THE REPUBLIC OF TURKEY
MUĞLA SITKI KOÇMAN UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF BIOINFORMATICS

UNDERSTANDING NEURODEGENERATION: PATHWAY ANALYSIS OF NEURODEGENERATIVE DISORDERS BY MERGED AND META-ANALYZED GENE EXPRESSION DATA

MASTER OF SCIENCE THESIS

ABDULAHAD BAYRAKTAR

JUNE 2019
MUĞLA

MUĞLA SITKI KOÇMAN UNIVERSITY
Graduate School of Natural and Applied Sciences

THESIS ACCEPTANCE CONFIRMATION

The thesis prepared by **ABDULAHAD BAYRAKTAR** titled **UNDERSTANDING NEURODEGENERATION: PATHWAY ANALYSIS OF NEURODEGENERATIVE DISORDERS BY MERGED AND META-ANALYZED GENE EXPRESSION DATA** has been accepted on 18/06/2019 in fulfillment of Master of Science Degree in the Department of Bioinformatics by the jury listed below.

JURY MEMBERS

Assist. Prof. Tuğba ÖNAL SÜZEK (**Head of Committee, Member**) Signature:
Department of Computer Engineering,
Muğla Sıtkı Koçman University, Muğla

Prof. Dr. Ömür BAYSAL (**Supervisor**) Signature:
Department of Molecular Biology and Genetics
Muğla Sıtkı Koçman University, Muğla

Prof. Dr. Gürkan ÖZTÜRK (**Member**) Signature:
Department of Physiology,
Istanbul Medipol University, Istanbul

DEPARTMENT CONFIRMATION

Assist. Prof. Tuğba ÖNAL SÜZEK Signature:
Head of Department of Bioinformatics
Muğla Sıtkı Koçman University, Muğla

Prof. Dr. Ömür BAYSAL (**Supervisor**) Signature:
Department of Molecular Biology and Genetics
Muğla Sıtkı Koçman University, Muğla

Date of Defense: 18/06/2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Abdulahad BAYRAKTAR

18/06/2019



ACKNOWLEDGEMENT

At the beginning, I would like to express my gratitude to thesis committee members, my supervisor Prof. Dr. Ömür BAYSAL, Prof. Dr. Gürkan ÖZTÜRK and Assist. Prof. Tuğba ÖNAL SÜZEK for their support, comments, guidance.

I would like to thank Istanbul Medipol University Regenerative and Restorative Medicine Research Center for supporting this work by providing the access for high performance computing cluster and workstation on which I was able to perform this study's intensive time-consuming computations. Once more, I would like to express my gratitude to the director of the center Prof. Dr. Gürkan ÖZTÜRK for giving me this opportunity.

Besides to my teachers mentioned above, I would like to thank Prof. Dr. Süleyman YILDIRIM and Assist. Prof. Barış SÜZEK for their guidance in my academic decisions including this thesis and future.

Furthermore, I would like to thank my friends in both universities cheering and pushing me through this process for their presence.

Lastly, I am grateful for every member of my family and dear relatives for their unending prays.

ÖZET

SİNİRBOZUNUMUNU ANLAMAK: SİNİR BOZUNUM HASTALIKLARININ BİRLEŞTİRİLMİŞ VE META-ANALİZ EDİLMİŞ GENE İFADE VERİSİ İLE YOLAK ANALİZİ

Abdulahad BAYRAKTAR

Yüksek Lisans Tezi
Fen Bilimleri Enstitüsü
Biyoinformatik Anabilim Dalı
Danışman: Prof. Dr. Ömür BAYSAL
Haziran 2019, 69 sayfa

Sinir bozunum hastalıkları, hastalığa özgü bölge ve/veya dokularda sinir hücrelerinin, örneğin nöron ve glial hücrelerin, ilerlemeli kaybı ile karakterize edilir. Pek çok tekil gen, örneğin Alzheimer hastalığında *APP* geni ve de biyolojik yolak sinir bozunum hastalıklarında çalışıldığı halde, sistem çapında bir hastalık gelişimi ortaya konamamıştır. Bu nedenle etkili tedavi stratejilerinden uzaktayız. Son yıllarda yüksek verimli dizileme çalışmaları, özellikle RNA dizileme (RNA-Seq), standart deneysel çalışmaların bulgularını doğrulayan umut verici sonuçlar vermiştir. Dahası, RNA-Seq verisinin kamuca erişilebilirliği pek çoğunun meta-analiz denen yöntemle paralel olarak yeniden analizini mümkün kılmıştır. Meta-analiz bu alanda laboratuvar koşullarının özgünlüğünden kaynaklı farklılıkları en aza indirmek ve çalışmalar arasında benzer şekilde değişen etkenlerin (gen ifadeleri gibi) belirlenmesindeki katkısı nedeniyle bu alanda iyi bir değerlendirme aracı olduğunu ispatlamıştır.

Bu tezde, üç sinir bozunum hastalığında, Alzheimer hastalığı, Parkinson hastalığı ve Amyotrofik Latereal Sklerosis (ALS), yapılmış RNA-Seq çalışmaları meta-analiz edildi. Bu amaçla, kamuca erişilebilir Alzheimer hastalığı, Parkinson hastalığı ve ALS çalışılan RNA-Seq veri setleri Sequence Read Archive'dan (SRA) indirildi. Öncelikle, bu çalışmalarda ortak ve belirgin olarak farklı ifade edilmiş genler tespit edildi ve hastalıkla ilişkilendirilebilecek yolak ve varyantlar arandı.

Sonrasında yapılan meta-analiz ısı şoku proteinleri ile ilişkilendirilmiş yolaklar ve GABA sentez yolağı genlerinin önemini doğruladı. GABA sentez yolağı ile ilintili aşağı yönlü düzenlenen genler (*GAD1* ve *GAD2* gibi) ve ısı stresine hücresel cevap ile alakalı yukarı yönlü düzenlenen genler (*DNAJB6* ve *HSP90AA1* gibi) varyasyon da içermektedir. Tespit edilen genlerin ve yolakların önemi sinir bozunum hastalıkları ile ilgili en güncel literatürde de doğrulanmıştır.

Anahtar Kelimeler: Sinir Bozunum, Meta-analiz, RNA-Seq, Farklı İfade Analizi, Yolak Analizi

ABSTRACT

UNDERSTANDING NEURODEGENERATION: PATHWAY ANALYSIS OF NEURODEGENERATIVE DISORDERS BY MERGED AND META-ANALYZED GENE EXPRESSION DATA

Abdulahad BAYRAKTAR

Master of Science (M.Sc.)
Graduate School of Natural and Applied Sciences
Department of Bioinformatics
Supervisor: Prof Dr Ömür BAYSAL
June 2019, 69 pages

Neurodegenerative diseases are characterized as the progressive loss of neural cells, e.g. neurons and glial cells, within disease-specific regions and/or tissues. While many individual genes, such as *APP* for Alzheimer disease, and biological pathways are studied for neurodegenerative diseases, system-wide pathogenesis could not be revealed; thus, we are far from effective treatment strategies. High throughput sequencing studies, especially RNA Sequencing (RNA-Seq), have delivered promising results in recent years, confirming findings of standard experimental studies. Moreover, the public availability of the RNA-Seq data made it possible to re-analyse many of them in parallel, namely meta-analysis. Meta-analysis is proven to be a good assessment tool in the field as it helps to minimize the batch effect derived differences and to identify similarly altered factors among studies.

In this thesis, a meta-analysis of RNA-Seq studies was carried out for three neurodegenerative diseases, namely Alzheimer's disease, Parkinson's disease and Amyotrophic Lateral Sclerosis (ALS). For this purpose, publicly available Alzheimer's disease, Parkinson's disease and ALS disease RNA-Seq datasets from Sequence Read Archive (SRA) database. The significantly differentially expressed genes common to these studies were first identified and looked for disease-associated pathways and variations.

A meta-analysis conducted in this thesis further validated the importance of genes in heat shock proteins (HSPs) associated pathways and GABA synthesis pathways. The downregulated genes linked to GABA synthesis (e.g. *GAD1* and *GAD2*), the upregulated genes related to the cellular response to heat stress (e.g. *DNAJB6* and HSP90AA1), also contain variations. The significance of genes and pathways identified corroborated by the recent literature on neurodegenerative diseases.

Keywords: Neurodegeneration, Meta-analysis, RNA-Seq, Differential Expression Analysis, Pathway Analysis

TABLE OF CONTENTS

ACKNOWLEDGEMENT ..... **iv**
ÖZET ..... **v**
ABSTRACT ..... **vi**
LIST OF TABLES ..... **viii**
LIST OF FIGURES ..... **ix**
SYMBOLS AND ABBREVATIONS ..... **x**
1. INTRODUCTION **1**
1.1. Alzheimer's Disease 2
1.2. Parkinson's Disease 7
1.3. Amyotrophic Lateral Sclerosis 10
1.4. Meta-Analysis 13
1.5. Objective 15
2. DATA **16**
3. MATERIALS AND METHODS **17**
3.1. Organization and Execution of Analysis on High-Performance Computing 17
3.2. Download Data by SRA Toolkit 19
3.3. Quality Control by FastQC 19
3.4. Adapter Trimming by Skewer 22
3.5. Contaminant Removal by BBDuk 22
3.6. Mapping Reads to Reference Genome by STAR 25
3.7. Variant Discovery by rnaseqmut 29
3.8. Changes on VCF prior to annotation 31
3.9. Annotation by ANNOVAR 32
3.10. Differential Expression Analysis by R (DESeq2) 33
3.11. Meta-Analysis by R (metaRNASeq) 34
3.12. Pathway Analysis or Functional Enrichment by Cytoscape 35
4. RESULTS **36**
5. DISCUSSION **48**
6. CONCLUSION **52**
REFERENCES **53**
CURRICULUM VITAE **66**

LIST OF TABLES

Table 3.1. Statistics of aligned reads of AD samples (PRJNA377568), which were extracted from 28

Table 3.2. rnaseqmut parameters used for each study 30

Table 4.1. The overrepresented pathways in Reactome for 200 significantly DEGs common for three diseases 37

Table 4.2. Overrepresented GO biological processes for 200 significantly DEGs produced by BiNGO 37

Table 4.3. Imported WikiPathways: their names, their IDs and their references….. 39

LIST OF FIGURES

Figure 1.1. Spread of tau carrying neurofibrillary lesions in the process of AD (left, green) and α-synuclein carrying lesions (Lewy bodies and neurites) in the process of PD 4

Figure 1.2. The "dying-forward" and "dying-back" hypotheses 11

Figure 1.3. Cellular and molecular processes mediating neurodegeneration in ALS 12

Figure 1.4. Conserved elements of neurodegeneration 15

Figure 3.1. 1st part of the work diagram 17

Figure 3.2. 2nd part of the work diagram 18

Figure 3.3. 3rd part of the work diagram 18

Figure 3.4. Per base sequence quality plot for SRR5305514 forward reads, before pre-processing 20

Figure 3.5. Per base sequence content plot for SRR5305514 forward reads, before pre-processing 21

Figure 3.6. Per base sequence quality plot for SRR5305514 forward reads, after pre-processing 24

Figure 3.7. Per base sequence content plot for SRR5305514 forward reads, after pre-processing 24

Figure 4.1. Network of 47 found by ReactomeFI plugin based on 200 significant DEG 38

Figure 4.2. Parkinson's Disease Pathway annotated with significant gene expression levels 40

Figure 4.3. Alzheimer's Disease Pathway annotated with significant gene expression levels 41

Figure 4.4. Amyotrophic Lateral Sclerosis Pathway annotated with significant gene expression levels 42

Figure 4.5. Cellular response to heat stress Pathway annotated with significant gene expression levels 43

Figure 4.6. GABA synthesis, release, reuptake, and degradation Pathway annotated with significant gene expression levels 44

Figure 4.7. Protein complexes present in the cellular response to heat stress 45

Figure 4.8. Protein complexes, dimers, and metabolites present in the pathways of GABA synthesis, release, reuptake and degradation 46

Figure 5.1. Drug binding sites on GABA receptors 51

SYMBOLS AND ABBREVATIONS

α	Alpha
γ	Gamma
H_2O	Water molecule
O_2	Oxygen molecule
P_e	Probability of error
Q_{PHRED}	Phred Quality Score
Aβ	Amyloid-beta
AD	Alzheimer's Disease
ALS	Amyotrophic Lateral Sclerosis
fALS	Familial ALS
sALS	Sporadic ALS
AMPA	α-Amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ATP	Adenosine TriPhosphate
BAM	Binary Alignment Map
DEA	Differential Expression Analysis
DEG	Differentially Expressed Gene
ENA	European Nucleotide Archive
FDR	False Discovery Rate
GABA	Gamma-aminobutyric acid
GAD	Glutamate Decarboxylase
GFF	Gene-Finding Format
GO	Gene Ontology
GTF	Gene Transfer Format
GWAS	Genome-Wide Association Study
HD	Huntington Disease
HSP	Heat Shock Protein
Indel	Insertion/Deletion
L2FC	Log2 (Logarithm at base 2) Fold Change
LMN	Lower Motor Neuron

MAPK	Mitogen-activated protein kinases
NCBI	National Center for Biotechnology Information
NDD	Neurodegenerative Disorder/Disease
NMDA	N-methyl-D-aspartate
PD	Parkinson Disease
RNA	RiboNucleic Acid
RNA-Seq	RNA-Sequencing
ROS	Radical Oxygen Species
TCGA	The Cancer Genome Atlas
SAM	Sequence Alignment Map
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
UMN	Upper Motor Neuron
VCF	Variant Call Format

1. INTRODUCTION

Neurodegenerative diseases are defined as loss of structure and function of specific neural cell groups resulting in mortality. These diseases are mostly associated with aging, due to the high prevalence in elder age groups (Feigin et al., 2017). Although the association of ageing has been shown experimentally to a degree (Kritsilis et al., 2018; Reeve, Simcox, & Turnbull, 2014), ageing is not the sole reason and many factors, such as familial background, stressful lifestyle and viral infection, are considered having causative effects (Fargo & Bleiler, 2014; Feigin et al., 2017).

For several neurodegenerative diseases, genetic factors are identified. A well-known example is the case of Huntington's disease. It was shown that an allele of the huntingtin gene (*HTT*) which has an extensive number of CAG repeats syntheses a form of huntingtin having polyglutamine repeats. This polymorphic form of the protein is folded abnormally; hence causing aggregation of the protein in cytosol and nucleus (Difiglia et al., 2016; Kaltenbach et al., 2007). These aggregates interact with a diverse range of proteins (S. Li & Li, 2006), cause apoptosis of neurons (Y. F. Liu, 1998) and drive the pathogenesis of Huntington's disease (Landles & Bates, 2004). Similar findings can be presented for other neurodegenerative diseases, such as amyotrophic lateral sclerosis (ALS) and Alzheimer Disease (Chen, Sayana, Zhang, & Le, 2013; Hung et al., 2016).

Although there are strong theories about the pathogenesis of the diseases, clinical trials of treatment by targeting molecules set by these theories did not become successful overall. Proposed treatments, e.g. vaccines (Schneeberger et al., 2009) and stem cell therapy (Ciervo, Ning, Jun, Shaw, & Mead, 2017), are mostly palliative (Fargo & Bleiler, 2014; Jankovic, 2008), besides many could not be able to pass beyond phase 3 level trials (Doody et al., 2013; Karran & Hardy, 2014)

The pathophysiologic overlap between different neurodegenerative diseases is widely studied. Neurodegenerative diseases share common biological processes (e.g. protein

aggregation, glial activation), chemical alterations (e.g. increase of the levels of cytokines and reactive oxygen species),(Wood, Winslow, Strasser, Wood, & Israel, 2015) and gene associations (Goedert, Clavaguera, & Tolnay, 2010). Protein aggregation stands out among common features. It is observed in most of the neurodegenerative diseases.

