FACULTY OF MEDICINE AND HEALTH SCIENCES

Regulatory networks in neuro-inflammatory disorders: Alzheimer's disease and major depressive disorder

Word count: 23739

Hanne Puype
Student number: 01705055

Supervisor(s): Prof. Dr. ir. Vanessa Vermeirssen

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Biomedical Sciences

Academic year: 2021 – 2022

GHENT UNIVERSITY

List of abbreviations

AD	Alzheimer's disease
APP	Amyloid precursor protein
ASD	Autism spectrum disorders
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
AUPR	Area under the precision-recall curve
AUROC	Area under the Receiver Operating Curve
Aβ	β-amyloid
BBB	Blood-brain barrier
BD	Bipolar disorder
BDNF	Brain-derived neurotrophic factor
CLR	Context Likelihood of Relatedness
CNS	Central nervous system
CRP	C-reactive protein
DAMs	Disease-associated microglia
DEGs	Differentially expressed genes
GABA	γ-amino butyric acid
GCD	Graphlet Correlation Distance
GO	Gene Ontology
GRN	Gene regulatory network
GWAS	Genome-wide association studies
KEGG	Kyoto Encyclopedia of Genes and Genomes
MDD	Major depressive disorder
MS	Multiple sclerosis
PCA	Principal components analysis
PCs	Principal components
PD	Parkinson's disease
PoLoBag	Polynomial Lasso Bagging
PPI	Protein-protein interaction
PSP	Progressive supranuclear palsy
RNA-seq	RNA sequencing
SCENIC	Single Cell rEgulatory Network Inference and Clustering
scGRNom	Single-cell Gene Regulatory Network prediction from multi-omics

scRNA-seq	Single-cell RNA sequencing
SCZ	Schizophrenia
snATAC-seq	Single-nucleus Assay for Transposase-Accessible Chromatin using sequencing
SNP	Single-nucleotide polymorphism/variant
snRNA-seq	Single-nucleus RNA sequencing
TF	Transcription factor
TLR	Toll-like receptor
TMM	Trimmed mean of M values
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
WGCNA	Weighted Gene Co-expression Network Analysis

TABLE OF CONTENTS

SUMMARY

Major depressive disorder and Alzheimer's disease are two prevalent and devastating disorders, which still lack effective treatment. Evidence is emerging that these two disorders are linked, with common pathophysiologies, such as neuroinflammation. In this master's dissertation, a systems biology approach was adopted to compare pathways and their regulators, affected in Alzheimer's disease and major depressive disorder. Gene regulatory networks were inferred with GENIE3, CLR and Lemon-Tree, from which an ensemble network was retrieved. This was achieved with publicly available RNA sequencing datasets, for each disease. The different networks inferred by the methods were compared to each other, confirming a small overlap between distinct network inference methodologies. In addition, the two ensemble networks were compared to one another. There is approximately a ten percent overlap between the edges, and the networks have a similar morphology. Further, publicly available single-cell RNA sequencing data was used to infer gene regulatory networks for Alzheimer's disease and depression, with SCENIC. With these results, it was possible to further characterize the cell types involved in the two diseases. Aberrations in immune pathways and microglial activation were found in both disorders, with several important regulators (IKZF1, IRF8, NFATC2, RUNX1, SPI1, and TAL1). All of these transcription factors have been implicated in Alzheimer's disease before, while only TAL1 has been associated with depression hitherto. Moreover, mitochondrial and proteasome dysfunctions were discovered in both disorders. These results can be used to prioritize targets for future therapy.

SOCIETAL IMPACT

Alzheimer's disease and major depressive disorder are two disorders that have a high impact on a large number of individuals, both those who are affected and their acquittances. Moreover, for Alzheimer's disease there are only disease-arresting medications, while for depression, a large number of the medications only work for some patients. Hence, new medications for these disorders are urgently needed. In addition, with the high-demanding society and increasing aging population, these disorders are still increasing in prevalence. An important reason for the lack of treatment is because of the limited understanding of the pathophysiology and risk factors for both Alzheimer's disease and depression. Thus, research is still needed to better understand these disorders. Moreover, an overlap of pathophysiology in different psychiatric and neurodegenerative disorders is emerging, indicating these disorders need to be studied together and not one disorder at a time. As such, a new target for the treatment of several diseases at a time might emerge. Thus, this research might influence the search for future medication and, as a consequence, the quality of life of patients suffering from Alzheimer's disease, depression, or both.

1. INTRODUCTION

1.1 The brain in health and disease

It is becoming more evident that the immune system is involved in the functioning of the brain, both in health and disease. Both the cells in the brain and peripheral immune cells are implicated. Microglia are brain-residing macrophages and are crucial for the development and functioning of the brain[1]. They phagocyte, protect against microorganisms, and play a crucial role in tissue maintenance and brain injury. On the other hand, microglia can release pro-inflammatory mediators, such as cytokines, causing neuronal damage[2]. Furthermore, these cytokines can damage the blood-brain barrier (BBB) and recruit pro-inflammatory immune cells, which exacerbates this neuroinflammation[2]. They can also exacerbate inflammation by interacting with astrocytes. Next to microglia, there are as well some other immune cells present within the central nervous system in physiological conditions, but only in the periphery[3]. More specifically, they are present in the blood vessels, in the meninges and at low levels in the cerebrospinal fluid. Next to microglia, the other glial cells are astrocytes and oligodendrocytes (see Figure 1). Oligodendrocytes develop and maintain the myelin sheath around neurons in the central nervous system. Astrocytes have numerous functions. They have fine processes that closely and dynamically enwrap synapses, neurons and blood vessels, and they help maintain the BBB[4]. Furthermore, astrocytes express a wide range of receptors that bind neurotransmitters and neuromodulators, and they can release gliotransmitters themselves. Hence, they monitor synaptic transmission and plasticity and are active players in information integration and processing[4]. These gliotransmitters also control metabolism, energy supply, development and inflammation. Hence, astrocytes have immunological functions as well. The different functions and characteristics of astrocytes are not constant, but depend on signals from neurons that actively coordinate and determine the molecular and functional properties of astrocytes[4]. Next to this, there are ependymal cells in the brain as well, which produce cerebrospinal fluid.

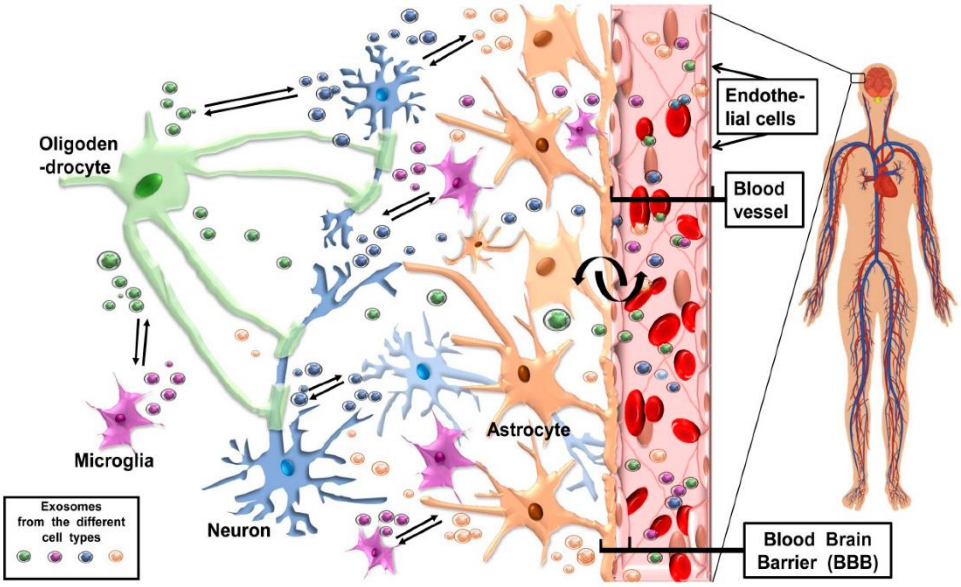


Figure 1. Overview of the cells in the central nervous system. Oligodendrocytes develop the myelin sheath, microglia have supportive and immunological functions and astrocytes have a pleiotropy of functions, such as maintaining the blood-brain barrier. (Adapted from Bavisotto et al.[5])

Neuroinflammation is a common characteristic in different neurodegenerative and neuropsychiatric disorders such as Alzheimer's and Parkinson's disease, multiple sclerosis, schizophrenia, major depressive disorder, bipolar disorder and autism spectrum disorders[1]. Neuroinflammation is characterized by infiltrating leukocytes in the central nervous system (CNS) and activation of microglia[3]. Alzheimer's and Parkinson's disease are the most and second most common neurodegenerative disorders. Alzheimer's disease is characterized by

β-amyloid plaques and hyperphosphorylated tau neurofibrillary tangles. However, it is becoming more obvious that microglia and neuroinflammation are key players in the pathogenesis[6]. Many risk genes are known for Alzheimer's disease, with the majority preferentially expressed in microglia. Parkinson's disease is characterized by the death of dopaminergic neurons in the substantia nigra and by Lewy bodies, which are deposits of α-synuclein[7]. Neuroinflammation and autoantibodies are also seen in Parkinson's patients. Major depressive disorder is characterized by decreased concentrations of serotonin, dopamine and noradrenaline in synaptic clefts[8]. Additionally, the hypothalamic-pituitary-adrenal axis is dysregulated in depression. Moreover, there is dysfunction of astrocytes and microglia, and neuroinflammation[8]. The etiology of schizophrenia is not yet well known, but it is seen that there is a strong genetic component[9]. Abnormalities in the development and differentiation of glial cells might contribute to the pathophysiology of schizophrenia. Immune activation of microglia during development might contribute to this deficit[9]. Bipolar disorder arises from genetic, environmental and epigenetic influences. Immunological alterations have been found regarding microglia, cytokines and T-cells[10]. Autism spectrum disorders are also caused by a combination of genetics, epigenetics and environmental factors[11]. They have a heterogeneous neurodevelopmental etiology. Immune system abnormalities are often seen in patients, next to other comorbidities[12]. Multiple sclerosis (MS) is a complex autoimmune disease. Polygenic risk and different environmental factors play an important role. In MS, the myelin sheath is attacked by immune cells, which results in damage to neurons and oligodendrocytes[13]. Additionally, microglial immune activation is seen. Next to autoreactive T-cells, pro-inflammatory T_H17 cells, regulatory T-cells and B-cells are as well involved[1].

In conclusion, different factors are shared by different disorders. Most are influenced by genetic and environmental factors. Microglia, astrocytes and oligodendrocytes all play a role and are as important as neurons. Autism and schizophrenia seem to have a neurodevelopmental basis. Moreover, both Parkinson's disease and MS have an autoimmune aspect. For some disorders, especially autism and schizophrenia, maternal immune activation also plays a role[1,9]. Despite being characterized by neuroinflammation, these shared pathophysiological mechanisms can result in entirely distinct disorders, indicating there are some specific processes as well (Figure 2).

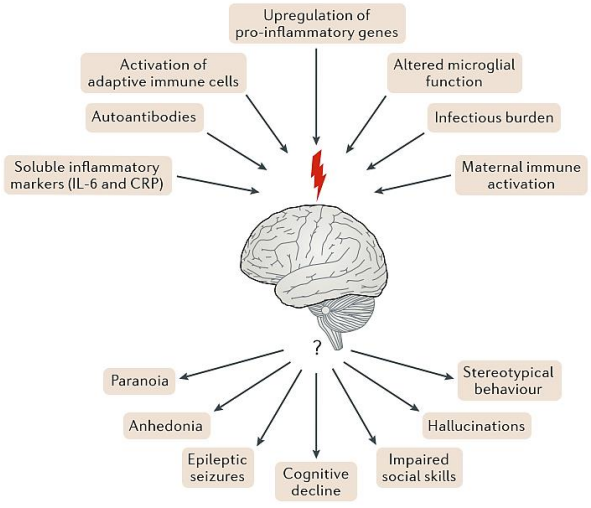
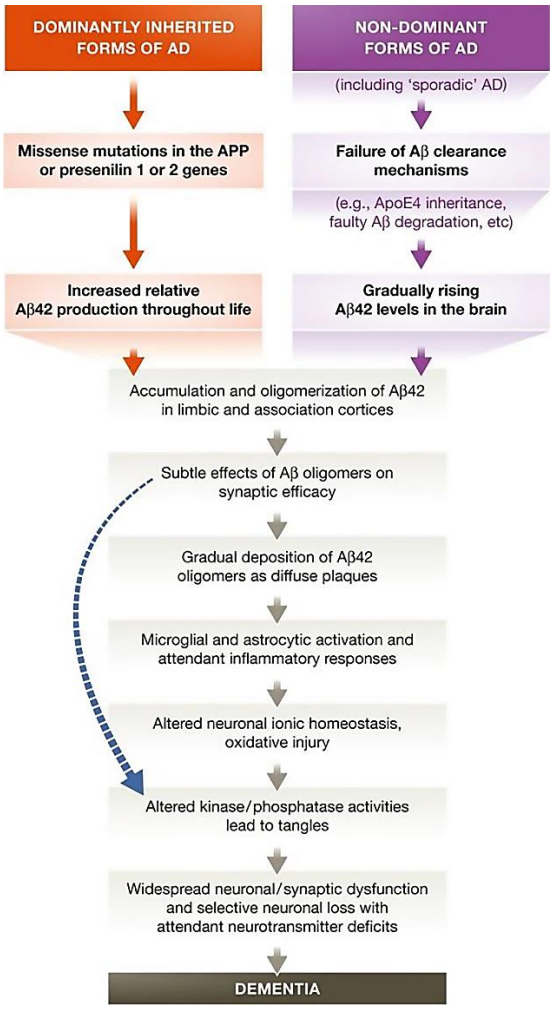


Figure 2. Common pathophysiological mechanisms of immune dysregulation in neuroinflammatory disorders and their distinct phenotypes emerging from this. (Adapted from Pape et al.[1])

1.2 Alzheimer's disease



Alzheimer's disease (AD) is the most prevalent cause of dementia[14,15]. It is mainly characterized by extracellular β-amyloid (Aβ) plaques and intracellular neurofibrillary tangles. B-amyloid is a cleaved peptide of the protein amyloid precursor protein (APP). The amyloid cascade hypothesis (Figure 3) states that by the accumulation of Aβ, through oligomers and amyloid fibrils, secondary events are induced, such as the hyperphosphorylation of tau, inflammation, excitotoxicity and oxidative stress[14]. Excitotoxicity is the overstimulation of excitatory neurons, resulting in toxicity in the post-synaptic neurons. This ultimately leads to neuronal loss and cognitive deficits. On the other hand, neurofibrillary tangles consist of hyperphosphorylated tau. Tau is a cytoskeletal protein and in its phosphorylated form, it can execute its function less efficiently. The protein mainly stabilizes microtubules and is involved in axonal transport and modulation of signaling pathways[14]. Hyperphosphorylated tau can lead to the impairment of signaling cascades, mitochondrial function and axonal transport. There is a clear link between tau accumulation and cognitive decline[14]. Even though mutations in APP and other genes associated with Aβ disposition (PSEN1 and PSEN2) are seen in familial forms of AD, Aβ disposition is not sufficient to cause AD as disposition is seen in healthy brains as well. Similarly, tau mutations alone do not cause AD[14]. However, there are many genes associated with a small increased risk to develop the disease[15].

Figure 3. An overview of the amyloid cascade hypothesis. Adapted from Lane et al.[15].

The accumulation of amyloid structures normally starts in the neocortex, before spreading to the allocortex[1] and eventually to the cerebellum[2]. Neurofibrillary tangles start in the superficial layer of the transentorhinal cortex and entorhinal cortex[14]. Next, it spreads to the hippocampus, then into the temporal region, and to the remaining cortex. Furthermore, there is symmetrical medio-temporal atrophy. An overview of different brain regions can be found in Figure 4. Alongside Aβ and tau and the associated consequences, loss of synaptic plasticity and synapses are also seen and lead to cognitive decline. In addition, there is BBB breakdown and vascular dysfunction[14]. One of the strongest risk genes for AD is the APOE gene. Apolipoprotein E is involved in lipid transport and metabolism, and in the transport of Aβ from the brain's extracellular matrix to the blood[14]. The APOEε4 allele results in the highest risk to develop AD, and lowers the age of onset, while the APOEε2 allele has a protective effect.

It is now clear that microglia also have a considerable influence on AD[6]. Normal functioning microglia actually protect against the development of the disease[6], as they are essential for

[1] The allocortex is small and includes the olfactory bulb, hippocampal formation, and entorhinal region[16].

the clearance of β-amyloid. Microglia eliminate synaptic connections with the help of complement. The complement system is normally involved in the innate immune system and excessive activation may induce neurodegeneration. It appears that complement acts downstream of Aβ. Moreover, β-amyloid aggregates can induce inflammation. Furthermore, complement activation seems to exacerbate tau pathology[6]. On the other hand, microglia can help spread neurofibrillary tangles from neuron to neuron. Disease-associated microglia (DAMs) have distinct transcriptional programs from homeostatic microglia. The expression of homeostatic genes is reduced, while the expression of neurodegenerative genes is induced[6]. DAMs localize to regions with Aβ deposition[17]. TREM2 is an important pattern-recognition receptor, involved in microglial phagocytosis, chemotaxis, survival, proliferation and inflammatory response[6,17]. Microglia lacking TREM2 are not able to fulfill these functions, resulting in exacerbation of the disease. Moreover, the majority of identified risk genes for AD are preferentially or selectively expressed in microglia and some of the proteins of these genes also bind to TREM2[6]. It is seen that the transition from homeostatic microglia to DAMs has an initial TREM2-independent phase and a secondary TREM2-dependent phase[17]. Further, it is seen that there is chronic inflammation in older brains and that they suffer from leaky BBB, resulting in the possible infiltration of immune cells[2]. Thus, aging is a risk factor by itself to develop AD. Besides the role of microglia, astrocytes are also implicated in AD. More specifically, there is astrogliosis. Activated microglia can induce a neurotoxic phenotype in astrocytes, leading to neurodegeneration[18]. This astrocyte subtype is induced by the secretion of IL-1α, TNFα and C1q, and cannot promote neuronal survival, outgrowth, synaptogenesis, and phagocytosis anymore, and induces the death of neurons and oligodendrocytes[18]. A large part of astrocytes in AD brains have this reactive phenotype, indicating they help drive neurodegeneration in the disease.

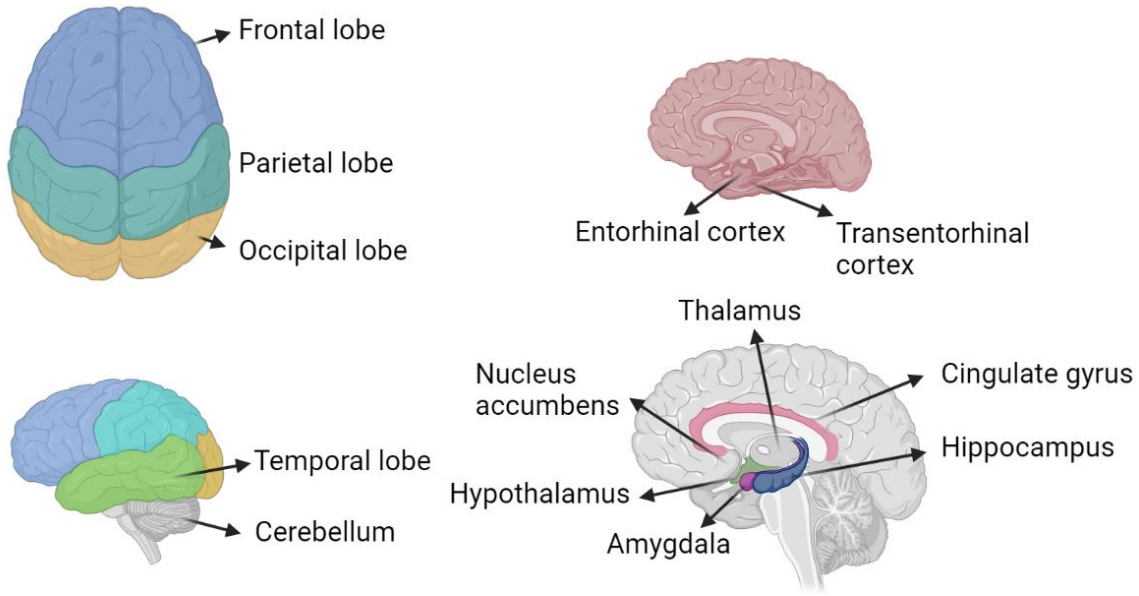


Figure 4. An overview of brain regions mentioned throughout this master's dissertation. Created with BioRender.

There is a difference in the prevalence of AD between men and women: two-thirds of patients are women, while one-third are men[14]. However, this is also driven by the longer life expectancy of women. Early symptoms of people with AD include mild cognitive impairment, with primary loss of episodic memory[14,15]. Next, topographical, language and multi-tasking difficulties arise[14,15]. Further in the disease progress, cognitive difficulties become more severe and widespread, ultimately interfering with activities of daily life. This is seen as dementia. In addition, changes in behavior, impaired mobility, hallucinations, and even seizures are possible[15]. In different neurodegenerative disorders, psychiatric comorbidities are also noticed.

In AD, for instance, depression and anxiety are frequently seen, next to sleep disturbances[14,15]. Moreover, aggression and psychosis are also observed, in later stages of AD[15].

1.3 Major depressive disorder

Major depressive disorder (MDD) is a neuropsychiatric disorder. The disorder manifests itself in several symptoms such as a low mood, anhedonia, loss of interest, decreased energy, suicidal thoughts and aches[8]. The pathophysiology is still largely unknown, but there are both genetic and environmental influences. There is large phenotypic heterogeneity in patients with depression. Moreover, it is seen that there is a gender difference in experiencing the disease: women are twice as likely to endure MDD, with an earlier onset, longer duration and higher severity[19].

There are several hypotheses for the etiology of depression. The first one is the monoamine hypothesis. It states that depression is caused by decreased concentrations of the neurotransmitters serotonin, noradrenaline and dopamine in synaptic clefts[8]. These neurotransmitters are also called monoamines, hence the monoamine hypothesis. Nowadays this hypothesis is seen as an oversimplification of the pathogenesis and it cannot explain why there is a latency period in the response of antidepressants[20]. Secondly, the neuroplasticity and neurogenesis hypotheses have their origin in the effect of stress on the hippocampus[20]. Stress activates the hypothalamic-pituitary-adrenal axis, which results in the secretion of glucocorticoids from the adrenal gland. When the levels of glucocorticoids are too high, there is negative feedback to the hippocampus to stop the secretion of glucocorticoids. However, in MDD, this negative feedback fails, resulting in atrophy of the hippocampus. The neuroplasticity and neurogenesis hypotheses agree up to this part, however, the neuroplasticity hypothesis states that glucocorticoids induce the atrophy of mature neurons in the hippocampus, while, on the other hand, the neurogenesis hypothesis states that there is a reduction of adult neurogenesis in the dentate gyrus of the hippocampus[20]. Both assumptions may be true and are interconnected. Noteworthy, depletion of noradrenaline and serotonin also reduces the proliferation of neural precursor cells in the dentate gyrus. Noradrenaline has a direct effect on proliferation, while serotonin affects neurogenesis in an indirect manner[20].

Next to these hypotheses, there are as well some disturbed pathways in the brain implicated in the pathophysiology of depression. The neurotransmitters glutamate and GABA (γ-amino butyric acid) are decreased in depressive patients[8]. Moreover, this can occur in specific brain regions. Astrocytes take up glutamate from the synaptic cleft and convert it into glutamine, which is then again transported to neurons. An increase in extracellular glutamate can be neurotoxic and is associated with inflammation and stress[8]. A reduction in the number of astrocytes can be caused by chronic stress and increases extracellular glutamate. Several studies support that in the case of MDD, there are significant reductions in the number and density of astrocytes in several brain regions[4]. Moreover, there is also astrocyte hypotrophy. Additionally, astrocytes promote adult hippocampal neurogenesis[20]. Another disturbed pathway is the catabolism of the amino acid tryptophan. Tryptophan is normally converted to serotonin. However, in depression, there is an increased conversion of tryptophan to kynurenine, which has pro-inflammatory effects[21]. Furthermore, brain-derived neurotrophic factor (BDNF) is an important factor in depression. Neurotrophins are important for the survival, growth, differentiation and plasticity of neurons[8]. Moreover, neurotrophic factors increase adult hippocampal neurogenesis[20]. proBDNF is its precursor protein, and this protein has several functions as well, which retrieve the opposite effect of BDNF. It is seen that the balance between proBDNF and BDNF is disturbed in MDD[8].

Next to the role of astrocytes, there is also mounting evidence for the implication of the immune system in depression[8]. Cytokines such as IL-1, IL-6 and TNF-α are overexpressed in the central nervous system and periphery of MDD patients. IL-1 and TNF-α lead to the activation of microglia and astrocytes, and activated microglia produce IL-6, which influences the process

of neuroprotection and neurodegeneration[8]. Activated microglia can lead to chronic inflammation. Moreover, levels of C-reactive protein (CRP) are higher in patients as well[19]. This protein is indicative of inflammation. Higher levels of T-cells and neutrophils are seen as well in the disorder[19,22]. Moreover, there is dysregulation in oxidative and nitrosative pathways, and mitochondrial dysfunction[19]. Increased cytokine levels and reactive oxygen species can lead to mitochondrial dysfunction[22]. Consequently, a neuroinflammation hypothesis is rising. It is seen that co-morbid MDD is prevalent in inflammatory conditions such as asthma, arthritis, Crohn's disease, diabetes and obesity[19,22]. This triggers a sickness behavior, which has similar features as MDD, such as anhedonia and fatigue. However, as mentioned before, depression can be highly heterogeneous in different patients, and this is the case as well regarding neuroinflammation[19], with higher inflammatory levels correlating with treatment resistance or a more severe phenotype[22]. An overview of different cells and molecules possibly implicated in depression can be seen in Figure 5.

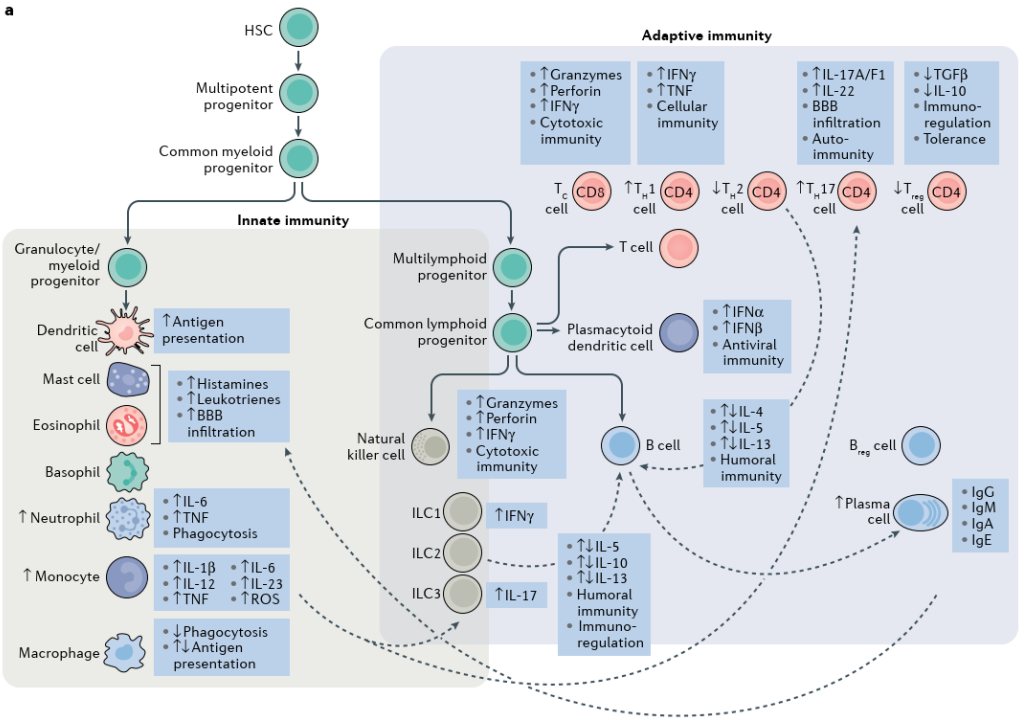


Figure 5. An overview of common alterations in the adaptive and innate immune system seen in patients with depression. Figure adapted from Drevets et al.[22].

