene expression**. *Nature genetics* 2007, **39**(10):1217-1224.
90. Elowitz MB, Levine AJ, Siggia ED, Swain PS: **Stochastic gene expression in a single cell**. *Science* 2002, **297**(5584):1183-1186.
91. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S: **Stochastic mRNA synthesis in mammalian cells**. *PLoS biology* 2006, **4**(10):e309.
92. Raser JM, O'Shea EK: **Control of stochasticity in eukaryotic gene expression**. *Science* 2004, **304**(5678):1811-1814.
93. Bahar R, Hartmann CH, Rodriguez KA, Denny AD, Busuttil RA, Dolle ME, Calder RB, Chisholm GB, Pollock BH, Klein CA *et al*: **Increased cell-to-cell variation in gene expression in ageing mouse heart**. *Nature* 2006, **441**(7096):1011-1014.
94. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S: **Transcriptome-wide noise controls lineage choice in mammalian progenitor cells**. *Nature* 2008, **453**(7194):544-547.

95. Hansen AS, O'Shea EK: **Promoter decoding of transcription factor dynamics involves a trade-off between noise and control of gene expression**. *Molecular systems biology* 2013, **9**:704.
96. Raj A, van Oudenaarden A: **Nature, nurture, or chance: stochastic gene expression and its consequences**. *Cell* 2008, **135**(2):216-226.
97. Kremsky I, Morgan TE, Hou X, Li L, Finch CE: **Age-changes in gene expression in primary mixed glia cultures from young vs. old rat cerebral cortex are modified by interactions with neurons**. *Brain, behavior, and immunity* 2012, **26**(5):797-802.
98. Goyal MS, Raichle ME: **Gene expression-based modeling of human cortical synaptic density**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(16):6571-6576.
99. Huang H, Jeon T, Sedmak G, Pletikos M, Vasung L, Xu X, Yarowsky P, Richards LJ, Kostovic I, Sestan N *et al*: **Coupling diffusion imaging with histological and gene expression analysis to examine the dynamics of cortical areas across the fetal period of human brain development**. *Cereb Cortex* 2013, **23**(11):2620-2631.
100. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH: **Transcriptomic analysis of autistic brain reveals convergent molecular pathology**. *Nature* 2011, **474**(7351):380-384.
101. Forabosco P, Ramasamy A, Trabzuni D, Walker R, Smith C, Bras J, Levine AP, Hardy J, Pocock JM, Guerreiro R *et al*: **Insights into TREM2 biology by network analysis of human brain gene expression data**. *Neurobiology of aging* 2013, **34**(12):2699-2714.
102. Thorisson GA, Smith AV, Krishnan L, Stein LD: **The International HapMap Project Web site**. *Genome research* 2005, **15**(11):1592-1593.
103. Colantuoni C, Henry G, Zeger S, Pevsner J: **SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis**. *Bioinformatics* 2002, **18**(11):1540-1541.
104. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology**. *Nature reviews Genetics* 2009, **10**(10):669-680.
105. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J: **Practical guidelines for the comprehensive analysis of ChIP-seq data**. *PLoS computational biology* 2013, **9**(11):e1003326.
106. Heger A, Webber C, Goodson M, Ponting CP, Lunter G: **GAT: a simulation framework for testing the association of genomic intervals**. *Bioinformatics* 2013, **29**(16):2046-2048.
107. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I: **Controlling the false discovery rate in behavior genetics research**. *Behavioural brain research* 2001, **125**(1-2):279-284.
108. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(16):9440-9445.
109. Narum SR: **Beyond Bonferroni: Less conservative analyses for conservation genetics**. *Conserv Genet* 2006, **7**(5):783-787.
110. Benner C: **HOMER**. In.
111. ArrayExpress: **Illumina Bodymap2 Transcriptome**. In.; 2011a.
112. ArrayExpress: **E-MTAB-513 - RNA-Seq of Human Individual Tissues and Mixture of 16 Tissues (Illumina Body Map)**. 2011b.
113. consortium G: **Genotype-Tissue Expression project** In.; 2014.
114. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nature biotechnology* 2010, **28**(5):511-515.
115. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets**. *Cell* 2005, **120**(1):15-20.

116. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nature genetics* 2000, **25**(1):25-29.
117. Supek F, Bosnjak M, Skunca N, Smuc T: **REVIGO summarizes and visualizes long lists of gene ontology terms**. *PloS one* 2011, **6**(7):e21800.
118. Kanehisa M: **Post-Genome Informatics**. In.: Oxofrd University Press; 2000.
119. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE: **The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse**. *Nucleic acids research* 2014, **42**(Database issue):D810-817.
120. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2009, **4**(1):44-57.
121. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions**. *Nature biotechnology* 2010, **28**(5):495-501.
122. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 2008, **9**:559.
123. Langfelder P, Luo R, Oldham MC, Horvath S: **Is my network module preserved and reproducible?** *PLoS computational biology* 2011, **7**(1):e1001057.
124. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells**. *Nature genetics* 2005, **37**(4):382-390.
125. Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A: **Reverse engineering cellular networks**. *Nat Protoc* 2006, **1**(2):663-672.
126. Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A: **Reverse engineering cellular networks**. *Nature protocols* 2006, **1**(2):662-671.
127. VanGilder RL, Huber JD, Rosen CL, Barr TL: **The transcriptome of cerebral ischemia**. *Brain research bulletin* 2012, **88**(4):313-319.
128. Hashimoto-Torii K, Kawasawa YI, Kuhn A, Rakic P: **Combined transcriptome analysis of fetal human and mouse cerebral cortex exposed to alcohol**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(10):4212-4217.
129. Rosell A, Vilalta A, Garcia-Berrocoso T, Fernandez-Cadenas I, Domingues-Montanari S, Cuadrado E, Delgado P, Ribo M, Martinez-Saez E, Ortega-Aznar A *et al*: **Brain perihematoma genomic profile following spontaneous human intracerebral hemorrhage**. *PloS one* 2011, **6**(2):e16750.
130. Li JZ, Bunney BG, Meng F, Hagenauer MH, Walsh DM, Vawter MP, Evans SJ, Choudary PV, Cartagena P, Barchas JD *et al*: **Circadian patterns of gene expression in the human brain and disruption in major depressive disorder**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(24):9950-9955.
131. Miller JA, Horvath S, Geschwind DH: **Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(28):12698-12703.
132. Twine NA, Janitz K, Wilkins MR, Janitz M: **Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease**. *PloS one* 2011, **6**(1):e16266.
133. Ponomarev I, Wang S, Zhang L, Harris RA, Mayfield RD: **Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence**. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 2012, **32**(5):1884-1897.