Nonetheless, these diseases are complex and heterogeneous. Therefore, the nature of the association of protein aggregates and the disease, whether neurodegeneration is triggered and driven by protein aggregation (causative pathophysiology) or whether protein aggregation is a physiological response to a pre-existing homeostatic dysregulation (responsive pathophysiology), is unclear (Wood et al., 2015)

Advancement of high-throughput sequencing technologies, particularly RNA-Seq led to increased availability of public data for several diseases including neurodegenerative diseases. This made it possible to conduct a meta-analysis, which is the co-interpretation of the results of multiple studies, on multiple and independent RNA-Seq studies to increase the statistical power of the results (Rau, Marot, & Jaffrézic, 2014; Sweeney, Haynes, Vallania, Ioannidis, & Khatri, 2017) and, even identify novel findings. Indeed, a meta-analysis of RNA-Seq of different types of diseases or similar diseases can unveil common transcripts. For instance, a large-scale RNA-Seq study put forth common genes of clusters of 21 cancer types using samples from TCGA (Peng et al., 2015).

A meta-analysis of RNA-Seq studies was carried out for three neurodegenerative diseases, namely Alzheimer's disease, Parkinson's disease and ALS in this work. The approach to RNA-Seq data collection and analysis, the gene expression, pathway and mutation load patterns for significantly DEGs in all three diseases were reported.

1.1. Alzheimer’s Disease

Alzheimer's disease (AD) is the most prevalent neurodegenerative disease in elder people. It is also the most prevalent neurodegenerative disease as a cause of death (Feigin et al., 2017; Organization, 2014) and the most common type of dementia (Fargo & Bleiler, 2014), which is the condition characterized by a decline in memory

or other cognitive skills. The disease is one of the best understood neurodegenerative disorders pathophysiologically.

The presence of neurofibrillary tangles and plaques, i.e. aggregates of abnormally folded proteins, is the most direct biomarker of the disease. Tau protein and amyloid β protein are the main constituents of these tangles and plaques, respectively (H. Braak et al., 1996; Selkoe, Hardy, Selkoe, & Hardy, 2016). Pathophysiologically, large projection neurons connecting the cerebral cortex to the hippocampus, which are placed in entorhinal cortex, are affected initially (De Calignon et al., 2012). Consequently, a worsening ability to remember new information is observed as the most initial symptom of AD (Fargo & Bleiler, 2014). The disease propagates toward the hippocampus, dentate gyrus, spreads to limbic and association cortices in later stages, progressively through years (De Calignon et al., 2012). Destructive effects of the disease on the hippocampus cause dementia and disruption of other related memory-related cognitive functions, such as poor judgment, confusion with time and space (Fargo & Bleiler, 2014).

AD is considered also as a chronic disease, developing as a result of multiple factors rather than a single cause. Advanced age is the greatest risk factor for AD (Kritsilis et al., 2018). Having an AD patient sibling or parent, carrying one or two copy of ε4 isoform of *APOE* gene as allele, having mild cognitive impairment, having cardiovascular disease factors such as obesity, high cholesterol and diabetes, being socially and cognitively less active, fewer years of education and experiencing a traumatic brain injury are considered as other risk factors for the disease (Fargo & Bleiler, 2014).

Disease progression is gradual and strict in AD. It has been studied through a series of studies for Alzheimer disease since the 1990s by Heiko Braak and his colleagues (H. Braak et al., 1996; Heiko Braak & Braak, 2008; Heiko Braak et al., 1994). Based on these studies, they classified Alzheimer disease into six developmental stages on account of neurofibrillary changes across the cerebral cortex. According to these, disease begins in transentorhinal cortex (clinically silent stages I and II), proceeds into adjoining cortical and subcortical components of the limbic system, including dentate gyrus, (stages III and IV – incipient AD) and spreads into association areas of the neocortex (stages V and VI - fully developed AD). There is a scarce amount of

neurofibrillary tangles in the transentorhinal cortex at stage I, while it turns to stage II if few tangles become many. At stage III and stage IV, transentorhinal and entorhinal regions become severely damaged. Entorhinal region projects sensory information into the dentate gyrus of the hippocampus. Hippocampus processes this information to produce memory. Loss of neurons in the entorhinal region triggers the destruction of neurons in the hippocampus. Therefore, the malfunction of this region causes the prevention of new memory formation (Fargo & Bleiler, 2014; Lazarov & Hollands, 2016)], which explains why worsening ability to remember new information is the most initial symptom of AD. In addition, disease symptoms may not be observable until stage IV. Moreover, the period between onset of pathophysiology and observable symptoms, which is named as pre-symptomatic stage, may take more than 20 years, if diagnosis was not set by cerebrospinal fluid and/or blood biomarkers (Bateman et al., 2012; Doecke et al., 2012; Reiman et al., 2012). Disease proceeds through the amygdala, thalamus, and hippocampus, finally reaching neocortex at stage V and VI and erodes nerve cells in tissues it passes. Patients show grave mental deterioration at these stages.

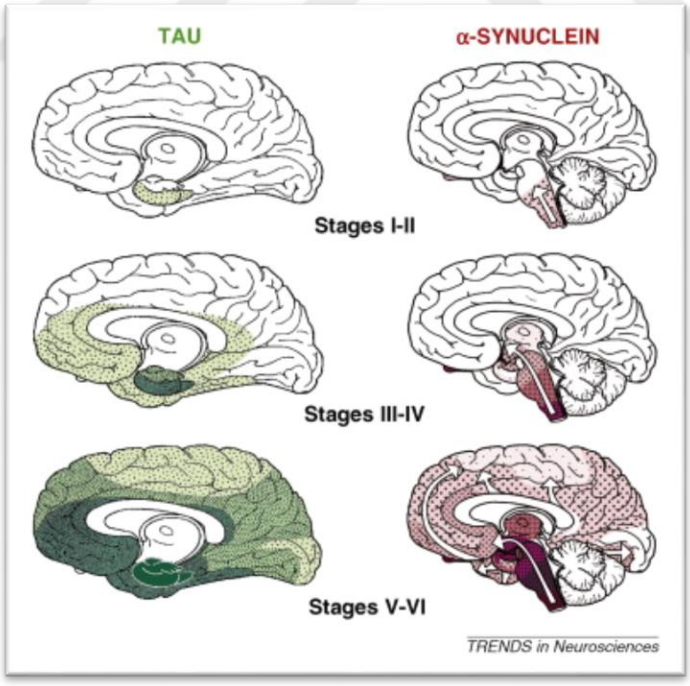

Figure 1.1. Spread of tau carrying neurofibrillary lesions in the process of AD (left, green) and α-synuclein carrying lesions (Lewy bodies and neurites) in the process of PD (right, red) (Goedert et al., 2010)

As given above, Alzheimer disease pathology is strongly defined by plaques and neurofibrillary tangles. Amyloid pathology and tau pathology theories are introduced to explain these structures (Goedert et al., 2010).

Amyloid pathology theory is oriented around the production and proteolysis of the amyloid precursor protein, and clearance of its products. Amyloid precursor protein (*APP*) gene synthesizes amyloid precursor protein, as its name implies. Overexpression of *APP* is directly linked to lifelong Aβ42 deposits, plaques, and AD. High prevalence of AD neuropathology in individuals with Down's syndrome, having three copies of chromosome 21, where *APP* gene is located, and observation of individuals carrying three wild-type copies of *APP* without Down's syndrome but AD conclusively support this finding (Rovelet-Lecrux et al., 2006; Selkoe et al., 2016). Amyloid precursor protein is proteolytically processed by two competing pathways: the non-amyloidogenic and amyloidogenic (Aβ producing) pathways. While transmembrane α-secretase complex cleaves within the Aβ domain of *APP* into soluble fragments in the non-amyloidogenic pathway, an intermembrane protein complex named γ-secretase cleaves the precursor protein to non-soluble/hydrophobic, self-aggregating Aβ oligomers, in which Aβ40 and Aβ42 are the major species in the latter (Krishnaswamy, Verdile, Groth, Kanyenda, & Martins, 2009). Presenilin 1 and presenilin 2 constitute the catalytic subunit of γ-secretase. At healthy individuals, these two pathways are at balance; thus, Aβ42 production is stabilized. Missense mutations of corresponding genes, *PSEN1* and *PSEN2*, decrease cleavage processivity and disrupts this balance in favour of Aβ42 and longer peptides, causing AD pathology (Selkoe et al., 2016).

A similar increase of Aβ42 is observed when genes responsible for clearance of them are problematic. Cholesterol/sterol metabolism, inflammation/brain's innate immune system, and endosomal vesicle recycling are evaluated as especially important biological processes that are associated with clearance of Aβ oligomers. Apolipoprotein E gene, which is a core component of plasma lipoproteins and involved in their production, conversion, and clearance (Gonzales, Gordts, Foley, & Esko, 2013), is considered as a major genetic risk factor for AD. Isoform 4 of APOE, ApoE ε4, was shown to decrease in mouse (Refolo et al., 2000) and human studies (Castellano et al., 2011; Strittmatter et al., 1993). Another cholesterol metabolism-

related gene, *ABCA7*, is also associated with the disease. Loss-of-function mutations in *ABCA7* have been revealed to increase AD risk by Stacy Steinberg and her colleagues (Steinberg et al., 2015). Brain innate immune system genes, e.g. *CR1*, *CD33*, and *TREM2*, and endosomal vesicle recycling regulatory genes, e.g. *SORL1*, *BIN1*, and *PICALM*, are shown as other important factors in amyloid pathogenesis in AD.

Tau pathology theory is based on conformational changes of tau protein and its catastrophic downstream effects. Tau protein is a natively unfolded protein (Cleveland, Hwo, & Kirschner, 1977), whose multiple isoforms are differentially expressed in different cell types, e.g. "big tau" in peripheral neurons (Couchie et al., 2006) or different developmental stages. Its phosphorylated form binds to tubulin protein of microtubules with its binding domains and stabilizes it. Besides, through its N-terminal domain, it binds to mitochondria and the neuronal plasma membrane. On the contrary, hyperphosphorylated tau protein loses its ability of binding to tubulin. Moreover, microtubule organization may be disrupted due to it. This affects axonal transport of intracellular membranous elements, such as mitochondria and exocytic vesicles, which may result in a decrease in glucose and lipid metabolism in axons, driving to loss of axon in a process known as "dying back". Upregulation of tau phosphorylating proteins (kinase) in the brain, e.g. GSK-3β, cdk5, PKA and CaMK-II, and/or downregulation of dephosphorylating proteins (phosphatase), e.g. PP1, PP2A, PP2B and PP2C, may lead to hyperphosphorylation of tau (Kolarova, García-Sierra, Bartos, Ricny, & Ripova, 2012). Besides this effect, hyperphosphorylated tau protein is prone to aggregation, ending in neurofibrillary tangles.

The general opinion about amyloid and tau hypothesis is that they cause AD pathogenesis simultaneously and cooperatively. Specific features of AD pathogenesis, such as neuron death and synaptic dysfunction are caused by both soluble forms Aβ and tau proteins (Bloom, 2014). Furthermore, there are in-vivo and in-vitro pieces of evidence for triggering of toxic conversion of tau by (Laferla, 2010; Seward et al., 2012) and vice versa (T. Liu et al., 2004).

On the other hand, there are counter-arguments against these hypotheses.

Fuyuki Kametani and Masato Hasegawa reviewed pieces of evidence for many genes in the case of AD, such as *APP*, *MAPT* and *PSEN* (Kametani & Hasegawa, 2018). Based on their inquisition, they asserted impairment of APP metabolism instead of Aβ accumulation as the trigger of the disease pathogenesis and supported the idea of progression by tau (Takahashi et al., 2015). As discussed in their review, a study showed that amyloid beta 42 inclusions do not cause memory impairments in mice (Kim et al., 2013). Failure of treatments targeting Aβ oligomers in increasing cognitive functions justifies this view (Doody et al., 2013; Dubois et al., 2014; Ostrowitzki et al., 2012).

1.2. Parkinson's Disease

Parkinson's disease is first described by James Parkinson, whose name was given later to disease, as "shaking palsy" in his 1817-dated article "An essay on the shaking palsy" (Goedert, Spillantini, Del Tredici, & Braak, 2013; Jankovic, 2008). In the article (Parkinson, 2002), J. Parkinson tells about the progressive nature of the disease, sense of weakness at the beginning, trembling at extremities, difficulty at conserving an upright posture, uncontrollable shaking, pain after posture changes, sleep disturbances, uncontrollable bowel movement observed in patients. While these symptoms are common among Parkinson's disease patients, rest tremor – shaking –, bradykinesia – slowness – and rigidity are accepted as major symptoms at current, formulated as TRAP with addition of postural instability (Ferrer, 2011; Goedert et al., 2013; Jankovic, 2008; T. Wu & Hallett, 2013).

Studies put forth that these symptoms are caused by loss of dopaminergic neurons and reduced dopaminergic input to corresponding brain regions at different stages (Heiko Braak, Ghebremedhin, Rüb, Bratzke, & Del Tredici, 2004; Ferrer, 2011). Temporary sedation of symptoms by dopamine-derived chemicals such as levodopa support this fact (Jankovic, 2008).

The general opinion is that neuropathological alterations begin at the dorsal motor nucleus (Heiko Braak et al., 2004; Del Tredici, Rüb, De Vos, Bohl, & Braak, 2002), even before symptoms occur, as mentioned by James Parkinson's original article as

medulla spinalis. Then these changes spread through the medulla oblongata, which is located in the brain stem (Parkinson, 2002). The disease affects substantia nigra, amygdala, thalamus and cerebellar cortex at later stages (Heiko Braak et al., 2004; T. Wu & Hallett, 2013).

Observation of Lewy bodies – protein aggregates – and Lewy neurites in disease-associated regions, followed by denervation, is accepted as the hallmark of disease progression (Elobeid, Libard, Leino, Popova, & Alafuzoff, 2016; Goedert et al., 2010; Jankovic, 2008; Recasens et al., 2014). On the other hand, some studies indicate alpha-synuclein accumulations, major component of Lewy bodies (Baba et al., 1998; Sai, Zou, Peng, & Dong, 2012; Wakabayashi, Matsumoto, Takayama, Yoshimoto, & Takahashi, 1997; T. Wu & Hallett, 2013), as cause of neurodegeneration in Parkinson's disease instead of Lewy bodies as a whole (Kramer & Schulz-Schaeffer, 2007). On another side, Walter J Schulz-Schaeffer reaches to a conclusion in his two reviews that presynaptic dysfunction is the cause of the disease instead of denervation (Schulz-Schaeffer, 2010, 2015). Independent of the true root of disease, the importance of alpha-synuclein within the pathology of the disease is obvious.

Immunohistochemical and genetic studies give us a good set of proteins associated with alpha-synuclein. For instance, tau protein is found to be an enhancer of alpha-synuclein aggregation (Badiola et al., 2011; Roy & Jackson, 2014), while it is also an aggregation protein (Binder, Guillozet-Bongaarts, Garcia-Sierra, & Berry, 2005; De Calignon et al., 2012). Genome-wide association studies (GWAS) made by two different strengthens this association and give polymorphism of alpha-synuclein gene *SNCA* and tau protein-encoding gene *MAPT* as common risk factors (Edwards et al., 2010; Simón-Sánchez et al., 2009), at least for European ancestry. Other GWAS associate *SNCA, PARK16,* and *LRRK2* with Parkinson's disease in both European and Asian ancestry populations, while it cannot replicate results for *MAPT* in the Asian population (X. L. Chang et al., 2011; Satake et al., 2009).

Another genome-wide association study confirms the association of *MAPT* with Parkinson's disease, while it also detects *GAK* as a candidate (Rhodes, Sinsheimer, Bordelon, Bronstein, & Ritz, 2011). If we shift our focus onto cytological and histochemical studies and look in protein aggregates, we encounter with tens of proteins, such as p62 and ubiquitin (Kuusisto, Parkkinen, & Alafuzoff, 2003)

(Kuusisto et al, 2003), DJ-1, Parkin, LRRK2, PINK-1 (Wakabayashi, Tanji, Mori, & Takahashi, 2007). A review of Beyer and Ariza focused on the general mechanism of how Lewy bodies form from these proteins and more, enriched with drawings (Beyer & Ariza, 2007).