1.4 Bulk network inference

In systems biology, cellular systems are often represented by networks. Networks consist of nodes and edges. Nodes are biomolecules such as genes, transcripts or proteins. Edges are the associations between the nodes, such as co-expression, regulation or binding[23]. Different kinds of networks exist. Protein-protein interaction (PPI) networks, also called the interactome, consist of proteins and edges that represent physical binding. Co-expression networks consist of transcripts or genes that are expressed at a similar time and in a similar context. Gene regulatory networks (GRNs) consist of transcription factors and target genes. The edges represent the regulation of the target genes by transcription factors (TFs) (Figure 6). This can be either activation or inhibition. Furthermore, networks can be directed or undirected. In directed networks, there is a clear flow of information from one node to another, such as in GRNs. In undirected networks, on the other hand, there is no causality. This is the case in PPI networks and co-expression networks. The edges can also be weighted. This indicates that there is a certain confidence that the edge truly exists or it can indicate the strength of the relationship[23]. In unweighted networks, the edge is either present or not. Modules are

subnetworks consisting of highly related and densely connected nodes. In GRNs, these are mostly co-expressed genes, regulated by the same TFs.

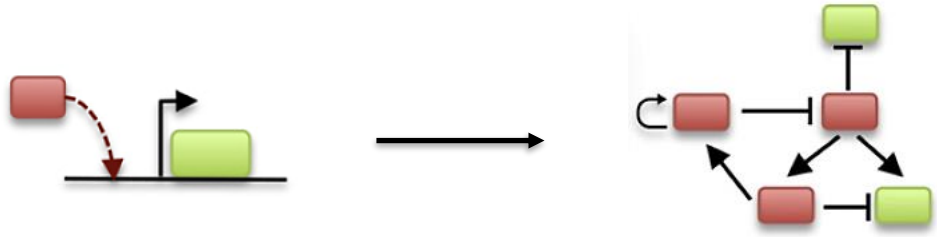


Figure 6. From transcription factor binding and activation of the expression of a target gene to a gene regulatory network. Adapted from Banf and Rhee[24].

GRNs are mostly inferred from transcriptome data. However, using transcriptome data alone has its limitations, especially for higher-order organisms. Other regulatory mechanisms are at play as well. This includes epigenetic modifications, post-translational modifications, protein interactions between different TFs and/or co-factors, non-coding RNA and enhancers. TFs need to be active, the transcriptional machinery needs to be active and chromatin needs to be accessible. Moreover, causal relationships cannot be defined based on transcriptome data alone[24]. Therefore, a multi-omics approach would be more feasible to infer a more accurate GRN, as additional layers of regulatory information, such as TF binding sites in the promotor of a possible target gene, are taken into account. However, little multi-omics data is available from the same individuals[25]. In addition, the inference of GRNs generally suffers from a high-dimensionality problem. There are far more genes than samples, which results in different possible solutions for the same data[24]. There are some properties that eukaryotic GRNs have in common[26]. Mostly, TFs regulate different genes and genes are regulated by different TFs. Moreover, TFs are regulated by their own regulators, which in turn are regulated by their TFs. GRNs are also modular and scale-free, meaning there are many genes with few edges and few hubs. Modular networks consist of highly connected clusters of nodes, with few edges connecting the different clusters (Figure 7).

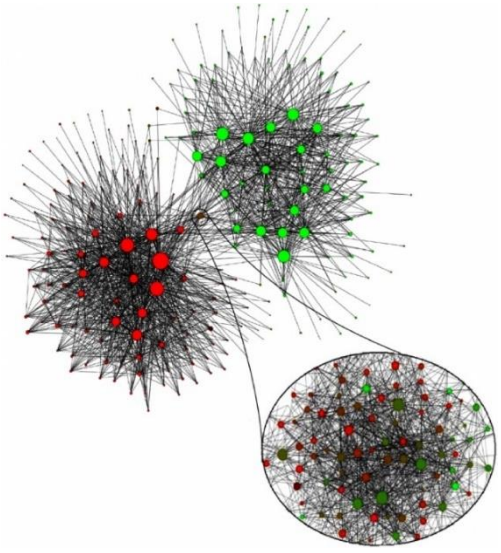


Figure 7. A modular network. Adapted from Blondel et al.[27].

It is commonly seen that transcriptional regulation and gene expression is altered in disease[28]. Network inference makes it thus possible to retrieve novel insights into disease mechanisms. There are several methods to infer GRNs, which are all based on different assumptions. Network inference can be done with correlation methods, information theory, Boolean network approaches, Bayesian network approaches, regression-based methods, differential-equation-based methods, and multi-omics integration methods[23,24]. They have their advantages and limitations, and none of them are perfect. It is actually favorable to combine several

approaches, as they complement each other and give better results combined[25,29]. In correlation methods the correlations between the expression profiles of genes are calculated[24]. This can be done for instance with the Pearson or Spearman correlation coefficient. There is no directionality, as the retrieved networks are gene co-expression networks[23,24]. One popular method is Weighted Gene Co-expression Network Analysis or WGCNA.

Information theory is based on mutual information. Mutual information can be defined as 'the amount by which the entropy of the joint distribution is reduced compared to the combined individual entropies'[23]. For these methods, discretization is mostly needed. Here again, no directionality can be retrieved[23]. This method is however able to detect non-linear interactions. Examples of algorithms using this method are ARACNE, CLR[30] and MRNET. CLR uses mutual information as a measure of similarity between expression profiles[30]. If the mutual information score between a regulator and a target gene is above a certain threshold, this is conceived as an association. CLR takes the network context into account, resulting in a better distinction between indirect and direct regulatory interactions. This is done by constructing a background normal distribution of the mutual information values for every gene pair[30]. Here it is not possible to specify the regulators beforehand.

In Boolean network approaches, genes are assumed to be either active or inactive, which results in information loss[24]. Boolean networks consist of nodes, where each node has a Boolean function. These functions indicate direction from one or more nodes to another node. Thus, the state of a node (active or inactive) depends on the states of the other nodes. A Boolean function is found for each gene. Additionally, Boolean networks are time-dependent[24]. An overview of this approach can be seen in Figure 8. Bayesian network methods are based on conditional probabilities[23,24]. The resulting network is a directed acyclic graph structure. Firstly, the structure of the model is learned and then the parameters are learned[24]. It can capture non-linear relationships as well[25]. This method is computationally expensive and is thus hard to implement for large networks[23]. Differential equation methods are based on rate equations; they quantify the rate of change of the gene expression of one gene as a function of the expression profiles of the other genes[24]. Non-linear interactions can be detected. Here an example is NonLinearODEs.

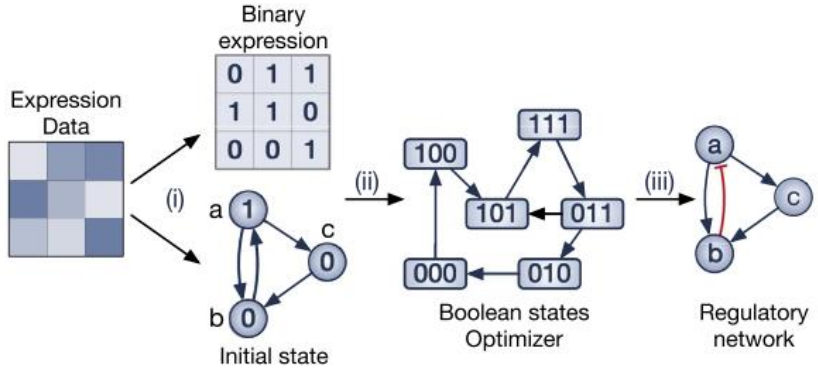


Figure 8. General workflow of Boolean network approaches. Gene expression is binarized and the initial state is inferred. The states of the different genes are then optimized and a regulatory network is retrieved with for every gene a Boolean function. Adapted from Nguyen et al.[31].

Regression-based methods look at the inference of a network as a feature selection problem[24]. They find the most predictive subset of TFs for each target gene. Here, directed edges can be retrieved[23]. There are different approaches for implementing regression. This can be for instance linear regression, logistic regression or tree-based ensemble methods. Examples here are LASSO and GENIE3[32]. GENIE3 is a tree-based ensemble method. It looks at the inferring problem as a feature selection problem: what are the genes that influence the expression profile of a specific target gene? As a result, a ranking of regulatory interactions for every gene is given[32]. Two different ensemble methods can be used with GENIE3: random

forests and extra-trees. Random forests and extra-trees are both ensemble methods based on decision trees. Different trees are grown and combined into one final tree. Bagging (bootstrap aggregation) is the method where deep decision trees are made, which are prone to overfitting and have a high variance. Moreover, in each iteration, only a part of the data (datapoints and features) is used to grow the tree. These iterations, or the forest of trees, are then averaged into the ensemble tree. The extra-trees method is based on random forests, however, in contrast to random forests, each tree is built from the original sample[32]. At each split, the tree is provided with a random sample of k features, without replacement[32]. Boosting, on the other hand, consists of growing weak learners, or shallow trees, into a strong ensemble. Shallow trees have a high bias. There are different methods on how to combine the different iterations into one final tree. GENIE3 can predict directionality, but only to some extent[32]. Other advantages are that there is no assumption about the nature of gene regulation, e.g. linear interactions, and the computation is relatively fast. Regulators can be specified when inferring the network. PoLoBag (Polynomial Lasso Bagging) is based on lasso regression and uses bagging as well[33]. In addition, polynomial features are incorporated to capture higher-order interactions (non-linear relationships).

Multi-omics approaches can be based on one of the methods mentioned above. Instead of only using transcriptomic data to infer the networks, multi-omics methods also use other data. Some methods use epigenetic and TF binding site data, other interactomics or genomic variants data. Examples are Lemon-Tree[34], MERLIN-P[35] and PANDA/SPIDER. Lemon-Tree is a module network inference method[34]. It separates the learning of modules and the assignment of regulators to modules. It can integrate different types of omics data. To find the modules, clustering is done with a model-based Gibbs sampler algorithm[34]. Different clustering permutations can be executed and are then combined into a final consensus, using a graph clustering algorithm[36]. The edge weights are equal to the frequency of the pairs of genes belonging to the same cluster in the different permutations[34]. This is done because every cluster step can be slightly different. By then taking the genes that consistently cluster together, a more robust cluster solution can be retrieved[36]. Genes that are not assigned to any cluster, are omitted. Next, regulators are assigned to each cluster. This is done by fitting an ensemble of decision trees. These can be different kinds of regulators, both continuous or discrete[34,36].

MERLIN-P, or Modular regulatory network learning with per gene information plus prior network, is a method that learns per-gene regulatory programs, but concurrently the network is constrained by a probabilistic graphical model that takes into account the modular structure of the network[26]. Thus, two genes in the same module have similar, but not identical regulators. In contrast with Lemon-Tree, the learning of the module membership and the assignment of regulators to these modules are not decoupled. In a probabilistic graphical model, there are two main components: the graph structure and the parameters[35]. They use a dependency network, where the expression levels of genes are each predicted as a function of its regulators (i.e. regression). The algorithm iterates between two steps. In the first step, the graph structure is updated, taking the current assignment of modules into account. The second step updates the module assignment, taking into account the current graph structure[35]. It starts with initial modules and iterates until convergence. Moreover, it is possible to integrate expression data with other types of regulatory data as structure priors[35]. The algorithm can integrate different types of prior networks. The prior networks can be weighted and are subsequentially combined to determine the prior probability of each edge[35]. The integrative network construction is based on a Bayesian framework.

Another multi-omics method that combines different methods to infer GRNs is KBoost[37]. It is a fast and scalable algorithm and uses kernel principal component analysis (PCA) regression, boosting and Bayesian model averaging. A prior network is included in the algorithm, which is based on ChIP-seq data. Different weights can be given to this prior network. For every gene a model is fit that predicts its expression, using the kernel PCA of the expression levels of a subset of TFs[37]. The boosting is implemented by fitting a new model to each gene expression's

residuals and selecting the TFs with the highest posterior probability per gene and then updating the predictions. Thus, in different iterations, different TFs are selected. With the Bayesian model averaging, different models are compared, and the probability that a TF regulates a gene is estimated. This model is then combined with the prior, and as output, the probability of each TF regulating each gene is given[37].

CLR and GENIE3 were both included in the benchmark study by Marbach et al. where they compared 35 different network inference methods[29]. This was done with the DREAM5 challenge. This is a well-known GRN inference challenge where researchers can run their algorithms on benchmark datasets. Combining different methods gave as good or better results than the top-performing methods in their benchmark[29]. This was as well more robust than using only one method. The performance increased with applying more methods, and with increasing the diversity of the used methods.

1.5 Single-cell network inference

Single-cell data is increasingly being used to answer research questions. Predominantly single-cell RNA sequencing (scRNA-seq) and ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) are performed, but other efforts are done as well, for example single-cell proteomics. Different cells in the same tissue have distinct functions and expression patterns. Thus, when doing bulk RNA sequencing (RNA-seq), for example, cell-specific signals are averaged and dominated by the bulk signals. When doing a single-cell analysis, cell-specific signals are picked up. In the brain, it is a large advantage to analyze cell-specific signals, as not only do the different glial cells have separate functions, but different neurons have distinct functions as well. Another advantage is that fewer patients are needed, as every cell is now seen as an individual sample. This can alleviate the high-dimensionality problem. A disadvantage of single-cell data is the fact that the signals are not strong, as they only come from a single cell. However, the signal can be increased by pseudo-bulk analysis or the aggregation of the reads of several cells from the same cell type or state. Because of the inherently different kinds of data, new methods had to be developed to analyze scRNA-seq data, as the bulk methods are not always convenient to use. Hence, new methods for network inference have been developed specifically for single-cell data. Multi-omics integration is as well increasingly done with single-cell data. Some researchers have benchmarked the performance of different single-cell network inference methods. A first paper has benchmarked GRN inference algorithms from scRNA-seq data[38]. They compared GENIE3, PPCOR, LEAP, SCODE, PIDC, SINCERITIES, SCNS, GRNVBEM, SCRIBE, GRNBoost2, GRISLI and SINGE. They tested the different methods on simulated datasets from synthetic networks, datasets from curated Boolean models from the literature and five experimental scRNA-seq datasets[38]. The results were examined using the area under the precision-recall curve (AUPR), early precision (fraction of true positives in the top-k edges), stability of the results, by analysis of the network motifs, and scalability. PIDC, GENIE3 and GRNBoost2 were the top-performing algorithms[38]. A substantial number of methods had a performance close to a random predictor. Another benchmark paper compared the following GRN inference methods: Boolean Pseudotime, BTR, SCNS, Inference Snapshot, SCODE, SCOUP, Empirical Bayes, Information Measures, NLNET, SINCERA, SCENIC, LEAP, SINCERITIES, SCIMITAR, and SCINGE[31]. The researchers used stimulation data and studied the AUROC (Area Under Receiver Operating Curve) of the different methods with different numbers of genes and with different levels of sparsity. Overall, SCENIC has the highest accuracy in most simulation studies, while LEAP and NLNET are the fastest methods, and SCOUP is the most stable method[31]. Here again, some methods only performed as well as a random predictor.

SCENIC (Single Cell rEgulatory Network Inference and Clustering) consists of a workflow in which the first step is to infer networks with GRNBoost2 or GENIE3[39]. Next, modules are identified in which the TF's binding motif is significantly enriched in the target genes, with RcisTarget. Lastly, AUCell scores the activity of these regulons in each cell, which is then binarized to be either active or inactive. Cell states are then predicted based on shared activity

between the cells. GRNBoost2 is based on GENIE3, but instead of using a bagging method, stochastic gradient boosting is used to train a strong model[39]. Stochastic gradient boosting indicates that at each iteration a randomly selected subsample of the data is used, increasing the accuracy of the model[40]. In gradient boosting, the current iteration is built using the error from the previous iteration. Thus, in contrast to random forests, the different iterations are not independent.

Another upcoming, promising technology is spatial single-cell analysis. Here, the RNA in the cells is sequenced and this can subsequently be coupled back to its position in the sample. More and more cells can be sequenced with scRNA-seq and the resolution of spatial analysis is still increasing. The single-cell omics field is rapidly evolving.

1.6 Comparison of networks

Tantardini et al. have made an overview of different methods that can be used to compare networks[41]. Network characteristics are often used to get a broad overview of the topological nature of a network. Examples of network characteristics are the degree, correlation coefficient, density, diameter, edge and node betweenness, number of connected components, and distance. The average degree indicates the average of the edges each node has. The correlation coefficient is a measure of how well the neighbors of a node are connected to one another. It is the number of edges between the neighbors of a node divided by the total number of possible edges[42]. The density is the ratio between the edges in the network and the total number of possible edges. The diameter indicates the longest shortest path of the network between any two nodes. The average edge betweenness indicates the average of all paths that pass through a certain edge. Similarly, the node betweenness signifies the number of paths that pass through a certain node. This is also called the betweenness centrality[42]. Connected components are the number of components or subgraphs in which each pair of nodes is connected with each other via a path. Lastly, the average path length or distance[42] is the average of all the path lengths between all the nodes in the network.

GCD-11 was the best performing method amongst the undirected methods[41]. The method is based on graphlets, which are 'small, connected, non-isomorphic subgraphs of large networks'[41]. Mostly, graphlets contain up to five nodes. Nodes in graphlets are called orbits, and each distinct orbit gets a number. In one graphlet, two or more orbits can be the same, i.e. there is no possible distinction between them. These are automorphism orbits. For instance, in a chain of three nodes, the two outer nodes are the same. The graphlets are numbered as well (G_0 to G_{29} for up to five-node graphlets). Moreover, some orbits are redundant, which means that their count in the network can be derived from the counts of the other orbits[43]. GCD (Graphlet Correlation Distance) gives the highest accuracy with up to four-node graphlets[43]. This method has eleven non-redundant orbits, and is thus called GCD-11. For each node in the network, a graphlet degree vector is created[43]. This is a vector containing the count for each of the possible orbits, for this node. This is then combined into a matrix with the number of rows equal to the number of nodes in the network, and the number of columns equal to eleven (possible orbits). Next, the Spearman's correlation coefficient is calculated between all pairs of columns[43]. This results in a symmetrical 11x11 matrix, called the Graphlet Correlation Matrix. As such, each network can be represented as an 11x11 matrix. With these matrices, the distance can be computed between two networks. This is done by taking the Euclidean distance of the upper triangle values of the two Graphlet Correlation Matrices[43]. This distance is termed the Graphlet Correlation Distance.

1.7 Network inference on omics data from neuroinflammatory disorders

There is still a lot of research necessary to improve the inference of GRNs, but they are useful to generate biological hypotheses and prioritize follow-up experiments. It is important to keep in mind that GRNs need to be experimentally validated, as in silico methods are not sufficient to prove a certain regulatory pathway. Researchers have already conducted network inference

on neuroinflammatory disorders. Chew and Petretto[2] have made an overview of transcriptional networks inferred from Alzheimer's samples. They describe several papers where network inference was done of gene co-expression networks or GRNs with microarray, RNA-seq, scRNA-seq, and some multi-omics data. Probably the most used method until now to infer networks is WGCNA. However, with this method co-expression networks are retrieved, not GRNs. Moreover, many researchers only infer networks on differentially expressed genes. By doing this, a substantial part of the data, that might be interesting, is not used. In another paper, researchers have used WGCNA to look at the degree of overlap of transcriptional dysregulation between autism (ASD), depression, schizophrenia (SCZ), bipolar disorder (BD) and alcoholism[44]. They used microarray datasets and generated RNA-seq for three of the five disorders. They found the largest transcriptome correlation between SCZ and BD. Next to doing a differentially expressed genes analysis and WGCNA, they also looked at single nucleotide variants (SNPs). They found significant correlations between SNP-based genetic correlations between diseases on the one hand and their corresponding transcriptome overlap on the other hand[44]. This indicates that the gene expression changes are partly coupled to genetic variation.

Another research group has implemented WGCNA as well, from samples of individuals with SCZ, ASD, Parkinson's Disease (PD), AD, BD, MDD, pathological aging, and progressive supranuclear palsy (PSP)[45]. They used bulk RNA-seq samples. The number of samples they used for each disorder was highly different; 906 for AD versus 29 for PD. They performed differential gene expression analysis. With the differentially expressed genes (DEGs), they executed gene enrichment analysis. Functions related to the immune response were the only recurring results among different disorders[45]. Here again, the largest overlap was between SCZ and BD. The researchers found different overlaps between conditions in different brain regions[45]. There was also a correlation between AD and ASD, and between AD and SCZ. With WGCNA, they found that neuronal modules were downregulated in AD, PD, ASD, SCZ, and BD. An oligodendrocyte module was upregulated in all conditions except for pathological aging and PSP. A microglia-associated module was upregulated in AD, PD, pathological aging, and autism. Moreover, it was enriched for genes involved in the immune response[45]. This is logical, as microglia are immune cells. An astrocyte module was upregulated in AD, PD, pathological aging, ASD, SCZ, and BD.

A system-level analysis of different neurodegenerative disorders was executed in another paper[46]. They used microarray data from AD, PD, Huntington's disease, SCZ, amyotrophic lateral sclerosis, and MS patients and control samples. The samples were taken from different brain regions for each disease. The number of samples they used was limited; around ten patients and ten control samples were used for every disease. They wanted to identify common pathways and factors involved in the development and progress of several neurodegenerative disorders[46]. They predicted the core GRNs in each disorder. Firstly, they identified DEGs between each disease and control samples. There were few common genes between diseases[46]. Next, they used the DEGs list to make PPI networks with the STRING database. Moreover, they used Enrichr to determine the TFs that regulate the DEGs. They constructed TF-gene networks and integrated data from STRING into these networks. Subsequently, they determined central genes and TFs and used these to construct the core GRNs. Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genomes) were used for functional enrichment analysis. The largest overlap of DEGs was observed between Huntington's and PD. After functional enrichment, they noticed that most genes were involved in cardiovascular and metabolic terms, followed by immune, neurological and pharmacogenomics terms[46]. ATF3, SOX2 and JUN were hub TFs for the DEGs of AD. UMPS and CDK1 were two hub DEGs in AD. SLC14A1 was found to be implicated in several diseases, both in this study and in other literature[46]. They found a large overlap in their genes and genes in literature, but they also found some genes that may be implicated in pathology, but have never been observed before.

Further, other researchers investigated several psychiatric and neurodegenerative disorders[28], but did this in another manner than in the previous paper. They reconstructed a GRN for the human brain by integrating brain-specific DNase footprinting and TF-gene co-expression[28]. For the co-expression, they utilized microarray expression profiles from the Allen Human Brain Atlas. Both Pearson correlation and lasso regression was adopted for co-expression. Then they retrieved the DEGs from transcriptomic data (RNA-seq and microarray) from SCZ, BD, depression, AD and autism patients and controls[28]. Here, the samples were all retrieved from the prefrontal cortex. Next, they identified TFs whose target genes were enriched in these DEGs. Their goal was to predict key TFs that regulate transcriptomic changes in the disorders, as well as to look at disease-associated SNPs that disrupt regulator binding sites[28]. They found no key regulators for MDD and 78 for AD. Some key regulators were associated with genetic risk for the same disease. These were MEF2C, GLIS3, TFEB, and NR3C2 for AD[28]. Target genes of MEFC2 were enriched for neuron-specific genes. Moreover, they saw that neuronal networks were often downregulated, while microglial networks were upregulated in AD[28].

1.8 Aims of this master's dissertation

The objective of this master dissertation is to find the distinct and common pathways between the neuroinflammatory disorders AD and MDD, and their regulators. AD can be seen as a representative of neurodegenerative disorders, while depression can be seen as a representative of psychiatric disorders. This has been done by inferring GRNs using different methodologies. For each method, the top 100 000 edges were retrieved. Using rank aggregation, an ensemble network per disorder was constructed from these networks. The ensemble networks of AD and MDD were compared to each other with different metrics, to see whether topologically similar networks were retrieved. Next, modules were constructed and functionally analyzed. In addition, single-cell RNA-seq data was used to infer GRNs with SCENIC and further characterize the cells that are implicated in the disorders, and to compare the two disorders.

The host lab of Prof. dr. ir. Vanessa Vermeirssen (Lab for Computational Biology, Integromics and Gene Regulation (CBIGR)) aims to acquire a functional understanding of gene regulation and signaling at a systems level in complex diseases. The lab is internationally recognized in GRNs and multi-omics data integration; developing and applying high-throughput methods for experimental GRN mapping, and benchmarking, and data integration methods for computational GRN inference. The host lab has shown that different network inference methods reveal complementary aspects of the underlying GRNs, and that integrating different omics data provides a more accurate, multi-modal view of gene regulation[47,48].

2. METHODS

2.1 Retrieval and preprocessing of data

Different criteria were taken into account when searching RNA-seq datasets. Firstly, the data had to originate from human post-mortem brain samples. Secondly, each dataset had to contain at least 20 samples and control samples had to be included as well. The samples may not have been enriched for certain cell types. The prefrontal cortex was the favored brain region, as a large amount of studies sample from the prefrontal cortex. Moreover, the prefrontal cortex has been implicated in several psychiatric, neurodevelopmental and neurodegenerative disorders[28]. Lastly, there needed to be a similar number of samples for each disorder.

The gene counts and metadata files of the different datasets were loaded into R (version 4.0.3). Different datasets of the same disease have been merged to make a compendium per disease. The *edgeR* package was used for preprocessing. Firstly, the features were filtered by only keeping the genes with more than one count in at least five samples. Then the trimmed mean of M values (TMM) normalization was done[49]. This method uses a weighted trimmed mean of the log expression ratios. For this, gene-wise log fold changes and 'absolute' expression levels are used. A trimmed mean indicates the mean after removing the upper and lower x% of the data[49]. The weights account for the mean-variance dependency. TMM normalization assumes that most genes are not differentially expressed[49]. Next, the normalized counts were logarithmically transformed (prior count of one). Batch effects and outliers were detected with multidimensional scaling and hierarchical clustering. For merged datasets, a batch correction has been executed, with the *removeBatchEffects* function from *edgeR*. Subsequently, highly variable genes were selected. Next to the selection of highly variable genes, only protein-coding genes were selected for further analysis. After the selection of the highly variable genes, regulators have been added again. Regulators indicate TFs that bind DNA and regulate the expression of their target genes. Lovering et al. have manually curated a list of human TFs, using several sources[50]. This list was used to define the regulators. It contains 1455 TFs in total. Lastly, scaling was done.

2.2 Bulk network inference methods

Two networks were made with every method, one with the AD dataset, and one with the MDD dataset. GENIE3[32] is implemented in the R package *GENIE3*. Random forest was used to infer the networks. The other parameters were set to default (number of regulators selected at each tree node: sqrt(total number of regulators); 1000 trees). CLR is implemented in the *minet* package in R (v.4.0.3)[51]. It is possible to use different estimators to calculate the mutual information. The empirical estimator was the default at the time the paper was written. However, they mention that this entropy estimator is biased. The Miller-Madow estimator reduces this bias, thus, this estimator was used to infer the networks. These estimators were designed to take discrete values. The equal frequency discretization was used for this[51]. Lemon-Tree is implemented in Java as a command-line program[34]. The latest version (v3.1.1) was used to infer the networks. The clustering was done for 100 permutations and there are a minimum of ten genes per module. Next, regulators are assigned to each cluster, which were TFs here, from the list from Lovering et al.[50]. MERLIN-P[35] was tried as well. It is implemented as a command-line program, with code written in C and C++, available on their GitHub. For MERLIN-P a prior network is needed, hence, a weighted directed network from Marbach et al.[52] has been utilized from the adult frontal lobe. They have constructed 394 cell- or tissue-specific regulatory networks and made them freely available (syn4956655). This region was chosen, as the expression datasets are from the prefrontal cortex.