134. Courtney E, Kornfeld S, Janitz K, Janitz M: **Transcriptome profiling in neurodegenerative disease**. *Journal of neuroscience methods* 2010, **193**(2):189-202.
135. Mills JD, Janitz M: **Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases**. *Neurobiology of aging* 2012, **33**(5):1012 e1011-1024.

136. Vidal M, Cusick ME, Barabasi AL: **Interactome networks and human disease**. *Cell* 2011, **144**(6):986-998.
137. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng ZP, Snyder M, Dermitzakis ET, Stamatoyannopoulos JA *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.
138. Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC *et al*: **A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers**. *Molecular systems biology* 2010, **6**:377.
139. Della Gatta G, Palomero T, Perez-Garcia A, Ambesi-Impiombato A, Bansal M, Carpenter ZW, De Keersmaecker K, Sole X, Xu L, Paietta E *et al*: **Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL**. *Nature medicine* 2012, **18**(3):436-440.
140. Lichtinger M, Ingram R, Hannah R, Muller D, Clarke D, Assi SA, Lie-A-Ling M, Noailles L, Vijayabaskar MS, Wu MC *et al*: **RUNX1 reshapes the epigenetic landscape at the onset of haematopoiesis**. *Embo J* 2012, **31**(22):4318-4333.
141. Norden DM, Godbout JP: **Review: microglia of the aged brain: primed to be activated and resistant to regulation**. *Neuropathology and applied neurobiology* 2013, **39**(1):19-34.
142. David S, Kroner A: **Repertoire of microglial and macrophage responses after spinal cord injury**. *Nature reviews Neuroscience* 2011, **12**(7):388-399.
143. Paolicelli RC, Bolasco G, Pagani F, Maggi L, Scianni M, Panzanelli P, Giustetto M, Ferreira TA, Guiducci E, Dumas L *et al*: **Synaptic pruning by microglia is necessary for normal brain development**. *Science* 2011, **333**(6048):1456-1458.
144. Schuitemaker A, van der Doef TF, Boellaard R, van der Flier WM, Yaqub M, Windhorst AD, Barkhof F, Jonker C, Kloet RW, Lammertsma AA *et al*: **Microglial activation in healthy aging**. *Neurobiology of aging* 2012, **33**(6):1067-1072.
145. Streit WJ, Sammons NW, Kuhns AJ, Sparks DL: **Dystrophic microglia in the aging human brain**. *Glia* 2004, **45**(2):208-212.
146. Hefendehl JK, Neher JJ, Suhs RB, Kohsaka S, Skodras A, Jucker M: **Homeostatic and injury-induced microglia behavior in the aging brain**. *Aging cell* 2014, **13**(1):60-69.
147. Conde JR, Streit WJ: **Microglia in the aging brain**. *Journal of neuropathology and experimental neurology* 2006, **65**(3):199-203.
148. VanGuilder HD, Bixler GV, Brucklacher RM, Farley JA, Yan H, Warrington JP, Sonntag WE, Freeman WM: **Concurrent hippocampal induction of MHC II pathway components and glial activation with advanced aging is not correlated with cognitive impairment**. *Journal of neuroinflammation* 2011, **8**:138.
149. Liao B, Zhao W, Beers DR, Henkel JS, Appel SH: **Transformation from a neuroprotective to a neurotoxic microglial phenotype in a mouse model of ALS**. *Experimental neurology* 2012, **237**(1):147-152.
150. Kumar A, Stoica BA, Sabirzhanov B, Burns MP, Faden AI, Loane DJ: **Traumatic brain injury in aged animals increases lesion size and chronically alters microglial/macrophage classical and alternative activation states**. *Neurobiology of aging* 2013, **34**(5):1397-1411.
151. Appel SH, Beers DR, Henkel JS: **T cell-microglial dialogue in Parkinson's disease and amyotrophic lateral sclerosis: are we listening?** *Trends in immunology* 2010, **31**(1):7-17.
152. Lampron A, Gosselin D, Rivest S: **Targeting the hematopoietic system for the treatment of Alzheimer's disease**. *Brain, behavior, and immunity* 2011, **25 Suppl 1**:S71-79.
153. Lee DC, Ruiz CR, Lebson L, Selenica ML, Rizer J, Hunt JB, Jr., Rojiani R, Reid P, Kammath S, Nash K *et al*: **Aging enhances classical activation but mitigates alternative activation in the central nervous system**. *Neurobiology of aging* 2013, **34**(6):1610-1620.

154. Nishioku T, Matsumoto J, Dohgu S, Sumi N, Miyao K, Takata F, Shuto H, Yamauchi A, Kataoka Y: **Tumor necrosis factor-alpha mediates the blood-brain barrier dysfunction induced by activated microglia in mouse brain microvascular endothelial cells**. *Journal of pharmacological sciences* 2010, **112**(2):251-254.
155. Sumi N, Nishioku T, Takata F, Matsumoto J, Watanabe T, Shuto H, Yamauchi A, Dohgu S, Kataoka Y: **Lipopolysaccharide-activated microglia induce dysfunction of the blood-brain barrier in rat microvascular endothelial cells co-cultured with microglia**. *Cellular and molecular neurobiology* 2010, **30**(2):247-253.
156. Ginhoux F, Greter M, Leboeuf M, Nandi S, See P, Gokhan S, Mehler MF, Conway SJ, Ng LG, Stanley ER *et al*: **Fate mapping analysis reveals that adult microglia derive from primitive macrophages**. *Science* 2010, **330**(6005):841-845.
157. Ponomarev ED, Veremeyko T, Weiner HL: **MicroRNAs are universal regulators of differentiation, activation, and polarization of microglia and macrophages in normal and diseased CNS**. *Glia* 2013, **61**(1):91-103.
158. Smith AM, Gibbons HM, Oldfield RL, Bergin PM, Mee EW, Faull RL, Dragunow M: **The transcription factor PU.1 is critical for viability and function of human brain microglia**. *Glia* 2013, **61**(6):929-942.
159. Aguzzi A, Barres BA, Bennett ML: **Microglia: scapegoat, saboteur, or something else?** *Science* 2013, **339**(6116):156-161.
160. Kierdorf K, Prinz M: **Factors regulating microglia activation**. *Frontiers in cellular neuroscience* 2013, **7**:44.
161. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, Xing Y, Lubischer JL, Krieg PA, Krupenko SA *et al*: **A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function**. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 2008, **28**(1):264-278.