Moreover, deformation of alpha-synuclein aggregation or clearance is another aspect of Lewy bodies and related mechanisms. Chaperones or heat shock proteins (HSPs) are responsible for these protective features. They are hypothesized as protectors of neurons against alpha-synuclein derived neurotoxicity. One reason is that they cause the stabilization of alpha-synuclein. Supporting this, Opazo and his colleagues showed by fluorescence staining at their study that Hsp70 chaperone protein does not only prevent α-synuclein aggregation but increases clearance (Opazo, Krenz, Heermann, Schulz, & Falkenburger, 2008). A similar effect has been shown also between Hsc70 and α-synuclein (Ebrahimi-Fakhari, Wahlster, & McLean, 2011; Pemberton et al., 2011). Another reason is that HSPs assist in folding of alpha-synuclein interacting proteins – e.g. Parkin, DJ-1, PINK-1 – in their natural conformation (Sai et al., 2012), preventing their sequestration into aggregates (Kalia et al., 2004; Rose, Novoselov, Robinson, & Cheetham, 2011); thus, HSPs also prevent disruption of healthy mechanisms, e.g. mitochondrial homeostasis (Sai et al., 2012), which may be caused by these proteins.

A final reason is that chaperones bind to protein aggregates, especially to alpha-synuclein monomers, converting them into less toxic fibril form (Ebrahimi-Fakhari et al., 2011; Pemberton et al., 2011). When we consider Parkinson's disease-associated genes and proteins based on their biological functions or molecular pathways, we meet with common results. Mitochondrial homeostasis and reactive oxygen species (ROS) clearance are among the most common associations. Mitochondrion is the organelle responsible for energy production by oxidative phosphorylation. The respiration process requires the involvement of many protein complexes. Dysfunction of these complexes may result in two situations: reduced ATP production and increased production of ROS. Since neurons consume a high amount of energy and ROS damages macromolecules, both are linked to Parkinsonism, and we take evidence directly from molecular biology. One of the earliest evidence showing such a link is found by JW Langston and colleagues (Langston, Ballard, Tetrud, & Irwin, 1983).

They observed that drug abusers who are exposed to MPTP, a mitochondrial complex I inhibitor, showed acute and irreversible parkinsonian syndrome.

1.3. Amyotrophic Lateral Sclerosis

ALS is known as motor neuron disease, which is a fast-progressing, lethal neurodegenerative disease, characterized by loss of motor neurons in the cerebral cortex, brainstem, and spinal cord. Depending on where the degeneration begins and which biomarker genes are mutated, the disorder progresses differently and the patient shows different symptoms (Kiernan et al., 2011).

There are three locations of neurodegeneration in ALS. Upper motor neurons (UMN) carry signals for voluntary movement from the cerebral cortex to the brainstem and spinal cord. They are regulated by glutamate. Loss of UMN causes spasticity, weakness and brisk deep tendon reflexes on limbs in ALS patients. In addition, Lower motor neurons (LMN) carry motor signals from UMN to muscle cells. Contrary to UMN, they are cholinergic. Loss of LMN causes fasciculations (involuntary frequent spasms in muscle), wasting, loss of reflexes and weakness (Javed & Daly, 2019; Kiernan et al., 2011). Bulbar UMN dysfunction, which refers to denervation of the medulla oblongata (lower section of brain stem), results in dysarthria (inability of speech) and dysphagia (difficulty in swallowing). Weight loss and malnutrition usually follow dysphagia. Respiratory failure stems from denervation of central respiratory centres and phrenic nerve – motor neuron of the diaphragm –, is the most prevalent cause of death in ALS. While most patients pass away in 2-3 years after the diagnosis, a significant percentage of them may survive more than 5 years (Kiernan et al., 2011).

While the denervation begins in one of these locations in general, it spreads to others in progress. To explain the spread mechanism, two hypotheses, i.e. dying-forward and dying-back, have been proposed (Figure 1.2.).

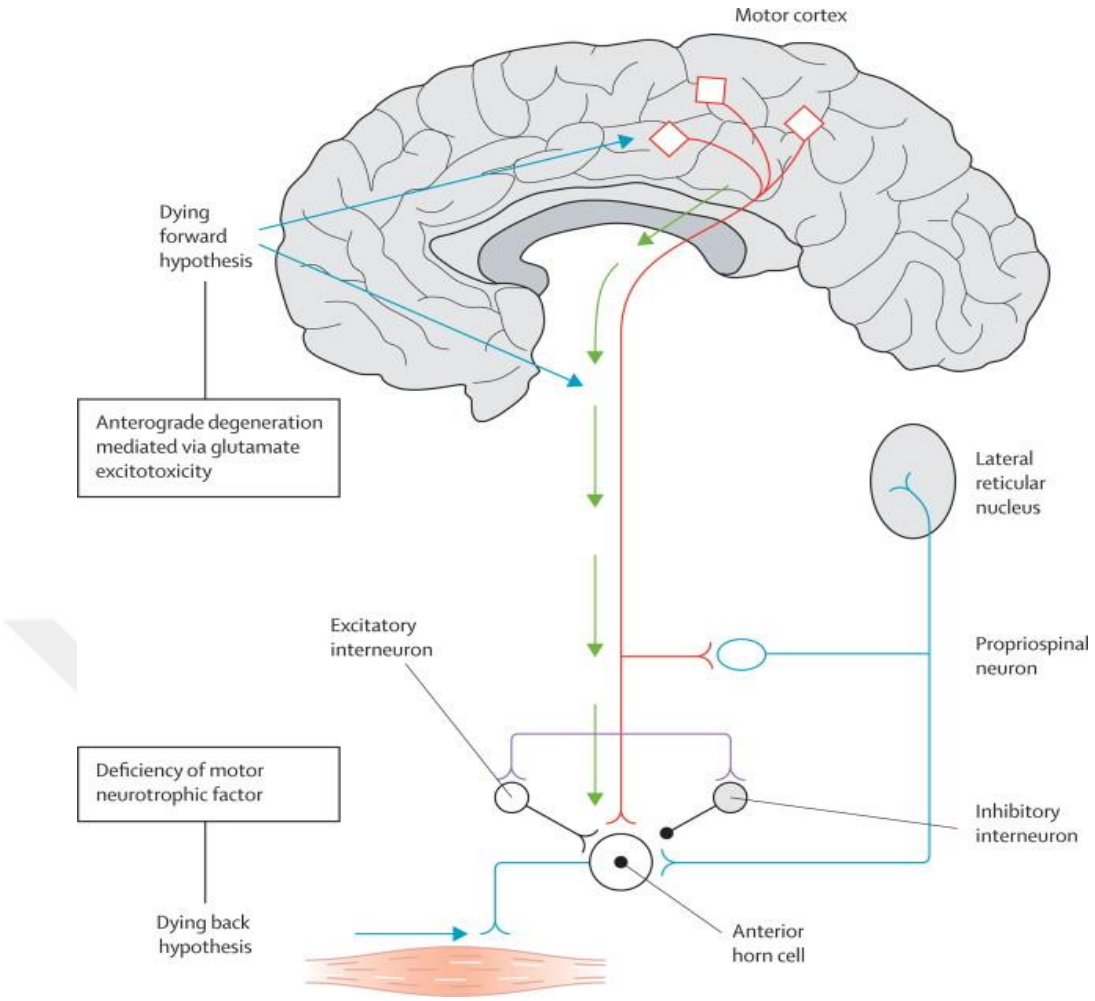

Figure 1.2. The "dying-forward" and "dying-back" hypotheses (Kiernan et al., 2011)

According to the dying-forward hypothesis, ALS affects UMN primarily and mediates the degeneration of anterior horn cells in the spinal cord (LMN), which they are synaptically connected, by glutamate excitotoxicity. That cortical hyperexcitability is an early feature in ALS and rarity of pure LMN forms of ALS are given as base of this hypothesis (Kiernan et al., 2011).

Glutamate excitotoxicity arises due to over-stimulation of glutamate receptors, namely NMDA and AMPA, by glutamate, which triggers calcium influx into postsynaptic neuron. An increase in the level of calcium disrupts mitochondrial activity and the production of radical oxygen species (ROS) climbs. This has two results. First, elevated ROS lead to neuronal death by causing protein aggregation, cytoskeletal and mitochondrial dysfunction. Second, ROS exit the cell and inactivate glutamate receptors, trigger the cycle forward (Barber, Mead, & Shaw, 2006).

According to the dying-back hypothesis, the primary target of ALS is either muscle cells or their junction with LMN. Axonal degeneration makes the basis of this hypothesis. Axonal degeneration mediated by glial cells that carry mutated SOD1 and produce ROS and neurotoxins has been speculated as triggering apoptosis (programmed cell death) on LMN (Yaron & Schuldiner, 2016). Similarly, a decline in skeletal muscle may trigger axonal degeneration. Studies of skeletal muscle decline associated with age (Chung et al., 2017) and high physical activity (Beard et al., 2016; Chio, Benzi, Dossena, Mutani, & Mora, 2005; Seals, Kioumourtzoglou, Gredal, Hansen, & Weisskopf, 2016). However, some oppose the effect of physical training. They show beneficiary effects of increased physical training on the health of motor neurons (Gardiner, Dai, & Heckman, 2006), and put forth head injuries instead of physical activity as a cause (McCrate & Kaspar, 2008; Schmidt, Kwee, Allen, & Oddone, 2010).

Apart from the direction of the spread of neurodegeneration, mechanisms associated with ALS are quite well studied. Protein aggregation, increased oxidative stress, mitochondrial dysfunction, glutamate excitotoxicity, disruption of axonal transport are among them (Figure 1.3.).

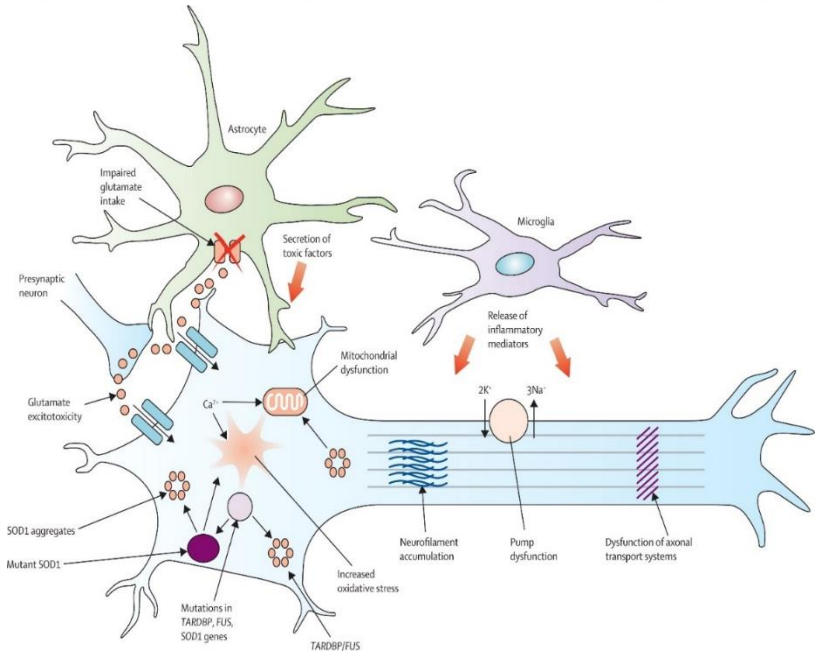

Figure 1.3. Cellular and molecular processes mediating neurodegeneration in ALS (Kiernan et al., 2011).

Based on whole-genome association studies and linkage analysis studies, mutations on many specific genes associated with these processes are linked to both familial ALS (fALS) and sporadic ALS (sALS). While sporadic cases constitute a large portion in ALS, the familial background is also a strong factor for ALS, according to them. As they revealed, *SOD1*, *TARDBP*, *SETX*, *FUS*, *ANG*, and *OPTN* are the most affected genes in the disease (Chen et al., 2013). Explaining every disease-associated gene will take a lot of space. Thus, SOD1, one of the major disease factors, is given as an example here.

SOD1 is the first gene linked to familial ALS. Nearly 20% of patients with familial background and less than 5% of sporadic ALS patients carry mutations on *SOD1*. *SOD1* gene encodes Cu/Zn ion-binding superoxide dismutase, which detoxifies by-products of oxidative phosphorylation. Mutations on *SOD1*, predictably, leads to increased ROS; therefore, mutant *SOD1* has been proposed as a major cause in ALS and mediates other pathophysiologic alterations mentioned previously (Hardiman et al., 2013; Yaron & Schuldiner, 2016).

1.4. Meta-Analysis

Meta-analysis is a quantitative method of synthesizing the results of similar but independent studies to create a consensus. By this definition, a meta-analysis takes place in a wide range of areas from medicine to business and social sciences. Moreover, it covers numerous methods, e.g. genome-wide association studies and systematic reviews, takes several approaches, i.e. effect size based and p-value based, into account. A meta-analysis is conducted to decide whether start new research or determine statistically significant, consistent and – in the case of biomedical studies – clinically important effects (Borenstein, Hedges, Higgins, & Rothstein, 2009, p. xxv).

For a while, meta-analyses of high throughput sequencing have been conducted in medicine to make robust clinical predictions. (Phan, Young, & Wang, 2012). These applications can reproduce the results of pre-analysed studies with higher statistical power. Truly, the increasing number of datasets has been showed to increase the accuracy of the meta-analysis (Sweeney et al., 2017).

Such as, a meta-analysis of publicly available 18 microarray datasets from metastatic and primary tumour samples which was executed by Marla H Daves and her colleagues revealed a common metastatic signature of 79 genes in different cancer types which were involved in known metastasis-associated pathways (Daves, Hilsenbeck, Lau, & Man, 2011). Another example study may be given from neurodegenerative disorders. A Borjabad and D Volsky showed in this study that immune responses were upregulated for HIV-Associated Neurocognitive Disorder (HAND), AD and multiple sclerosis based on microarray datasets (Borjabad & Volsky, 2012).

Meta-analysis can work well on large datasets from diverse samples with multiple factors. An inspiring study was performed by Hamel Patel, Richard J.B Dobson, Stephen J Newhouse, is a good example of this (Patel et al., 2018). In this study, more than 2500 microarray datasets mapping to four brain regions samples from multiple AD, HD, schizophrenia, bipolar disorder, major depressive disorder, and PD studies were meta-analysed. Differential expression analysis was performed on each of them as in this thesis. A different p-value combining method, which is suitable for microarray data, were applied. By this way, they determined AD specific significant differentially expressed genes (DEGs) in different brain regions.

In another study, researchers focused on common features, i.e. genes and ontologies, of neurodegenerative disorders rather than differences. They studied 1270 post-mortem central nervous system tissue samples, microarray data, from 13 patient cohorts of four neurodegenerative disorders (NDDs), namely AD, HD, ALS and PD (M. D. Li, Burns, Morgan, & Khatri, 2014). In this, however, they used an effect size combination method. Based on their findings, they suggested the impaired mitochondrial function as the most reliable specific feature of neurodegeneration (Figure 1.4.).