PoLoBag[33] is implemented in Python. A script must be run, where only the file names have to be changed, and if desired, different parameters can be changed as well. The default parameters were used, except the Lasso regularization parameter was changed to 0.2 instead of 0.1. KBoost[37] is implemented in the R package K*Boost* and uses a prior network as well.

This prior network is included in the function *KBoost_human_symbol*, which was used for the network inference. In addition, TFs are defined beforehand, but another resource is used than the list from Lovering et al. VIPER infers the activity of proteins with the expression pattern of its target genes[53]. VIPER is implemented in the R package *viper*. In the first step, a prior network must be made, which the developers have made with ARACNE (R package *minet*), which was tried as well. In the next step, this prior network has to be changed to a regulon. Two groups of samples are compared to each other. Then enrichment of each regulon on the gene expression signature of these groups is calculated, using the analytic Rank-based Enrichment Analysis algorithm[53]. This enrichment is then compared to a null model to determine statistical significance.

2.3 Ensemble networks

From the different methods, an ensemble network was created. This was done by rank aggregation. Rank aggregation can be performed with different methods. In the paper from the benchmarking with the DREAM5 dataset[29], they used average rank aggregation to make the ensemble networks. The *TopKLists* R (v.4.1.3) package was used with the *Borda* function for average rank aggregation, using the results from the different methods. Next, modules were retrieved with the Jaccard similarity index and k-medoids clustering. The overlap in the predicted regulators of all the genes was calculated using the Jaccard index. These indices were then used to allocate each gene to a module with k-medoids. The R package *cluster* was utilized for this, with the *pam* function. K-medoids clustering is more robust than k-means clustering as the median is used instead of the mean of the clusters. Hence, one gene is used as a representative of its cluster. In k-means/k-medoids, firstly, the k cluster centroids are randomly assigned. Then for every gene, the nearest centroid is calculated and the gene is assigned to this cluster. Next, the centroids are updated to be the average/median of the assigned cluster points. These steps are repeated until convergence. Further, the regulators were assigned to each cluster. A maximum of ten regulators was chosen to be allocated to each module. TFs were ordered by the number of genes they regulate in the module. In addition, the TFs had to regulate at least half of the genes in the module.

The modules from the Lemon-Tree networks and the ensemble networks were visualized in Module Viewer[47]. Module Viewer is a Java program in which the expression of different modules can be visualized, together with its regulators and annotation data. Network visualizations were created with the *igraph* package in R or with Cytoscape[54].

2.4 Functional characterization

Functional enrichment analysis was done with the *enrichR* package in R. The databases used for the enrichment of the Lemon-Tree modules were Gene Ontology Biological Process (2018), GO Molecular Function (2018) and KEGG (2019). For the functional enrichment analysis of the ensemble networks, the same databases were used as above, however, the most recent versions were used this time. These were from 2021. Moreover, some additional databases have been used as well: Reactome_2016 and WikiPathway_2021_Human.

2.5 Comparison of networks

The different networks retrieved from the different methods were compared with Venn diagrams and network characteristics. This was done with the *BioVenn* and *igraph* R (v.4.0.3) packages, respectively. For each disease, the networks were compared that were retrieved from the different methods. The *igraph* package was used as well for the network characteristics of the ensemble network. Here, the network characteristics were used to compare the ensemble networks of AD and MDD. Next to this, the ensemble networks were compared with distance measures. The first distance measure used was the Jaccard distance. The Jaccard similarity index is calculated as the intersection of the edges, divided by the union

of the edges[41]. The intersection indicates the common elements, while the union indicates the edges present in either of the two networks. The Jaccard distance is then calculated as one minus the similarity. Hence, this is a distance measure that gives a broad overview of the number of edges that are shared between two networks. Next to this, the GCD was used, more specifically, GCD-11. This was done by utilizing the Python scripts and the orca.exe from the authors. However, some changes had to be made to the Python scripts, probably because the scripts were written in an older Python version. The networks need to be in Leda format, which was done with the *write_graph* function from the *igraph* package in R. GCD11 is an undirected method. Hence, the 'duplicated' edges had to be omitted first (A-B versus B-A), which was done in R as well. In the first step, the Graphlet Degree Vector matrix is computed for each network. In the next step, the GCD is computed between the networks.

2.6 Further characterization with single-cell RNA-seq data

Next to using bulk RNA-seq datasets, scRNA-seq datasets were used as well. These had to be originating from post-mortem brain samples. Preferably, these had to be from the same region as the bulk datasets, i.e. the prefrontal cortex. Both neurons and glial cells had to be included, as all cells can be implicated in the pathology. Similarly as in the bulk datasets, control samples need to be included and there have to be a similar number of samples (cells) for each disorder. The publicly available files were preprocessed in R (v.4.0.3). The *Seurat* package was used for quality control, preprocessing and visualizations. Cells with too low or too high UMI (unique molecular identifier) counts were omitted. Too few counts are mostly due to empty droplets, while too many counts can indicate there were two cells in one droplet. Next to this, only protein-coding genes were selected. Subsequently, the dataset was logarithmically transformed (*NormalizeData* function) and the highly variable genes were selected. Next, the regulators were added again. Here again, the list of Lovering et al.[50] was used. Lastly, scaling was executed to make plots (*ScaleData* function).

To visualize the cells in a low-dimensional space, firstly PCA was done. This was done with fewer genes than the genes selected for further analysis with SCENIC. Next, the number of principal components (PCs) to keep was evaluated with an elbow plot. With the selected PCs, the cells were visualized with UMAP (Uniform Manifold Approximation and Projection) and t-SNE (t-distributed Stochastic Neighbor Embedding) plots. Both methods are frequently used with single-cell datasets. UMAP is better able to keep the global structure of the dataset. This was done by running *FindNeighbors*, *FindClusters*, *RunUMAP,* and *RunTSNE*. Firstly, a K-nearest neighbor graph is constructed, based on the Euclidian distance in the PCA space[55]. Next, modularity optimization techniques are used to cluster the cells[55]. This is done with the Louvain algorithm by default.

In addition, SCENIC was used to infer GRNs from the single-cell data. SCENIC can be implemented in both R and Python. GRNBoost2 is faster than GENIE3[39] and can only be used in Python for inference. GENIE3 can be run in both Python and R. It was run by Joke Deschildre, a PhD student in the lab of Prof. Vermeirssen, as she has experience with this method. It was run in Python (*pySCENIC*). GRNBoost2 was used for the network inference, with the *arboreto* package.

3. RESULTS

3.1 Overview of the expression data

A GitHub repository was made for the scripts executed in this master's dissertation (see Addendum 2). RNA-seq data were mostly retrieved from the NCBI Gene Expression Omnibus. GSE174367/syn22130832 contains bulk RNA-seq samples from 48 healthy controls and 47 patients with AD[56]. The samples were taken from the prefrontal cortex. In addition, GSE101521 and GSE80655 contain samples from 53 people with depression and 53 healthy controls. They

both contain samples from the dorsolateral prefrontal cortex. GSE101521 contains 30 samples of patients with MDD and 29 controls[57]. GSE80655 contains 23 samples from people with MDD and 24 control people[58]. An overview of the datasets can be seen in Table 1.

Table 1. Overview of the bulk and single-cell datasets.

Disorder	Accession	Data type	Brain region	Number of disease samples	Number of control samples
AD	GSE174367 syn22130832	RNA-seq	Prefrontal cortex	47	48
MDD	GSE101521	RNA-seq	Dorsolateral prefrontal cortex	30	29
MDD	GSE80655	RNA-seq	Dorsolateral prefrontal cortex	23	24
Healthy	syn4956655	Prior network	Frontal lobe		
AD	GSE174367 syn22130832	snRNA-seq	Prefrontal cortex	8	11
MDD	GSE144136	snRNA-seq	Prefrontal cortex	17	19

It is hard to find readily available multi-omics data of the same individuals. Up to today, there are still more microarray than RNA-seq datasets available. Subsequently, scRNA-seq datasets are even more sparse. Most available epigenetic data, such as ATAC-seq or DNase-seq, has been retrieved from healthy individuals and not patient samples. As such, multi-omics data were not found for AD and MDD.

In the AD dataset, six outliers out of 95 samples were omitted. These consisted of four AD samples and two controls. Similarly, in the first MDD dataset, one outlier sample was deleted. This was a depressive sample. After the selection of highly variable genes and protein-coding genes, 9210 genes were left in the AD dataset and 8660 in the combined MDD dataset. 514 regulators were added again to the Alzheimer dataset and 557 to the depression dataset. These TFs were expressed in the samples and the list of Lovering et al.[50] was used to know which genes are seen as TFs. Even though these TFs are not highly variable over the different samples, they can still have an important function in the network, for instance by being constitutively active. This ultimately resulted in 9724 and 9217 genes in the AD and MDD dataset, respectively. There are a total of 1045 regulators in the AD dataset and 1101 in the MDD dataset.

3.2 Bulk networks inferred through different methodologies

Firstly, GENIE3 was used to infer the two networks, one for AD and one for MDD. The list of Lovering et al.[50] was used to define the regulators. Here, a total of 10.121.643 edges were found in the AD network and 10.146.816 edges in the MDD network. Next, the edges were filtered to retrieve the top 100 000 edges, because too many edges are not feasible to work with. Moreover, these are the most significant edges. Next, CLR was used to again infer two networks. After inferring the networks with CLR, the edges were filtered to contain at least one regulator. If the edge was between the gene of a TF and a non-TF gene, then only the edge from the regulator to its target gene was kept. On the other hand, if an edge connects two TFs, then both edges were kept, as the direction is not known. Edges where the weight was zero,

were also omitted. Ultimately, this resulted in 5.592.222 regulatory edges in the AD network and 5.382.600 in the MDD network. Here the top 100 000 edges were selected as well for further analysis.

In Lemon-Tree, the TF list of Lovering et al.[50] was used again to define the regulators. The expression of these TFs in the Alzheimer and depression datasets was used as regulatory information. In the AD network, 155 modules were found, similar to the 156 modules found in the MDD network. The top one percent regulators were used to visualize the modules and their regulators with Module Viewer. Next, the modular output of Lemon-Tree was converted to an edge list, similar to those retrieved with GENIE3 and CLR. This was done by assigning each regulator to each target gene in the respective module. As a default in Lemon-Tree, the top one percent regulators are retrieved in the networks. However, when utilizing the top one percent regulators (687 for AD, 818 for MDD), there were only 45 161 edges in the AD network and 44 653 in the MDD network. Moreover, Lu et al. have dissected Lemon-Tree by assessing the network performance using different parameters and data[59]. In the paper, they recommend not using the top one percent regulators, which results in information loss, but to use at least the top 30%. As the top 100 000 edges were selected in the previous methods, this was done here as well. There were a total of 68 698 (AD) and 81 792 (MDD) regulators assigned to all modules. Lemon-Tree also assigns regulators randomly to modules. This can be used to compare these scores to the scores of the assigned regulators. However, as Lu et al. mention, this random list still has a reasonable performance[59]. When filtering with the highest random score (3.1 for AD and 3.2 for MDD), there were still less than 100 000 edges. As such, a score of 2.98 and 2.58 was used to filter the weight in respectively the networks of AD and MDD. There were some auto-regulatory edges in the network, because the expression of the regulators was used as well when retrieving the modules. As such, some regulators were assigned to the module and assigned as a regulator of the same module. Thus, these edges were deleted from the network, before selecting the top edges. The other networks have no auto-regulatory edges.

MERLIN-P was tried as well. However, there was only one module made and as such, only one regulator was assigned to different target genes. The method was first executed without an initial module assignment. Subsequently, an initial module assignment was made with k-medoids clustering with the R package *cluster*. The number of clusters was set to the number of clusters retrieved by Lemon-Tree (155 for AD, 156 for MDD). However, with the initial cluster assignment, the output of MERLIN-P was still not as desired, as it was the same as before. There were 5018 genes assigned to one module for AD and only 28 genes to one module for MDD. Perhaps the prior network and the datasets were too large for MELRIN-P.

Next to the methods described above, PoLoBag, KBoost and VIPER were tried as well. PoLoBag was too computationally expensive; it was still running after 72 hours. VIPER is outdated, many people don't get it running and there was uncertainty about the output. A prior network with ARACNE was made, but it was not possible to convert it to a regulon to use in VIPER. KBoost is a fast method. However, the number of retrieved edges was too little compared to the other methods. For AD there were 5425 edges inferred and for MDD 3551 edges. Other possible methods to infer GRNs were either not feasible or were too similar to the already used methods. Some were not feasible due to not having enough data (e.g. epigenetic data), being implemented in MATLAB, or being too hard to implement e.g. because of not enough information provided about how to run the method.

3.3 Analysis of method-specific networks

The overlap between the networks retrieved by the different methods was investigated by making Venn diagrams of the edges. Next to this, Venn diagrams were made to see the overlap in the top 100 regulators of the different networks, for both disorders. These were the top 100 regulators with the most 'out' edges (out-degree). The Venn diagrams were made with the R package *BioVenn*. As seen in the diagrams, there is not much overlap between the edges of

the different methods (Figure 9). There are as well some differences between the two disorders. It must be noted that the weights of the edges were not taken into account here. There is, however, a substantive overlap between the top 100 regulators (Figure 10). There are 45 top regulators in common for the AD networks and 40 for the MDD networks. These are 170 total regulators of all methods for AD and 171 for MDD. Of these, there are 62 regulators in common between AD and MDD. To see whether there are any substantive differences between the number of nodes and the number of TFs for each network, these were as well verified. An overview of the values can be seen in Tables 2 and 3, for AD and MDD respectively. There are more regulators and fewer target genes in the MDD networks compared to the AD networks. This is because of the input expression data (see 3.1). Lemon-Tree has a substantive lower number of nodes, due to the modular output. Not every regulator could be assigned to regulate a module and not every gene could be assigned to be part of a module.

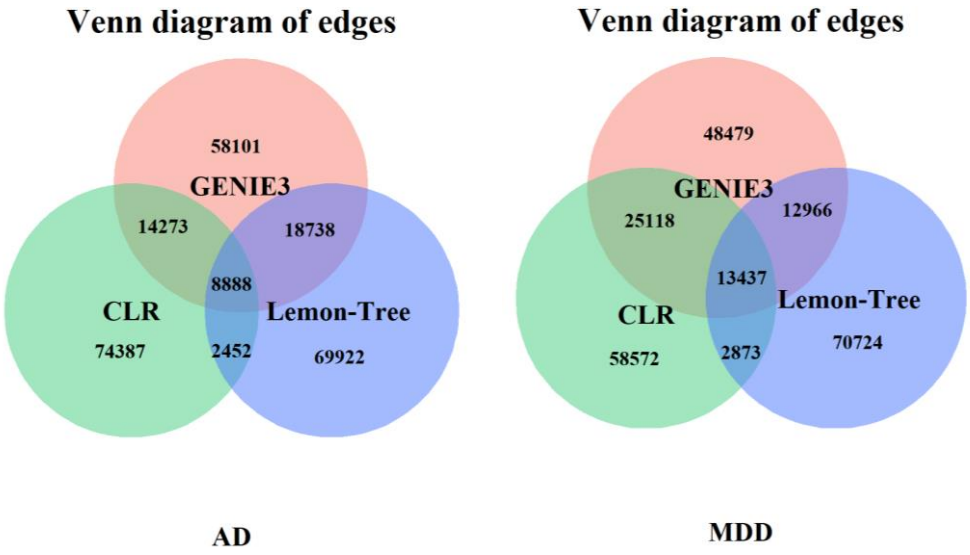


Figure 9. Venn diagrams of the overlap of the edges for the networks retrieved by CLR, GENIE3 and Lemon-Tree, for Alzheimer's disease (left, AD) and depression (right, MDD).

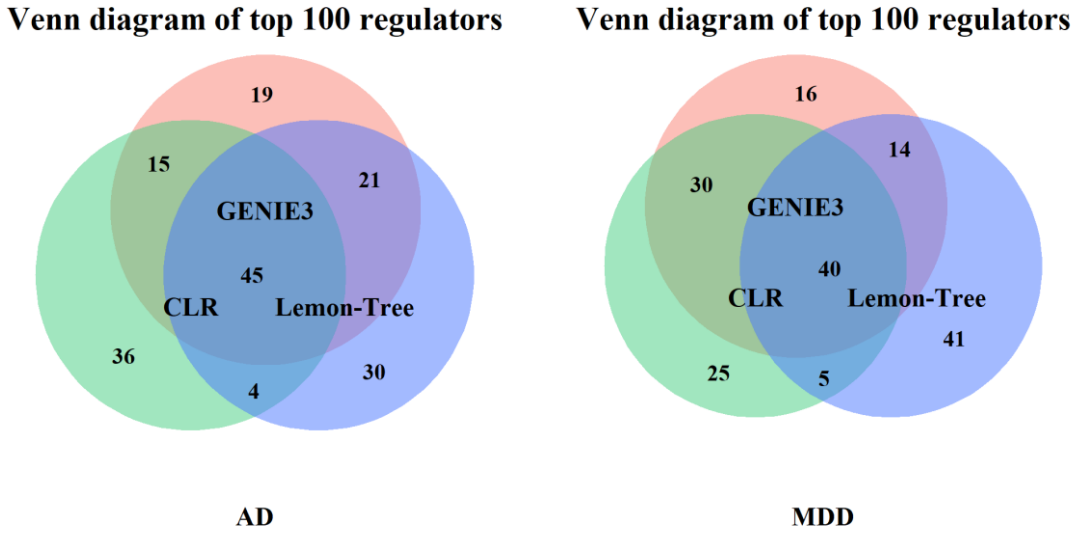


Figure 10. Venn diagrams of the top 100 regulators for each network retrieved by CLR, GENIE3 and Lemon-Tree, for Alzheimer's disease (AD) and major depressive disorder (MDD).

Table 2. Overview of the number of nodes for the three Alzheimer's disease networks.

	CLR	GENIE3	LEMON-TREE
TOTAL NODES	9715	9384	8736
TARGET GENES	9714	9340	8674
TRANSCRIPTION FACTORS	1041	994	618

Table 3. Overview of the number of nodes of the three networks for depression.

	CLR	GENIE3	LEMON-TREE
TOTAL NODES	9183	8871	8012
TARGET GENES	9182	8859	7934
TRANSCRIPTION FACTORS	1101	1069	763

Moreover, each network, for each disease, was inspected by topological measures. These were the average degree, correlation coefficient, density, diameter, edge betweenness, node betweenness, directed edge/node betweenness, connected components, and the (directed) average path length. In Tables 4 and 5, the values are indicated for each network characteristic, for Alzheimer's and depression, respectively. These network measures were calculated with the *igraph* package in R. The directed measures are similar to the undirected measures, with the only difference that only directed paths are considered. The diameter and edge betweenness were normalized for the number of nodes in the network.

Table 4. Network characteristics of the three networks made by CLR, GENIE3 and Lemon-Tree of the Alzheimer's disease dataset.

Measure	*CLR*	*GENIE3*	*Lemon-Tree*
Average degree	0,002119	0,002271	0,002621
Clustering coefficient	0,063749	0,410084	0,438643
Density	0,001060	0,001136	0,001310
Node betweenness	10892,8	10813,3	8459,6
Directed node betweenness	2432,6	2664,3	1462,2
Edge betweenness	0,000231	0,000246	0,000222
Directed edge betweenness	2.5782e-05	3,0266e-05	1,9165e-05
Diameter	6	0,1358	7
Connected components	1	1	1
Average path length	3,2427	3,2393	2,9370
Directed average path length	3,3395	4,7104	3,8552

Table 5. Network characteristics of the three networks made by CLR, GENIE3 and Lemon-Tree of the depression dataset.

Measure	*CLR*	*GENIE3*	*Lemon-Tree*
Average degree	0,002372	0,002542	0,003117
Clustering coefficient	0,096489	0,355748	0,443298
Density	0,001186	0,001271	0,001558
Node betweenness	10828,6	10936,9	8200,1
Directed node betweenness	2685,5	2877,3	1925,2
Edge betweenness	0.000257	0,000278	0,000256
Directed edge betweenness	3.1857e-05	3,6576e-05	3,0003e-05
Diameter	6	0,1073	7
Connected components	1	3	1
Average path length	3,3587	3,3754	3,0472
Directed average path length	3,4419	4,1439	3,8248

Most values are similar. However, there is one striking difference, namely in the diameter. For CLR and Lemon-Tree, the diameter is two times six and seven, while for GENIE3, it is 0.1. Normally the diameter is an integer. This indicates that all nodes are connected in the GENIE3 network. All the networks consist of one connected component, except for the MDD network of GENIE3. This is surprising, as here, the diameter is lower than one. The clustering coefficients of the CLR networks are lower than the networks of the other two methods. This indicates that the nodes are less clustered together in this network. The node and directed node betweenness are lower in the Lemon-Tree networks. These results again illustrate - next to the small overlap in edges - that different methods retrieve different networks. A representation of the networks can be seen in Supplementary figure S1 (Addendum 3), and the degree distribution for every network can be found in Supplementary figure S2.

For Lemon-Tree, functional enrichment analysis was performed as well. This was most convenient with this method, as modules are already constructed by the algorithm. This was done with the *enrichR* package in R. The terms were filtered to have an adjusted p-value equal to or smaller than 0.05. Several modules were enriched for immunological functions in the Gene Ontology Biological Process terms. In particular, modules sixteen and fourteen from AD and MDD respectively were of interest, because of an extensive overlap. An overview of the terms can be seen in Figure 11. In module sixteen there are 153 genes, in module fourteen 111, of which there are 68 in common. An overview of the gene expression of these modules can be seen in Addendum 3, Figures S3 and S4. There are eight (AD) (ATOH8, IKZF1, IRF8, NFATC2, RUNX1, RUNX2, TAL1, TCF3) and nine (MDD) (FOS, IKZF1, IRF8, MAF, RHOXF2, SPI1, TAL1, TFEC, ZNF551) regulators of these modules, with three shared regulators (IRF8, IKZF1, TAL1) (see Figure 12). All three are implicated in hematopoietic cell differentiation[60]. Moreover, IRF8 plays a regulatory role in immune cells and is involved in interferon response[60]. TAL1 is highly expressed in microglia[61]. IRF8 is implicated in microglial activation and neuroinflammation in AD mice models[62]. Several studies have found mutations that interrupt the binding of TAL1 in patients with AD[61]. Further, TAL1 was found to be implicated in MDD in two studies[63,64], and according to DisGeNET, FOS has been implicated with depression as well[65]. Lastly, TCF3 and SPI1 have been associated with AD as well[65]. As mentioned in the

introduction, there are several cytokines overexpressed in depressive patients. In Figure 4, some of these cytokines are indicated. Moreover, neutrophils and T-cells have as well already been implicated in the disease[22]. Neutrophils have also been implicated in the pathology of AD[66]. β-amyloid could be a possible chemoattractant and attract neutrophils and microglia to the deposits. Moreover, neutrophils secrete reactive oxygen species, which are harmful to the brain.

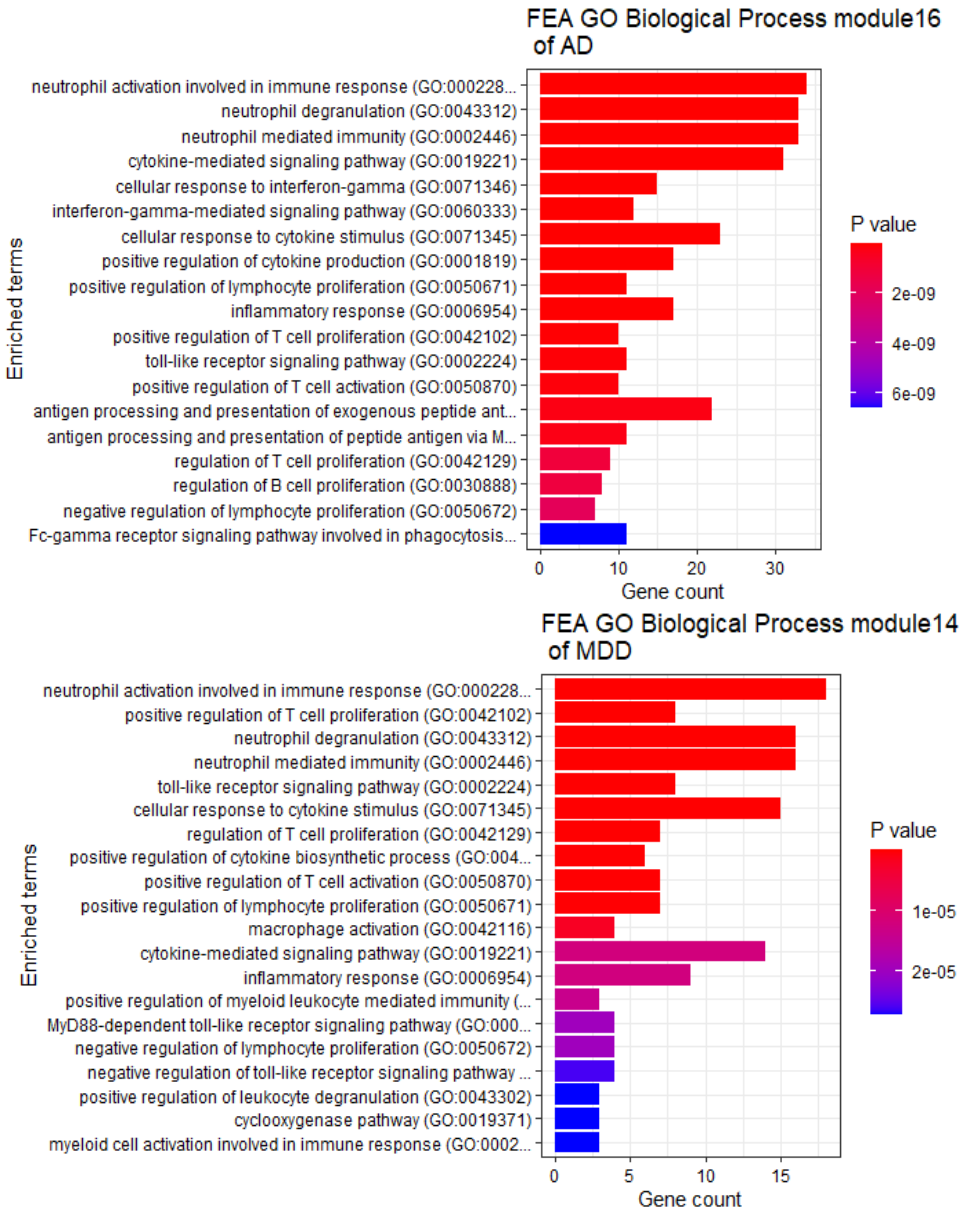


Figure 11. Representation of the top twenty terms of functional enrichment analysis by Gene Ontology Biological Process of modules 16 (AD) and 14 (MDD), ordered by increasing p-value. The gene count is represented on the x-axis.