162. Butovsky O, Jedrychowski MP, Moore CS, Cialic R, Lanser AJ, Gabriely G, Koeglsperger T, Dake B, Wu PM, Doykan CE *et al*: **Identification of a unique TGF-beta-dependent molecular and functional signature in microglia**. *Nature neuroscience* 2014, **17**(1):131-143.
163. Varnum MM, Ikezu T: **The Classification of Microglial Activation Phenotypes on Neurodegeneration and Regeneration in Alzheimer's Disease Brain**. *Arch Immunol Ther Ex* 2012, **60**(4):251-266.
164. Harry GJ: **Microglia during development and aging**. *Pharmacol Therapeut* 2013, **139**(3):313-326.
165. Wynne AM, Henry CJ, Huang Y, Cleland A, Godbout JP: **Protracted downregulation of CX(3)CR1 on microglia of aged mice after lipopolysaccharide challenge**. *Brain Behavior and Immunity* 2010, **24**(7):1190-1201.
166. Saijo K, Glass CK: **Microglial cell origin and phenotypes in health and disease**. *Nat Rev Immunol* 2011, **11**(11):775-787.
167. Kierdorf K, Erny D, Goldmann T, Sander V, Schulz C, Perdiguero EG, Wieghofer P, Heinrich A, Riemke P, Holscher C *et al*: **Microglia emerge from erythromyeloid precursors via Pu.1- and Irf8-dependent pathways**. *Nature neuroscience* 2013, **16**(3):273-280.
168. Jin H, Li L, Xu J, Zhen FH, Zhu L, Liu PP, Zhang MJ, Zhang WQ, Wen ZL: **Runx1 regulates embryonic myeloid fate choice in zebrafish through a negative feedback loop inhibiting Pu.1 expression**. *Blood* 2012, **119**(22):5239-5249.
169. Le Clech M, Chalhoub E, Dohet C, Roure V, Fichelson S, Moreau-Gachelin F, Mathieu D: **PU.1/Spi-1 binds to the human TAL-1 silencer to mediate its activity**. *Journal of molecular biology* 2006, **355**(1):9-19.
170. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, Cotsapas C, Daly MJ: **Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology**. *Plos Genet* 2011, **7**(1):e1001273.

171. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic acids research* 2005, **33**(Database issue):D501-504.
172. Ajami B, Bennett JL, Krieger C, Tetzlaff W, Rossi FMV: **Local self-renewal can sustain CNS microglia maintenance and function throughout adult life**. *Nature neuroscience* 2007, **10**(12):1538-1543.
173. Parakalan R, Jiang B, Nimmi B, Janani M, Jayapal M, Lu J, Tay SSW, Ling EA, Dheen ST: **Transcriptome analysis of amoeboid and ramified microglia isolated from the corpus callosum of rat brain**. *Bmc Neurosci* 2012, **13**.
174. Zusso M, Methot L, Lo R, Greenhalgh AD, David S, Stifani S: **Regulation of Postnatal Forebrain Amoeboid Microglial Cell Proliferation and Development by the Transcription Factor Runx1**. *Journal of Neuroscience* 2012, **32**(33):11285-11298.
175. Shin DM, Lee CH, Morse HC: **IRF8 Governs Expression of Genes Involved in Innate and Adaptive Immunity in Human and Mouse Germinal Center B Cells**. *PloS one* 2011, **6**(11).
176. Kubosaki A, Lindgren G, Tagami M, Simon C, Tomaru Y, Miura H, Suzuki T, Arner E, Forrest ARR, Irvine KM *et al*: **The combination of gene perturbation assay and ChIP-chip reveals functional direct target genes for IRF8 in THP-1 cells**. *Mol Immunol* 2010, **47**(14):2295-2302.
177. Cribbs DH, Berchtold NC, Perreau V, Coleman PD, Rogers J, Tenner AJ, Cotman CW: **Extensive innate immune gene activation accompanies brain aging, increasing vulnerability to cognitive decline and neurodegeneration: a microarray study**. *Journal of neuroinflammation* 2012, **9**.
178. Terry RD, Deteresa R, Hansen LA: **Neocortical Cell Counts in Normal Human Adult Aging**. *Ann Neurol* 1987, **21**(6):530-539.
179. Staber PB, Zhang P, Ye M, Welner RS, Nombela-Arrieta C, Bach C, Kerenyi M, Bartholdy BA, Zhang H, Alberich-Jorda M *et al*: **Sustained PU.1 Levels Balance Cell-Cycle Regulators to Prevent Exhaustion of Adult Hematopoietic Stem Cells**. *Molecular cell* 2013, **49**(5):934-946.
180. !!! INVALID CITATION !!!
181. Dey S, Curtis DJ, Jane SM, Brandt SJ: **The TAL1/SCL Transcription Factor Regulates Cell Cycle Progression and Proliferation in Differentiating Murine Bone Marrow Monocyte Precursors**. *Molecular and cellular biology* 2010, **30**(9):2181-2192.
182. Streit WJ, Miller KR, Lopes KO, Njie E: **Microglial degeneration in the aging brain - bad news for neurons?** *Front Biosci* 2008, **13**:3423-3438.
183. Furukawa S, Fujita T, Shimabukuro M, Iwaki M, Yamada Y, Nakajima Y, Nakayama O, Makishima M, Matsuda M, Shimomura I: **Increased oxidative stress in obesity and its impact on metabolic syndrome**. *J Clin Invest* 2004, **114**(12):1752-1761.
184. Adamo L, Naveiras O, Wenzel PL, McKinney-Freeman S, Mack PJ, Gracia-Sancho J, Suchy-Dicey A, Yoshimoto M, Lensch MW, Yoder MC *et al*: **Biomechanical forces promote embryonic haematopoiesis**. *Nature* 2009, **459**(7250):1131-U1120.
185. Wang Y, Cao YY, Jia XM, Cao YB, Gao PH, Fu XP, Ying K, Chen WS, Jiang YY: **Cap1p is involved in multiple pathways of oxidative stress response in Candida albicans**. *Free Radical Bio Med* 2006, **40**(7):1201-1209.
186. Millar T, Walker R, Arango JC, Ironside JW, Harrison DJ, MacIntyre DJ, Blackwood D, Smith C, Bell JE: **Tissue and organ donation for research in forensic pathology: the MRC Sudden Death Brain and Tissue Bank**. *The Journal of pathology* 2007, **213**(4):369-375.
187. Trabzuni D, Ryten M, Walker R, Smith C, Imran S, Ramasamy A, Weale ME, Hardy J: **Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies**. *Journal of neurochemistry* 2011, **119**(2):275-282.

188. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.
189. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities**. *Molecular cell* 2010, **38**(4):576-589.