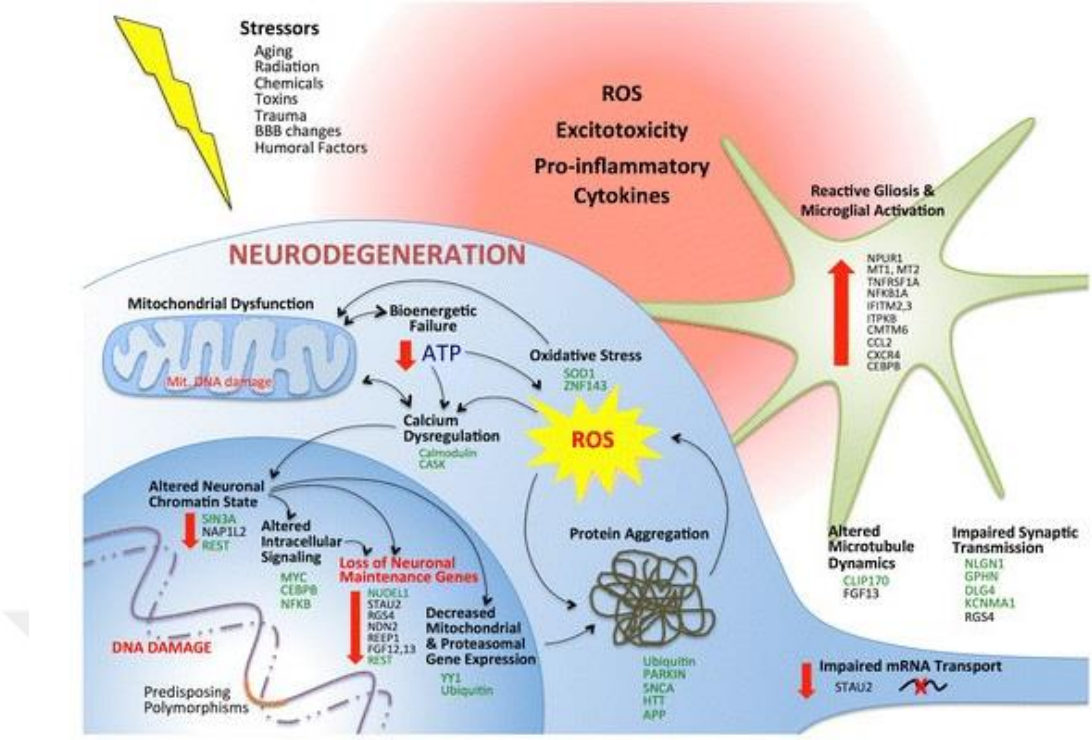

Figure 1.4. Conserved elements of neurodegeneration. Common genes they found are shown in black text. Their prediction of hub genes and upstream transcription factors shown in green text (M. D. Li et al., 2014).

1.5. Objective

The study aims to find shared genes and pathways for the diseases introduced above by meta-analysis of differential expression analyses and co-evaluating with pathway and variant discovery analyses to make interpretations about neurodegeneration as a whole.

2. DATA

The Sequence Read Archive (SRA) (Leinonen, Sugawara, & Shumway, 2011) of National Center for Biotechnology Information (NCBI) and European Nucleotide Archive (ENA) (Harrison et al., 2019) databases were searched for RNA sequence data, with the terms "parkinson", "alzheimer" and "amyotrophic". The findings were narrowed to publicly available, human, post mortem, non-blood tissue studies. In the end, 4 studies were identified fitting with the selection criteria in the SRA database.

PRJNA279249 (Prudencio et al., 2015) and PRJNA306659 were identified for ALS. PRJNA279249 has 33 fALS and sALS and 17 control samples, taken from the cerebellum and prefrontal cortex sequenced by Illumina HiSeq 2000. The library used is paired and the average sequence length (forward + reverse) is 202. PRJNA306659 has 13 sALS and 8 control samples, sequenced by Illumina Genome Analyzer II. Motor neurons dissected from the spinal cord were studied. Library prepared was single, 50-nucleotide was the average sequence length.

PRJNA283498 (Dumitriu et al., 2016) was identified for PD. It has 44 PD and 29 control samples, sequenced by Illumina HiSeq 2000. Prefrontal cortex was used as sample. Library used is paired and average sequence length (forward + reverse) was 202. Due to low quality, one PD and one control sample were excluded from the PD accession table, SRR2015769, and SRR2015725, respectively. PRJNA377568 (Friedman et al., 2018) was identified for AD. It has 84 patients and 33 controls. Samples were dissected from fusiform gyrus, sequenced by Illumina HiSeq 2500. Library used was paired and average sequence length (forward + reverse) was 200.To study on them, accession tables including sample IDs and symbols of disease status were prepared for each study. Disease statuses provided in run information tables were accepted as it is except for PRJNA279249, in which we considered both *C9orf72*-associated fALS and sALS samples as ALS. The FASTQ files containing both forward and reverse pairs are split as separate files for forward and reverse reads for further analysis. In total, 259 samples (173 patients + 86 control) were studied.

3. MATERIALS AND METHODS

3.1. Organization and Execution of Analysis on High-Performance Computing

High performance computing cluster (HPCC) is a computer system having extensive computational resources, which enables operation of high-demanding and complex jobs. Cause the tools used demand high resources, i.e. CPU and memory, besides 259 FASTQ samples were included in this study, a desktop computer was not reliable. Therefore, analyses were designed on an HPCC. The workflow was composed of one loop for differential expression analysis (DEA) (Figure 3.1.), one loop for variant discovery analysis (Figure 3.2.), DEA by DESeq2, a meta-analysis of DEA result files and integration of disease-associated genes and variants lists into a network analysis tool for pathway analysis (Figure 3.3.).

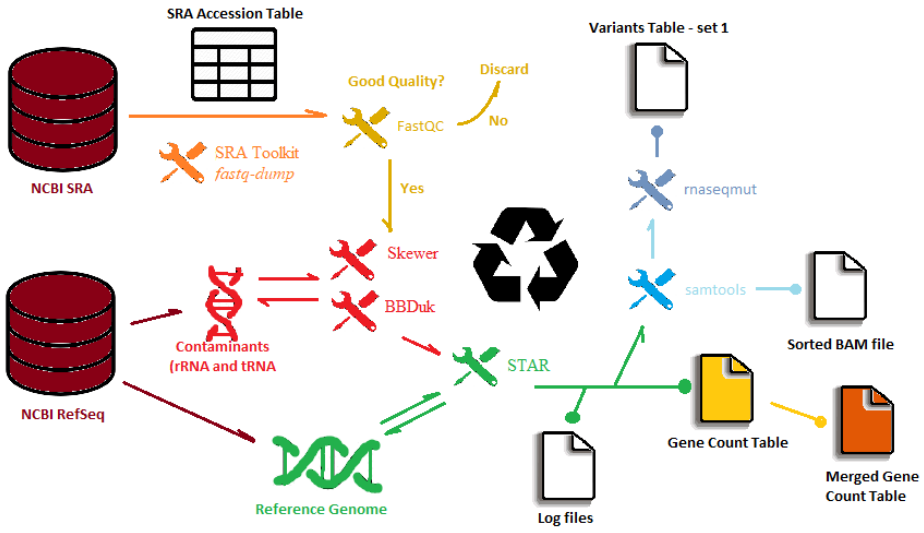

Figure 3.1. 1st part of the work diagram. It is simply a loop for differential expression analysis. Loop downloads samples from NCBI SRA database, according to the SRA accession table, makes appropriate operations using references from NCBI RefSeq.

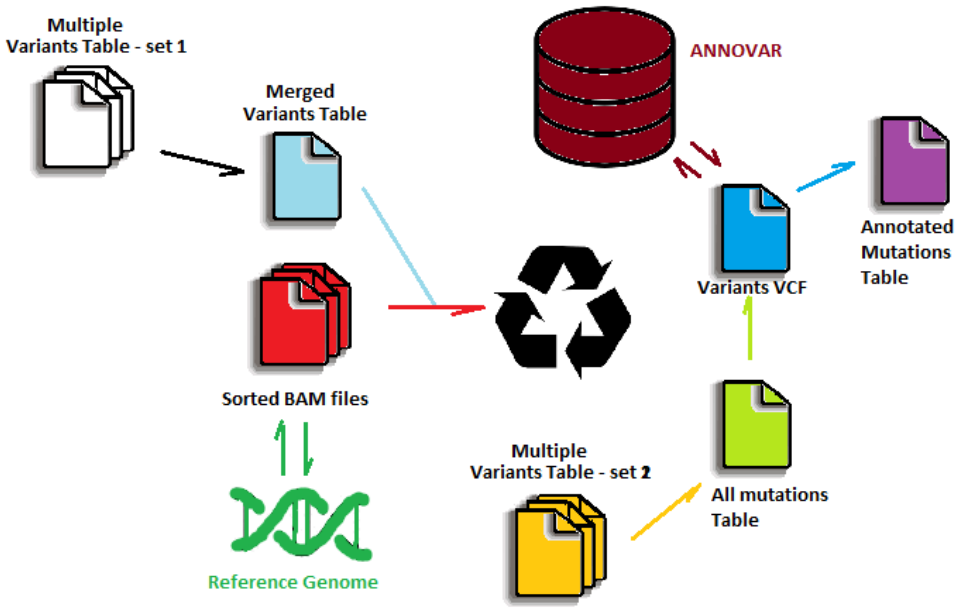

Figure 3.2. 2nd part of the work diagram. In this part, a loop works on sorted BAM files and variant tables produced for each sample, for every disease separately. At final, it gives annotated variant tables whose variants are unique to patient samples.

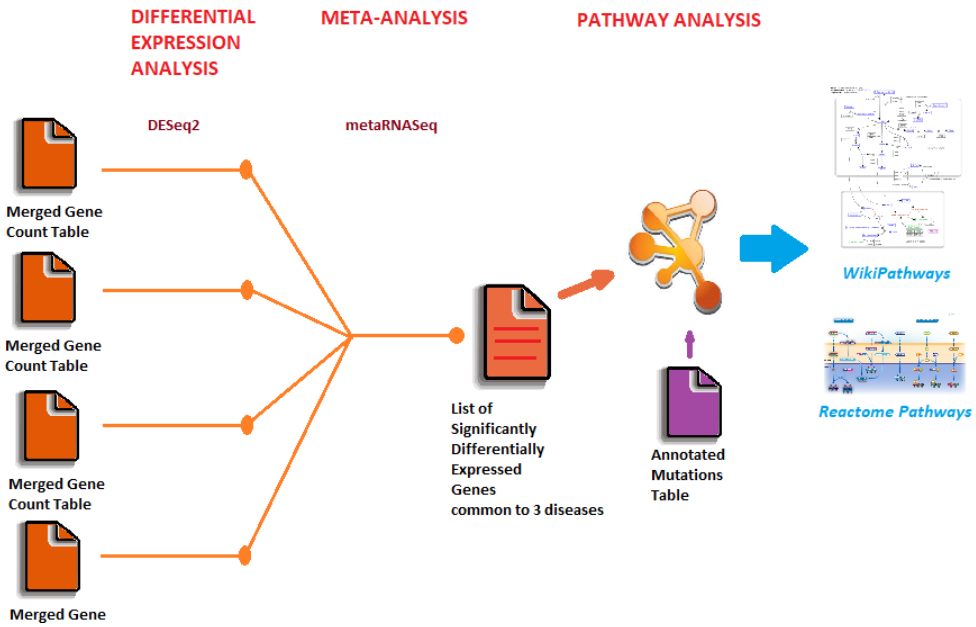

Figure 3.3. 3rd part of the work diagram. Identification of significantly differentially expressed genes after a meta-analysis of all single studies. Later, this list is integrated into pathway analysis together with disease-associated variants list.

3.2. Download Data by SRA Toolkit

The Sequence Read Archive (SRA) is an international public database to store high throughput sequencing data (Leinonen et al., 2011). Within NCBI infrastructure, it is integrated with other databases, which allows finding data associated information and files, such as sequencing library, organism, project where data is used, article if the study was published. Owing to its advanced query system, it is possible to find relevant data and exclude irrelevant data using filters.

SRA Toolkit ("SRA-Tools - NCBI," n.d.) is a collection of tools and libraries to use data stored in SRA. It allows faster download of data compared to manual download, besides to easiness of implementing it in a script.

In the thesis, the fastq-dump tool in the SRA toolkit was used to download FASTQ files. Sample accession IDs from accession tables written for each study were given as input. Example command was given below:

$ *fastq-dump SRR1234567*

3.3. Quality Control by FastQC

FASTQ (Cock, Fields, Goto, Heuer, & Rice, 2010) is a data format, created as an extension to the FASTA format (Pearson & Lipman, 1988), which stores a numeric quality score for each nucleotide (base-call) in a sequence. This score, namely the Phred score, represents the possibility of an error in read of nucleotide. To clarify, a high score, up to 40 with current technology, points out that read is less likely erroneous and more likely the device detected this nucleotide as correct. Phred score is formulated as below, where QPHRED is Phred quality score and Pe is the probability of error:

$$Q_{PHRED} = -10 \times \log_{10}(P_e)$$

In this regard, FASTQ score has a critical role in the decision about data, since data with bad quality score affects downstream analysis and causes misinterpretations.

Hence, a strategy must be developed to handle data, e.g. discarding or trimming, by pre-processing tools based on this information.

FastQC (Andrews, 2010) is a tool developed to convert this quality score information to human-readable statistics and metrics. Average quality score of positions of sequences, percentage of duplicated sequences and presence of primer and adapter sequences are among features of the program.

In the thesis, FastQC was applied to random chosen non-processed FASTQ to measure the overall quality of sequences. Three metrics were considered at most. “Per base sequence quality” shows the distribution of the Phred quality score for each position in box plots. Score above 30 (> Q30) are evaluated as ideal (Illumina, 2011). Regarding this, it seems that overall sequence quality was good for the samples (Figure 3.4.).

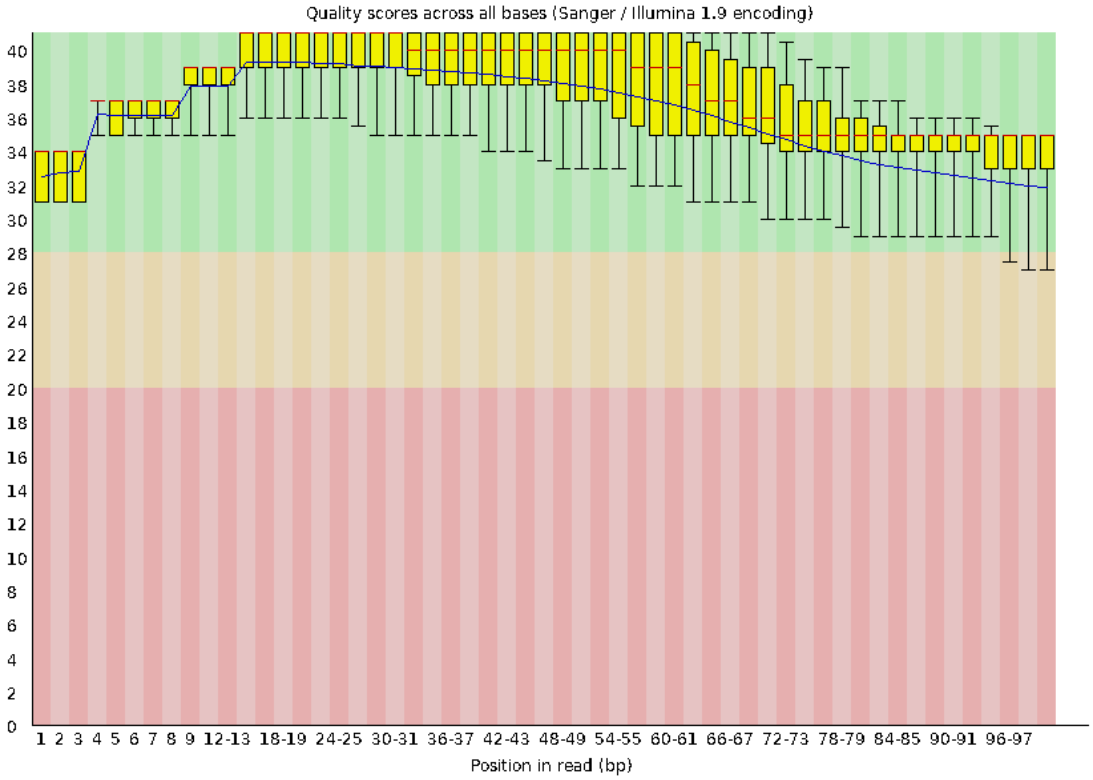

Figure 3.4. Per base sequence quality plot for SRR5305514 forward reads, before pre-processing

On the other hand, the requirement of pre-processing seems obvious when we look at "Per base sequence content". As given in example Figure 3.5., frequencies of

nucleotides toward the 5′ end deviate from the standard distribution. This indicates the repeats of 10-18 nucleotide length sequence/s, namely adapters.