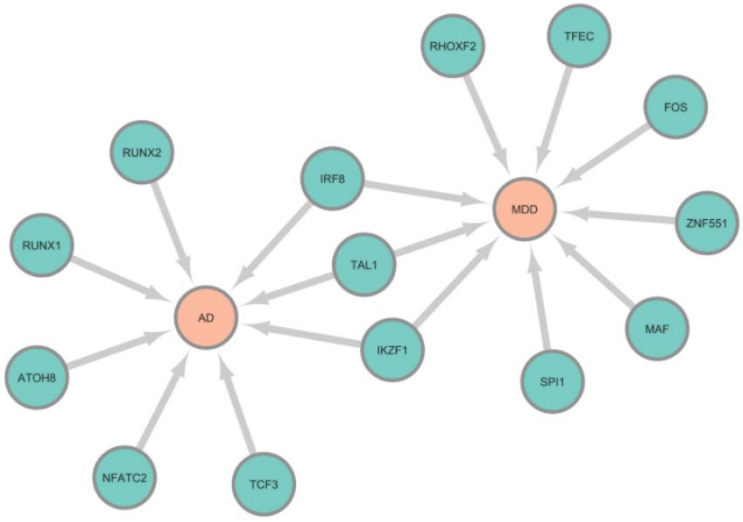


Figure 12. Representation of the regulators of modules 14 from the depression (MDD) network and 16 from the Alzheimer's disease (AD) network. IRF8, TAL1 and IKZF1 are common regulators. The nodes 'AD' and 'MDD' represent the target genes in these modules. Created with Cytoscape.

3.4 Consensus regulatory programs for AD and MDD

The top 100 000 edges were retrieved from the average rank aggregation for the ensemble networks. Thereafter, the overlap in edges between the ensemble networks and the networks retrieved by each method was compared. The overlap between the ensemble network and the initial networks was around 50 000 edges each time. The largest overlap was found between the ensemble network and the networks retrieved by GENIE3, both for Alzheimer's (53029 edges) and depression (58810). For depression, the overlap of the networks was each time higher, compared to the networks of AD. There are 1041 TFs and 9675 target genes generating a total of 9685 nodes in the ensemble AD network. In the ensemble MDD network, there are 1101 regulators, 9006 target genes and a total of 9021 nodes. Of these, there are 7042 nodes in common between the AD and MDD networks. For the ensemble networks, the same network characteristics were calculated as before, with the *igraph* package. In Table 6, there is an overview of all the characteristics of the two ensemble networks. All values are similar to each other, and to the values of the different network inference methods. The directed node betweenness is somewhat higher in the ensemble networks, meaning there are on average more paths that pass through a node. The diameter of the ensemble MDD network is one path longer than the diameter of the AD network. Both ensemble networks consist of one connected component. All values are higher for the MDD network, except for the directed average path length. Of the top 100 regulators of the ensemble networks, there are 34 in common. These can be found in Supplementary table 1 (Addendum 3).

In addition, the networks were compared with distance measures. Firstly, the Jaccard similarity index was calculated between the two networks. The similarity was low, i.e. 0.0554. The Jaccard distance is then calculated as one minus similarity, thus 0.9446. This indicates that there are few edges in common between the two networks. This was visualized as a Venn diagram, see Figure 13. There are only 10500 common edges. However, looking at the number of edges each network contained, this is about ten percent of the edges the networks have in common. To see where these common edges are situated in the two networks, histograms were made from the rank of these common edges in both networks (Figure 14). Most of the edges are situated in higher ranks, thus with higher confidence. The ranks are higher in the AD network, with about 3500 shared edges in the highest 10 000 ranks. In the MDD network, there are about 2900 shared edges in the highest 10 000 ranks.

Table 6. An overview of the network characteristics of the ensemble networks of Alzheimer's disease (AD) and major depressive disorder (MDD).

Measure	*Ensemble AD*	*Ensemble MDD*
Average degree	0,002132	0,002458
Clustering coefficient	0,313963	0,350065
Density	0,001066	0,001229
Node betweenness	10033,5	10182,4
Directed node betweenness	2916,2	2955,6
Edge betweenness	0,000214	0,000250
Directed edge betweenness	3,1020e-05	3,6331e-05
Diameter	6	7
Connected components	1	1
Average path length	3,0722	3,2577
Directed average path length	3,9606	3,8846

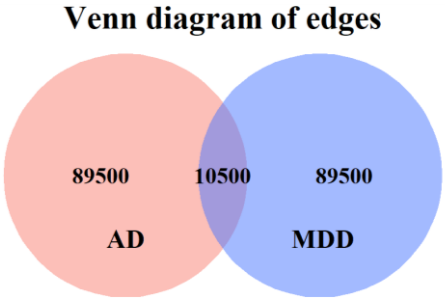


Figure 13. Venn diagram of the edges of the two ensemble networks of Alzheimer's disease (AD) and depression (MDD). There are 10500 edges in common.

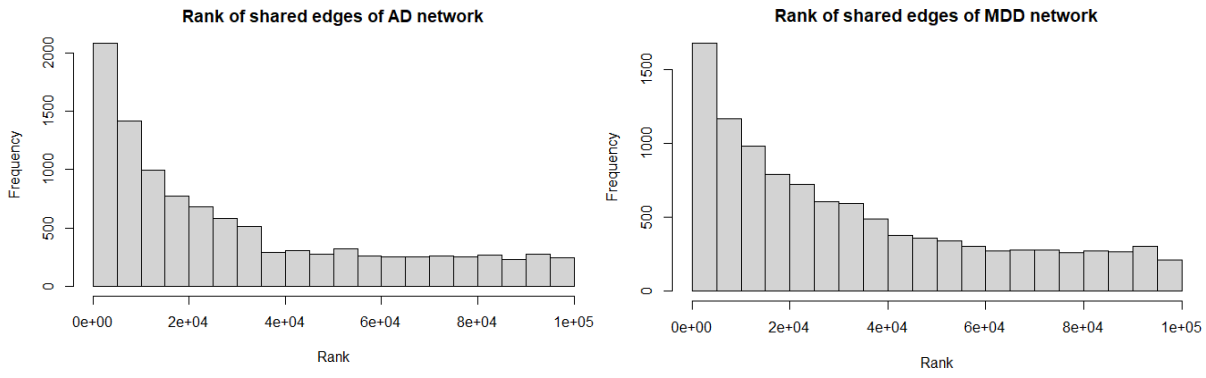


Figure 14. Histograms of the rank of the shared edges (10500) between the two ensemble networks, from the Alzheimer's disease network (left), and the depression network (right).

Further, the GCD[43] was calculated. The GCD between the AD and MDD ensemble networks was 1.063. This distance measure is not that informative, as only two networks are compared. It is more informative if more than two networks are compared between each other. Looking at supplemental Figure S8[43], a measure of around one is quite different, but not excessively different. Additionally, heatmaps were made with the Graphlet Correlation Matrixes. This was done by using the output of the first step (see methods) and calculating the Spearman

correlation between the eleven non-redundant orbits[43]. In Figures 15 and 16, the heatmaps are plotted for AD and depression, respectively.

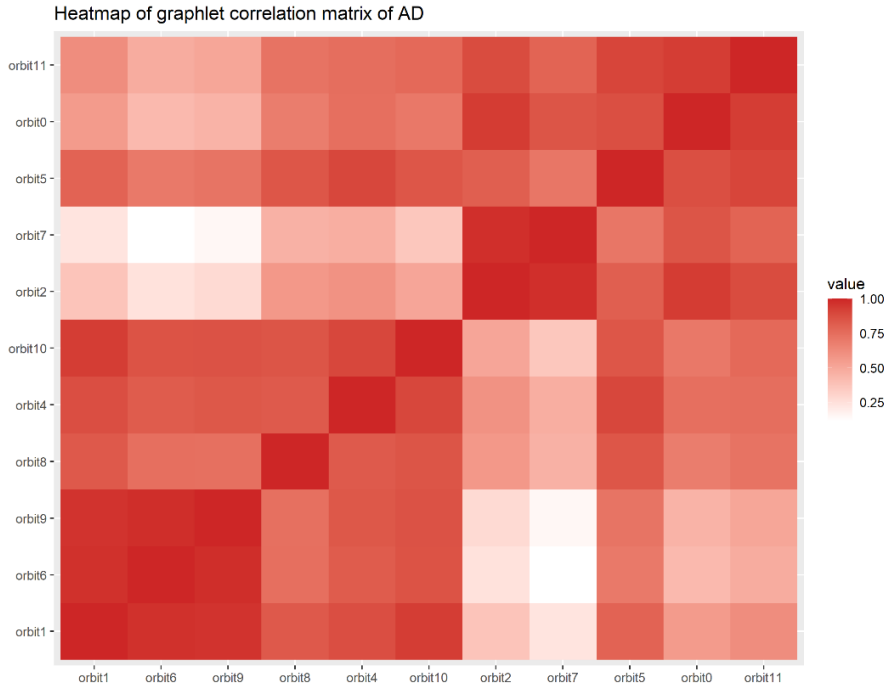


Figure 15. Heatmap of the graphlet correlation matrix of the Alzheimer's disease ensemble network. Only non-redundant orbits from up to four-node graphlets are depicted. Dark red indicates a high Spearman correlation, while white indicates no correlation.

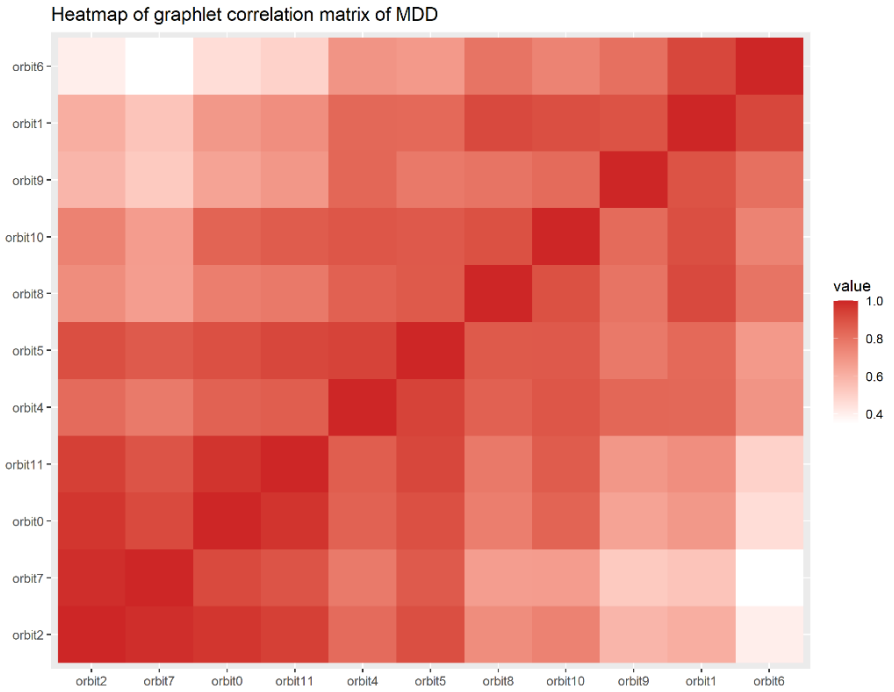


Figure 16. Heatmap of the graphlet correlation matrix of the depression ensemble network. Only non-redundant orbits from up to four-node graphlets are depicted. Dark red indicates a high Spearman correlation, while white indicates no correlation.

Orbits one, six and nine are characteristic of the existence of many degree-one nodes[43]. These orbits are clustered together in the AD network. Orbits two and seven, just like zero and five

are characteristic of hubs[43]. These are seen to cluster together in both the AD and the MDD networks. In the AD heatmap, there is no correlation between the orbits two and seven, and six and nine (white tiles). Moreover, the correlation between the two orbits two and seven and the orbits one, four, six, eight, nine and ten is lower than the correlation between these two orbits and the orbits zero, five and eleven. This indicates that nodes are either peripheral nodes or clustered nodes[43]. In the MDD heatmap, there is as well a low correlation between orbit six, and orbits zero, two, seven, and eleven. This is reasonable, as hubs have a large degree (not degree-one nodes). Orbits one and six also form a small cluster in the MDD network (peripheral nodes). Hence, similar clusters are found between the different orbits within the two ensemble networks. This indicates they have a similar structure, but there are still some differences, looking at the GCD. The two ensemble networks were as well visualized with *igraph*, see Supplementary figure S5 (Addendum 3).

3.4.1 Module generation with k-medoids

As mentioned in the methods, modules were retrieved with the Jaccard similarity index and k-medoids clustering. The number of clusters to choose with k-medoids can be hard, especially for a large dataset. One possible method is to keep the number of clusters that were retrieved with Lemon-Tree, i.e. 155 for AD and 156 for MDD. However, it is possible that more clusters would be a better fit for the data. With 155 and 156 clusters for k-medoids, there was for both the AD and the MDD network one cluster with more than 1000 genes, and nine modules with more than 100 genes. As such, the optimal number of clusters was calculated using different indices, which were found in the paper by Saelens et al.[67]. In the supplementary material of the paper, they refer to the *NbClust* package in R. In the package, the function *NbClust* was used to retrieve the optimal number of clusters. The average silhouette width, Calinski-Harabasz index and Davis-Bouldin index were used to estimate the optimal number of clusters with the scaled counts datasets. Firstly, the median method was used to cluster the data and find the optimal number of clusters. Next, k-means was used as a method. It was not possible to use k-medoids in the *NbClust* function. For the median method, the optimal number of clusters for the AD dataset was 189 for the Calinski-Harabasz index and Davis-Bouldin index. For the average silhouette width, this was 156. For the MDD dataset, all indices indicated 150 as the optimal number. 150 was the lower border value for which the indices were calculated. When using k-means as a method, the Calinski-Harabasz index and average silhouette width indicated 150 and the Davis-Bouldin index signifies 243, for the AD dataset. Similarly, for MDD, the Calinski-Harabasz index pinpointed 150, average silhouette width 152, and Davis-Bouldin index 246. As the Davis-Bouldin index retrieved around 245 modules for both datasets with the k-means method, the clustering with k-medoids – with the Jaccard index – was done with k equal to 243 for AD and 246 for MDD. Here again, there was for each network a module with more than 800 genes and one module with more than 100 genes. As most indices indicated the optimal number of clusters around 150, and a similar large module was retrieved in the two clustering solutions, the solution with 155 and 156 modules was selected. The two large modules were briefly inspected and then omitted. They both contained genes of which the significant (adjusted p-value ≤ 0.05) functional enrichment terms indicated regulation of transcription and DNA binding, indicating the modules contained mostly TFs and co-factors. An explanation for this large module could be that these TFs were themselves regulated by few or by very different TFs in the networks, as the modules were based on the Jaccard similarity between the shared regulators for every gene. Hence, it would be hard to add them to a module. Another explanation could be that all these genes are regulated by the same or similar TFs, but this would be rare.

In addition, regulators were added to each module. Each module had at least one regulator in the AD network and at least two in the MDD network, except for the large modules in both networks. The fact that these large modules have no regulator assigned, confirms the hypothesis that few and/or very different TFs were assigned to these genes. Next, the modules and regulators were visualized in Module Viewer, as was done with the Lemon-Tree network

before. The two large modules were not visualized. The figures can be found on the GitHub repository (see Addendum 2).

3.4.2 Functional enrichment analysis

All modules were inspected for functional enrichment using GO Biological Process, GO Molecular Function, KEGG, Reactome and WikiPathways. Here again, the terms were filtered to have an adjusted p-value of 0.05 or lower. For AD, there were some modules with predominantly immune functions, while some other modules were enriched for the 'Alzheimer's disease' term from KEGG. For the MDD network, there were as well some modules with enriched terms related to the immune system. In addition, there were also some modules with the 'Alzheimer's disease' term from KEGG, indicating an etiological overlap. Moreover, there was one module with nervous system development terms such as 'nervous system development', 'glial cell development', and 'astrocyte differentiation' from GO Biological Process. However, none of these modules had a very clear difference in expression between patients and controls. Thus, many of the retrieved modules are unchanged in AD and MDD, compared to controls. Moreover, when there is a difference, this difference is not pronounced, perhaps because of the heterogeneity of the disorders.

Module 22 from the AD network contains immune-related terms (see Figure 17). Even though there is not a substantive difference between the expression in the Alzheimer's patients and the controls, there seems to be a trend in which the control individuals have a lower expression (see Figure 17). The TFs of this module are IKZF1, NFATC2 and RUNX1. These three regulators were also regulators of module 16 from the Lemon-Tree AD network. All three genes are regulators of immune functions. IKZF1 is involved in hematopoietic cell differentiation[60]. Altered binding sites of IKZF1 have been implicated in risk genes for AD[68]. This has been seen for the ABCA7 and INPP5D genes, of which the latter is also a part of this module. INPP5D encodes for the SHIP1 protein, which is involved in several pathways. It acts as a negative regulator of myeloid cell proliferation/survival and chemotaxis, immune cells homeostasis, it regulates macrophage programming, phagocytosis and activation, and neutrophil migration[60]. The gene has been implicated in several disorders, such as AD, cancers, systemic lupus erythematosus, and inflammatory bowel disease[65]. NFATC2 induces the expression of cytokines in T-cells[60]. Interestingly, this gene was found to be implicated in the activation of microglia in a mouse model of AD[69]. RUNX1 is essential for normal hematopoiesis and is mainly involved in T-cell functioning[60]. This gene has as well been associated with AD before[70].

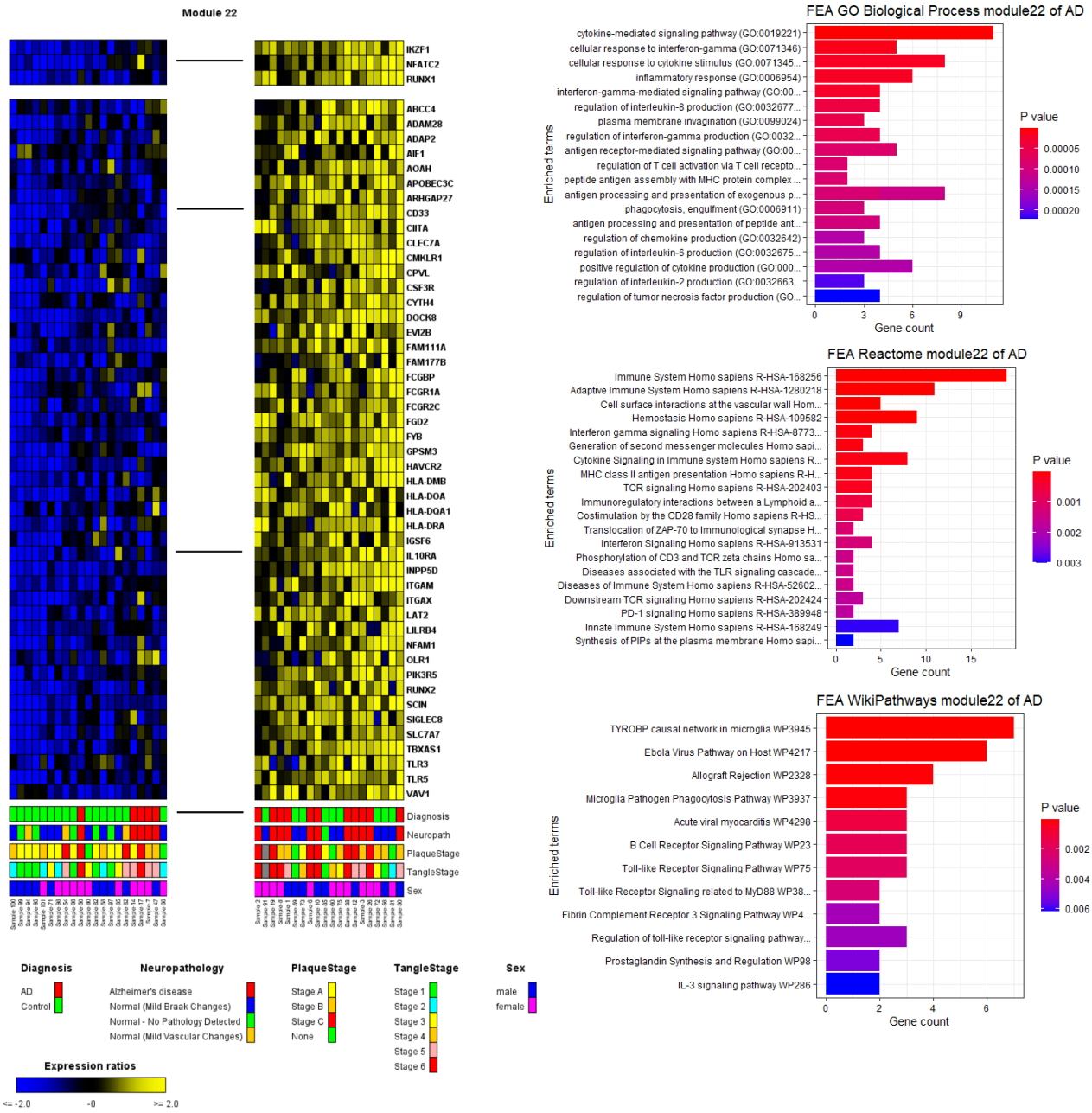


Figure 17. Overview of module 22 from the AD network, which contains 47 genes. The expression of the module was visualized with Module Viewer, with the corresponding annotation data diagnosis, neuropathology, plaque stage, tangle stage, and sex (see legend). The upper panel represents the regulators of this module. In addition, the functional enrichment plots from the databases GO Biological Process, Reactome and WikiPathways are pictured. The top twenty terms are depicted, ordered by increasing p-value. The number of genes from the module belonging to the terms is represented on the x-axis.

Similarly, module 24 of the MDD network contains immune-related terms (see Figure 18). However, here the expression of the depressive patients seems to be lower than the expression of the controls (see Figure 18). The regulators of this module are IKZF1, RUNX1 and TFEC. Hence, there are two regulators in common between this module and module 22 from the AD network. However, NFATC2 is part of this module 24. In addition, IKZF1 and TFEC were as well regulators of module 14 of the Lemon-Tree MDD network. Despite having similar regulators, there are only six genes in common between the two ensemble modules: ABCC4, CPVL, FCGR1A, GPSM3, IL10RA, and SLC7A7. ABCC4 and FCGR1A have been implicated in AD, while CPVL has been implicated in psychosis and AD[65]. CPVL is a protease that is

involved in the cleaving of phagocytosed particles in the lysosome, in an inflammatory protease cascade, and in pruning of peptides for antigen presentation[60]. Further, IL10RA has been associated with schizophrenia and multiple sclerosis[65]. In addition, IL10 has been associated with a worsening plaque load and reduced Aβ phagocytosis by microglia in mouse models of AD[71]. Looking at the functional enrichment analysis, there are some similar terms, and some differences between the two immune-related modules, so it is not entirely the same pathway. Both modules do have the 'TYROPB causal network in microglia' as the most enriched term, and Microglia Pathogen Phagocytosis Pathway' as the fourth and third term, respectively, from WikiPathways. TYROBP is an adaptor protein that associates with activating receptors of immune cells[60], and has been implicated in AD before[65]. It associates with TREM2, and both are required for phagocytosis in microglia. However, both TYROBP and TREM2 are not part of these modules. As both modules contain immune-related terms and terms related to microglia, these modules are probably active in microglia and the majority of the reads are probably coming from these cells. As the gene expression tends to be higher in Alzheimer's patients, there is more activation of microglia or there are more microglia in AD, compared to control. On the other hand, as the expression of these genes tends to be lower in depressive patients, there is less activation or less microglia in this disease, compared to the controls. However, there are some terms with 'negative regulation of' in the MDD module, which might indicate that inhibitors of immune functions are downregulated in depression.

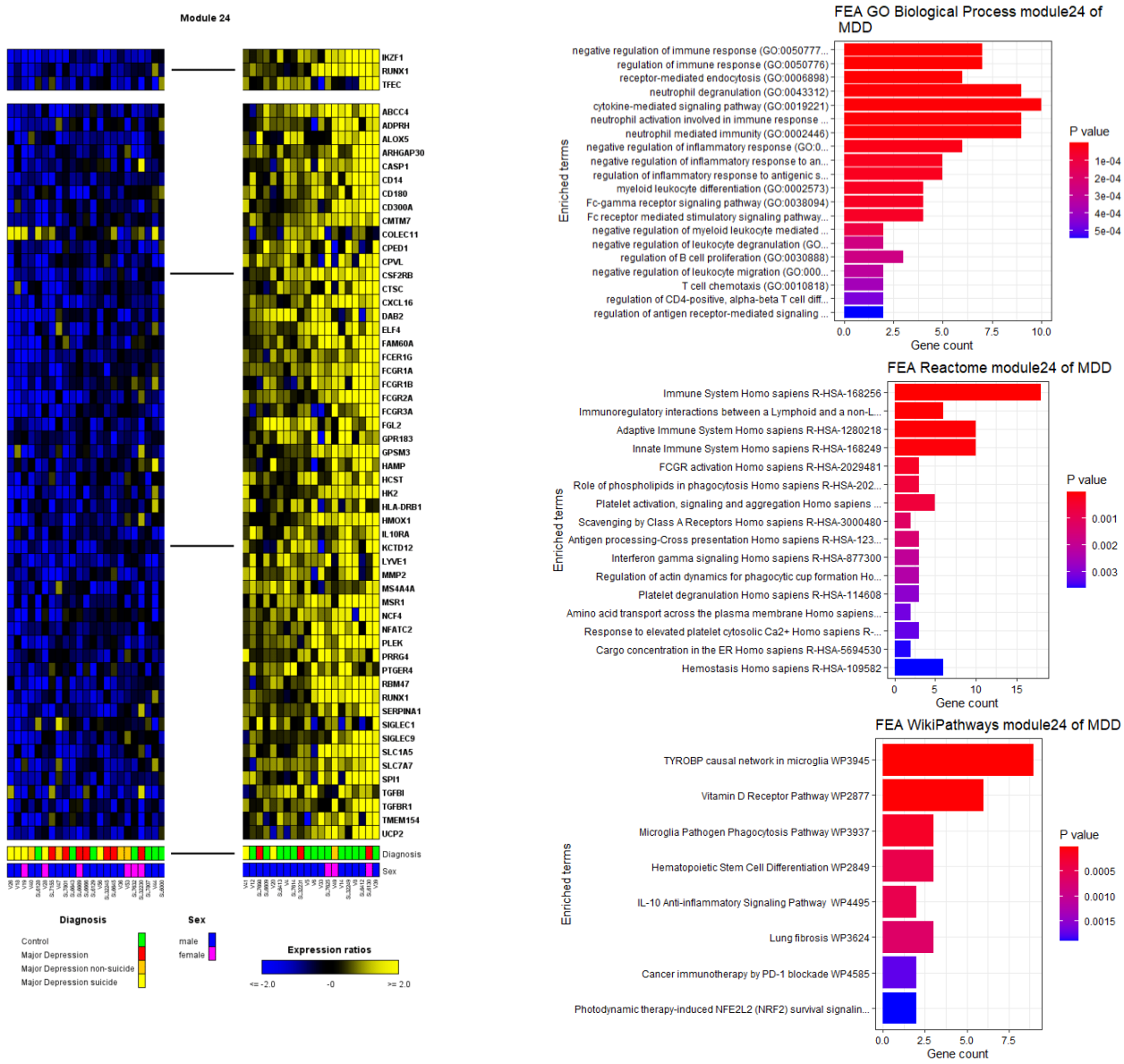


Figure 18. Overview of module 24 from the MDD network, which contains 54 genes. The expression of the module is visualized in Module Viewer, with the corresponding annotation data diagnosis and sex (see legend). The division between 'Major depression non-suicide' and 'Major depression suicide' is from the dataset GSE101521, while 'Major depression' is from the dataset GSE80655. 'Control' is from both datasets. The upper panel represents the regulators of this module. In addition, the functional enrichment plots from the databases GO Biological Process, Reactome and WikiPathways are pictured. The top twenty terms are depicted, ordered by increasing p-value. The number of genes from the module belonging to the terms is represented on the x-axis.