190. Ptasinska A, Assi SA, Mannari D, James SR, Williamson D, Dunne J, Hoogenkamp M, Wu M, Care M, McNeill H *et al*: **Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding**. *Leukemia* 2012, **26**(8):1829-1841.
191. Sanda T, Lawton LN, Barrasa MI, Fan ZP, Kohlhammer H, Gutierrez A, Ma W, Tatarek J, Ahn Y, Kelliher MA *et al*: **Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia**. *Cancer cell* 2012, **22**(2):209-221.
192. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution**. *Nature reviews Genetics* 2009, **10**(4):252-263.
193. Wingender E, Schoeps T, Donitz J: **TFClass: an expandable hierarchical classification of human transcription factors**. *Nucleic acids research* 2013, **41**(Database issue):D165-170.
194. Patricio P, Mateus-Pinheiro A, Sousa N, Pinto L: **Re-cycling paradigms: cell cycle regulation in adult hippocampal neurogenesis and implications for depression**. *Molecular neurobiology* 2013, **48**(1):84-96.
195. Kodama M, Fujioka T, Duman RS: **Chronic olanzapine or fluoxetine administration increases cell proliferation in hippocampus and prefrontal cortex of adult rat**. *Biological psychiatry* 2004, **56**(8):570-580.
196. Gould E, Reeves AJ, Graziano MS, Gross CG: **Neurogenesis in the neocortex of adult primates**. *Science* 1999, **286**(5439):548-552.
197. Goncalves L, Silva R, Pinto-Ribeiro F, Pego JM, Bessa JM, Pertovaara A, Sousa N, Almeida A: **Neuropathic pain is associated with depressive behaviour and induces neuroplasticity in the amygdala of the rat**. *Experimental neurology* 2008, **213**(1):48-56.
198. Bernier PJ, Bedard A, Vinet J, Levesque M, Parent A: **Newly generated neurons in the amygdala and adjoining cortex of adult primates**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(17):11464-11469.
199. Fowler CD, Johnson F, Wang Z: **Estrogen regulation of cell proliferation and distribution of estrogen receptor-alpha in the brains of adult female prairie and meadow voles**. *The Journal of comparative neurology* 2005, **489**(2):166-179.
200. Kokoeva MV, Yin H, Flier JS: **Neurogenesis in the hypothalamus of adult mice: potential role in energy balance**. *Science* 2005, **310**(5748):679-683.
201. Fowler CD, Liu Y, Ouimet C, Wang Z: **The effects of social environment on adult neurogenesis in the female prairie vole**. *Journal of neurobiology* 2002, **51**(2):115-128.
202. Dayer AG, Cleaver KM, Abouantoun T, Cameron HA: **New GABAergic interneurons in the adult neocortex and striatum are generated from different precursors**. *The Journal of cell biology* 2005, **168**(3):415-427.
203. Bedard A, Gravel C, Parent A: **Chemical characterization of newly generated neurons in the striatum of adult primates**. *Experimental brain research* 2006, **170**(4):501-512.
204. Zhao M, Momma S, Delfani K, Carlen M, Cassidy RM, Johansson CB, Brismar H, Shupliakov O, Frisen J, Janson AM: **Evidence for neurogenesis in the adult mammalian substantia nigra**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(13):7925-7930.

205. Yoshimi K, Ren YR, Seki T, Yamada M, Ooizumi H, Onodera M, Saito Y, Murayama S, Okano H, Mizuno Y *et al*: **Possibility for neurogenesis in substantia nigra of parkinsonian brain**. *Ann Neurol* 2005, **58**(1):31-40.
206. Jiao J, Chen DF: **Induction of neurogenesis in nonconventional neurogenic regions of the adult central nervous system by niche astrocyte-produced signals**. *Stem Cells* 2008, **26**(5):1221-1230.
207. Bonfanti L, Ponti G, Luzzati F, Paola Crociara, Parolisi R, Armentano M: **Parenchymal Neuro-Glio-Genesis Versus Germinal Layer- Derived Neurogenesis: Two Faces of Central Nervous System Structural Plasticity**. In: *Neural Stem Cells - New Perspectives.* Edited by Bonfanti L: InTech; 2013.
208. Emsley JG, Mitchell BD, Kempermann G, Macklis JD: **Adult neurogenesis and repair of the adult CNS with neural progenitors, precursors, and stem cells**. *Progress in neurobiology* 2005, **75**(5):321-341.
209. Ernst A, Alkass K, Bernard S, Salehpour M, Perl S, Tisdale J, Possnert G, Druid H, Frisen J: **Neurogenesis in the striatum of the adult human brain**. *Cell* 2014, **156**(5):1072-1083.
210. Bonfanti L, Peretto P: **Adult neurogenesis in mammals--a theme with many variations**. *The European journal of neuroscience* 2011, **34**(6):930-950.
211. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K *et al*: **Transcriptional landscape of the prenatal human brain**. *Nature* 2014.
212. Stiles J, Jernigan TL: **The basics of brain development**. *Neuropsychology review* 2010, **20**(4):327-348.
213. Lui JH, Hansen DV, Kriegstein AR: **Development and evolution of the human neocortex**. *Cell* 2011, **146**(1):18-36.
214. Niu W, Zang T, Zou Y, Fang S, Smith DK, Bachoo R, Zhang CL: **In vivo reprogramming of astrocytes to neuroblasts in the adult brain**. *Nature cell biology* 2013, **15**(10):1164-1175.
215. Genin EC, Caron N, Vandenbosch R, Nguyen L, Malgrange B: **Forkhead pathway in the control of adult neurogenesis**. *Stem Cells* 2014.
216. Kozlenkov A, Roussos P, Timashpolsky A, Barbu M, Rudchenko S, Bibikova M, Klotzle B, Byne W, Lyddon R, Di Narzo AF *et al*: **Differences in DNA methylation between human neuronal and glial cells are concentrated in enhancers and non-CpG sites**. *Nucleic acids research* 2014, **42**(1):109-127.
217. Ro SH, Liu D, Yeo H, Paik JH: **FoxOs in neural stem cell fate decision**. *Archives of biochemistry and biophysics* 2013, **534**(1-2):55-63.
218. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO *et al*: **Identification of genes periodically expressed in the human cell cycle and their expression in tumors**. *Mol Biol Cell* 2002, **13**(6):1977-2000.
219. Liu DZ, Ander BP, Sharp FR: **Cell cycle inhibition without disruption of neurogenesis is a strategy for treatment of central nervous system diseases**. *Neurobiology of disease* 2010, **37**(3):549-557.