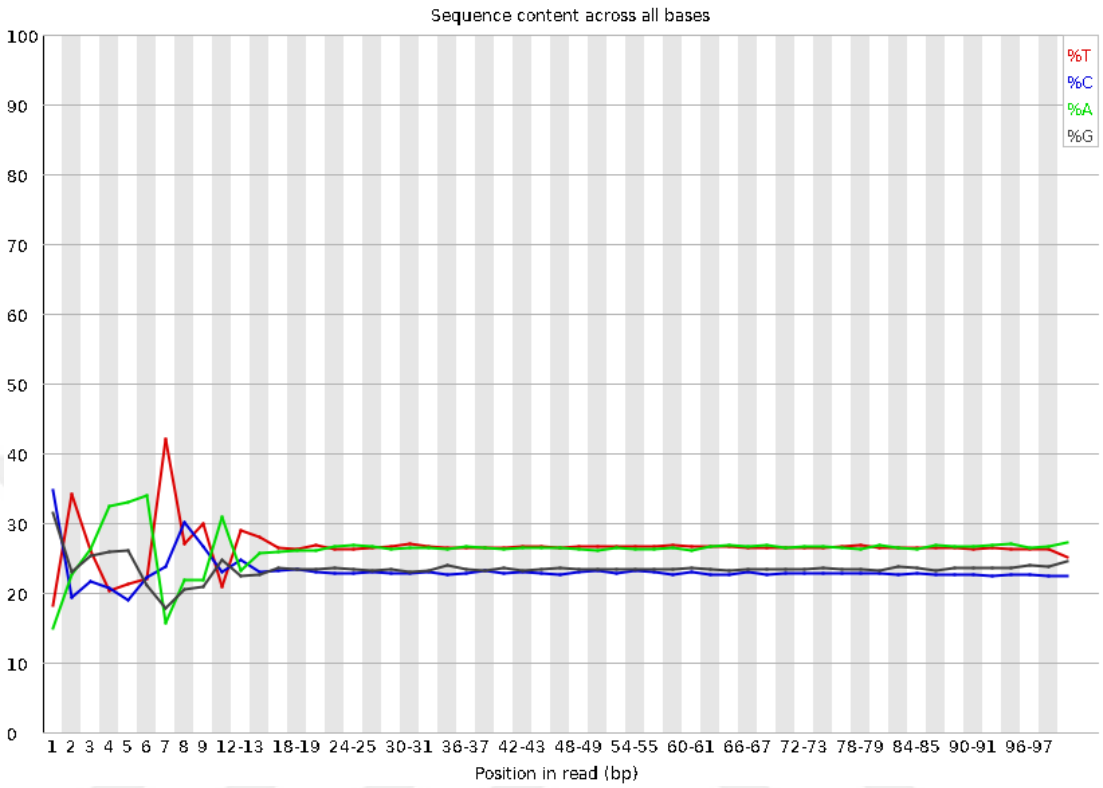

Figure 3.5. Per base sequence content plot for SRR5305514 forward reads, before pre-processing

Moreover, the "Overrepresented sequences" statistic was used to support this information. It gives overrepresented sequences, their counts, the percentage in all sequences and possible biological source. As an example, SRR5305514, an AD sample, has 37721 repeats of a TruSeq Adapter sequence, which constitutes nearly 12 % of all. Another example is SRR1927051 of ALS study, PRJNA279249. SRR1927051 has repeats of 4 nucleotide sequences with no possible source. Since they have no source and their percentage in total is high (> 70%), it is possible that they are contaminants.

Considering trials on each study, it is understood that samples are in good quality to be used in differential expression and variant discovery analyses. Nonetheless, they must be treated by a trimmer program to remove adapters and a contaminant removal program for some. The reason is given below.

3.4. Adapter Trimming by Skewer

Adapter sequences are short oligonucleotides ligated onto the ends of DNA fragments to allow for amplification of them. Usage of adapters in next-generation sequencing is a common practice in this regard. However, their presence in reads decreases the ratio of true alignment of the sequences to the genome dramatically; as a consequence, expression levels of genes are miscalculated in differential expression analysis. A similar situation is observed if reads are bad in quality (having a low Q score). Indeed, consecutive nucleotides with low quality, which may be a reflection of contamination, erroneous sequencing or base-calling (machine-read of respective nucleotide), cause false alignments, thus, alter downstream differential expression analysis drastically (Williams, Baccarella, Parrish, & Kim, 2016).

Trimming or clipping is the removal of these adapters and/or primer sequences and nucleotides with bad quality scores from the ends of reads. Skewer (Hongshan Jiang, Rong Lei, Shou-Wei Ding, & Shuifang Zhu, 2014) is a tool created for adapter trimming. It was preferred in this thesis due to its straightforwardness and speed. An example command is given below:

```
$ skewer --mode ap -c 10,10 -x ~/refs/mine/rRNA_tRNA_GRCh38.gtf
SRR1234567.F.fastq SRR1234567.R.fastq -t 16 -o "run"
```

Under this command, skewer runs using 16 threads (-t 16) in amplicon mode (--mode ap), trims 10 nucleotides of 5′ end from forward and reverse pairs. Then, it removes contaminant sequences of rRNA and tRNA given in file (-x ~/refs/mine/rRNA_tRNA_GRCh38.gtf), give output with given prefix (-o “run”). Based on its default parameters, skewer also discards bad quality bases on ends and masks if they are in the middle of the read.

3.5. Contaminant Removal by BBDuk

While adapter trimming is a prerequisite before continuing to downstream analysis, it is not the only step. Besides to adapters and primers, contaminants also pull down the performance of the further steps. ALS studies, PRJNA279249 and PRJNA306659, had

issues of contamination, which resulted in poor alignment. Although the skewer can remove contaminants from a FASTQ file to a degree, as done in the section above, tools designed for this specific purpose performs better. Therefore, another tool was required for contaminant removal. In this regard, BBDuk (Bushnell, 2015) was used in this thesis. It yielded good results, which validated in the alignment step. The command used is given below:

> $ *~/tools/bbmap/bbduk.sh in1=run-trimmed-pair1.fastq in2=run-trimmed-pair2.fastq out1=run-trimmed-pair1.clean.fastq out2=run-trimmed-pair2.clean.fastq ref=~/refs/mine/ /rRNA_tRNA_GRCh38.gtf ktrim=l k=21 mink=11 hdist=2 tpe tbo*

This command treated read pairs together (in1 and in2), list of rRNA and tRNA sequences from NCBI RefSeq (O'Leary et al., 2016) as reference (ref), gave outputs separately (out1 and out2). Other parameters were chosen to more sensitive trimming. Trimming was performed from left to right (ktrim=l). kmers longer than 21-nucleotides were evaluated as a contaminant (k=21), with maximum Hamming distance 2 (hdist=2). Adapters were trimmed based on where paired reads overlap (tbo).

After, quality metrics were measured again by FastQC. Results show successful pre-processing. Based on them, adapters were trimmed and contaminants (overrepresented sequences) were removed. Examples are given in Figure 3.6. and Figure 3.7.

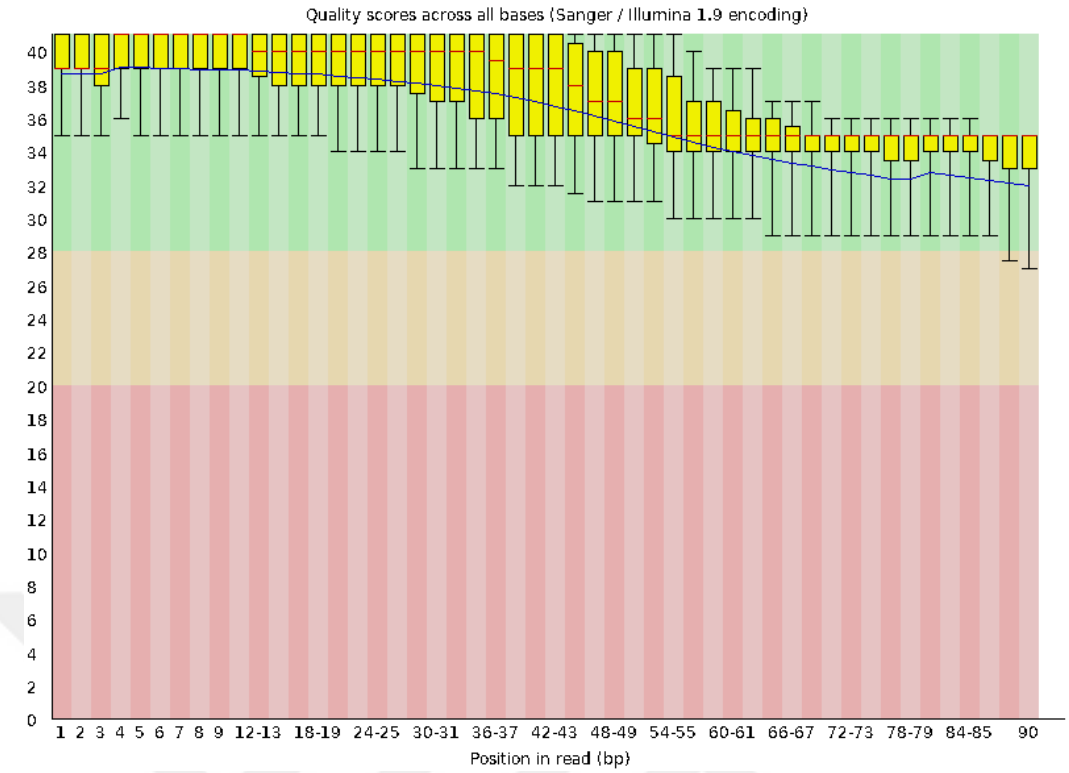

Figure 3.6. Per base sequence quality plot for SRR5305514 forward reads, after pre-processing

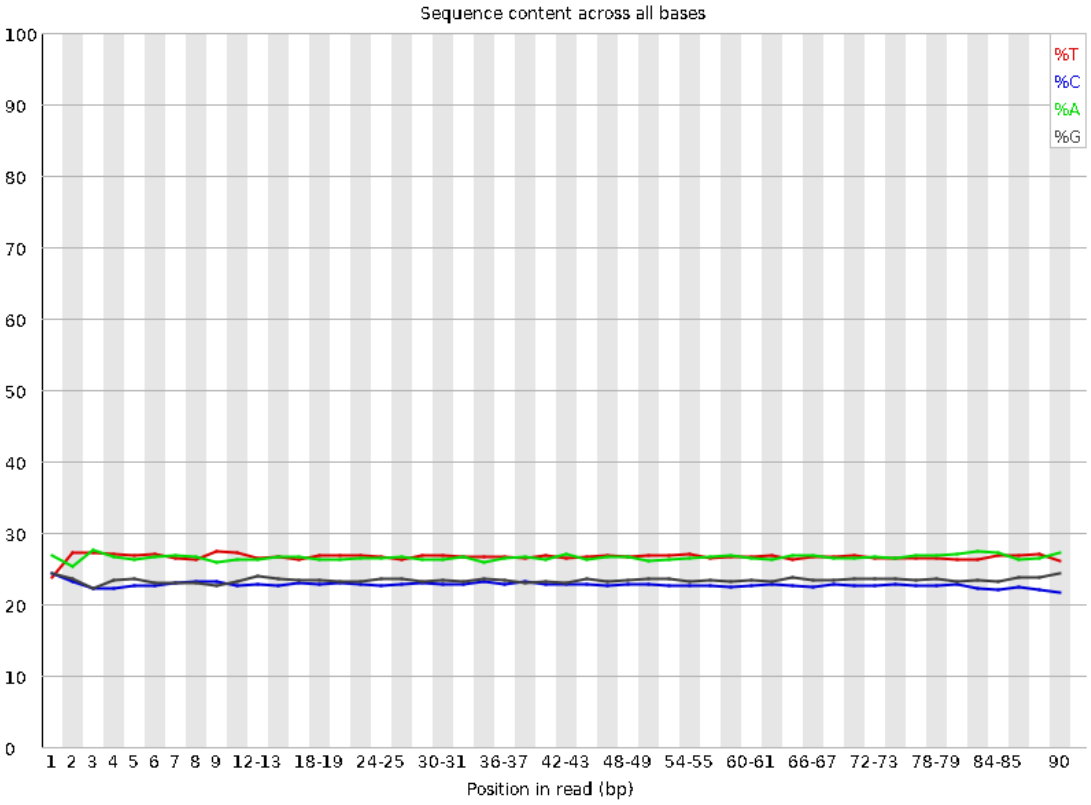

Figure 3.7. Per base sequence content plot for SRR5305514 forward reads, after pre-processing

3.6. Mapping Reads to Reference Genome by STAR

STAR (Dobin et al., 2013) is a highly accurate and ultra-fast alignment tool, based on C++, enables mapping of RNA-Seq reads onto the reference genome, detects annotated and novel splice junctions. Its accuracy besides its high speed verified by benchmark studies (Baruzzo et al., 2017; Teng et al., 2016). Another benefit of STAR over other aligners is that its newer versions, for example, v2.6.0c which was used in this thesis, counts reads corresponding to genes itself if the program run in the correct mode. Also, the program comes with a lot of parameters, which allows for the setting of values best suited for the purpose. At the downside, the program forces usage of at least 30 GB RAM for human genome alignments and works on Unix/Linux and Mac OS X operating systems, not on Windows. On a Linux based computing cluster, however, its advantages become more prominent.

Alignment using STAR is composed of two steps. At first step, STAR is run to produce genome associated information into tabular files or genome indexes files called in its documentation, such as chromosomal positions of splice junctions, using reference genome in FASTA format and annotation file in GTF/GFF format. While usage of the GTF file to annotate the genome is not necessary at this step, it greatly reduces the time for annotation of reads. The average read length of samples must be considered here before alignment. Adapter trimming must be included in this consideration. For instance, to align a paired-end sample with an average of 200 nucleotide length – forward and reverse read lengths are 100 nucleotides then -, and after a 10 nucleotide-length adapter trimming, necessary code is given below. sjdbOverhang represents the length of the donor/acceptor sequence on each side of the splice junctions and ideally 1 nucleotide shorter than average read length.

```
$ mkdir ~/references/length90 && cd ~/references/length90

$ STAR --runMode genomeGenerate --genomeDir . --sjdbOverhang 89 --
genomeFastaFiles GCF_000001405.38_GRCh38.p12_genomic.fna --
sjdbGTFfile  GCF_000001405.38_GRCh38.p12_genomic.gtf
```

At the second step, reads are mapped to the reference genome. Two parameters are especially important for this step. The first parameter is twopassMode as Basic. In two-

pass mode, STAR maps reads twice. In the 1st read or pass, STAR detects novel junctions and add them into genome indices. In the 2nd pass, it re-maps reads using annotated and novel junctions. The detection of novel junctions significantly reduces the misalignment of reads to similar adjacent genomic regions (Dobin et al., 2013); thus, reads are aligned more correctly. The second parameter is quantMode as GeneCounts. This parameter makes STAR count reads aligned for each gene. By this parameter, we get a read count table beside an aligned reads SAM file. An example code is:

> $ *STAR --runThreadN 32 --genomeDir ~/references/length90 --readFilesIn SRR1234567.F-trimmed.fastq SRR1234567.R-trimmed.fastq --sjdbOverhang 89 --twopassMode Basic --quantMode GeneCounts --sjdbGTFfile GCF_000001405.38_GRCh38.p12_genomic.gtf*

The code above produces an aligned reads file in SAM (Sequence Alignment Mapping) format (Aligned.out.sam), a read counts table (for example SRR1234567.out.tab), a table of splice junctions (SJ.out.tab) and three report files (Log.out, Log.progress.out, Log.final.out). SAM file is further used for variant discovery analysis. Furthermore, the SAM file and its binary form BAM file can be used for differential expression analysis (Anders, Pyl, & Huber, 2015; Liao, Smyth, & Shi, 2014).