Further, module 40 from the MDD network contained the term 'Alzheimer's disease' from the KEGG database. The gene expression of the control individuals seems to be lower than the gene expression of the depressive individuals (see Figure 19). Terms from GO Biological Process include 'mitochondrial respiratory chain complex assembly', 'NADH dehydrogenase complex assembly', 'mitochondrial electron transport, NADH to ubiquinone', 'aerobic electron transport chain', and 'mitochondrial ATP synthesis coupled electron transport', which are all terms related to the mitochondrial respiratory chain. As mentioned in the introduction, there is mitochondrial dysfunction in MDD[19,22] and AD[14]. As the genes seem to be upregulated in depression here, this might indicate a homeostatic reaction to counteract the dysfunction. The regulators of this module are CREB3, HEY1, IKZF4, PLAGL2, THAP11, USF1, ZNF532, and ZNF576. CREB3 and IKZF4 are as well involved in immune functions[60]. The genes that are enriched in the 'Alzheimer's disease' term are NDUFA8, NDUFA6, PSMA1, NDUFA12,

PSMA2, NDUFS4, NDUFB3, NDUFS3, NDUFB2, and UQCRFS1. The NDUF genes and UQCRFS1 are involved in the mitochondrial respiratory chain, and the PSMA genes encode parts of the proteasome[60].

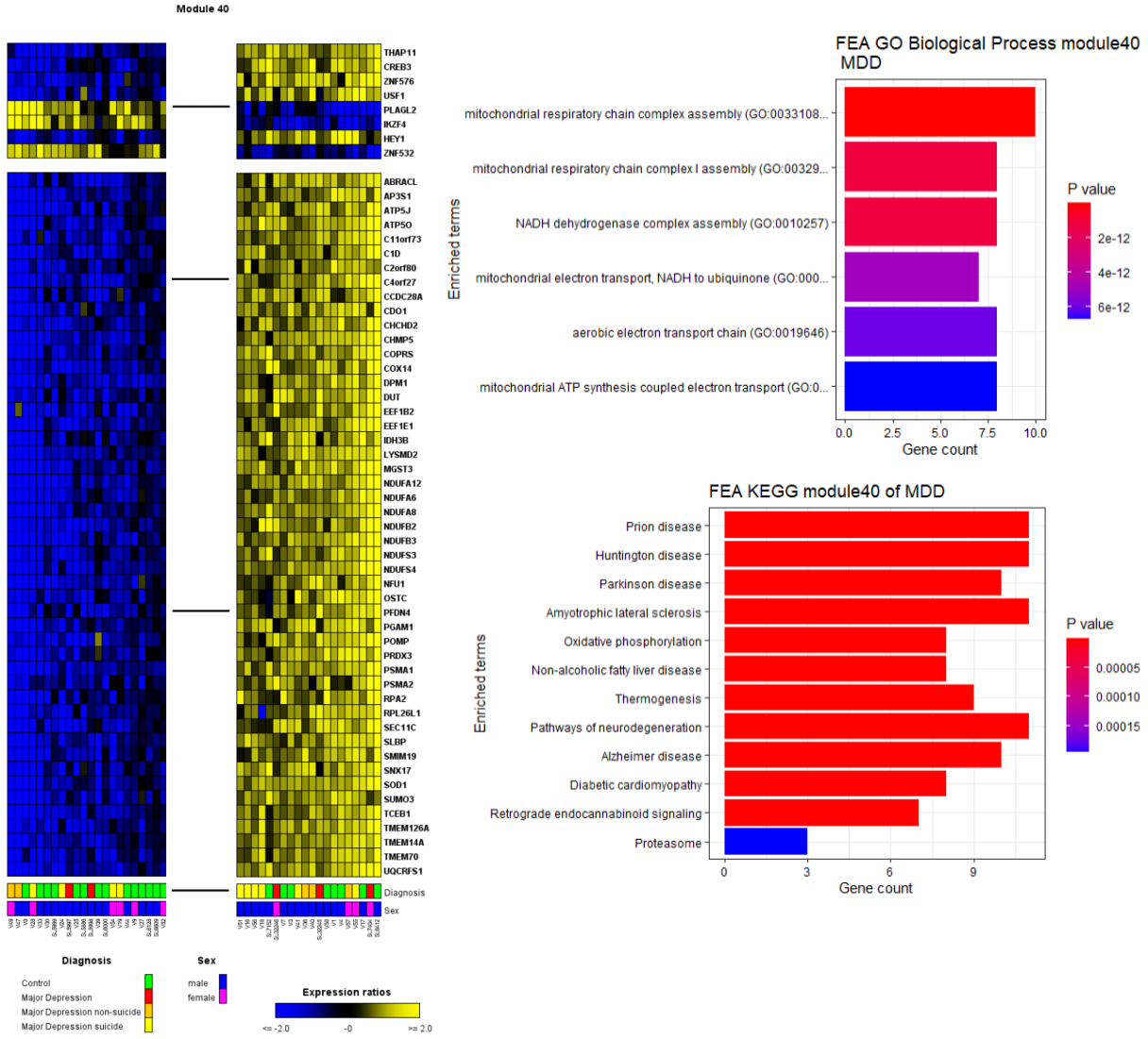


Figure 19. Overview of module 40 from the MDD network, which contains 49 genes. The expression of the module is visualized in Module Viewer, with the corresponding annotation data diagnosis and sex (see legend). The division between 'Major depression non-suicide' and 'Major depression suicide' is from the dataset GSE101521, while 'Major depression' is from the dataset GSE80655. 'Control' is from both datasets. The upper panel represents the regulators of this module. In addition, the functional enrichment plots from the databases GO Biological Process and KEGG are pictured. The top twenty terms are depicted, ordered by increasing p-value. The number of genes from the module belonging to the terms is represented on the x-axis.

In addition, there are some other modules with a subtle difference in the gene expression between diseased patients and healthy controls. The visualizations of these modules can be found in the supplementary figures (Addendum 3). According to the functional enrichment analysis, module 26 (Figure S6) of AD is involved in cholesterol biosynthesis and intracellular protein transport, and module 60 (Figure S7) of the AD network is involved in mitochondrial pathways, such as mitochondrial protein import and the citric acid cycle, and metabolism of proteins. There are as well seven genes of this module (PSMB7, PSMC6, MAPT, CYC1, COX6A1, GAPDH, RTN4) enriched in the 'Alzheimer's disease' term of the KEGG database. PSMB7 and PSMC6 are part of the proteasome, and MAPT encodes the tau protein. CYC1 is part of the Cytochrome C family, COX6A1 is involved in the mitochondrial respiratory chain,

while GADPH is involved in glycolysis. Lastly, RTN4 is a neurite outgrowth inhibitor. Module 93 of the MDD ensemble network (Figure S8) has terms related to oligodendrocytes, GABA receptor signaling, 'mBDNF and proBDNF regulation of GABA neurotransmission', and kinase activity.

3.5 Single-cell analysis

3.5.1 Single-cell RNA-seq datasets and preprocessing

Similarly to the bulk RNA-seq datasets, single-cell RNA-seq datasets were retrieved from the prefrontal cortex. For depression, the single-nucleus dataset GSE144136 was found[72]. However, all samples were male. The dataset consists of seventeen and nineteen depressive patients and control individuals, respectively (see Table 1). This dataset contains 78 886 cells and 30 062 genes. For Alzheimer's, the single-nucleus dataset GSE174367 was utilized[56]. Here, there are eleven patients with AD and 8 controls, and nine female and ten male individuals. In this dataset, there are 61 472 cells and 36 114 genes. The raw counts were preprocessed with the *Seurat* package in R (see methods). Firstly, quality control was done. As such, it was decided for which value to filter the UMI counts in a cell. For the AD dataset, this was set to be between 200 and 7500, while for MDD it was set between 200 and 4000. Thereafter, 59 968 cells were remaining in the AD dataset and 73371 cells in the MDD dataset. The top 8000 genes were retrieved as highly variable features. After filtering for highly variable genes, adding the regulators again, and filtering for protein-coding genes, there were 8654 genes left in the AD dataset and 8631 genes in the MDD dataset. There were 654 and 631 regulators added again, to the AD and MDD dataset, respectively.

In addition, UMAP and t-SNE plots were made. Firstly, PCA was executed with the top 2000 highly variable genes and the function *RunPCA* from *Seurat*. Next, the elbow plot was used to determine the best number of PCs to work further with. This was thirteen for the AD dataset and twenty for the MDD dataset. These PCs were then used to plot the UMAP and t-SNE plots (Figures 20 and 22). In addition to the retrieved clusters visualized on the plots, the cell types that the researchers annotated were visualized as well. The clusters are not identical, as there are some clusters with other cell types in them (see Figures 21 and 23). The clusters from the AD dataset are more similar to the ones called by the researchers than the clusters from the MDD dataset. The fact that the cell types don't overlap the clusters here completely indicates that a different clustering method can retrieve different results and annotate some cells to a different cell type. In the MDD paper, they used unsupervised graph-based clustering to identify the different cell types[72]. They refer to a previous version of *Seurat*. So the principle was the same, but the method was perhaps somewhat different. In the pbmc3k_tutorial from *Seurat*, they state themselves that 'our approach to partitioning the cellular distance matrix into clusters has dramatically improved' since the previous version. Moreover, the researchers used the top 50 PCs, while here the first twenty were used. In the AD paper, they used the Leiden algorithm for clustering[56]. In the t-SNE plot of AD with the cell types (Figure 21), the excitatory neurons are divided into several clusters, while in the UMAP plot this is approximately one cluster. This confirms that UMAP is better able to keep the global structure of the data.

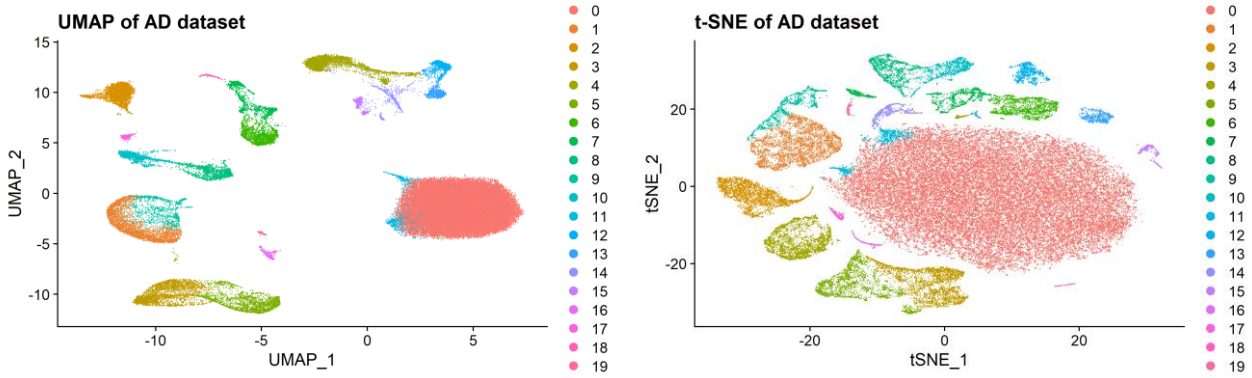


Figure 20. UMAP and t-SNE plots of the single-nucleus Alzheimer's disease dataset, with twenty clusters. The clusters were retrieved with the Louvain algorithm. The plots were made with the first thirteen principal components, retrieved from the top 2000 highly variable genes.

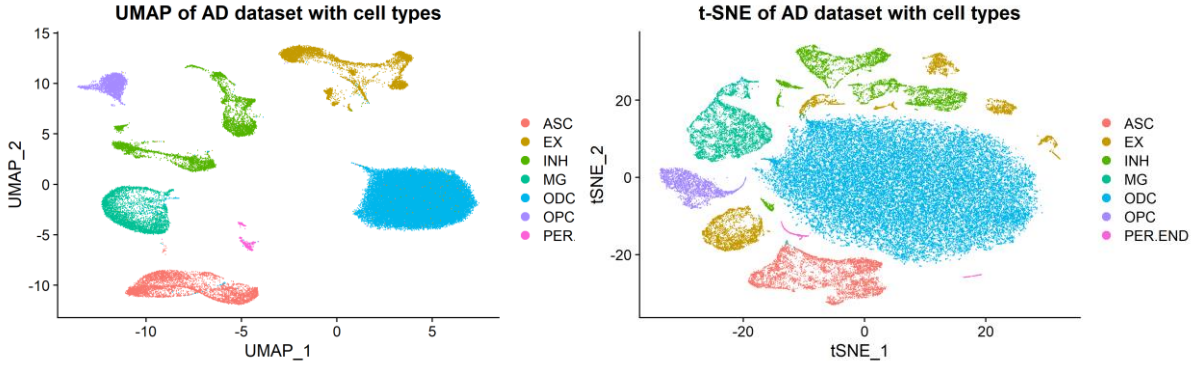


Figure 21. UMAP and t-SNE plots of the single-nucleus AD dataset. The plots are the same as above, however, the annotation is different. Here, the cell types that were annotated by the researchers are plotted[56]. Abbreviations: ASC astrocytes; EX excitatory neurons; INH inhibitory neurons; MG microglia; ODC oligodendrocytes; OPC oligodendrocyte precursor cells; PER.END pericytes/endothelial cells.

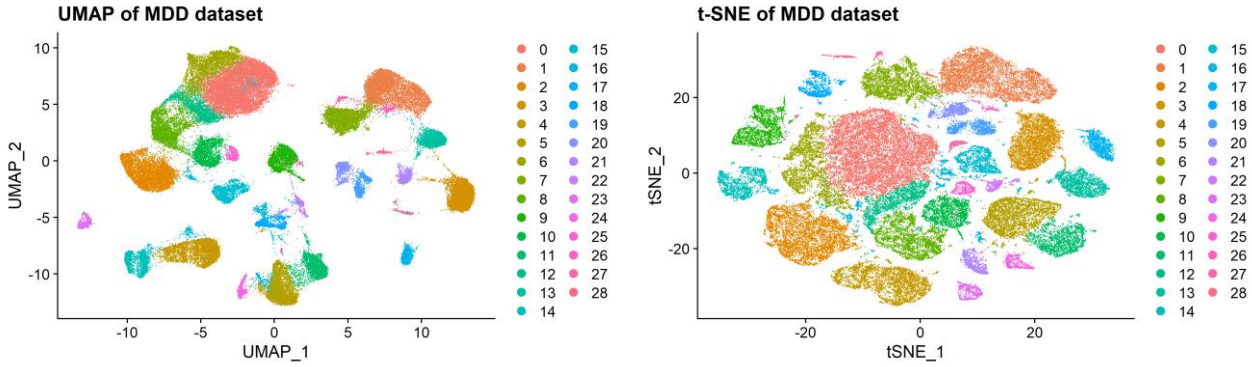


Figure 22. UMAP and t-SNE plots of the single-nucleus major depressive disorder dataset, with 29 clusters. The clusters were retrieved with the Louvain algorithm. The plots were created with the first twenty principal components, retrieved from the top 2000 highly variable genes.

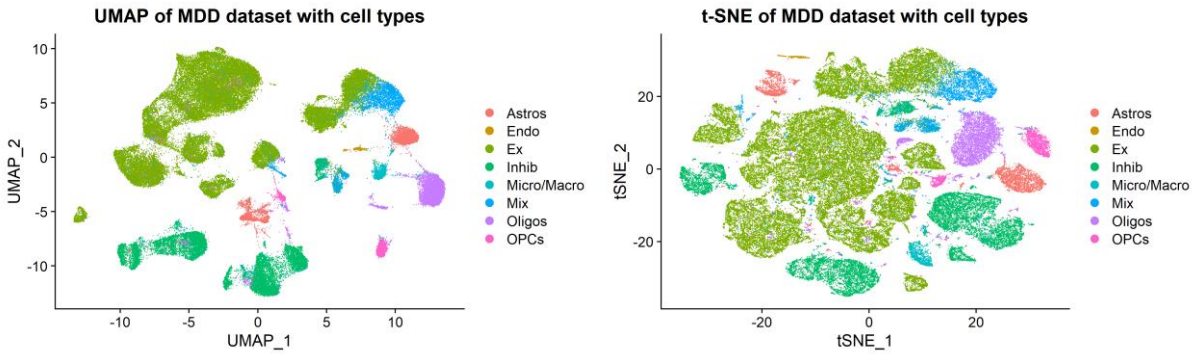


Figure 23. UMAP and t-SNE plots of the single-nucleus MDD dataset. The plots are the same as above, however, the annotation is different. Here, the cell types that were annotated by the researchers are plotted[72]. Abbreviations: Astros astrocytes; Ex excitatory neurons; Inhib inhibitory neurons; Micro/Macro microglia/macrophages; Oligos oligodendrocytes; OPCs oligodendrocyte precursor cells; Endo endothelial cells; Mix mix of cells.

3.5.2 Network inference with SCENIC

SCENIC was run in Python, with help from Joke Deschildre. As output, a table of the regulons, together with the activity scores of the different regulons in different cell types, and boxplots indicating the number of regulons per cell and number of cells per regulon were retrieved. A regulon signifies a TF together with its target genes. The table indicates which TF motif is found in association with which target gene, from RcisTarget. The boxplots are depicted in Supplementary figure S9 (see Addendum 3). The heatmaps with the activity scores of the regulons for each cell type can be seen on GitHub (see Addendum 2), as they are too large to paste into this document. In Table 7, an overview of the highly active regulons in each cell type, from the AD and MDD networks, can be found. A large part of the regulons that are highly active in a certain cell type of one of the two networks, is also active in the same cell type in the other network (dark red in Table 7).

Table 7. Overview of the highly active regulons in different cell types of the brain. Transcription factors in dark red are also active in the other network, in the same cell type.

Cell type	*Alzheimer's disease*	*Major depressive disorder*
Excitatory neurons	AHR, HLF, MEF2C, ZEB1, ZMAT4	FOXP1, NFAT5, PBX1, RFX3, TEAD4, ZMAT4, ZNF282, ZNF699
Inhibitory neurons	HLF, MEF2C, PKNOX2, ZEB1, ZMAT4	DLX1, DLX2, DLX5, FOXN3, HIVEP3, LHX6, MAFB, NR2F2, TCF4, ZMAT4, ZNF282
Microglia	ELF1, ELK3, ETS2, ETV6, FLI1, IKZF1, IRF8, MAF, NFATC2, RUNX1, SPI1, STAT6	FOXN3, IRF8, RUNX1
Oligodendrocytes	MXI1, NFIX, SREBF2, ZNF536	FOXN2, SOX10
Oligodendrocyte precursor cells	PRRX1, PRRX2, SOX6, VSX1, ZEB1, ZNF227	FOXN3, PBX3, PRRX1, SOX13, SOX4, ZEB1
Astrocytes	FOXO1, RFX2, SOX5, TCF7L1, TCF7L2	FOXO1, PAX6, RARG, RXRA, SOX2, SOX9, TCF7L2

3.5.3 Further analysis of the single-cell networks

The regulons of RUNX1, NFATC2 and IKZF1 are highly active in the microglia in the single-cell AD network, which confirms the hypothesis that the immune-related bulk ensemble modules (module 22 from AD and 24 from MDD) represent modules from microglia. The regulon of RUNX1 is highly active and the regulon of IKZF1 is active in the microglia in the single-cell MDD network. The regulon of TFEC is not present in both single-cell networks, and the regulon of NFATC2 is not present in the single-cell MDD network. In addition, when looking at the regulators of the two selected modules from Lemon-Tree, the regulons of IRF8 and TAL1 are most operative in the microglia in both single-cell networks. Moreover, the MAF regulon is most viable in the microglia of the single-cell AD network, while on the other hand, the TCF3 regulon is most active in the microglia of the MDD network. The regulon of SPI1 is most functional in the microglia in the single-cell MDD network, while it is more active in the oligodendrocytes of the AD network.

As the regulons of IKZF1, IRF8, NFATC2, RUNX1, and TAL1 are highly active in microglia and there is a difference between the activation of these cells in Alzheimer's and depression in the bulk networks, these were further investigated. In the single-cell networks, IKZF1 has 1520 target genes in the AD network, while it has 227 target genes in the MDD network. IRF8 has 938 target genes in the AD network and 142 in the MDD network. NFATC2 has 541 target genes in the AD network, while it is no regulator in the MDD network. RUNX1 has 1016 target genes in the AD network and 314 in the MDD network. Lastly, TAL1 has 401 target genes in the AD network and 89 in the MDD network. As the number of target genes is too large to represent in a network, a core GRN has been retrieved with only these five TFs (Figure 24). They are all connected to each other, however, there are more edges in the AD network, compared to the MDD network. IRF8 only regulates itself and the other regulators, but is not regulated by any of the other TFs. In addition, in the MDD network, NFATC2 is only regulated by other TFs, but does not regulate any target genes, as this gene was not retrieved as a regulator in the single-cell MDD network. As the regulon of TAL1 of the MDD network was the smallest, this regulon was visualized as well (see figure 25). In addition, functional enrichment analysis with GO Biological Process, Molecular Function, KEGG, Reactome and WikiPathways was performed on this regulon. Most terms are related to microglia and cytokine response. As mentioned above, TAL1 has been implicated before in both AD and MDD.

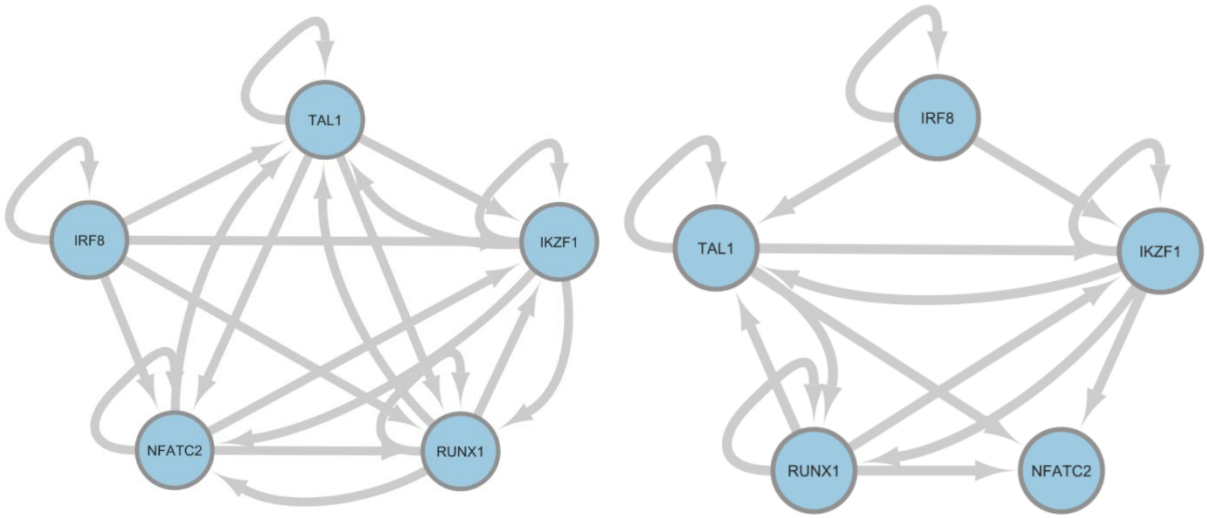


Figure 24. Microglial core gene regulatory networks with five regulators. The single-cell Alzheimer's disease network is depicted on the left, the single-cell depression network is depicted on the right. This figure was created with Cytoscape.

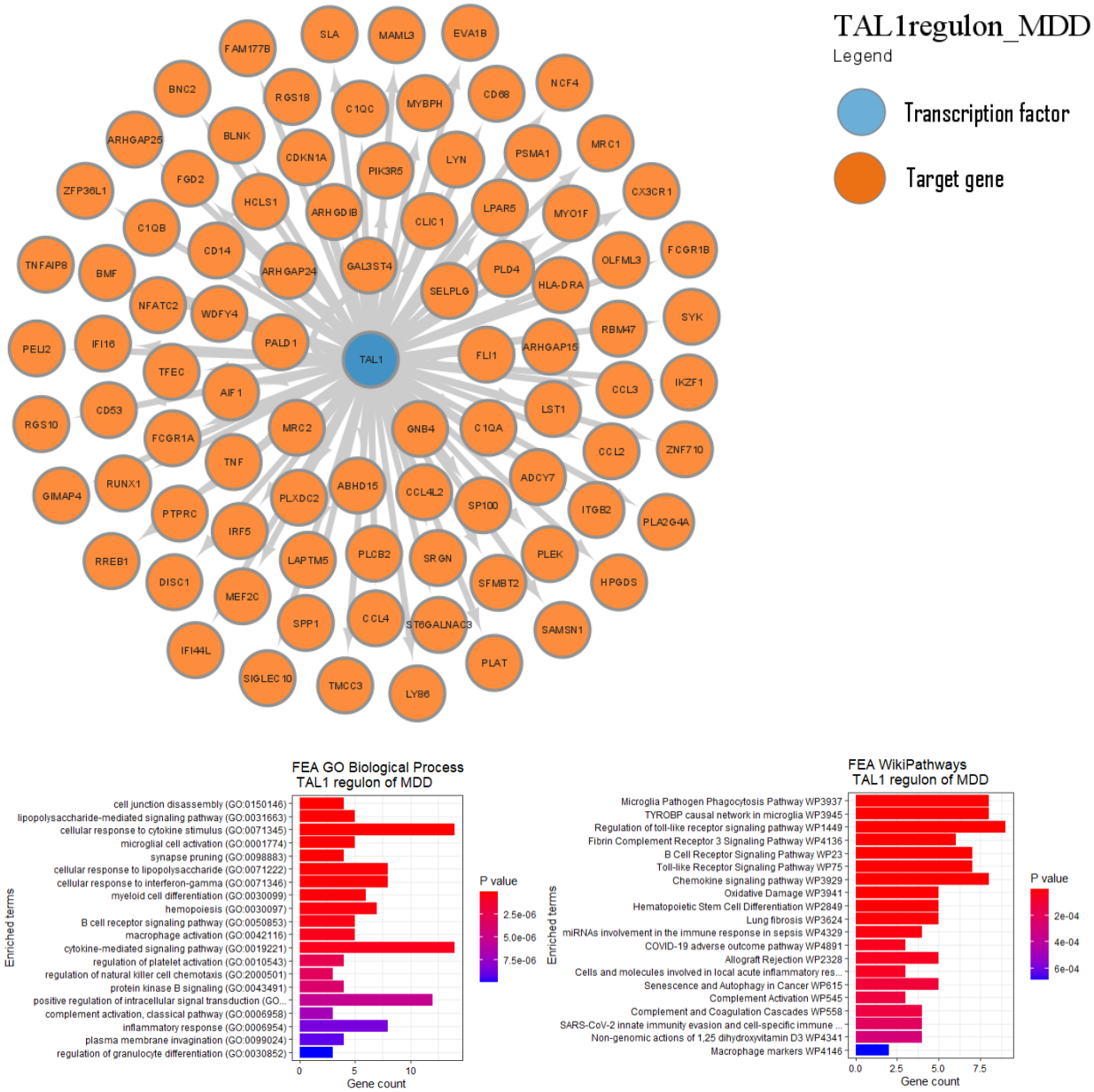


Figure 25. Visualization of the TAL1 regulon in the single-cell depression network, together with the functional annotation from Gene Ontology Biological Process and WikiPathways. The top twenty terms are depicted, ordered by increasing p-value. The number of genes from the regulon belonging to the terms is represented on the x-axis. Some of the target genes are transcription factors as well, such as RUNX1, NFATC2, TFEC and NKZF1. The graph was created with Cytoscape.