220. Wang X, Kiyokawa H, Dennewitz MB, Costa RH: **The Forkhead Box m1b transcription factor is essential for hepatocyte DNA replication and mitosis during mouse liver regeneration**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(26):16881-16886.
221. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching**. *Nucleic acids research* 2009, **37**(Web Server issue):W202-208.
222. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K *et al*: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes**. *Nucleic acids research* 2006, **34**(Database issue):D108-110.

223. McLeay RC, Bailey TL: **Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data**. *BMC bioinformatics* 2010, **11**:165.
224. Endesfelder S, Zaak I, Weichelt U, Buhrer C, Schmitz T: **Caffeine protects neuronal cells against injury caused by hyperoxia in the immature brain**. *Free radical biology & medicine* 2014, **67**:221-234.
225. Alvarez-Buylla A, Lim DA: **For the long run: maintaining germinal niches in the adult brain**. *Neuron* 2004, **41**(5):683-686.
226. Halbach OvBu: **Immunohistological markers for proliferative events, gliogenesis, and neurogenesis within the adult hippocampus**. *Cell and Tissue Research* 2011, **345**(1).
227. Lee HK, Barbarosie M, Kameyama K, Bear MF, Huganir RL: **Regulation of distinct AMPA receptor phosphorylation sites during bidirectional synaptic plasticity**. *Nature* 2000, **405**(6789):955-959.
228. Krawetz S: **Bioinformatics for Systems Biology**. In.: Springer Science & Business Media; 2008: 750.
229. Bishop KM, Goudreau G, O'Leary DDM: **Regulation of area identity in the mammalian neocortex by Emx2 and Pax6**. *Science* 2000, **288**(5464):344-349.
230. Stolt CC, Lommes P, Sock E, Chaboissier MC, Schedl A, Wegner M: **The Sox9 transcription factor determines glial fate choice in the developing spinal cord**. *Genes & development* 2003, **17**(13):1677-1689.
231. Ivanovski S, Hamlet S, Retzepi M, Wall I, Donos N: **Transcriptional profiling of "guided bone regeneration" in a critical-size calvarial defect**. *Clin Oral Implan Res* 2011, **22**(4):382-389.
232. Bae SK, Bessho Y, Hojo M, Kageyama R: **The bHLH gene Hes6, an inhibitor of Hes1, promotes neuronal differentiation**. *Development* 2000, **127**(13):2933-2943.
233. Nagase T, Kikuno R, Ishikawa KI, Hirosawa M, Ohara O: **Prediction of the coding sequences of unidentified human genes. XVI. The complete sequences of 150 new cDNA clones from brain which code for large proteins in vitro**. *DNA Res* 2000, **7**(1):65-73.
234. Malik TH, Von Stechow D, Bronson RT, Shivdasani RA: **Deletion of the GATA domain of TRPS1 causes an absence of facial hair and provides new insights into the bone disorder in inherited tricho-rhino-phalangeal syndromes**. *Molecular and cellular biology* 2002, **22**(24):8592-8600.
235. Konishi A, Shimizu S, Hirota J, Takao T, Fan Y, Matsuoka Y, Zhang L, Yoneda Y, Fujii Y, Skoultchi AI *et al*: **Involvement of histone H1.2 in apoptosis induced by DNA double-strand breaks**. *Cell* 2003, **114**(6):673-688.
236. Aruga J, Yokota N, Hashimoto M, Furuichi T, Fukuda M, Mikoshiba K: **A novel zinc finger protein, zic, is involved in neurogenesis, especially in the cell lineage of cerebellar granule cells**. *Journal of neurochemistry* 1994, **63**(5):1880-1890.
237. Bylund M, Andersson E, Novitch BG, Muhr J: **Vertebrate neurogenesis is counteracted by Sox1-3 activity**. *Nature neuroscience* 2003, **6**(11):1162-1168.
238. Fiorenza MT, Mukhopadhyay M, Westphal H: **Expression screening for Lhx3 downstream genes identifies Thg-1pit as a novel mouse gene involved in pituitary development**. *Gene* 2001, **278**(1-2):125-130.
239. Fletcher CF, Okano HJ, Gilbert DJ, Yang Y, Yang C, Copeland NG, Jenkins NA, Darnell RB: **Mouse chromosomal locations of nine genes encoding homologs of human paraneoplastic neurologic disorder antigens**. *Genomics* 1997, **45**(2):313-319.
240. Lotem J, Levanon D, Negreanu V, Groner Y: **The False Paradigm of RUNX3 Function as Tumor Suppressor in Gastric Cancer**. *Journal of Cancer Therapy* 2013, **4**(1A).
241. Hemmi K, Ma D, Miura Y, Kawaguchi M, Sasahara M, Hashimoto-Tamaoki T, Tamaoki T, Sakata N, Tsuchiya K: **A homeodomain-zinc finger protein, ZFHX4, is expressed in neuronal differentiation manner and suppressed in muscle differentiation manner**. *Biol Pharm Bull* 2006, **29**(9):1830-1835.

242. Li XA, Kokame K, Okubo K, Shimokado K, Tsukamoto Y, Miyata T, Kato H, Yutani C: **Cloning and characterization of a novel human gene encoding a zinc finger protein with 25 fingers**. *Biochimica et biophysica acta* 1999, **1489**(2-3):405-412.
243. Jacquemin P, Lannoy VJ, Rousseau GG, Lemaigre FP: **OC-2, a novel mammalian member of the ONECUT class of homeodomain transcription factors whose function in liver partially overlaps with that of hepatocyte nuclear factor-6**. *The Journal of biological chemistry* 1999, **274**(5):2665-2671.
244. Li C, Lusis AJ, Sparkes R, Tran SM, Gaynor R: **Characterization and chromosomal mapping of the gene encoding the cellular DNA binding protein HTLF**. *Genomics* 1992, **13**(3):658-664.
245. Jeyaraj D, Haldar SM, Wan X, McCauley MD, Ripperger JA, Hu K, Lu Y, Eapen BL, Sharma N, Ficker E *et al*: **Circadian rhythms govern cardiac repolarization and arrhythmogenesis**. *Nature* 2012, **483**(7387):96-99.
246. Fu H, Cai J, Clevers H, Fast E, Gray S, Greenberg R, Jain MK, Ma Q, Qiu M, Rowitch DH *et al*: **A genome-wide screen for spatially restricted expression patterns identifies transcription factors that regulate glial development**. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 2009, **29**(36):11399-11408.
247. Sandberg M, Kallstrom M, Muhr J: **Sox21 promotes the progression of vertebrate neurogenesis**. *Nature neuroscience* 2005, **8**(8):995-1001.