Read count table has 4 columns. Counts for unmapped reads, multi mapped reads, reads with no feature, ambiguous reads are placed in first 4 rows. On following rows, gene IDs and its corresponding counts for unstranded RNA-Seq, the 1st/forward read and 2nd/reverse read strand aligned with RNA are given. Counts on 2nd column in general equal to sum of count on 3rd and 4th columns. Besides counts on 3rd were not much higher than counts on 4th column, vice versa. Therefore, 2nd column was selected which reflects gene counts better. While making count table with gene symbols, this column was kept besides to column of gene IDs. The respective code is:

> $ *join -1 1 -2 1 <(sort -k1 ~/refs/mine/genes.names.GRCh38.p12.table) <(awk '{print $1,$2}' "SRR1234567.out.tab" | grep -Pv "^N_" | sort -k1) > "SRR1234567.out.counts.tab"*

After run of all samples, count tables produced by STAR were merged into a single read count table for each study for downstream analyses. The code below was used to merge tables:

```
$ while read -r line; do join -1 1 -2 1 <(sort -k1
~/refs/mine/genes.names.GRCh38.p12.table) <(awk '{print $1,$2}'
$line/$line.out.tab | grep -Pv "^N_" | sort -k1) > $line/$line.read.counts.tab
;done< <(ls | grep SRR)

$ awk '{print $1,$2}' SRR*/SRR*read.counts.tab | sort -uk1 > a.merge.out

$ while read -r line; do join --nocheck-order -a1 -1 1 -2 1 -e"0" <(sort -k1
a.merge.out) <(awk '{print $1,$3}' $line | sort -k1) > a.merge.out.tmp &&
mv a.merge.out.tmp a.merge.out; done< <(ls SRR*/SRR*read.counts.tab)
```

As mentioned above STAR gives three output files. Log.out has information about executed command, default and executed parameters. Log.progress.out shows aligned read count, various mapping status of reads (unique region mapping, multi region mapping, unmapped reads) with 1 minute intervals. Log.final.out shows mapping status of reads with end of alignment and how many reads aligned with these statuses. Ideally, more than 90% of reads should be aligned uniquely and unmapped reads should not constitute more than 5% of all reads, mismatch and insertion/deletion rates should be as low as possible. Higher ratio of unmapped reads indicates overall bad alignment, while high percentages of reads mapped to multiple loci shows unsuccessful removal ribosomal RNA from sample in case of RNA-Seq. Command below is to extract statistics of aligned reads:

```
$ echo -e "Uniq\tMulti\t Ratio (Multi / Uniq)\t Sum (Multi +
Sum)\tSample\tUniqReadCount" \

&& \

paste \

<(grep "Uniquely" SRR*/Log.final.out | grep "%" | grep -Po "\d+\.\d+") \

<(grep "multiple" SRR*/Log.final.out | grep "%" | grep -Po "\d+\.\d+") \

<(ls SRR*/Log.final.out | grep "number" | awk '{print $1}') \
```

```
<(grep "Uniquely" SRR*/Log.final.out | grep "number" | awk '{print $1,$7}') \
| awk '{print $1,$2,$2/$1,$1+$2,$3,$4,$5}' | sort -k4 | sed "s/ /\t/g" \
&&\
paste \
<(grep "Uniquely" SRR*/Log.final.out | grep "%" | grep -Po "\d+\.\d+") \
<(grep "multiple" SRR*/Log.final.out | grep "%" | grep -Po "\d+\.\d+") \
<(ls SRR*/Log.final.out | grep "number" | awk '{print $1}') \
<(grep "Uniquely" SRR*/Log.final.out | grep "number" | awk '{print $1,$7}') \
| awk '{print $1,$2,$2/$1,$1+$2,$3,$4,$5}' | sort -k4 | sed "s/ /\t/g" | \
awk '{c+=1;s1+=$1;s2+=$2;s3+=$3;s4+=$4;s6+=$6}END{print s1/c,s2/c,s3/c,s4/c,"average",s6/c}'
```

An example table (Table 3.1.) is given below. According to this, average unique read percentage was close to 90% and average unique read counts was above 30 million for AD samples. This implies that gene expression counts were reliable for differential expression analysis.

Table 3.1. Statistics of aligned reads of AD samples (PRJNA377568), which were extracted from

Uniq	Multi	Ratio (Multi / Uniq)	Sum (Multi + Sum)	Sample	UniqReadCount
90.76	8.35	0.0920009	99.11	SRR5305522/Log.final.out:	23228079
91.28	8	0.0876424	99.28	SRR5305533/Log.final.out:	27884097
90.99	8.4	0.0923178	99.39	SRR5305573/Log.final.out:	26930739
91.15	8.31	0.0911684	99.46	SRR5305518/Log.final.out:	33817698
82.22	17.24	0.209681	99.46	SRR5305529/Log.final.out:	19539913

Table 3.2. cont.

89.13	10.39	0.116571	99.52	SRR5305557/Log.final.out:	32921670
91.66	7.84	0.0855335	99.5	SRR5305553/Log.final.out:	34777191
91.46	8.15	0.08911	99.61	SRR5305588/Log.final.out:	31629803
…	…	…	…	…	…
90.64	9.16	0.101059	99.8	SRR5305506/Log.final.out:	37357199
90.83	8.97	0.0987559	99.8	SRR5305525/Log.final.out:	46954260
92.09	7.71	0.0837224	99.8	SRR5305530/Log.final.out:	22862416
85.7	14.1	0.164527	99.8	SRR5305541/Log.final.out:	18692349
92.2	7.6	0.0824295	99.8	SRR5305548/Log.final.out:	23499672
89.72	10.08	0.11235	99.8	SRR5305566/Log.final.out:	29130541
90.68	9.12	0.100573	99.8	SRR5305578/Log.final.out:	22981712
90.1692	9.58479	0.106788	99.754	average	3.03E+07

Log.final.out files. Four columns from left give percentages of uniquely mapped and multi loci mapped reads, the ratio of multi to unique and percentage of unique and multi in total, respectively. The 5th column represents a respective log file. The 6th column gives uniquely mapped read counts.

3.7. Variant Discovery by rnaseqmut

rnaseqmut (W. Li, 2016) is a C++ program created to detect variants from RNA-Seq derived BAM files. Source package provides a core rnaseqmut program, a Bash script with a one-line command to merge multiple VCF outputs and two Python scripts for merging RNA-Seq variant scans in VCF and filtering them on given criteria. This program is capable of detecting both SNPs and indels. On the other hand, it assumes

that most RNA-Seq are performed by Illumina sequencing, which is true for this study; thus, it skips indels.

The core program requires a sorted BAM file with an index file. Therefore, before the rnaseqmut run, BAM files were sorted by samtools sort and indexed by samtools index (H. Li et al., 2009).

After that, rnaseqmut run with the reference genome for each sorted BAM with default parameters. The maximum number of mismatches in a read is 1, the minimum read count for the mutation to output is 1 and the minimum distance of the mutation to the beginning (end) of the read is 4, by default. The output is a variant list of the sample.

$ *rnaseqmut "$line.sorted.bam" -r $genome > rnaseqmut.vars.txt*

Each list only contains variants of its corresponding sample. These lists were merged into a single file. rnaseqmut run again, but only variants in merged variants list were detected this time. These lists were merged in the VCF format (Danecek et al., 2011) by merge2ndvcf.py. filtermut.py is used to filter out disease-specific variants from this VCF. filtermut.py requires labels of each sample separated by a comma - e.g. cont1,cont2,alt1,alt2 – and Python-style index of control samples separated by a comma. Besides to them, --min-recurrent parameter was altered, which allows mutations occurring in at least this number disease sample, value to numbers given in Table 3.2.

Table 3.3. rnaseqmut parameters used for each study

Project Name	PD_PRJNA283498	ALZ_PRJNA377568	ALS_PRJNA279249	ALS_PRJNA306659
-r / --ref_fasta	GCF_000001405.38_GRCh38.p12_genomic.fna			
-i / --min_read	1 (default)	1 (default)	1 (default)	1 (default)
-t / --min-recurrent	1 (default)	1 (default)	1 (default)	1 (default)

$ *~/tools/rnaseqmut-master/script/merge1stfile SRR*/rnaseqmut.vars.txt > merged.vars.txt*

```
$ mkdir vars
$ while read -r line; do ~/tools/rnaseqmut-master/bin/rnaseqmut -r
/gpfs/user/abayraktar/refs/mine/GCF_000001405.38_GRCh38.p12_genomic.
fna -l merged.vars.txt "$line"/"$line".sorted.bam > vars/"$line".vars.txt &
done< <(cat "$(pwd)/sra_accessions.csv" | cut -d , -f 1)
$ samples=$(cat "$(pwd)/sra_accessions.csv" | cut -d , -f 1 | tr '\n' ',' | sed
"s/,$//g")
$ control_index=$(cat "$(pwd)/sra_accessions.csv" | cut -d , -f 2 | grep -n C |
cut -d : -f 1
| awk '{print $1=$1-1}' | tr '\n' ',' | sed "s/,$//g")
$ mkdir annotation
$ wait
$ python ~/tools/rnaseqmut-master/script/merge2ndvcf.py -l "$samples"
vars/*vars.txt > annotation/all.mutations.txt
$ wait
$ python ~/tools/rnaseqmut-master/script/filtermut.py -c "$control_index" -l
"$samples" < annotation/all.mutations.txt > annotation/variants.vcf
```

3.8. Changes on VCF prior to annotation

Two changes were made prior to annotation. First, VCF files were converted to correct form. In previous step, disease specific variants were collected in VCF files. On the other hand, these files are not in proper VCF format. Second, chromosome names were converted from RefSeq format (NC_000001.11, NC_000001.12 …), which was the format of the reference genome for alignment, to GENCODE format (chr1, chr2 …) (Harrow et al., n.d.), which is the format needed for used annotation tool.

```
$ echo "##fileformat=VCFv4.2" > annotation/variants.tmp.vcf
$ echo
"##reference=///gpfs/user/abayraktar/refs/mine/GCF_000001405.38_GRCh3
8.p12_genomic.fna" >> annotation/variants.tmp.vcf
```

```
$ grep -Pv "^# " annotation/variants.vcf | sed "s/pos/pos\tid/g" | awk '{ print $1,$2,$3,$4,$5 }' | \
tr ' ' '\t' |tr [:lower:] [:upper:] | sed "1s/$/\tQUAL\tFILTER\tINFO/g;s/$/\t100\tPASS\t\./g" >> annotation/variants.tmp.vcf
$ mv annotation/variants.tmp.vcf annotation/variants.vcf
$ while read -r line; do old=$(echo $line | cut -d " " -f 1); new=$(echo $line | cut -d " " -f 2); sed -i "s/$old/$new/g" annotation/variants.vcf; done < chr_conversion.csv
```

3.9. Annotation by ANNOVAR

In the previous step, variants specific to each disease were discovered. On the other hand, variants in VCF only have information on chromosomal positions, read counts and related quality scores. To gather human-readable information, i.e. annotations, ANNOVAR was used. ANNOVAR (Wang, Li, & Hakonarson, 2010)matches chromosomal positions of variants on multiple cross-validated databases and inserts information of gene and frequency related information.

```
$ perl ~/tools/annovar/convert2annovar.pl -format vcf4 annotation/variants.vcf -outfile  annotation/outfile.avinput -allsample -withfreq -include
$ perl ~/tools/annovar/table_annovar.pl annotation/outfile.avinput ~/tools/annovar/humandb_hg38/ -buildver hg38 -out annotation/annotated.variants.txt -remove -protocol refGene,knownGene,esp6500siv2_all,1000g2015aug_all,exac03,avsnp150,dbnsfp33a,clinvar_20170905,gnomad_exome,kaviar_20150923 -operation g,g,f,f,f,f,f,f,f -nastring .
```

3.10. Differential Expression Analysis by R (DESeq2)

As mentioned previously, alignment/mapping reads produced gene count tables and SAM files. Both file types can be used for differential expression analysis. Working on Rstudio was chosen due to its easiness. Therefore, a package for differential expression analysis was selected, namely DESeq2. Since DESeq2 works on count tables, the gene count tables provided by STAR were used.

DESeq2 package (Love, Huber, & Anders, 2014) is one of the best tools developed for RNA-Seq data for the identification of differentially expressed genes. In addition to its wide acceptance by the genetics community, its statistical power has been demonstrated by simulation-based, comparative studies (Costa-Silva, Domingues, & Lopes, 2017; Schurch et al., 2016). It estimates log2-fold change (LFC), standard deviation of LFC, base mean count of each gene or isoform, calculates p-values and adjusted p-values for these findings.

R projects were conducted for each study using DESeq2 to obtain result tables for them, which have measurements given above. Later, they saved as CSV files. Code for PD study, PRJNA283498, is given below:

```
$ library(DESeq2)
$ sra_list <- read.csv("sra_accessions.csv", header = F)
$ pd <- read.csv("a.merge.out", sep = " ", header = F, row.names = 1)
$ colnames(pd) <- c("gene symbol",as.vector(rev(sra_list[,1])))
$ pd <- na.omit(pd)
$ dset.pd <- DESeqDataSetFromMatrix(countData =
na.omit(pd[,2:ncol(pd)]), colData = sra_list, design = ~V1+V2)
$ ds2.pd <- DESeq(dset.pd)
$ res.pd <- results(ds2.pd, contrast = c("V2","C","P"), alpha = 0.05,
pAdjustMethod = "fdr")
$ res.pd$gene <- pd[which(rownames(res.pd) %in% rownames(pd)),1]
$ res.pd <- na.omit(res.pd)
$ write.csv(res.pd, file = "res.pd.PRJNA283498.csv")
```

It should be noted that setting a p-value (alpha = 0.05) and an adjustment method (pAdjustMethod = "fdr") is important in standard method. On the other hand, upcoming meta-analysis does not require them.

3.11. Meta-Analysis by R (metaRNASeq)

After completing differential expression analysis for the studies, a meta-analysis project in R was created. The metaRNASeq package was used for meta-analysis in this project. metaRNASeq (Rau et al., 2014) implements two p-value combination methods, i.e. Fisher and inverse normal, which are successful methods for microarray studies (L.-C. Chang, Lin, Sibille, & Tseng, 2013; Hong & Breitling, 2008; Tseng, Ghosh, & Feingold, 2012), to investigate multiple related RNA-Seq studies.

At first, four results were inserted table into the project. After, they were intersected into a single table. Then, Fisher and inverse normal p-value combination methods of the metaRNASeq package were applied onto the table based on p-values of each study, while setting p-value as 0.05. In other words, the methods consider raw p-values of each study, - inverse normal considers sample sizes too -, calculates new p-values for genes and apply Benjamini-Hochberg correction on p-values to control false discovery rate at 0.05 at default. Later, two lists were made by union and intersection of the methods.

Genes that are either upregulated or downregulated in all four studies were kept. There were two reasons behind it. First, it is related to the scope of this thesis. In this thesis, neurodegeneration is hypothesized as a common phenomenon among NDDs. A gene showing different expressional changes indicates a disease-specific modification for this gene. Therefore, changes in the same direction were considered. Second, it would increase the statistical power of the test by skipping suspicious genes. Some genes can be set as significant DEG, although it is not significant in one or more studies. A gene may be upregulated in real disease cases, while it was measured as downregulated. Though, it was not worthy to take this risk; hence, it should be omitted.