In addition to the regulators per cell type that were compared to some modules from the bulk ensemble networks, the single-cell networks were also compared to the bulk ensemble networks. Firstly, the edges were compared between the networks of the same disease with Venn diagrams (see Figure 26). The two disorders have a similar number of edges in common, but the single-cell network of MDD has fewer edges than the single-cell network of AD. More specifically, the single-cell network of AD consists of 37 070 edges, while the network of MDD has 22 544 edges. Secondly, the networks were compared through the Jaccard index. The Jaccard similarity index of the bulk and single-cell AD networks is 0.0138, resulting in a Jaccard distance of 0.9862. On the other hand, the Jaccard similarity index between the single-cell and bulk MDD networks equals 0.0153, resulting in a Jaccard distance of 0.9847. Thirdly, the number of regulators shared between the bulk and single-cell networks for the same disease was verified. There are 258 regulators in common between the AD networks, and 238 regulators in common between the MDD networks.

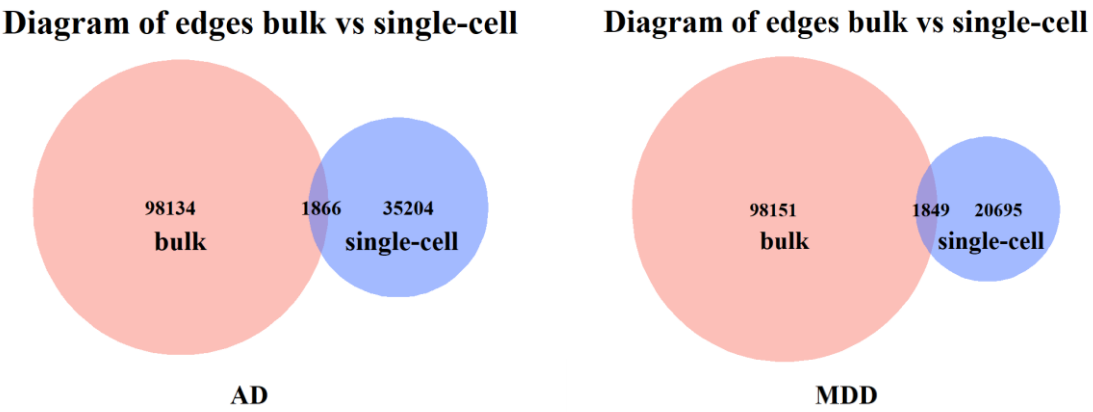


Figure 26. Venn diagrams for the comparison between the edges of the ensemble bulk networks and the single-cell networks of Alzheimer's disease (AD) and major depressive disorder (MDD).

Furthermore, the number of nodes of the single-cell networks was determined. An overview of the number of nodes, target genes and regulators of the single-cell networks can be seen in Table 8. The number of nodes is substantially less in the MDD network compared to the AD network. The single-cell networks were as well compared to each other through the number of overlapping edges (see Figure 27) and the Jaccard index. There are 3663 edges in common, which is about ten percent of the total edges of the networks. The Jaccard similarity index equals 0.0655, resulting in a Jaccard distance of 0.9345. Thus, the Jaccard similarity index between the two single-cell networks is larger than the Jaccard similarity indices between the bulk and single-cell networks of the same disorder. This could be explained by the fact that the single-cell networks were retrieved by a different method than the ensemble bulk networks. Moreover, the number of regulators is substantially less in the single-cell networks compared to the bulk networks. SCENIC makes use of RcisTarget, which cannot predict regulons for a TF with an unknown motif. Further, the number of edges is not equal in the bulk and single-cell networks. In comparison, the Jaccard similarity index of the ensemble networks was 0.0554.

Table 8. Overview of the number of target genes, regulators and the total number of nodes of the single-cell networks of Alzheimer's disease and depression.

	ALZHEIMER'S DISEASE	DEPRESSION
TOTAL NODES	6360	4814
TARGET GENES	6001	4549
TRANSCRIPTION FACTORS	359	265

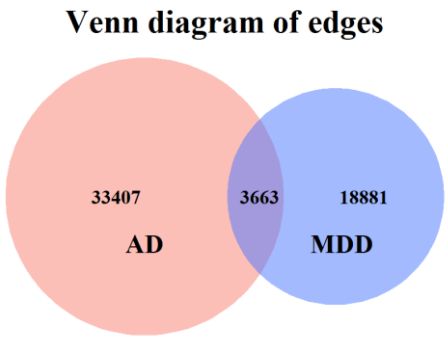


Figure 27. Venn diagram of the edges of the single-cell networks of Alzheimer's disease (AD) and major depressive disorder (MDD). There are 3663 edges in common.

4. DISCUSSION

The purpose of this master's dissertation was to find common and distinct pathways and regulators between AD and MDD. This was done by inferring GRNs with different methods (GENIE3, CLR and Lemon-Tree). There was only a small overlap in the results retrieved from the different methods, highlighting the different underlying assumptions and algorithms. In the next step, ensemble networks were made from each disorder, from which modules were retrieved. Given the small overlap in the different network inference methods, creating an ensemble provides a broader view of the true underlying network. The modules were used to perform functional enrichment analysis with GO, KEGG, Reactome, and WikiPathways. Some modules were prioritized, with a subtle difference in gene expression between patients and controls. Two of these modules contained immune-related terms (module 22 from the AD network and module 24 from the MDD network). Interestingly, despite having a small overlap in genes between these modules, there was an overlap in the regulators. IKZF1 and RUNX1 were common regulators, while one of the regulators of the AD module, NFATC2, was also a member of the MDD module. Similar modules had as well been found in the networks inferred by Lemon-Tree. Here again, there was overlap between the regulators of the Lemon-Tree AD and MDD modules: IRF8, KZF1 and TAL1 were shared. Interestingly, TAL1 has been associated with both AD and MDD before. Furthermore, the gene expression of module 22 was increased in AD patients compared to controls, while it was decreased in MDD patients (module 24) compared to controls, indicating a distinct disruption of these pathways in the two disorders. This is as well confirmed by the small overlap in genes in the two modules.

Further, networks were inferred with SCENIC, with scRNA-seq data. As such, cell-type-specific regulons were retrieved for both AD and MDD. By also inferring GRNs with single-cell data, it was possible to further characterize the results. The main drawback of bulk RNA-seq is the fact that all reads are averaged over the cells, and that the results are influenced by the number of cells of each cell type present in the sample. To investigate the regulators of the immune-related modules from the bulk network inference further, the activity of these regulators (IKZF1, IRF8, NFATC2, RUNX1 and TAL1) was investigated in the different cell types. Most of the TFs had regulons that were most active in microglia, in both AD and depression. These regulons were further inspected in the single-cell networks, and it was seen that all TFs regulate each other (see Figure 24). Thus, a large part of the genes in the immune-related bulk modules are probably coming from reads from microglia. It is already known that microglia play a pivotal role in AD. As mentioned above, the module of the AD network had a higher gene expression in patients compared to controls, indicating activation of microglia in AD. In a mouse model of AD, NFATC2 was found to be implicated in the activation of microglia[69]. The researchers have crossed AβPP/PS1 mice with NFATc2$^{-/-}$ mice, which resulted in mice with diminished cytokine levels, reduced microgliosis and reduced astrogliosis, but with no effect on plaque load, compared to AβPP/PS1 mice. In addition, IRF8 contributes to microglial activation by regulating microglial immune responses and chemotaxis[62]. The expression of IRF8 was found to be increased in the brains and microglia of an AD mouse model[62]. Moreover, Aβ was discovered to promote the expression of IRF8, while overexpression of IRF8 aggravated microglial activation.

In module 22 of the AD network (Figure 17), there was an enrichment of several terms related to cytokines, with several specific terms of cytokine production of IFN-γ, IL-2, IL-6, IL-8, and TNF. NFATC2 is a regulator of cytokine release[60]. In addition, there were terms related to phagocytosis. TYROBP is required in microglia for phagocytosis, together with TREM2. There were as well some terms related to T-cells, such as 'regulation of T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell', 'TCR signaling', 'Costimulation by the CD28 family' and 'Phosphorylation of CD3 and TCR zeta chains'. NFATC2 and RUNX1 are both involved in T-cell functioning[60]. Mutations in RUNX1 have been associated with increased risk for AD[70]. Lastly, there were some terms related to Toll-like receptor (TLR) signaling. On the other hand, in module 24 of the MDD network (Figure

18), the cytokine response (IFN-γ, IL-10) was not dominant, but rather neutrophil activation stood out. Neutrophils and IFN-γ are increased in depression, while IL-10 is decreased (Figure 5)[22]. The α-subunit of the IL-10 receptor (IL10RA) was one of the genes shared in the two modules. In addition, there were some enriched terms related to B- and T-cells. B-, T_H2 and T_{reg} cells are decreased, while T_H1 and T_H17 cells are increased in depression[22]. As there were some terms with negative regulation, it is hard to determine which cells and/or pathways are decreased or increased here. Hence, this needs further research. Phagocytosis was recurring in this module as well. Despite TYROBP not being a part of modules 22 and 24 in the bulk ensemble networks, it was regulated by the TFs IKZF1, IRF8 and RUNX1 in the AD single-cell network and by IKZF1 and IRF8 in the single-cell MDD network. TYROBP plays an important role in signal transduction in microglia and has been found to be significantly upregulated in the brain of patients with AD[73]. TYROBP is important for the phagocytic activity of microglia, together with TREM2, which is a well-known risk gene for AD. In addition, TYROBP can also suppress cytokine production and secretion. Thus, TYROBP deficiency could to lead to aberrant phagocytosis and increased cytokine release, which the AD brain tries to counteract by increasing the expression levels of TYROBP. Further, altered binding sites for IKZF1 have been found in AD[68]. IKZF1 is an important TF in hematopoietic cell differentiation[60]. It is known to regulate INPP5D, which is involved in several pathways and is as well an AD risk gene[68]. This gene was part of module 22 of the AD network. INPP5D is a negative regulator of myeloid cell proliferation and of chemotaxis, it is involved in immune cell homeostasis and regulates macrophage phagocytosis and activation, and neutrophil migration. Functional enrichment analysis of the TAL1 regulon in the single-cell MDD network (Figure 25) was as well performed, where cytokine response and microglial activation and phagocytosis were recurring themes. TAL1 is as well a regulator of hematopoietic differentiation[60]. It was found to regulate six circular miRNAs dysregulated in MDD and was associated with a higher risk to develop MDD[63,64]. In addition, disrupted binding sites for TAL1 were found in AD[61].

In addition to the immune-related modules, one module (module 40, Figure 19) of the MDD network was prioritized as well. This module was regulated by the TFs CREB3, HEY1, IKZF4, PLAGL2, THAP11, USF1, ZNF532, and ZNF576. All of these TFs were part of the top 100 regulators of the MDD module (Supplementary table 1). THAP11 was the top one, USF1 the second, CREB3 the sixth, and HEY1 the seventh regulator. This module contained genes related to the mitochondrial respiratory chain complex. Mitochondrial dysfunction has been implicated in both AD and MDD before[14,19,22]. In addition, there were some genes of which the proteins are part of the proteasome complex, which has been implicated in AD and treatment-resistant MDD[74]. Moreover, the proteasome system has been associated with psychosis, BD and SCZ as well. Module 60 of the AD network (Supplementary figure S7) also contained genes involved in mitochondrial pathways and parts of the proteasome. In addition, the gene encoding tau (MAPT) was part of this AD module.

Of the top 100 regulators of the AD and MDD ensemble networks, MEF2C and TFEB were two common regulators. Interestingly, in Pearl et al. these two regulators were found as well as hub regulators in AD[28] (see introduction). In the AD network, TFEB was a regulator of modules 12, 18 and 126, while MEFC2 was a regulator of 34 modules, including module 26 (Supplementary figure S6). MEFC2 was the second top regulator in this network, indicating this TF had a broad function in the AD network. In the MDD network, MEFC2 was a regulator of 13 modules, while TFEB was a regulator of modules 15, 38, 127, and 140. Here, MEF2C was the fifth top regulator. In the AD network, module 12 contained genes involved in myelination by oligodendrocytes, similarly to module 15 from the MDD network. Modules 38 and 127 from the MDD network were related to endocytosis, cytoskeleton reorganization, extracellular matrix organization, and Notch signaling. TFEB is involved in lysosomal degradation and autophagy. It is also involved in the immune response against bacteria and in T-cell-mediated antibody responses[60]. MEF2C is essential for hippocampal-dependent learning and memory by suppressing the number of excitatory synapses, and for normal neuronal development[60]. The regulon of MEF2C was highly active in both the excitatory and

inhibitory neurons in the single-cell network of AD, while MEF2C was not inferred as a TF in the MDD network. TFEB was not a regulator in both single-cell networks. In addition, in this paper from Pearl et al., microglial networks were upregulated as well[28].

The paper from which the AD bulk and scRNA-seq datasets were utilized, has profiled the chromatin accessibility and transcriptome of AD patients and controls[56]. They performed bulk RNA-seq, single-nucleus (sn) RNA-seq, and snATAC-seq from the prefrontal cortex of postmortem brains. snATAC-seq and snRNA-seq data was integrated. They identified several TF motifs with an increased enrichment in AD, in astrocytes, excitatory neurons and microglia[56]. Multiple neuronal and glial subpopulations were found and the composition of each cluster was examined in the context of AD. Moreover, differentially accessible chromatin regions and DEGs were identified in AD for each cell cluster. The researchers indicate this may signify the dysregulation of particular biological pathways in distinct cell populations in AD[56]. Furthermore, they constructed cell-type-specific GRNs for microglia and oligodendrocytes. They identified candidate target genes of a given TF by looking at the accessibility of the promoters and other cis-regulatory elements and whether they contain the TF's binding motif, in the cell type of interest. In these GRNs, they found multiple genes located at known AD genome-wide association studies (GWAS) loci. They also introduced scWGCNA to infer gene co-expression networks with both snRNA and bulk RNA-seq data. Additionally, they performed pseudo-time trajectory analysis in oligodendrocytes, microglia and astrocytes. The researchers have found both cis- and trans-gene regulation disruption in AD. SPI1 was prioritized as a transcriptional repressor in microglia in AD[56]. Interestingly, the regulon of SPI1 was highly activated in the microglia of the single-cell AD network (see Table 7). Disruption of SPI1 binding sites were also found in AD patients in another paper[68].

Pantazatos et al.[57] have profiled the transcriptome of individuals with depression and controls. They made a distinction between suicides and non-suicides. The data comes from the dorsolateral prefrontal cortex of postmortem brains. This dataset was used for bulk network inference in this master's dissertation. Additionally, small RNA-seq was also performed to examine miRNA expression. They choose the dorsolateral prefrontal cortex because it is 'involved in the regulation of impulsivity, decision-making, cognitive control of mood and other executive functions related to suicidal behavior'[57]. None of the individuals took any recent psychotropic medication. DEG analysis was executed between controls, non-suicide depressive patients and suicides. Thus, they looked at both the effects of depression and suicide. In addition, they also looked at differential exon usage. Humanin-like 8 (MTRNR2L8 / HN8) was overexpressed in depression and suicide[57]. This gene has neuroprotective, anti-apoptotic and anti-inflammatory effects. This could be a compensation for the chronic stress in MDD[57]. Furthermore, they executed GO functional enrichment analysis. The expression of genes involved in immune-related and microglial cellular functions was decreased in both the depression and suicide groups. Further, they examined some specific pathways. As such, they found lower expression of genes related to 'oligodendrocyte differentiation' in depression and to 'astrocyte cell migration' in depression and suicide[57]. Lower expression of transcripts involved in 'regulation of synaptic transmission, glutamatergic' was found as well. Their findings of lowered immune-related functions contradict other studies and indicate that MDD is a highly heterogeneous disorder. However, as they mentioned themselves, this lower expression might also be indicative of lower levels of microglia and astrocytes[57]. Brain region differences may also have an effect. The researchers indicated the limitations of their study, which are the modest sample sizes, the examination of only one brain region, and not making a distinction between neurons and glial cells[57]. The first limitation was attempted to overcome in this master's dissertation by using two different datasets. The latter limitation can be overcome with scRNA-seq.

The second dataset for depression originates from a paper where they analyzed the transcriptome of patients diagnosed with SCZ, BD and MDD, and controls[58]. The RNA originates from the anterior cingulate cortex, the dorsolateral prefrontal cortex and the nucleus

accumbens. These regions are often associated with mood alterations, cognition, impulse control, motivation, reward, and pleasure[58]. The data from depressive patients and control individuals from the dorsolateral prefrontal cortex was used in this dissertation. Next to transcriptional profiling, they also did metabolic profiling of the anterior cingulate cortex with mass spectrometry. Similarly as in the previous paper, they executed DEG analysis within each brain region. No genes were significantly differentially expressed between MDD and control samples, in any brain region. The largest number of DEGs were found in SCZ samples in the anterior cingulate cortex. GO enrichment analysis was performed on the genes, and altered metabolites and genes were analyzed for enrichment with KEGG[58]. Moreover, they deconvoluted the expression data into cell types. As such, they found a significant decrease in neuron-specific gene expression and an increase in astrocyte-specific expression in the anterior cingulate cortex, in SCZ and BD compared to controls[58]. Metabolite levels of MDD were similar to control samples. As most changes were found for SCZ, the study largely focused on this disease. The limitations of this study were that women are underrepresented and that there is no information about the smoking status or patient drug use. Some toxicology reports were positive in patients. Another limitation they mention is the RNA quality coming from post-mortem brains[58].

The snRNA-seq data for depression originates from a paper where they used snRNA-seq from MDD cases and psychiatrically healthy controls to identify cell-type-specific DEGs[72]. All samples were from male individuals and all MDD patients died from suicide. The toxicological reports were positive for some of the samples. Three control samples and eight MDD samples contained residues of alcohol/drug use, while three MDD individuals had taken psychiatric medication[72]. The samples originate from the dorsolateral prefrontal cortex. The researchers found 96 DEGs in sixteen different cell clusters. This signifies the complex interplay between different cell types in the pathophysiology of MDD. Most DEGs were found in the clusters of immature oligo-precursor cells and deep layer excitatory neurons[72]. Oligodendrocyte precursor cells are believed to function as a distinct cell type in the brain and have as well been implicated in depressive-like behavior[72]. Functional enrichment analysis of the DEGs implicated terms related to synaptic plasticity, 'kinesins', 'HSP90 chaperone cycle for steroid hormone receptors' and the 'innate immune system'[72]. Next, they also did a network analysis with STRING. From this, dysregulated pathways including cytoskeletal function, immune system function and chaperone cycling were found. In addition, they also used WGCNA with averaged gene expression profiles and the percentages of the contributions of the cell types as correlates[72]. They found five modules significantly associated with depression, of which four were strongly associated with the deep layer excitatory neurons. The researchers indicate some limitations of their study, which were the use of only male individuals, technical limitations of droplet-based snRNA-seq of human brain samples and retrieving a much greater proportion of neurons compared to glial cells[72].

Interestingly, a recent study investigated the shared genetic etiology between AD and MDD using GWAS datasets[75]. The researchers found a moderate level of polygenic overlap between the two disorders. Noteworthy, they found a stronger overlap when the SNPs in the APOE region were excluded from the analysis. This is because there is a very strong association between SNPs in this region and AD, with a change in p-value of up to 227 orders of magnitude, compared to the other SNPs[75]. An enrichment of SNPs was found on chromosome eleven, which were linked to expression regulation in myeloid cells, such as microglia. The associated SNPs were mapped to 40 genes (in closest proximity) when studying enrichment for AD, given MDD GWAS association, of which nine genes (BIN1, CELF1, CR1, FERMT2, MS4A6A, PICALM, PTK2B, SORL1, and SPI1.) were already known as AD risk genes[75]. These nine genes are involved in two major pathways; immune response (CELF1, CR1, MS4A6A, and SPI1) and regulation of endocytosis (BIN1, FERMT2, PICALM, PTK2B, and SORL1). This indicates that indeed MDD and AD both have an immunological compound, and that there is some genetic overlap in immune aberrations. Next, they also did a gene set enrichment analysis, where one interesting term was 'leukocyte transendothelial migration'[75]. SPI1 was

part of module 24 from the ensemble MDD network (Figure 18), and its regulon was highly active in microglia in the single-cell AD network (Table 7). This gene was as well prioritized in the single-cell analysis of AD in a paper above[56]. Similarly, in a new GWAS of AD, the most significantly enriched gene sets were related to tau and amyloid, lipids, endocytosis and to immunity[76]. The endocytosis pathway might be related to microglial phagocytosis, which is impaired in AD[76]. In addition, they found several genes implicated in the TNF-α signaling pathway.

To predict cell-type disease genes and GRNs of neurodegenerative and neuropsychiatric diseases, researchers have developed scGRNom (single-cell Gene Regulatory Network prediction from multi-omics)[77]. It consists of integrating multi-omics data for predicting GRNs and identifying disease genes and regulatory elements. In the first step chromatin interactions are discovered with, for instance, Hi-C data, between enhancers, promotors and genes. Then, TF binding sites are inferred in the interacting regions, outputting a reference GRN linking these TFs, enhancers and promotors of the target genes. In the last step, the final GRN is predicted by TF-target gene expression relationships inferred through elastic net regression[77]. It is also possible to provide open chromatin regions inferred from ATAC-seq to refine the network. To determine the cell-type-specific disease genes and regulatory elements, the cell-specific GRN is needed together with a list of SNPs associated with the disease. The interruption of TF binding sites by the SNPs is investigated and the linked genes and enhancers/promotors are given as output. The researchers applied the pipeline to SCZ and AD and found that enhancers in cell-type-specific GRNs are enriched with GWAS SNPs. When linking these SNPs to genes, they found cross-disease and disease-specific functions at the cell-type level[77]. They used healthy brain multi-omics data with disease-specific SNPs.

In the following paragraph, some limitations of the used method in this master's dissertation will be indicated. Firstly, only TFs were used as regulators, without taking into account co-factors or regulatory RNA molecules such as lncRNA and miRNA. As such, GRNs were inferred, and not regulatory networks. In addition, only transcriptomics data was used to infer the networks, without additional information. However, Merlin-P and KBoost were tried as well, but the output was not as desired. In both methods, a prior network was used, integrating additional data. SCENIC uses additional data as well from RcisTarget. Further, k-medoids was used to assign the genes into modules from the ensemble network. As a consequence, a gene can only belong to one module. In reality, however, genes mostly belong to different pathways. Next, by using publicly available data, there is less control over the samples that are used. For instance, it is not always known if the patients took medication, which might influence the results. Additionally, two different datasets were used for the depressive samples, for which batch effects had to be accounted for. This will always influence the results. Another limitation is the fact that, overall, research is still focused on white individuals. As a consequence, the available samples are mostly from people of European descent. This might influence the results as well, as the genetic diversity is limited. Using samples from individuals of African or Asian descent, for instance, would result in a larger genetic diversity and a better possibility to extrapolate findings to the whole human population, instead of a minority. Lastly, post-mortem brain samples are prone to degradation, which might result in the degradation of RNA molecules[78]. In addition, these samples only represent one snapshot of the patient's lifetime[78].

In conclusion, different pathways have been found that are dysregulated in both AD and MDD. These are the activation of microglia, the mitochondrial respiratory chain and the proteasome system. Especially regarding the activation of microglia, different regulators were found that were implicated in both AD and MDD. Of particular interest were the TFs IKZF1, IRF8, NFATC2, TAL1, RUNX1, and SPI1. All of these TFs have been implicated with AD, but only TAL1 has been implicated with MDD before. In addition, INPP5D, TYROBP and TREM2 have been linked with the activation of microglia in the AD bulk network, and they have been implicated in AD before as well. TYROBP and TREM2 have been linked with the MDD bulk network as well. Moreover, TFEB and MEF2C were two hub regulators in AD and MDD that

were also found to be hub regulators in AD before. All these genes and pathways need to be further investigated, especially in MDD, and can be prioritized when searching for new medication.

4.1 Future perspectives

In future research, more neuroinflammatory disorders should be compared to each other using regulatory network inference. Ideally, this would include different kinds of methods and multi-omics data. These different networks can be combined into an ensemble network and by using distinct methods, a more comprehensive network will be retrieved, for each disorder. As the brain is a heterogeneous organ, not only different brain regions should be studied, but also different cell types and even cell states. Hence, single-cell data should be used to determine the cells and regulators in these cells that play a key role in the disease. With the increasing amount of data becoming available, this should be feasible in the near future. scGRNom could be utilized as one of the methods, using disease-specific single-cell omics data. This will result in a better understanding of the pathophysiology of different psychiatric and neurodegenerative disorders, and might lead to new medication.

5. REFERENCES

1. Pape, K., Tamouza, R., Leboyer, M. & Zipp, F. Immunoneuropsychiatry — novel perspectives on brain disorders. *Nature Reviews Neurology* **15**, 317–328 (2019).
2. Chew, G. & Petretto, E. Transcriptional Networks of Microglia in Alzheimer's Disease and Insights into Pathogenesis. *Genes (Basel)* **10**, (2019).
3. Schwartz, M. & Deczkowska, A. Neurological Disease as a Failure of Brain-Immune Crosstalk: The Multiple Faces of Neuroinflammation. *Trends Immunol* **37**, 668–679 (2016).
4. Wang, Q., Jie, W., Liu, J. H., Yang, J. M. & Gao, T. M. An astroglial basis of major depressive disorder? An overview. *Glia* **65**, 1227–1250 (2017).
5. Caruso Bavisotto, C. *et al.* Extracellular Vesicle-Mediated Cell–Cell Communication in the Nervous System: Focus on Neurological Diseases. *Int J Mol Sci* **20**, (2019).
6. Hansen, D. V., Hanson, J. E. & Sheng, M. Microglia in Alzheimer's disease. *J Cell Biol* **217**, 459–472 (2018).
7. De Virgilio, A. *et al.* Parkinson's disease: Autoimmunity and neuroinflammation. *Autoimmunity Reviews* **15**, 1005–11 (2016).
8. Pitsillou, E. *et al.* The cellular and molecular basis of major depressive disorder: towards a unified model for understanding clinical depression. *Mol Biol Rep* **47**, 753–770 (2020).
9. Dietz, A. G., Goldman, S. A. & Nedergaard, M. Glial cells in schizophrenia: a unified hypothesis. *Lancet Psychiatry* **7**, 272–281 (2020).
10. Benedetti, F., Aggio, V., Pratesi, M. L., Greco, G. & Furlan, R. Neuroinflammation in Bipolar Depression. *Front Psychiatry* **11**, 71 (2020).
11. Huang, Y. *et al.* Integrated multifactor analysis explores core dysfunctional modules in autism spectrum disorder. *Int J Biol Sci* **14**, 811–818 (2018).
12. Masi, A., DeMayo, M. M., Glozier, N. & Guastella, A. J. An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options. *Neurosci Bull* **33**, 183–193 (2017).
13. Dobson, R. & Giovannoni, G. Multiple sclerosis - a review. *Eur J Neurol* **26**, 27–40 (2019).
14. Soria Lopez, J. A., González, H. M. & Léger, G. C. Alzheimer's disease. *Handb Clin Neurol* **167**, 231–255 (2019).
15. Lane, C. A., Hardy, J. & Schott, J. M. Alzheimer's disease. *Eur J Neurol* **25**, 59–70 (2018).
16. Del Tredici, K. & Braak, H. Idiopathic Parkinson's Disease: Staging an α-Synucleinopathy with a Predictable Pathoanatomy. *Molecular Mechanisms in Parkinson's Disease* (2000).
17. Troutman, T. D., Kofman, E. & Glass, C. K. Exploiting dynamic enhancer landscapes to decode macrophage and microglia phenotypes in health and disease. *Mol Cell* **81**, 3888–3903 (2021).
18. Liddelow, S. A. *et al.* Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).
19. Liu, C. S., Adibfar, A., Herrmann, N., Gallagher, D. & Lanctôt, K. L. Evidence for Inflammation-Associated Depression. in *Inflammation-Associated Depression: Evidence, Mechanisms and Implications* (eds. Dantzer, R. & Capuron, L.) 3–30 (Springer International Publishing, 2017). doi:10.1007/7854_2016_2.
20. Boku, S., Nakagawa, S., Toda, H. & Hishimoto, A. Neural basis of major depressive disorder: Beyond monoamine hypothesis. *Psychiatry Clin Neurosci* **72**, 3–12 (2018).
21. Zhang, Q. *et al.* Kynurenine regulates NLRP2 inflammasome in astrocytes and its implications in depression. *Brain Behav Immun* **88**, 471–481 (2020).
22. Drevets, W. C., Wittenberg, G. M., Bullmore, E. T. & Manji, H. K. Immune targets for therapeutic development in depression: towards precision medicine. *Nat Rev Drug Discov* 1–21 (2022) doi:10.1038/s41573-021-00368-1.
23. Saint-Antoine, M. M. & Singh, A. Network inference in systems biology: recent developments, challenges, and applications. *Current Opinion in Biotechnology* **63**, 89–98 (2020).
24. Banf, M. & Rhee, S. Y. Computational inference of gene regulatory networks: Approaches, limitations and opportunities. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1860**, 41–52 (2017).
25. Yan, J., Risacher, S. L., Shen, L. & Saykin, A. J. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics* **19**, 1370–1381 (2018).
26. Roy, S. *et al.* Integrated Module and Gene-Specific Regulatory Inference Implicates Upstream Signaling Networks. *PLoS Computational Biology* **9**, (2013).
27. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
28. Pearl, J. R. *et al.* Genome-Scale Transcriptional Regulatory Network Models of Psychiatric and Neurodegenerative Disorders. *Cell Syst* **8**, 122-135.e7 (2019).

29. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* **9**, 796–804 (2012).
30. Faith, J. J. *et al.* Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLOS Biology* **5**, e8 (2007).
31. Nguyen, H., Tran, D., Tran, B., Pehlivan, B. & Nguyen, T. A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data. *Briefings in bioinformatics* (2020) doi:10.1093/bib/bbaa190.
32. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS One* **5**, e12776 (2010).
33. Ghosh Roy, G., Geard, N., Verspoor, K. & He, S. PoLoBag: Polynomial Lasso Bagging for signed gene regulatory network inference from expression data. *Bioinformatics* **36**, 5187–5193 (2021).
34. Bonnet, E., Calzone, L. & Michoel, T. Integrative Multi-omics Module Network Inference with Lemon-Tree. *PLOS Computational Biology* **11**, e1003983 (2015).
35. Siahpirani, A. F. & Roy, S. A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Res* **45**, e21 (2017).
36. Erola, P., Bonnet, E. & Michoel, T. Learning Differential Module Networks Across Multiple Experimental Conditions. in *Gene Regulatory Networks: Methods and Protocols* (eds. Sanguinetti, G. & Huynh-Thu, V. A.) 303–321 (Springer, 2019). doi:10.1007/978-1-4939-8882-2_13.
37. Iglesias-Martinez, L. F., De Kegel, B. & Kolch, W. KBoost: a new method to infer gene regulatory networks from gene expression data. *Sci Rep* **11**, 15461 (2021).
38. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* **17**, 147–154 (2020).
39. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086 (2017).
40. Friedman, J. H. Stochastic gradient boosting. *Computational Statistics & Data Analysis* **38**, 367–378 (2002).
41. Tantardini, M., Ieva, F., Tajoli, L. & Piccardi, C. Comparing methods for comparing networks. *Scientific reports* **9**, 17557 (2019).
42. Koutrouli, M., Karatzas, E., Paez-Espino, D. & Pavlopoulos, G. A. A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology* **8**, 34 (2020).
43. Yaveroğlu, Ö. N. *et al.* Revealing the Hidden Language of Complex Networks. *Scientific reports* **4**, 4547 (2014).
44. Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).
45. Sadeghi, I. *et al.* Brain transcriptomic profiling reveals common alterations across neurodegenerative and psychiatric disorders. *bioRxiv* 2021.08.16.456345 (2021) doi:10.1101/2021.08.16.456345.
46. Godini, R., Fallahi, H. & Ebrahimie, E. A comparative system-level analysis of the neurodegenerative diseases. *Journal of Cellular Physiology* **234**, 5215–5229 (2019).
47. Vermeirssen, V., De Clercq, I., Van Parys, T., Van Breusegem, F. & Van de Peer, Y. Arabidopsis Ensemble Reverse-Engineered Gene Regulatory Network Discloses Interconnected Transcription Factors in Oxidative Stress. *The Plant Cell* **26**, 4656 (2014).
48. Defoort, J., Van de Peer, Y. & Vermeirssen, V. Function, dynamics and evolution of network motif modules in integrated gene regulatory networks of worm and plant. *Nucleic Acids Res* **46**, 6480–6503 (2018).
49. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, R25 (2010).
50. Lovering, R. C. *et al.* A GO catalogue of human DNA-binding transcription factors. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 194765 (2021) doi:10.1016/j.bbagrm.2021.194765.
51. Meyer, P. E., Lafitte, F. & Bontempi, G. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinformatics* **9**, 461 (2008).
52. Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods* **13**, 366–370 (2016).
53. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* **48**, 838–847 (2016).

54. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–504 (2003).
55. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
56. Morabito, S. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nature genetics* **53**, 1143–1155 (2021).
57. Pantazatos, S. P. *et al.* Whole-transcriptome brain expression and exon-usage profiling in major depression and suicide: evidence for altered glial, endothelial and ATPase activity. *Molecular Psychiatry* **22**, 760–773 (2017).
58. Ramaker, R. C. *et al.* Post-mortem molecular profiling of three psychiatric disorders. *Genome Medicine* **9**, 72 (2017).
59. Lu, Y., Zhou, X. & Nardini, C. Dissection of the Module Networks Implementation "LemonTree": Enhancements towards Applications in Metagenomics and Translation in Autoimmune Maladies. *Mol. BioSyst.* **13**, (2017).
60. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2021).
61. Schwartzentruber, J. *et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nature Genetics* **53**, 392–402 (2021).
62. Zeng, Q. *et al.* IRF-8 is Involved in Amyloid-β1-40 (Aβ1-40)-induced Microglial Activation: a New Implication in Alzheimer's Disease. *Journal of Molecular Neuroscience* **63**, 159–164 (2017).
63. Samaan, Z. *et al.* Obesity genes and risk of major depressive disorder in a multiethnic population: a cross-sectional study. *The Journal of Clinical Psychiatry* **76**, e1611-1618 (2015).
64. Rasheed, M. *et al.* A Systematic Review of Circulatory microRNAs in Major Depressive Disorder: Potential Biomarkers for Disease Prognosis. *International Journal of Molecular Sciences* **23**, 1294 (2022).
65. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **48**, D845–D855 (2020).
66. Katayama, H. Anti-interleukin-17A and anti-interleukin-23 antibodies may be effective against Alzheimer's disease: Role of neutrophils in the pathogenesis. *Brain Behav* **10**, e01504 (2020).
67. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat Commun* **9**, 1090 (2018).
68. Rosenthal, S. L., Barmada, M. M., Wang, X., Demirci, F. Y. & Kamboh, M. I. Connecting the Dots: Potential of Data Integration to Identify Regulatory SNPs in Late-Onset Alzheimer's Disease GWAS Findings. *PLOS ONE* **9**, e95152 (2014).
69. Manocha, G. D. *et al.* NFATc2 Modulates Microglial Activation in the AβPP/PS1 Mouse Model of Alzheimer's Disease. *J Alzheimers Dis* **58**, 775–787 (2017).
70. Kimura, R. *et al.* The DYRK1A gene, encoded in chromosome 21 Down syndrome critical region, bridges between β-amyloid production and tau phosphorylation in Alzheimer disease. *Human Molecular Genetics* **16**, 15–23 (2007).
71. Lobo-Silva, D., Carriche, G. M., Castro, A. G., Roque, S. & Saraiva, M. Balancing the immune response in the brain: IL-10 and its regulation. *Journal of Neuroinflammation* **13**, 297 (2016).
72. Nagy, C. *et al.* Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat Neurosci* **23**, 771–781 (2020).
73. Ma, J., Jiang, T., Tan, L. & Yu, J.-T. TYROBP in Alzheimer's disease. *Mol Neurobiol* **51**, 820–826 (2015).
74. Minelli, A. *et al.* Proteasome system dysregulation and treatment resistance mechanisms in major depressive disorder. *Transl Psychiatry* **5**, e687 (2015).
75. Lutz, M. W., Sprague, D., Barrera, J. & Chiba-Falek, O. Shared genetic etiology underlying Alzheimer's disease and major depressive disorder. *Translational Psychiatry* **10**, (2020).
76. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* 1–25 (2022) doi:10.1038/s41588-022-01024-z.
77. Jin, T. *et al.* scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. *Genome Med* **13**, 95 (2021).
78. Dong, X., Liu, C. & Dozmorov, M. Review of multi-omics data resources and integrative analysis for human brain disorders. *Brief Funct Genomics* **20**, 223–234 (2021).

ADDENDUM 1: POSTER

FACULTY OF MEDICINE AND HEALTH SCIENCES

CBIGR

LAB FOR COMPUTATIONAL BIOLOGY, INTEGROMICS AND GENE REGULATION (CBIGR)

REGULATORY NETWORKS IN NEUROINFLAMMATORY DISORDERS

Hanne Puype°, Vanessa Vermeirssen*

° Second master student Biomedical Sciences (major Systems Biology)

*** Department of Biomedical Molecular Biology (WE14); Department of Biomolecular Medicine (GE31); Cancer Research Institute Ghent (CRIG)**

Introduction

Alzheimer's disease (AD) and major depression disorder (MDD) have a high burden on global health. Both disorders still lack effective treatment. Recent insights indicate that several brain disorders are characterized by similar pathways, of which one is neuroinflammation. However, there are some specific processes, as these shared pathophysiological mechanisms can result in entirely distinct disorders (see fig. 1). Finding a common pathway that could be targeted in different disorders, would possibly shine a new light on the treatment of neuroinflammatory disorders.

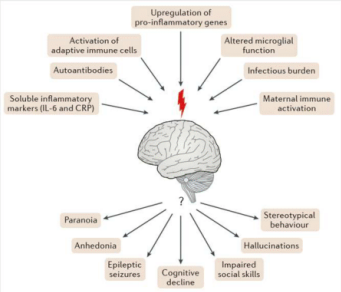


Figure 1: Common pathophysiological mechanisms of immune dysregulation in neuroinflammatory disorders and their distinct phenotypes emerging from this. Adapted from Pape *et al.*[1]

Objectives

The objective of this project is to find distinct and common pathways between Alzheimer's disease and depression, and their regulators. As such, gene regulatory networks will be inferred with publicly available RNA sequencing data. Therefor, different methods will be adopted. The different networks will then be compared between each other and between the disorders.

Methods

- RNA-seq from Gene Expression Omnibus (GEO) and Synapse
- Prefrontal cortex post-mortem brain samples from AD, MDD and controls
- Networks inferred with GENIE3, CLR and Lemon-Tree
- Top 100 000 edges selected
- Overlap between different methods
- Functional characterization with GO and KEGG

Results

- AD dataset[2]: 47 AD samples, 48 control samples
- MDD datasets[3,4]: 53 MDD samples, 53 control samples
- Directed edges (regulator → target gene)
- The three methods retrieved different outcomes (see fig. 2)
- Top 100 regulators: 45 and 40 shared regulators in the three networks from the three methods, for AD and MDD respectively (see fig. 3)
- Functional enrichment of Lemon-Tree networks: module from AD and module from MDD enriched for immunological terms and large overlap (see fig. 4)

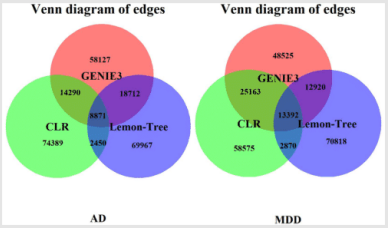


Figure 2: Venn diagrams of the edges in the networks retrieved by CLR, GENIE3 and Lemon-Tree, for AD and MDD.

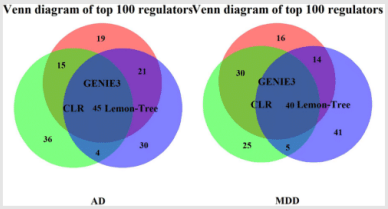


Figure 3: Venn diagrams of the top 100 regulators of the networks retrieved by the three methods, for AD and MDD.

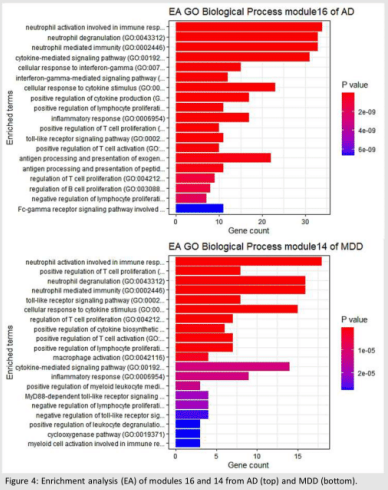


Figure 4: Enrichment analysis (EA) of modules 16 and 14 from AD (top) and MDD (bottom).

Discussion

Three different methods were used to infer GRNs of AD, MDD and control samples. These networks brought about different results, with minimal overlap (see fig. 3).

Functional enrichment analysis

Functional enrichment analysis was executed on the modules of the Lemon-Tree networks. Different modules were enriched for immunological functions in the Gene Ontology Biological Process terms. In particular, modules 16 and 14 from AD and MDD respectively were of interest, because of an extensive overlap. Figure 4 shows the top 20 enriched terms according to p-value. In module 16 there are 153 genes, in module 14 111. Of these, there are 68 in common (see table 1). There are eight and nine regulators of these modules, with three shared regulators (IRF8, IKZF1, TAL1). All three are implicated in hematopoietic cell differentiation[5]. Moreover, IRF8 plays a regulatory role in immune cells and is involved in IFN response[5]. TAL1 is highly expressed in microglia[6]. IRF8 is implicated in microglial activation and neuroinflammation in AD mice models[7]. Several studies have found mutations which interrupt binding of TAL1 in patients with AD[6]. Next to this, TAL1 has been found to be implicated in MDD in two studies[8,9].

Table 1: Overview of the shared genes in the two modules and of the module regulators

	GENES IN MODULES	TRANSCRIPTION FACTORS
ALZHEIMER'S DISEASE	153	ATOH8, IKZF1, IRF8, NFATC2, RUNX1, RUNX2, TAL1, TCF3
DEPRESSION	111	FOS, IKZF1, IRF8, MAF, RHOXF2, SP11, TAL1, TFEC, ZNF551
IN COMMON	ADAM28, ADAP2, AIF1, ALOX5AP, AOAH, APBB1IP, ARHGDIB, BLNK, C1QA, C1QB, C1QC, C3, C3AR1, CCR1, CD33, CD4, CD53, CD84, CD86, CLEC7A, CSF1R, CYTH4, DOCK2, DOCK8, FAM177B, FYB, HCK, HLA-DMB, HLA-DOA, HLA-DPA1, IGSF6, IKZF1, IL18, IRF8, ITGAM, ITGB2, LAIR1, LAPTM5, LILRB4, LPAR5, LPAR6, LPCAT2, MNDA, NCKAP1L, PIK3AP1, PIK3R5, PLCG2, PTAFR, PTGS1, PTPRC, RASAL3, SAMSN1, SIGLEC8, SLA, SLC2A5, SYK, TAL1, TBXAS1, TFEC, TLR1, TLR3, TLR5, TLR6, TMEM119, TMEM156, TYROBP, VAV1, WDFY4	IKZF1, IRF8, TAL1

Future directions

A fourth method will be used to infer gene regulatory networks. This will be MERLIN-P. Next, the networks will be combined into an ensemble network with rank aggregation, for each disease. These ensemble networks will then be further characterized with functional enrichment and the networks of AD and MDD will be compared to each other. Next to this, networks will be inferred from single cell RNA-seq data, using SCENIC.

References

1. Pape, K., Tamouza, R., Leboyer, M. & Zipp, F. Immunoneuropsychiatry — novel perspectives on brain disorders. *Nature Reviews Neurology* **15**, 317–328 (2019).
2. Morabito, S. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nature Genetics* **53**, 1143–1155 (2021).
3. Pantazatos, S. P. *et al.* Whole-transcriptome brain expression and exon-usage profiling in major depression and suicide: evidence for altered glial, endothelial and ATPase activity. *Molecular Psychiatry* **22**, 760–773 (2017).
4. Ramaker, R. C. *et al.* Post-mortem molecular profiling of three psychiatric disorders. *Genome Medicine* **9**, 72 (2017).
5. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2021).
6. Schwartzentruber, J. *et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nature Genetics* **53**, 392–402 (2021).
7. Zeng, Q. *et al.* IRF-8 is Involved in Amyloid-β1-40 (Aβ1-40)-induced Microglial Activation: a New Implication in Alzheimer's Disease. *Journal of Molecular Neuroscience* **63**, 159–164 (2017).
8. Samaan, Z. *et al.* Obesity genes and risk of major depressive disorder in a multiethnic population: a cross-sectional study. *The Journal of Clinical Psychiatry* **76**, e1611-1618 (2015).
9. Rasheed, M. *et al.* A Systematic Review of Circulatory microRNAs in Major Depressive Disorder: Potential Biomarkers for Disease Prognosis. *International Journal of Molecular Sciences* **23**, 1294 (2022).

Contact

Hanne.Puype@ugent.be
www.irc.ugent.be/index.php?id=vermeirssenhome
Universiteit Gent
@ugent
Hanne Puype

GHENT UNIVERSITY

CBIGR

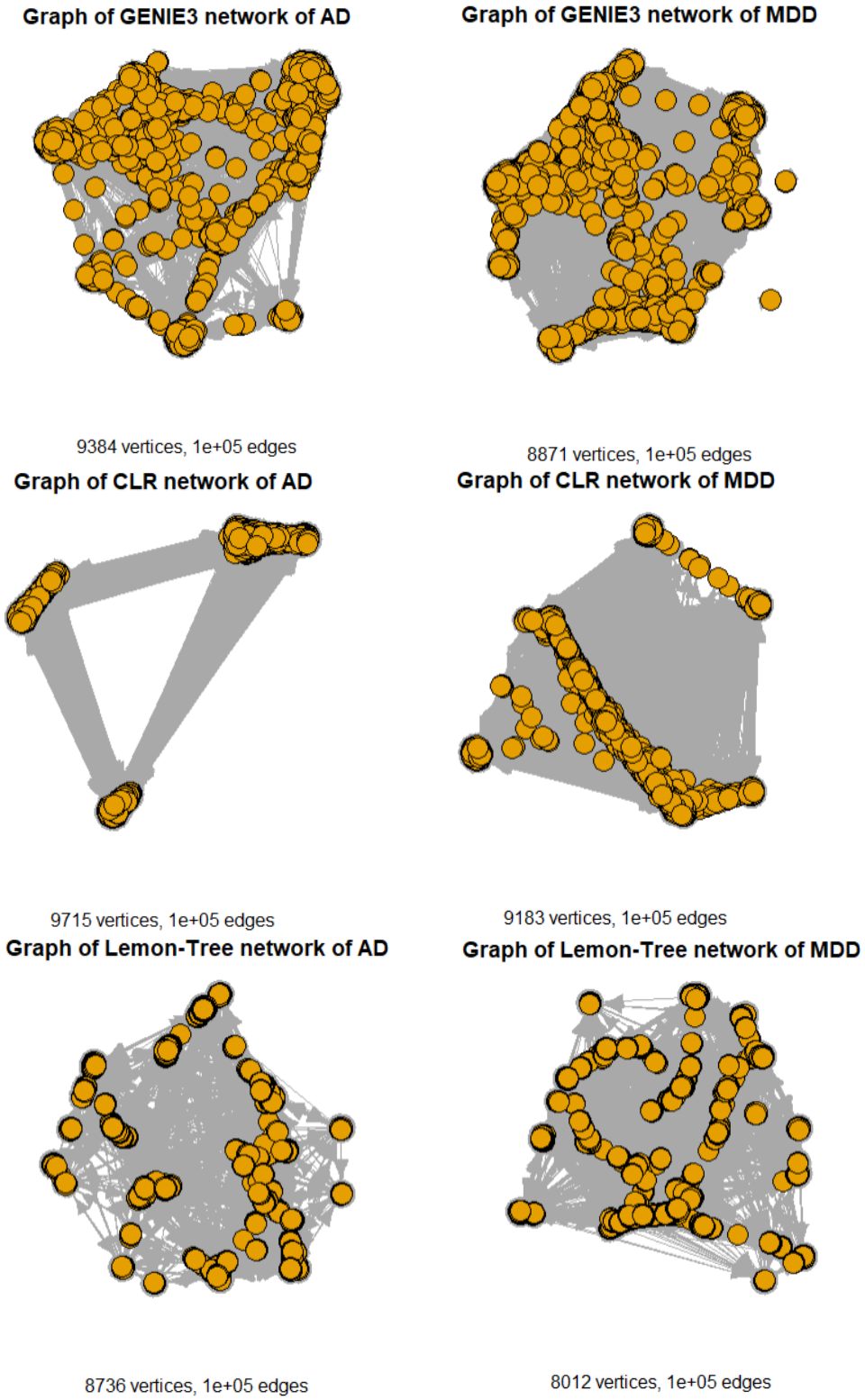
CRIG CANCER RESEARCH INSTITUTE GHENT

GUTBRAINAXIS

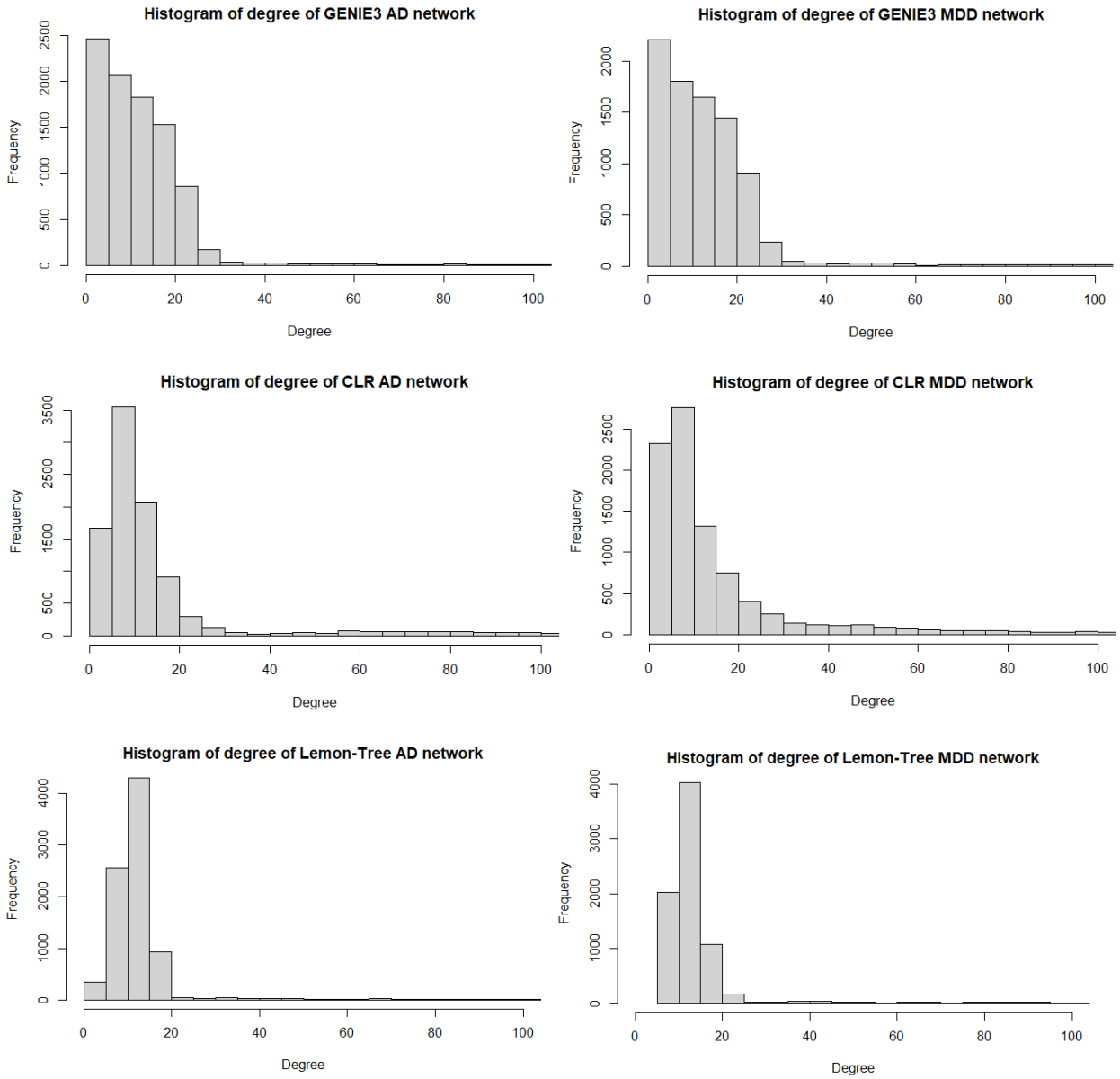
ADDENDUM 2: LAB NOTEBOOK

For the lab notebook, I used GitHub: https://github.ugent.be/vermeirssenlab/neuro_Hanne. More specifically, the scripts and images are saved in the *Code* tab, while the *Wiki* tab was used to write everything out. The same directory as last year was used. The repository is part of the page of the lab of Prof. dr. ir. Vanessa Vermeirssen. The repository is made public, so everyone with a UGent account can reach it. The lab notebook on the *Wiki* tab starts at 2.0 for the Master's dissertation. The pages are in chronological order and there is frequently referred to the code script, which can be reached by clicking the links. I have tried to make the titles of the *Code* folders as informative as possible, but I sometimes included some additional information with a README file. For R Markdown (Rmd) files, the HTML and PDF files were added as well. I recommend reading the *Wiki* pages first and then looking at the code and additional figures of the respective page.