248. Johnson RL, Tabin CJ: **Molecular models for vertebrate limb development**. *Cell* 1997, **90**(6):979-990.
249. Hsu YC, Kao CY, Chung YF, Chen MS, Chiu IM: **Ciliogenic RFX transcription factors regulate FGF1 gene promoter**. *Journal of cellular biochemistry* 2012, **113**(7):2511-2522.
250. Ferbus D, Le Chalony C, Prosperi MT, Muleris M, Vincent-Salomon A, Goubin G: **Identification, nuclear localization, and binding activities of OZF, a human protein solely composed of zinc-finger motifs**. *European journal of biochemistry / FEBS* 1996, **236**(3):991-995.
251. Tews B, Felsberg J, Hartmann C, Kunitz A, Hahn M, Toedt G, Neben K, Hummerich L, von Deimling A, Reifenberger G *et al*: **Identification of novel oligodendroglioma-associated candidate tumor suppressor genes in 1p36 and 19q13 using microarray-based expression profiling**. *Int J Cancer* 2006, **119**(4):792-800.
252. Wehrspaun CC, Ponting CP, Marques AC: **Brain-expressed 3'UTR extensions strengthen miRNA cross-talk between ion channel/transporter encoding mRNAs**. *Frontiers in genetics* 2014, **5**:41.
253. Shyu AB, Wilkinson MF, van Hoof A: **Messenger RNA regulation: to translate or to degrade**. *EMBO J* 2008, **27**(3):471-481.
254. Wang ET, Sandberg R, Luo SJ, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.
255. Kuersten S, Goodwin EB: **The power of the 3 ' UTR: Translational control and development**. *Nat Rev Genet* 2003, **4**(8):626-637.
256. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM: **Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution**. *Cell* 2005, **123**(6):1133-1146.
257. Chen K, Rajewsky N: **Natural selection on human microRNA binding sites inferred from SNP data**. *Nature genetics* 2006, **38**(12):1452-1456.
258. Kawase-Koga Y, Otaegi G, Sun T: **Different Timings of Dicer Deletion Affect Neurogenesis and Gliogenesis in the Developing Mouse Central Nervous System**. *Dev Dynam* 2009, **238**(11):2800-2812.
259. Wang Y, Liang Y, Lu Q: **MicroRNA epigenetic alterations: predicting biomarkers and therapeutic targets in human diseases**. *Clin Genet* 2008, **74**(4):307-315.

260. Wu D, Raafat A, Pak E, Clemens S, Murashov AK: **Dicer-microRNA pathway is critical for peripheral nerve regeneration and functional recovery in vivo and regenerative axonogenesis in vitro**. *Exp Neurol* 2012, **233**(1):555-565.
261. Nelson PT, Wang WX, Rajeev BW: **MicroRNAs (miRNAs) in neurodegenerative diseases**. *Brain Pathol* 2008, **18**(1):130-138.
262. Hughes TA: **Regulation of gene expression by alternative untranslated regions**. *Trends Genet* 2006, **22**(3):119-122.
263. Iseli C, Stevenson BJ, Souza SJd, Samaia HB, Camargo AA, Buetow KH, Strausberg RL, Simpson AJ, Bucher P, Jongeneel CV: **Long-range heterogeneity at the 3′ ends of human mRNAs**. *Genome research* 2002, **12**(7):1068-1074.
264. Smibert P, Miura P, Westholm JO, Shenker S, May G, Duff MO, Zhang D, Eads BD, Carlson J, Brown JB *et al*: **Global patterns of tissue-specific alternative polyadenylation in Drosophila**. *Cell reports* 2012, **1**(3):277-289.
265. Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, Sive H, Bartel DP: **Extensive alternative polyadenylation during zebrafish development**. *Genome research* 2012, **22**(10):2054-2066.
266. Zhang HB, Lee JY, Tian B: **Biased alternative polyadenylation in human tissues**. *Genome Biol* 2005, **6**(12).
267. Ji Z, Lee JY, Pan ZH, Jiang BJ, Tian B: **Progressive lengthening of 3 ' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(17):7028-7033.
268. Hilgers V, Perry MW, Hendrix D, Stark A, Levine M, Haley B: **Neural-specific elongation of 3' UTRs during Drosophila development**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(38):15864-15869.
269. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC: **Widespread and extensive lengthening of 3 ' UTRs in the mammalian brain**. *Genome research* 2013, **23**(5):812-825.
270. Ramskold D, Wang ET, Burge CB, Sandberg R: **An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data**. *Plos Comput Biol* 2009, **5**(12):e1000598.
271. Coy JF, Sedlacek Z, Bächner D, Delius H, Poustka A: **A complex pattern of evolutionary conservation and alternative polyadenylation within the long 3′-untranslated region of the methyl-CpG-binding protein 2 gene (MeCP2) suggests a regulatory role in gene expression**. *Human molecular genetics* 1999, **8**(7):1253-1262.
272. Balmer D, Goldstine J, Rao YM, LaSalle JM: **Elevated methyl-CpG-binding protein 2 expression is acquired during postnatal human brain development and is correlated with alternative polyadenylation**. *Journal of molecular medicine* 2003, **81**(1):61-68.
273. Han K, Gennarino VA, Lee Y, Pang KF, Hashimoto-Torii K, Choufani S, Raju CS, Oldham MC, Weksberg R, Rakic P *et al*: **Human-specific regulation of MeCP2 levels in fetal brains by microRNA miR-483-5p**. *Gene Dev* 2013, **27**(5):485-490.
274. Ala U, Karretha FA, Bosiab C, Pagnanib A, Taullia R, Léopolda V, Taya Y, Proverod P, Zecchinab R, Pandolfia PP: **Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments**. *PNAS* 2013, **110**(18):7154–7159.
275. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V *et al*: **The landscape of C. elegans 3'UTRs**. *Science* 2010, **329**(5990):432-435.
276. Chi SW, Zang JB, Mele A, Darnell RB: **Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps**. *Nature* 2009, **460**(7254):479-486.
277. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution**. *Nat Rev Genet* 2009, **10**(4):252-263.

278. Holtmaat A, Svoboda K: **Experience-dependent structural synaptic plasticity in the mammalian brain**. *Nature reviews Neuroscience* 2009, **10**(9):647-658.
279. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: **Molecular Biology of the Cell**, 4 edn. New York: Garland Science; 2002.
280. Jentsch TJ, Hubner CA, Fuhrmann JC: **Ion channels: Function unravelled by dysfunction**. *Nat Cell Biol* 2004, **6**(11):1039-1047.
281. Schulz DJ, Temporal S, Barry DM, Garcia ML: **Mechanisms of voltage-gated ion channel regulation: from gene expression to localization**. *Cell Mol Life Sci* 2008, **65**(14):2215-2231.