3.12. Pathway Analysis or Functional Enrichment by Cytoscape

Pathway analysis has a broad meaning in biological sciences. By this term, analysis of gene ontological terms, physical interaction networks, biochemical networks or molecular mechanisms covering these terms to some degree such as MAPK signalling pathway and exocytosis (Khatri, Sirota, & Butte, 2012). The major purpose of the thesis is to observe changes in pathways present pathway databases, e.g. Reactome (Croft et al., 2011; Fabregat et al., 2018); hence this is what is meant by pathway analysis here. Similarly, functional enrichment was used for annotation of common gene ontologies.

In this regard, Cytoscape (Shannon et al., 2003) version 3.7.1, which an open-source widely-used software for integration of biological molecular interaction networks, was used. Significant expression lists were investigated in Cytoscape by different plugins (tools to pull information from databases). ReactomeFI (G. Wu & Haw, 2017) was used to find out statistically significant pathways having these genes. BiNGO (Maere, Heymans, & Kuiper, 2005) was performed to look for common Gene Ontology (Gene Ontology Consortium, 2004) biological processes. Most significant pathways besides disease pathways were selected to be merged with significant DEGs. Pathways were downloaded by WikiPathways (Kutmon, Lotia, Evelo, & Pico, 2014). Genes having an L2FC score were chosen and the colours of respective nodes were changed: reds for upregulation, blues for downregulation, light purple for close to zero-fold change. Pathway views were saved in JPEG format.

4. RESULTS

Differential expression analysis of four datasets had 13603 genes in common. The application of Fisher and inverse normal p-value combination methods passed 2954 and 1365 significant genes whose p-value is lower than 0.05, respectively. Union of the methods had 358 significant DEGs whose expression levels changed in the same direction. Whereas, the intersection of the methods had 200 significant DEGs in the same direction. Through the rest of the thesis, these 200 significant DEGs in the same direction is going to be called as 200 genes. In R, 200 genes were compared with disease-associated gene lists which downloaded from NCBI RefSeq (O'Leary et al., 2016). 12 of the 200 genes (*CALN1, VGF, ADAMTSL1, FKBP4, PTGES3, AHSA1, HSP90AA1, GAS7, PKN1, C22orf34, SPR, RBMS1*) have known associations with AD, PD or ALS based on previous studies curated for RefSeq.

Reactome enrichment of these 200 genes revealed overrepresentation of pathways mainly related to cellular responses to stress (e.g. attenuation phase, HSF-1 dependent transactivation pathways, regulation of HSF1-mediated heat shock response, HSP90-chaperone cycle for steroid hormone receptors) and GABA synthesis (e.g. MECP2 regulates transcription of genes involved in GABA signalling) (Table 4.1.). Genes overrepresented in the top three pathways were *HSP90AA1, PTGES3, FKBP4, DNAJB1, DNAJB6, DEDD2* and *HSPB1*.

GO biological process enrichment validates Reactome enrichment results and further explains the significance of some pathways. The genes associated with protein folding, refolding and responding to unfolded protein are also the ones represented in the cellular response to heat stress pathways, except *DEDD2* and *HSPB1*. In addition, *GAD1* and *GAD2*, as they are involved in glutamate decarboxylation to succinate, are represented in the GABA synthesis pathway (Table 4.2.).

Table 4.1. The overrepresented pathways in Reactome for the 200 significantly DEGs common for three diseases

Pathway name	Entities				Reactions	
	found	ratio	p-value	FDR*	found	ratio
Attenuation phase	9 / 47	0.003	7.96e-08	4.12e-05	4 / 5	4.13e-04
HSF1-dependent transactivation	9 / 59	0.004	5.24e-07	1.36e-04	5 / 8	6.60e-04
MECP2 regulates transcription of genes involved in GABA signaling	4 / 6	4.24e-04	2.95e-06	5.08e-04	4 / 4	3.30e-04
HSF1 activation	7 / 43	0.003	6.47e-06	8.35e-04	4 / 7	5.78e-04
TFAP2A acts as a transcriptional repressor during retinoic acid induced cell differentiation	4 / 9	6.36e-04	1.44e-05	0.001	4 / 7	5.78e-04
HSP90 chaperone cycle for steroid hormone receptors (SHR)	8 / 70	0.005	1.81e-05	0.002	12 / 12	9.91e-04
Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors	7 / 52	0.004	2.17e-05	0.002	9 / 44	0.004
Cellular response to heat stress	10 / 135	0.01	6.51e-05	0.004	13 / 29	0.002
Regulation of HSF1-mediated heat shock response	7 / 113	0.008	0.002	0.128	4 / 14	0.001
Cellular responses to stress	17 / 514	0.036	0.003	0.17	33 / 184	0.015
GABA synthesis	2 / 7	4.95e-04	0.006	0.262	2 / 2	1.65e-04

Table 4.2. Overrepresented GO Biological Processes for 200 significant DEGs. Produced by BiNGO.

p-value	Description	Corrected p-value	Cluster frequency	Total frequency	Genes
8.4556E-5	Glutamate decarboxylation to succinate	2.4487E-2	2/132 1.5%	2/14301 0.0%	GAD1 GAD2
3.8739E-6	Protein folding	1.8698E-3	10/132 7.5%	170/14301 1.1%	DNAJB2 DNAJA1 DNAJB1 HSP90AA1 DNAJB6 PTGES3 AHSA1 FKBP4 PPID HSPD1

Table 4.2. cont.

8.4307E-7	Protein refolding	1.2208E-3	4/132 3.0%	9/14301 0.0%	DNAJB1 HSP90AA1 PTGES3 HSPD1
1.2493E-5	Response to protein stimulus	4.5223E-3	8/132 6.0%	117/14301 0.8%	DNAJB2 DNAJA1 DNAJB1 HSP90AA1 DNAJB6 HSPB1 BCL2L1 HSPD1
2.4088E-6	Response to unfolded protein	1.7440E-3	7/132 5.3%	66/14301 0.4%	DNAJB2 DNAJA1 DNAJB1 HSP90AA1 DNAJB6 HSPB1 HSPD1

Gene set analysis was performed by ReactomeFI without linker genes option using 200 genes list. In this way, genes were connected via pathway interactions pulled from Reactome. Figure 4.1. illustrates the components of the network produced. Red nodes represent upregulated genes, while blue nodes represent the downregulated genes. Interaction network presents that *HSP90AA1*, *PTGES3*, *FKBP4*, *AHSA1*, *DNAJA1*, *DNAJB1*, *PPID*, and *HSPD1* are connected and up-regulated in the largest connected component of the cluster. Downregulated *GAD1* and *GAD2* also make a pair.

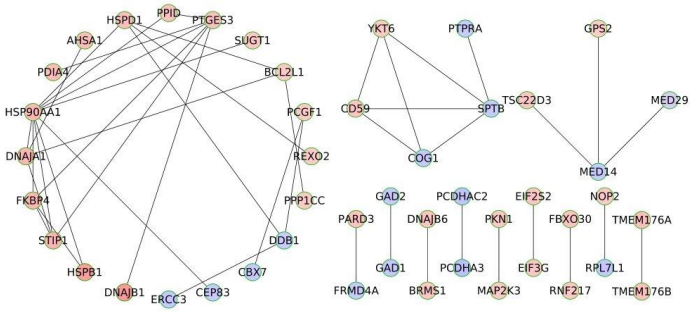

Figure 4.1. Network of 47 found by ReactomeFI plugin based on 200 significant DEG

Findings up to this point exposed that two pathways, namely cellular response to heat stress and GABA synthesis, affected in these three neurodegenerative disorders. Terms "protein folding", "attenuation phase" and others can be summarised under the header of "Cellular response to heat stress", which is the appropriate pathway in WikiPathways. Hence, it was imported. The closest term for "GABA synthesis" in WikiPathways was “GABA synthesis, release, reuptake and degradation”; thus, it was imported. Pathway names, their IDs and references are listed in Table 4.3. Figures of pathways were given below.

Table 4.3. **Imported WikiPathways: their names, their IDs and their references**

Pathway Name	WikiPathways ID	Reference
AD (Homo sapiens)	WP2059	(Salomonis et al., n.d.)
ALS (Homo sapiens)	WP2447	(Riutta et al., n.d.)
PD Pathway (Homo sapiens)	WP2371	(Pico, Hanspers, Willighagen, Fehrhart, & Kutmon, n.d.)
Cellular response to heat stress (Homo sapiens)	WP3395	(Reactome Team & Kutmon, n.d.)
GABA synthesis, release, reuptake and degradation	WP2685	(Reactome Team, Bohler, & Kutmon, n.d.)

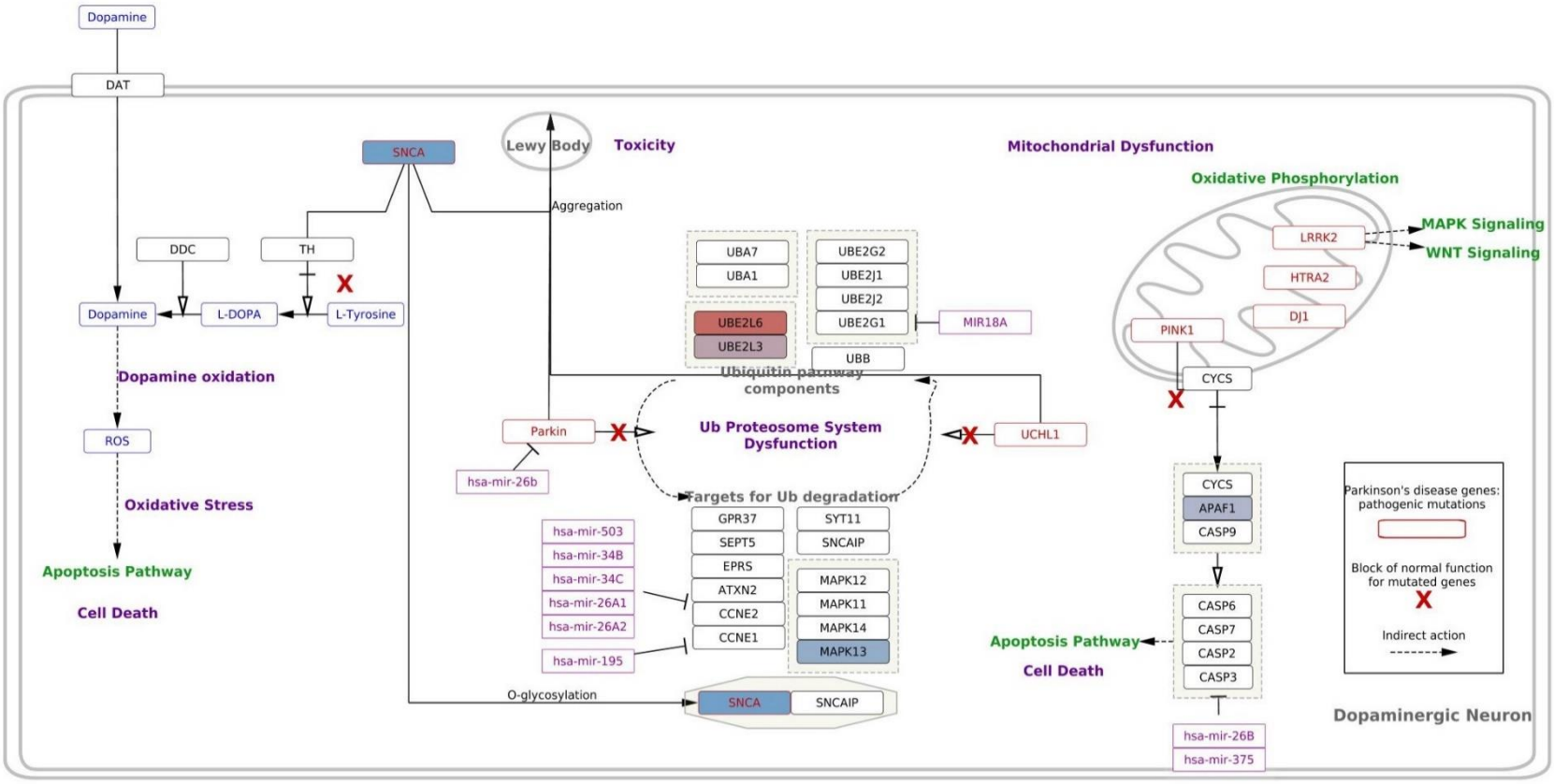

Figure 4.2. Parkinson's Disease Pathway annotated with significant gene expression levels. *UBE2L3* (up) was the only significant DEG in the pathway. Found genes whose expression has changed in the same direction, but not significant: *SNCA, APAF1, MAPK3* (down), *UBE2L6* (up).

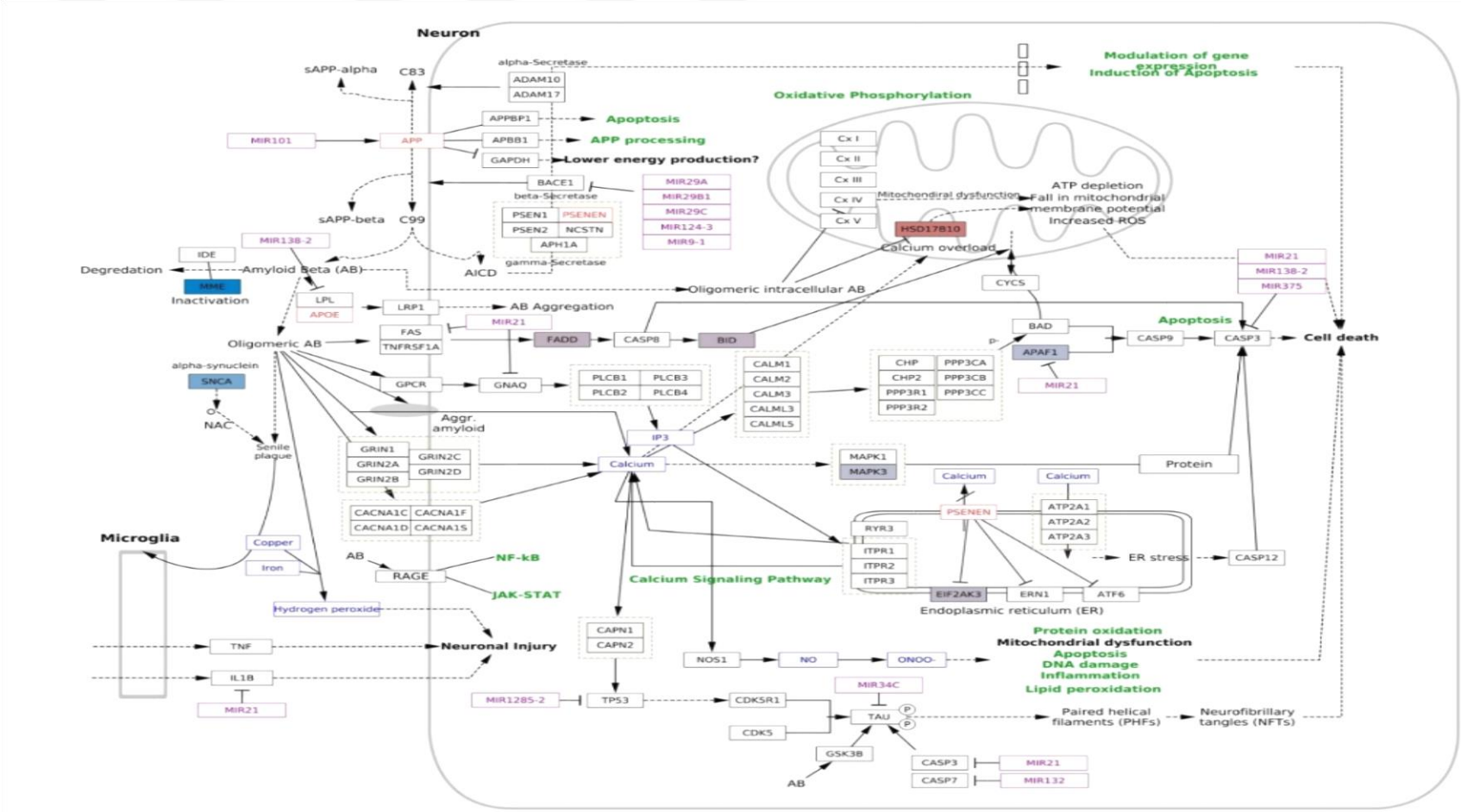

Figure 4.3. Alzheimer's Disease Pathway annotated with significant gene expression levels. No significant DEG was found in the pathway. Found genes whose expression has changed in the same direction, but not significant: *SNCA, MME, APAF1, MAPK3* (down), *HSD17B10, FADD, BID, EIF2AK3* (up).