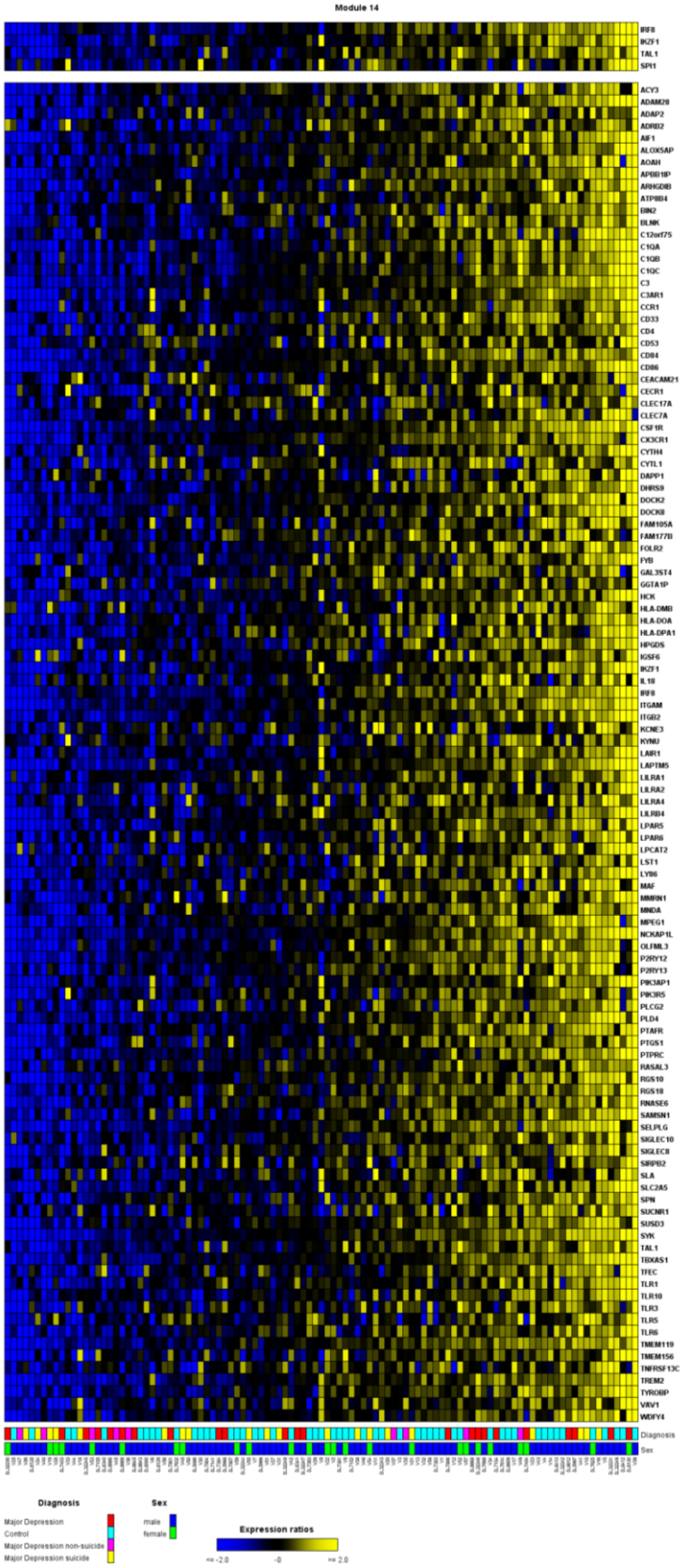
ADDENDUM 3: SUPPLEMENTARY FIGURES



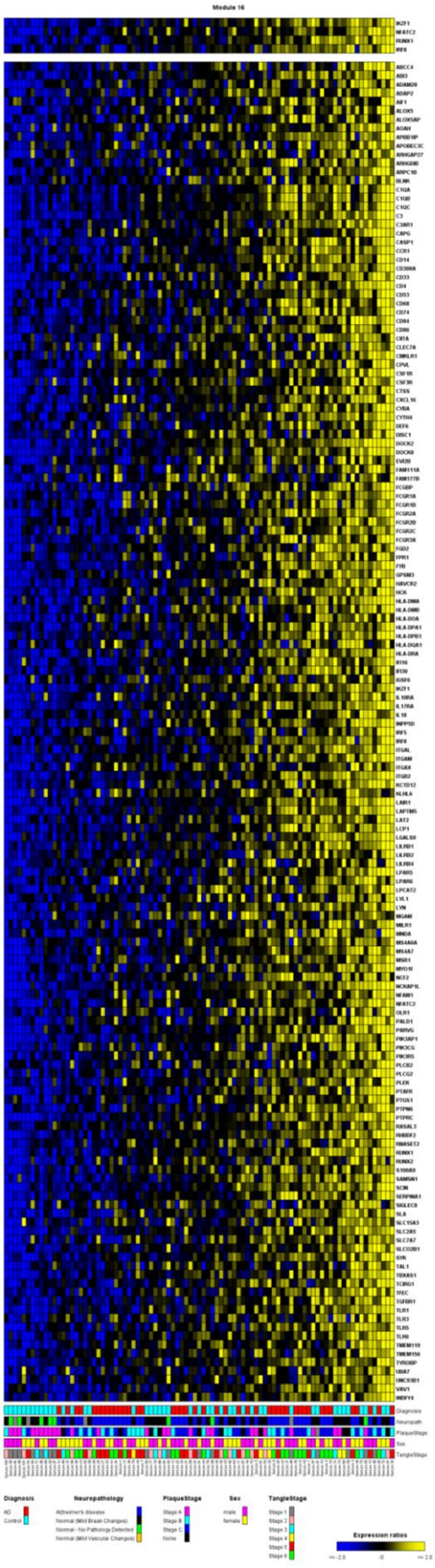
Supplementary figure S1. Representation of the different method-specific networks of Alzheimer's disease (AD) and major depressive disorder (MDD). The graphs were made with the *igraph* package in R. The broad topological characteristics are distinct in the networks from the different methodologies.



Supplementary figure S2. Degree distributions of the different method-specific networks, for Alzheimer's disease (AD) and depression (MDD). The x-axis represents the degree and was set to a limit of 100 because of visualization purposes. Most nodes have a degree of less than twenty, while few nodes have a large degree. The maximal degree of all networks were: GENIE3 AD 2003, GENIE3 MDD 992, CLR AD 562, CLR MDD 478, Lemon-Tree AD 1954, and Lemon-Tree MDD 1077. The maximal degree of the AD networks were consistently higher than the maximal degree of the MDD networks.



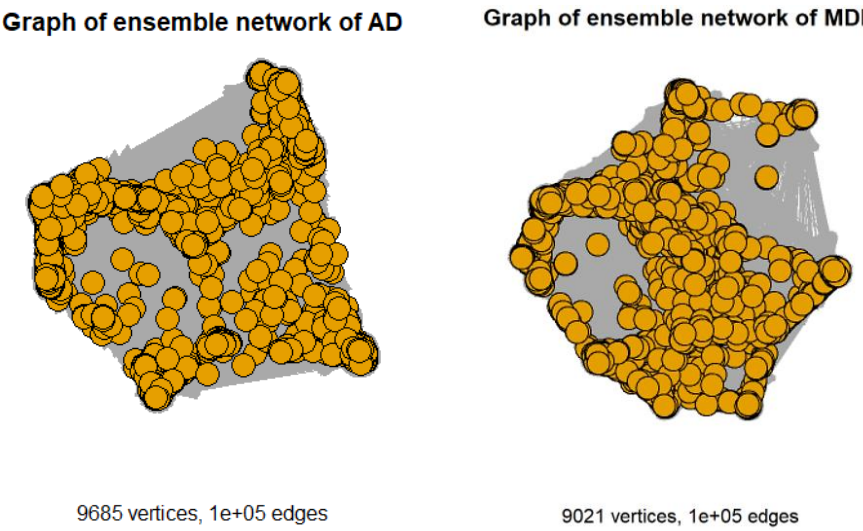
Supplementary figure S3. Module Viewer figure of the gene expression of Module 14 from the MDD Lemon-Tree network. In the upper panel are the regulators of this module. The annotation data diagnosis and sex are included as well (see legend). The division between 'Major depression non-suicide' and 'Major depression suicide' is from the dataset GSE101521, while 'Major depression' is from the dataset GSE80655. 'Control' is from both datasets.


Module 16
IKZF1
NFATC2
RUNX1
PPP
ABCC4
AIM3
ADAM28
ADAP2
AIF1
ALOX5
ALOX5AP
AOAH
APOBEC3C
ARHGAP27
ARHGEF6
ARPC1B
BLNK
C1QA
C1QB
C1QC
C3
C3AR1
CAPG
CASP1
CCR1
CD14
CD300A
CD33
CD4
CD53
CD68
CD74
CD84
CD86
CBTA
CLEC7A
CMKLR1
CPVL
CSF1R
CSF3R
CTSS
CXCL16
CYBA
CYTH4
DEF6
DISC1
DOCK2
DOCK8
EVI2B
FAM115A
FAM177B
FCGBP
FCGR1A
FCGR1B
FCGR2A
FCGR2B
FCGR2C
FCGR3A
FGD2
FPR1
FYB
GPSM3
HAVCR2
HCK
HLA-DMA
HLA-DMB
HLA-DOA
HLA-DPB1
HLA-DPB1
HLA-DQA1
HLA-DRA
IER5
IF30
IGSF6
IKZF1
IL10RA
IL17RA
IL18
INPP5D
IRF5
IRF8
ITGAL
ITGAM
ITGAX
ITGB2
KCTD12
KLHL6
LAIR1
LAPTM5
LAT2
LCP1
LGALS9
LILRB1
LILRB2
LILRB4
LPAR5
LPAR6
LPCAT2
LY6.1
LYN
MGAM
MILR1
MNDA
MS4A6A
MS4A7
MSR1
MYO1F
NCF2
NCKAP1L
NFAM1
NFATC2
OLR1
P4LD1
PARVG
PIK3AP1
PIK3CG
PIK3R5
PLCB2
PLCG2
PLEK
PTAFR
PTGS1
PTPN6
PTPRC
RASAL3
RHBDF2
RNASET2
RUNX1
RUNX2
S100A9
SAMSN1
SCIN
SERPINA1
SIGLEC8
SLA
SLC15A3
SLC2A5
SLC7A7
SLCO2B1
SYK
TBL1
TBXAS1
TCIRG1
TEC
TGFBI1
TLR1
TLR3
TLR5
TLR6
TMEM119
TMEM156
TYROBP
UBA7
UNC93B1
VAV1
WDFY3
Diagnosis
Neuropath
PlaqueStage
Sex
TangleStage
Diagnosis
AD
Control
Neuropathology
Alzheimer's disease
Normal (Mild Braak Changes)
Normal - No Pathology Detected
Normal (Mild Vascular Changes)
PlaqueStage
Stage A
Stage B
Stage C
None
Sex
male
female
TangleStage
Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Stage 6
Expression ratios
<= -2.0
-0
>= 2.0

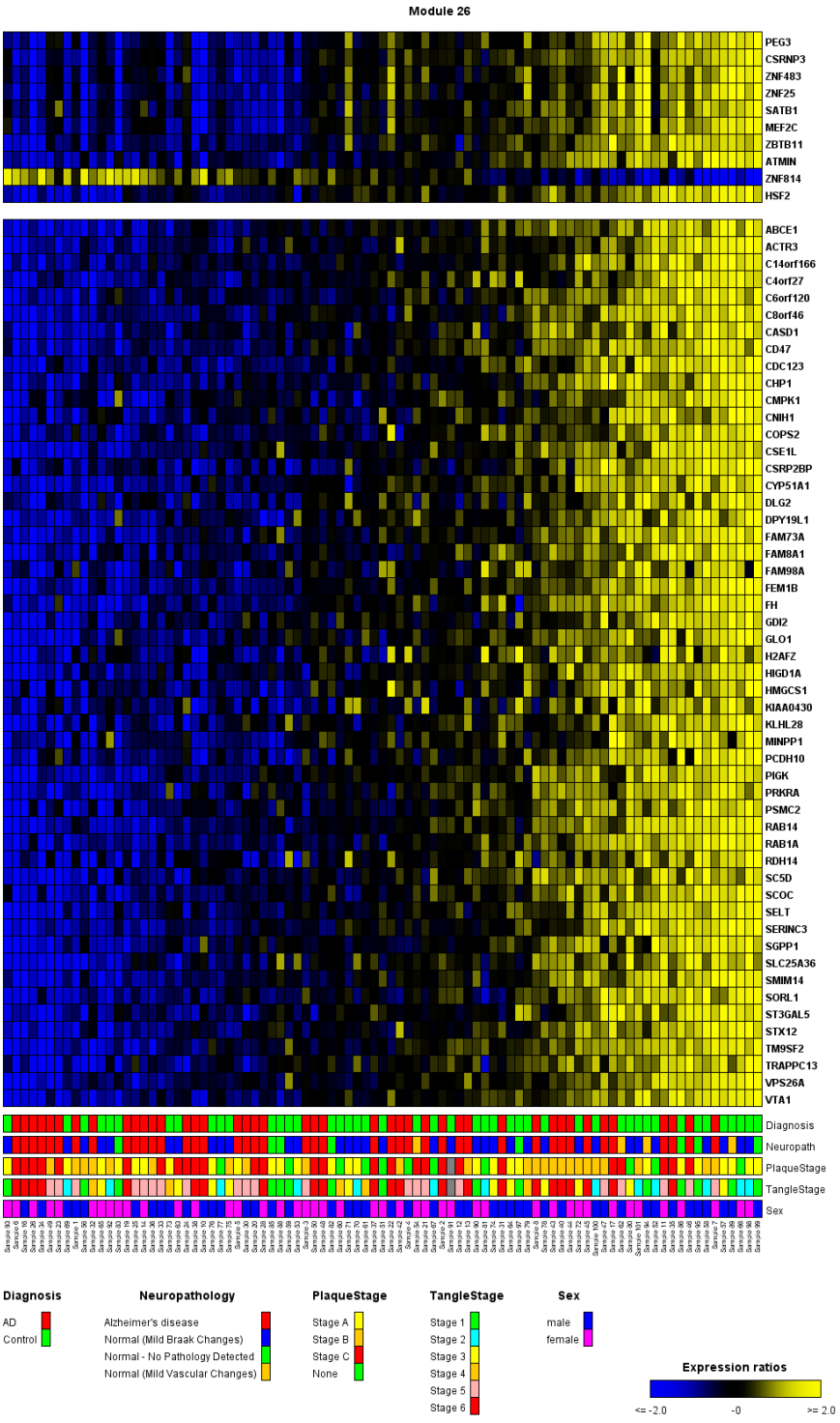
Supplementary figure S4. Module Viewer figure of the gene expression of Module 16 from the AD Lemon-Tree network. In the upper panel are the regulators of this module. The annotation data diagnosis, neuropathology, tangle stage, plaque stage and sex are included as well (see legend).

Supplementary table 1. Top 100 regulators of the ensemble Alzheimer's disease and depression networks. The transcription factors are ordered according to their out-degree.

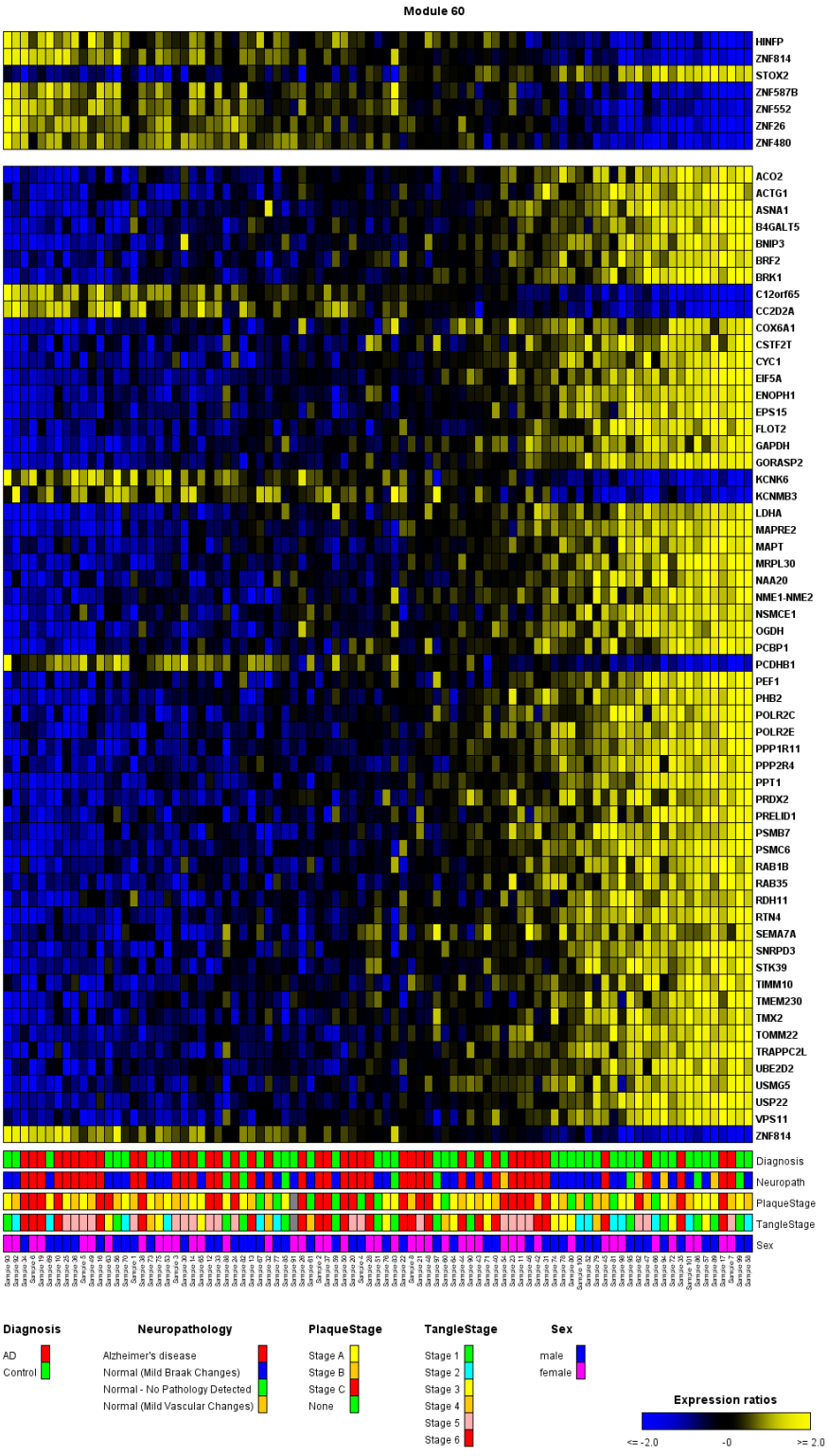
ALZHEIMER'S DISEASE NETWORK	CSRNP3, MEF2C, PEG3, HLF, SATB1, PRDM2, THRB, ZNF25, NKRF, MYT1L, ZNF814, DACH2, ZNF483, STOX2, ZNF26, MEF2D, ZNF480, EGR3, ZBTB11, HIVEP2, CDC5L, ZSCAN30, ZNF777, HINFP, ZNF552, ZNF512, STAT4, BCL11A, POU5F2 TCF7, ZNF280B, ZNF621, ATMIN, ZNF587B, ZNF204P, ZBTB37, ZNF471, ZNF711, ZNF184, ATF2, MEF2A, CREB5, SOX10, PROX1, SOX8, ST18, NFIX, MYRF, ATF7, ZNF528, ZNF536, NFIA, LHX6, ZNF345, CTCF, BBX, TBR1, ARNT2, NEUROD6, ZBTB45, NFE2L3, STAT3, TCF12, RBPJ, NKX6-2, ZNF833P, POU2F1, ZNF317, PLSCR1, FOXJ3, TFEB, CREB3L2, ZBTB4, FOXN2, ZNF652, RFX4, ZBED3, OLIG1, SOX9, ZNF382, ZNF99, EPAS1, ZNF322, VEZF1, ELF1, ZNF217, KLF6, RUNX1, SP1, PPARA, ZGPAT, ZNF692, CREB3, ZBTB17, NR2C2, ZFP1, ZNF562, ZBED6, ZNF492, IKZF3
MAJOR DEPRESSIVE DISORDER NETWORK	THAP11, USF1, MLXIP, ATF2, MEF2C, CREB3, HEY1, HLF, CSRNP3, ZFHX2, ZNF576, ZBTB18, HSF1, NR2F6, ZNF787, ENO1, ZNF316, ZBTB22, SOX9, ATF7, ZNF25, HIVEP2, ATF4, ZNF711, PEG3, THRB, ST18, CIC, IRF3, MESP1, WIZ, RXRB, MAFG, MYRF, NPAS3, PPARA, SALL1, GLI3, RFX1, MEF2A, PLAGL2, SOX10, RXRA, NR2E1, ZHX2, MLXIPL, NKX6-2, DEAF1, KLF15, CC2D1A, ZNF532, SKI, TFEB, NR1H2, ZNF552, IKZF4, PAX6, SOX21, ZBTB11, ZNF518B, ZNF660, RFX4, NFKB2, ZNF536, POU5F2, TRPS1, KLF16, ZFP57, TEAD1, FOXO1, ZNF579, ATMIN, SREBF1, PLSCR1, RREB1, SOX2, SOX8, STOX1, SATB1, GLI4, VEZF1, USF2, ZFP1, ZNF204P, ZNF322P1, ZNF282, ZNF512B, MEIS3, RELA, ZNF444, CDC5L, MAZ, OLIG2, NKRF, ZNF783, ZNF865, HIVEP3, RARB, TCF12, HSF4
SHARED TRANSCRIPTION FACTORS	CSRNP3, MEF2C, PEG3, HLF, SATB1, THRB, ZNF25, NKRF, ZBTB11, HIVEP2, CDC5L, ZNF552, POU5F2, ATMIN, ZNF204P, ZNF711, ATF2, MEF2A, SOX10, SOX8, ST18, MYRF, ATF7, ZNF536, TCF12, NKX6-2, PLSCR1, TFEB, RFX4, SOX9, VEZF1, PPARA, CREB3, ZFP1



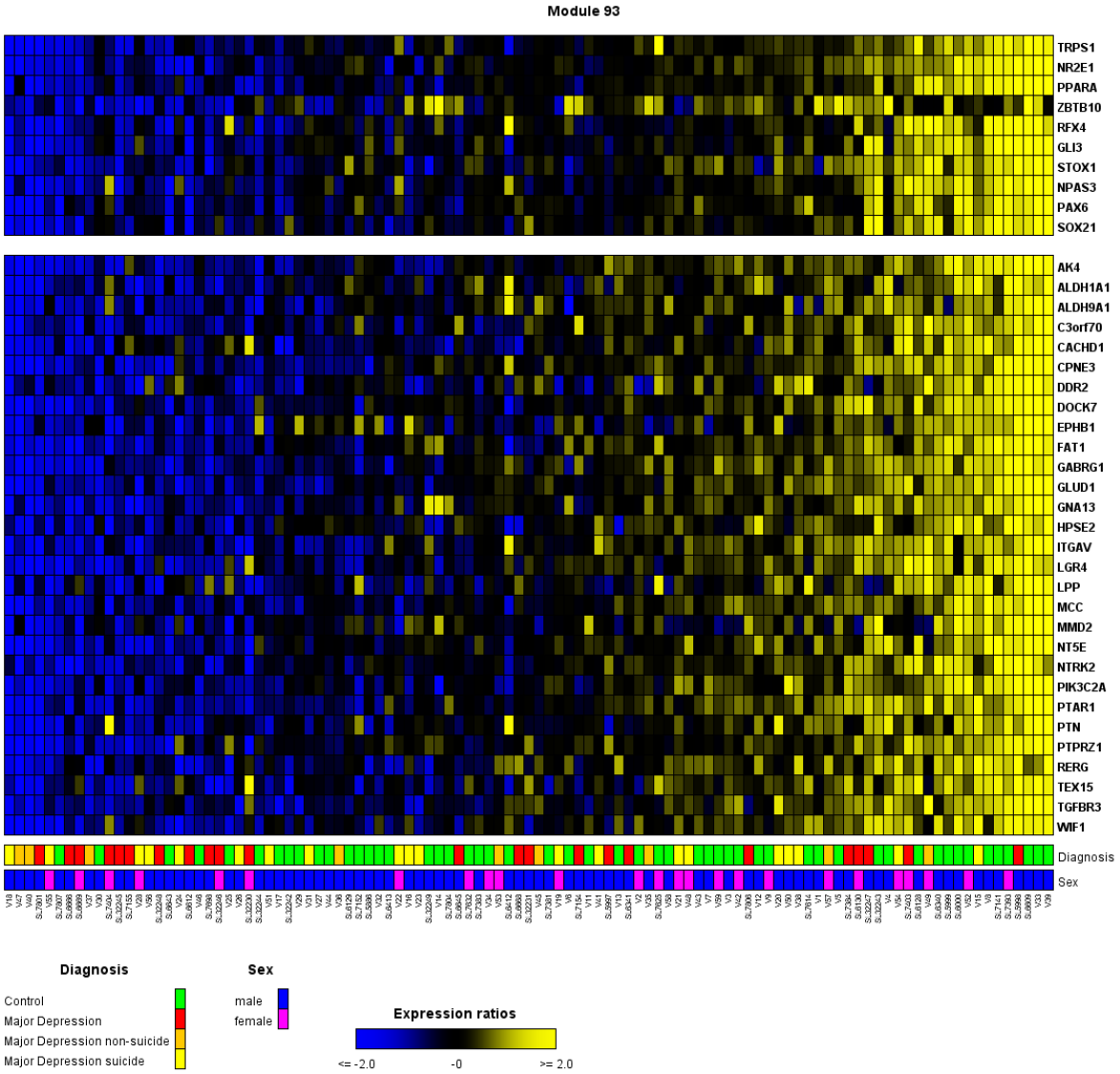
Supplementary figure S5. Visualization of the ensemble graphs of Alzheimer's disease (AD) and depression (MDD). Created with *igraph*.



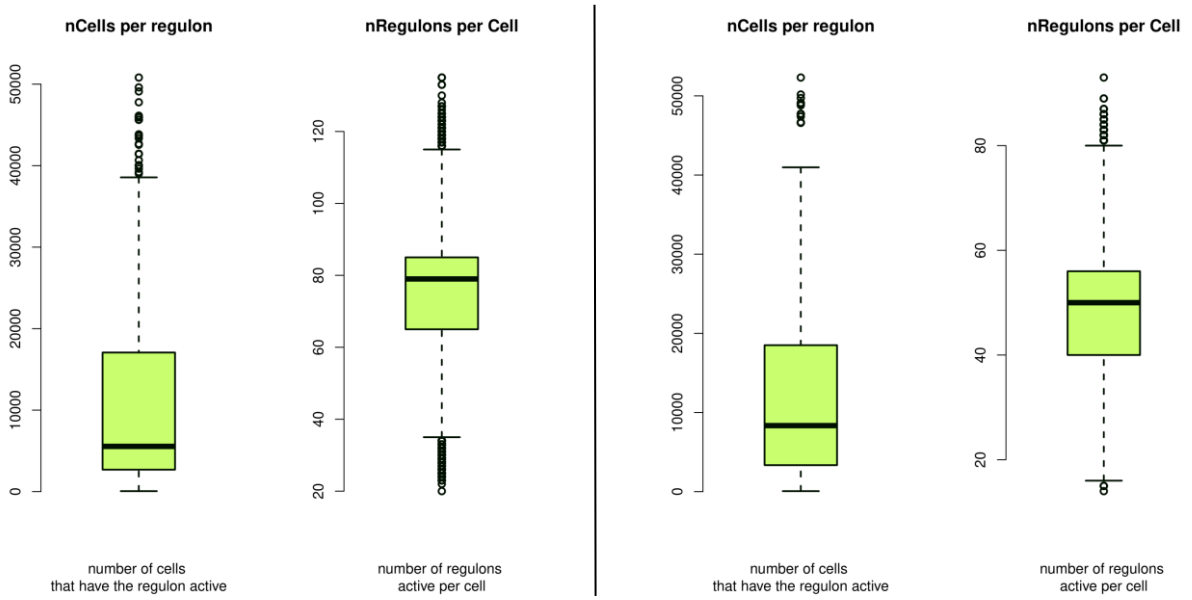
Supplementary figure S6. Module Viewer figure of the gene expression of module 26 from the Alzheimer's disease ensemble network. The upper panel represents the regulators of this module. The annotation data diagnosis, neuropathology, tangle stage, plaque stage and sex are included as well (see legend). There is a trend of higher gene expression in control individuals.



Supplementary figure S7. Module Viewer figure of module 60 from the Alzheimer's disease ensemble network. The upper panel represents the regulators of this module. The annotation data diagnosis, neuropathology, tangle stage, plaque stage and sex are included as well (see legend). There is a trend of increased gene expression in the control individuals.



Supplementary figure S8. Module Viewer figure of the gene expression of module 93 of the depression ensemble network. The upper panel depicts the regulators of this module. The annotation data diagnosis and sex are included as well (see legend). The division between 'Major depression non-suicide' and 'Major depression suicide' is from the dataset GSE101521, while 'Major depression' is from the dataset GSE80655. 'Control' is from both datasets. The gene expression tends to be lower in depressive patients, compared to controls.



Supplementary figure S9. Boxplots from SCENIC with the number of cells per regulon and the number of active regulons per cell. The left figure is from the single-cell Alzheimer's disease network, while the right figure comes from the single-cell major depressive disorder network. A regulon is a transcription factor with all its inferred target genes.