282. Favereaux A, Thoumine O, Bouali-Benazzouz R, Roques V, Papon M-A, Salam SA, Drutel G, Léger C, Calas A, Nagy F *et al*: **Bidirectional integrative regulation of Cav1. 2 calcium channel by microRNA miR-103: role in pain**. *The EMBO journal* 2011, **30**(18):3830-3841.
283. Splawski I, Timothy KW, Priori SG, Napolitano C, Bloise R: **Timothy Syndrome**. In: *GeneReviews*^{TM}. Edited by Pagon RA, Adam MP, Bird TD, Dolan CR, Fong CT, Stephens K. http://www.ncbi.nlm.nih.gov/books/NBK1403/; 2006.
284. Antzelevitch C, Pollevick GD, Cordeiro JM, Casis O, Sanguinetti MC, Aizawa Y, Guerchicoff A, Pfeiffer R, Oliva A, Wollnik B *et al*: **Loss-of-function mutations in the cardiac calcium channel underlie a new clinical entity characterized by ST-segment elevation, short QT intervals, and sudden cardiac death**. *Circulation* 2007, **115**(4):442-449.
285. Haerty W, Ponting CP: **Mutations within lncRNAs are effectively selected against in fruitfly but not in human**. *Genome Biol* 2013, **14**(R49).
286. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, Vaughan B, Preuss D, Leinonen R, Shumway M *et al*: **The 1000 Genomes Project: data management and community access**. *Nature methods* 2012, **9**(5):459-462.
287. Lunter G, Ponting CP, Hein J: **Genome-wide identification of human functional DNA using a neutral indel model**. *Plos Comput Biol* 2006, **2**(1):2-12.
288. Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, Colantuoni EA, Elkahloun AG, Herman MM, Weinberger DR *et al*: **Temporal dynamics and genetic control of transcription in the human prefrontal cortex**. *Nature* 2011, **478**(7370):519-523.
289. Merico D, Isserlin R, Stueker O, Emili A, Bader GD: **Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation**. *Plos One* 2010, **5**(11).
290. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M *et al*: **A mammalian microRNA expression atlas based on small RNA library sequencing**. *Cell* 2007, **129**(7):1401-1414.
291. Whiting PJ, Bonnert TP, McKernan RM, Farrar S, Le Bourdelles B, Heavens RP, Smith DW, Hewson L, Rigby MR, Sirinathsinghji DJS *et al*: **Molecular and functional diversity of the expanding GABA-A receptor gene family**. *Ann Ny Acad Sci* 1999, **868**:645-653.
292. Shen H, Sabaliauskas N, Sherpa A, Fenton AA, Stelzer A, Aoki C, Smith SS: **A Critical Role for alpha 4 beta delta GABA(A) Receptors in Shaping Learning Deficits at Puberty in Mice**. *Science* 2010, **327**(5972):1515-1518.
293. Roberts DS, Raol YH, Bandyopadhyay S, Lund IV, Budreck EC, Passini MJ, Wolfe JH, Brooks-Kayal AR, Russek SJ: **Egr3 stimulation of GABRA4 promoter activity as a mechanism for seizure-induced up-regulation of GABA(A) receptor alpha 4 subunit expression (vol 102, pg 11894, 2005)**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(37):13351-13351.
294. Pongs O: **Voltage-gated potassium channels: from hyperexcitability to excitement**. *Febs Lett* 1999, **452**(1-2):31-35.
295. Park KS, Mohapatra DP, Misonou H, Trimmer JS: **Graded regulation of the Kv2.1 potassium channel by variable phosphorylation**. *Science* 2006, **313**(5789):976-979.

296. Patel N, Itakura T, Gonzalez JM, Jr., Schwartz SG, Fini ME: **GPR158, an orphan member of G protein-coupled receptor Family C: glucocorticoid-stimulated expression and novel nuclear role**. *Plos One* 2013, **8**(2):e57843.
297. Voglis G, Tavernarakis N: **The role of synaptic ion channels in synaptic plasticity**. *EMBO reports* 2006, **7**(11):1104-1110.
298. Cajigas IJ, Tushev G, Will TJ, Fuerst N, Schuman EM: **The local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging**. *Neuron* 2012, **74**(3):453-466.
299. Andreassi C, Riccio A: **To localize or not to localize: mRNA fate is in 3′UTR ends**. *Trends in Cell Biology* 2009, **19**(9):465-474.
300. Matsumoto M, Setou M, Inokuchi K: **Transcriptome analysis reveals the population of dendritic RNAs and their redistribution by neural activity**. *Neurosci Res* 2007, **57**(3):411-423.
301. Loohuis NFMO, Kos A, Martens GJM, Van Bokhoven H, Kasri NN, Aschrafi A: **MicroRNA networks direct neuronal development and plasticity**. *Cell Mol Life Sci* 2012, **69**(1):89-102.
302. Manakov SA, Grant SGN, Enright AJ: **Reciprocal regulation of microRNA and mRNA profiles in neuronal development and synapse formation**. *Bmc Genomics* 2009, **10**.
303. Smalheiser NR, Lugli G: **microRNA regulation of synaptic plasticity**. *Neuromolecular medicine* 2009, **11**(3):133-140.
304. Park CS, Tang SJ: **Regulation of microRNA Expression by Induction of Bidirectional Synaptic Plasticity**. *J Mol Neurosci* 2009, **38**(1):50-56.
305. Jiang X, Zhang JT, Chan HC: **Ion channels/transporters as epigenetic regulators? — a microRNA perspective**. *Science China Life Sciences* 2012, **55**(9):753-760.
306. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.
307. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S *et al*: **Ensembl 2012**. *Nucleic Acids Res* 2012, **40**(Database issue):D84-90.
308. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.
309. Marques AC, Ponting CP: **Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness**. *Genome Biol* 2009, **10**(11).
310. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.
311. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG *et al*: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. *Genome research* 2012, **22**(9):1775-1789.
312. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**. *Genes & development* 2011, **25**(18):1915-1927.
313. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP *et al*: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals**. *Nature* 2009, **458**(7235):223-227.
314. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS: **Specific expression of long noncoding RNAs in the mouse brain**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(2):716-721.
315. Kapsimali M, Kloosterman WP, de Bruijn E, Rosa F, Plasterk RHA, Wilson SW: **MicroRNAs show a wide diversity of expression profiles in the developing and mature central nervous system**. *Genome biology* 2007, **8**(8).
316. Fliers E, Unmehopa UA, Alkemade A: **Functional neuroanatomy of thyroid hormone feedback in the human hypothalamus and pituitary gland**. *Molecular and cellular endocrinology* 2006, **251**(1-2):1-8.

317. Kuersten S, Goodwin EB: **The power of the 3' UTR: translational control and development**. *Nature reviews Genetics* 2003, **4**(8):626-637.
318. Miura P, Sanfilippo P, Shenker S, Lai EC: **Alternative polyadenylation in the nervous system: to what lengths will 3' UTR extensions take us?** *BioEssays : news and reviews in molecular, cellular and developmental biology* 2014, **36**(8):766-777.
319. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC: **Widespread and extensive lengthening of 3' UTRs in the mammalian brain**. *Genome research* 2013, **23**(5):812-825.
320. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S *et al*: **Ensembl 2014**. *Nucleic acids research* 2014, **42**(D1):D749-D755.
321. Wang D, Lu M, Miao J, Li T, Wang E, Cui Q: **Cepred: predicting the co-expression patterns of the human intronic microRNAs with their host genes**. *Plos One* 2009, **4**(2):e4421.
322. Herzer S, Silahtaroglu A, Meister B: **Locked nucleic acid-based in situ hybridisation reveals miR-7a as a hypothalamus-enriched microRNA with a distinct expression pattern**. *Journal of neuroendocrinology* 2012, **24**(12):1492-1504.
323. Sangiao-Alvarellos S, Pena-Bello L, Manfredi-Lozano M, Tena-Sempere M, Cordido F: **Perturbation of hypothalamic microRNA expression patterns in male rats after metabolic distress: impact of obesity and conditions of negative energy balance**. *Endocrinology* 2014, **155**(5):1838-1850.
324. Baulieu EE, Schumacher M: **Progesterone as a neuroactive neurosteroid, with special reference to the effect of progesterone on myelination**. *Steroids* 2000, **65**(10-11):605-612.
325. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL *et al*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription**. *Science* 2007, **316**(5830):1484-1488.
326. Nagano T, Fraser P: **No-nonsense functions for long noncoding RNAs**. *Cell* 2011, **145**(2):178-181.
327. Maziere P, Enright AJ: **Prediction of microRNA targets**. *Drug discovery today* 2007, **12**(11-12):452-458.
328. Garber M, Grabherr MG, Guttman M, Trapnell C: **Computational methods for transcriptome annotation and quantification using RNA-seq**. *Nature methods* 2011, **8**(6):469-477.
329. Vinnikov IA, Hajdukiewicz K, Reymann J, Beneke J, Czajkowski R, Roth LC, Novak M, Roller A, Dorner N, Starkuviene V *et al*: **Hypothalamic miR-103 protects from hyperphagic obesity in mice**. *J Neurosci* 2014, **34**(32):10659-10674.
330. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
331. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M *et al*: **A mammalian microRNA expression atlas based on small RNA library sequencing**. *Cell* 2007, **129**(7):1401-1414.
332. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S *et al*: **GENCODE: the reference human genome annotation for The ENCODE Project**. *Genome research* 2012, **22**(9):1760-1774.
333. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C *et al*: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs**. *Nature biotechnology* 2010, **28**(5):503-510.
334. Newman MEJ: **Networks: An Introduction**. USA: Oxford University Press; 2010.
335. Conaco C, Bassett DS, Zhou H, Arcila ML, Degnan SM, Degnan BM, Kosik KS: **Functionalization of a protosynaptic gene expression network**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109 Suppl 1**:10612-10618.

336. Rubinov M, Sporns O: **Complex network measures of brain connectivity: Uses and interpretations**. *Neuroimage* 2010, **52**(3):1059-1069.
337. Gonzalez RC, Woods RE, Eddins SL: **Digital Image Processing Using MATLAB**. In. New Jersey: Prentice Hall; 2003.
338. Rubinov M, Sporns O: **Weight-conserving characterization of complex functional brain networks**. *Neuroimage* 2011, **56**(4):2068-2079.
339. Coifman RR, Wickerhauser MV: **Entropy-Based Algorithms for Best Basis Selection**. *Ieee T Inform Theory* 1992, **38**(2):713-718.
340. Newman MEJ: **Assortative mixing in networks**. *Phys Rev Lett* 2002, **89**(20).
341. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R: **One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering**. *Cell* 2013, **153**(4):910-918.
342. Cheng AA, Ding H, Lu TK: **Enhanced killing of antibiotic-resistant bacteria enabled by massively parallel combinatorial genetics**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(34):12462-12467.
343. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER: **The next-generation sequencing revolution and its impact on genomics**. *Cell* 2013, **155**(1):27-38.
344. Mutz KO, Heilkenbrinker A, Lonne M, Walter JG, Stahl F: **Transcriptome analysis using next-generation sequencing**. *Current opinion in biotechnology* 2013, **24**(1):22-30.
345. Shapiro E, Biezuner T, Linnarsson S: **Single-cell sequencing-based technologies will revolutionize whole-organism science**. *Nature reviews Genetics* 2013, **14**(9):618-630.
346. Blainey PC, Quake SR: **Dissecting genomic diversity, one cell at a time**. *Nature methods* 2014, **11**(1):19-21.
347. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL *et al*: **Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma**. *Science* 2014, **344**(6190):1396-1401.
348. Purcell O, Jain B, Karr JR, Covert MW, Lu TK: **Towards a whole-cell modeling approach for synthetic biology**. *Chaos* 2013, **23**(2):025112.
349. Deisseroth K: **Optogenetics**. *Nature methods* 2011, **8**(1):26-29.
350. Aston-Jones G, Deisseroth K: **Recent advances in optogenetics and pharmacogenetics**. *Brain research* 2013, **1511**:1-5.
351. Stirman JN, Crane MM, Husson SJ, Wabnig S, Schultheis C, Gottschalk A, Lu H: **Real-time multimodal optical control of neurons and muscles in freely behaving Caenorhabditis elegans**. *Nature methods* 2011, **8**(2):153-158.
352. Yosef N, Regev A: **Impulse control: temporal dynamics in gene transcription**. *Cell* 2011, **144**(6):886-896.
353. Raser JM, O'Shea EK: **Noise in gene expression: origins, consequences, and control**. *Science* 2005, **309**(5743):2010-2013.
354. Kandel ER, Schwartz JH, Jessell TM, Siegelbaum, A.; S, Hudspeth AJ: **Principles of Neural Science**, 5 edn. New York: McGraw-Hill; 1981.
355. Levin JE, Miller JP: **Broadband neural encoding in the cricket cercal sensory system enhanced by stochastic resonance**. *Nature* 1996, **380**(6570):165-168.