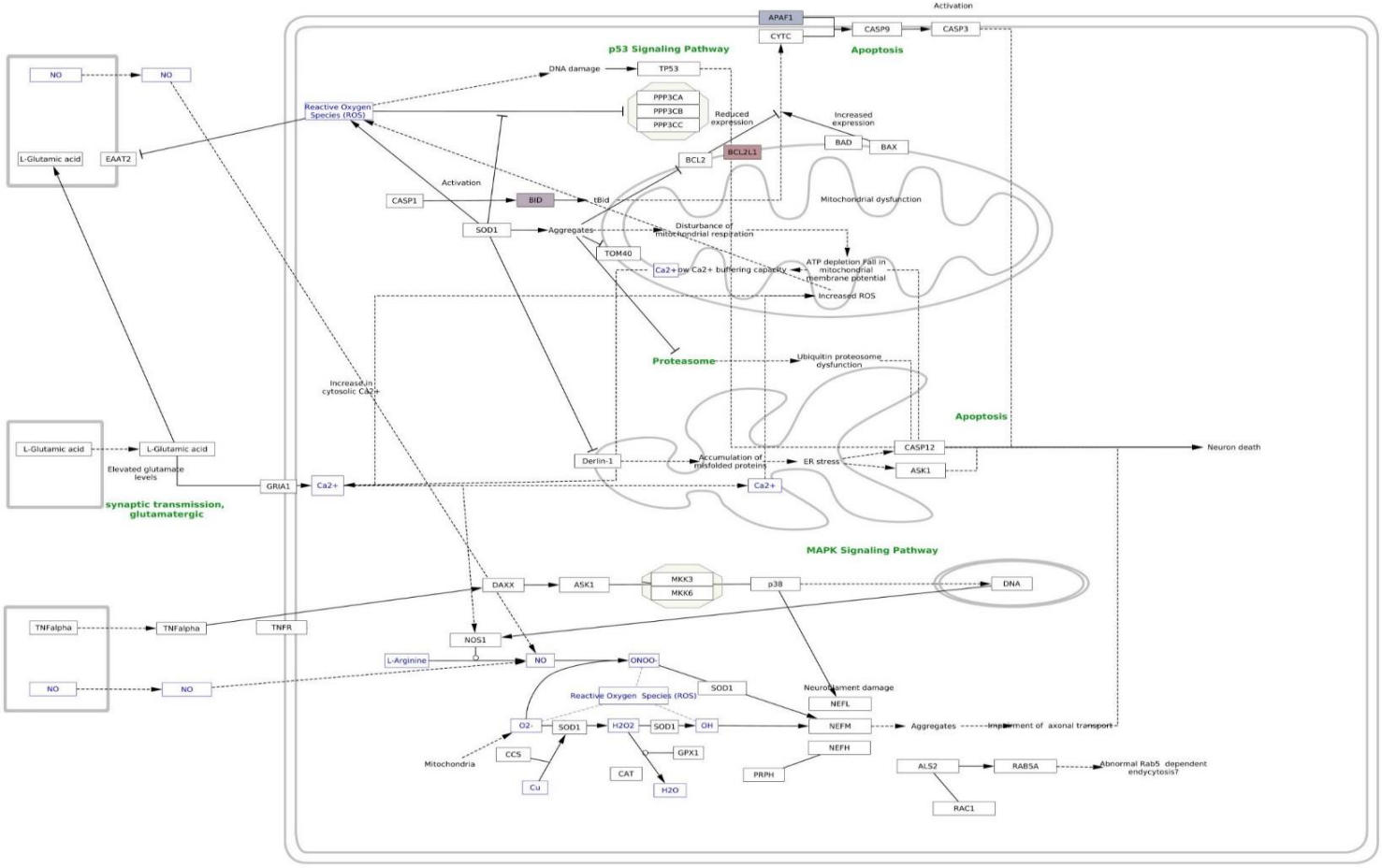

Figure 4.4. Amyotrophic Lateral Sclerosis Pathway annotated with significant gene expression levels. *BCL2L1* (up) was the only significant DEG in the pathway. Non-significant genes found: *BID* (up) and *APAF1* (down). Others are *DBC1* and *p-T454-DBC1*, both downregulated.

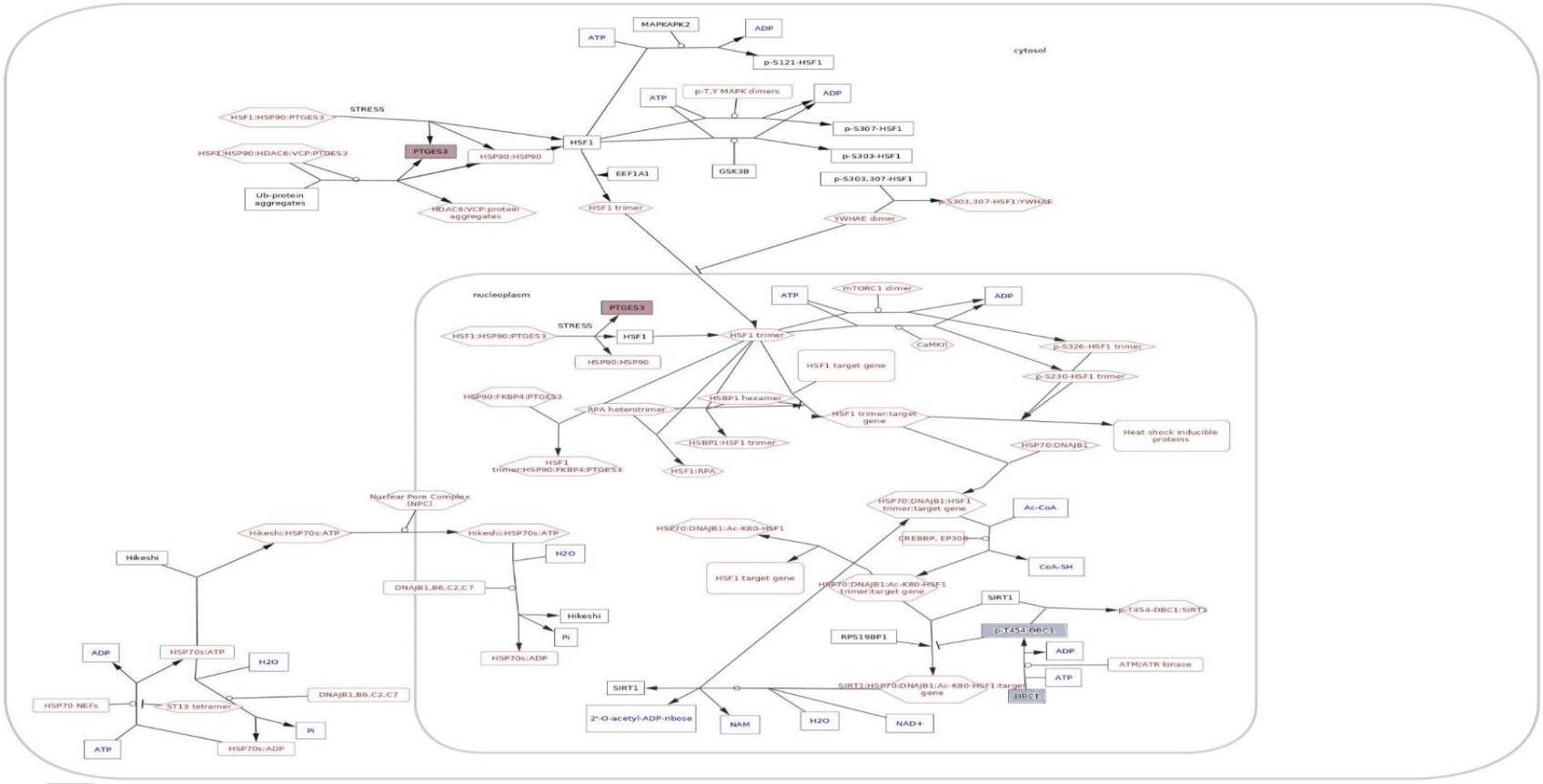

Figure 4.5. Cellular response to heat stress Pathway annotated with significant gene expression levels. *PTGES3* (up) was the only significant DEG in the pathway. *DBC1* (down) is the other gene found, but not significant.

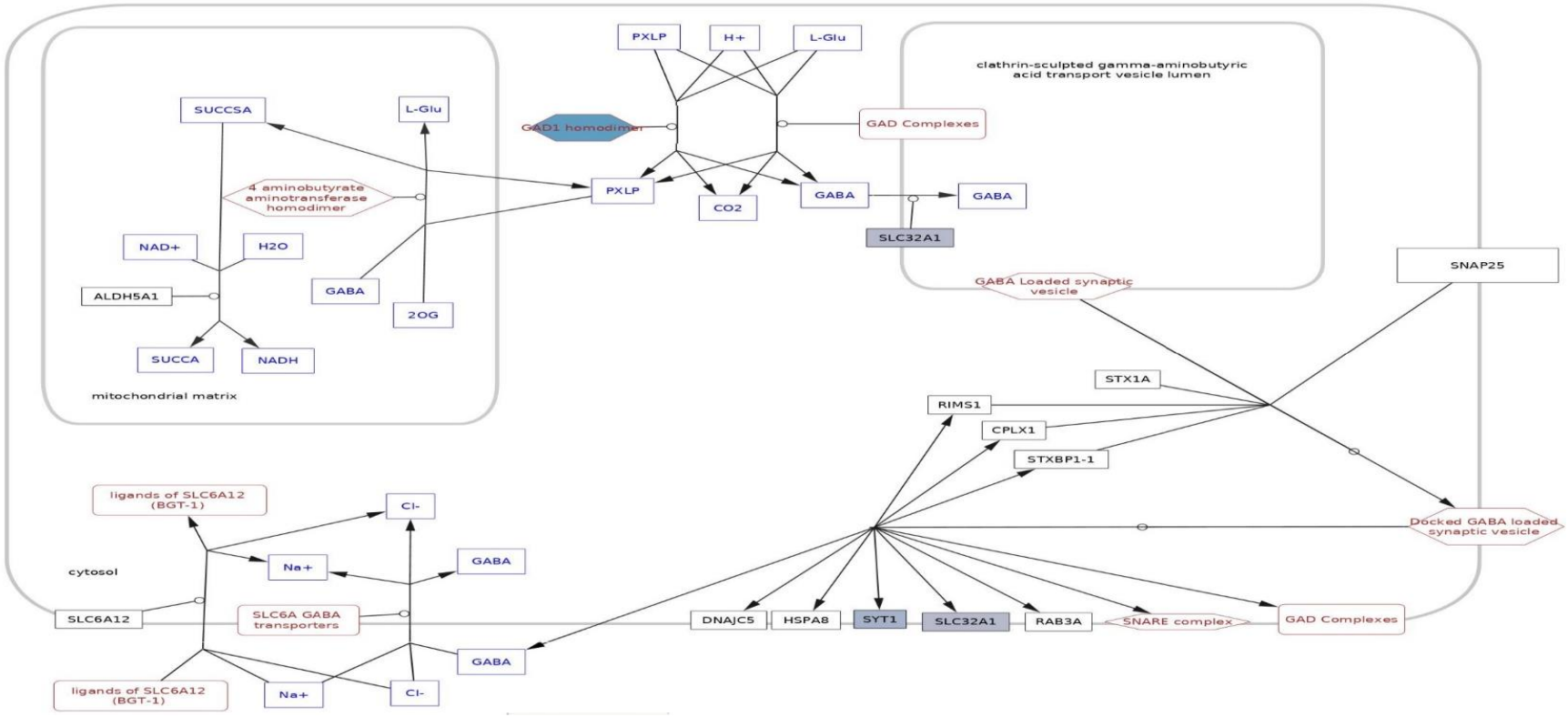

Figure 4.6. GABA synthesis, release, reuptake, and degradation Pathway annotated with significant gene expression levels. *GAD1* and *SLC32A1* were significant DEGs, downregulated in this pathway. *GAD1* is present as a homodimer and part of GAD complexes, which was not annotated here. *SLC32A1* is significant for the Fisher method. Another found gene is *SYT1*, downregulated also.

Node tables of downloaded pathways were merged with a list of genes regulated in the same direction, having 1712 genes and including 200 significant DEGs. Among significant DEGs, *UBE2L3* for PD, *BCL2L1* for ALS, *PTGES3* for cellular response to heat stress and *GAD1* and *SLC32A1* for GABA synthesis were identified. Considering these findings, it can be said that there is not a significant alteration in the pathway despite the previous results.

Nonetheless, changes in protein complexes support the previous ones. *DEDD2*, *DNAJB1*, *DNAJB2*, *DNAJB6*, *HSP90AA1*, *PTGES3* and *FKBP4* are among the components of protein complexes which take place in the cellular response to heat stress. When they were highlighted, which are all upregulated, it was observed that a serious portion of the protein complexes in this pathway was upregulated in all three neurodegenerative diseases (Figure 4.7.). Similarly, more than half of the protein complexes in GABA pathways were downregulated (Figure 4.8.).

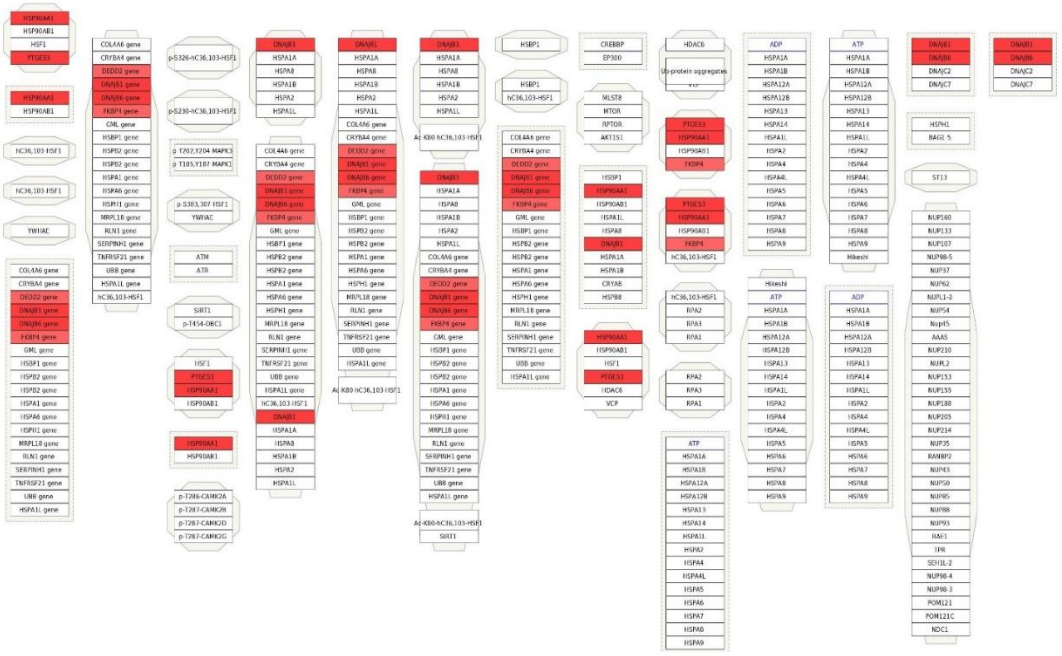

Figure 4.7. Protein complexes present in the cellular response to heat stress. Complexes having *DEDD2, DNAJB1, DNAJB2, DNAJB6, HSP90AA1, PTGES3,* and *FKBP4* were highlighted.

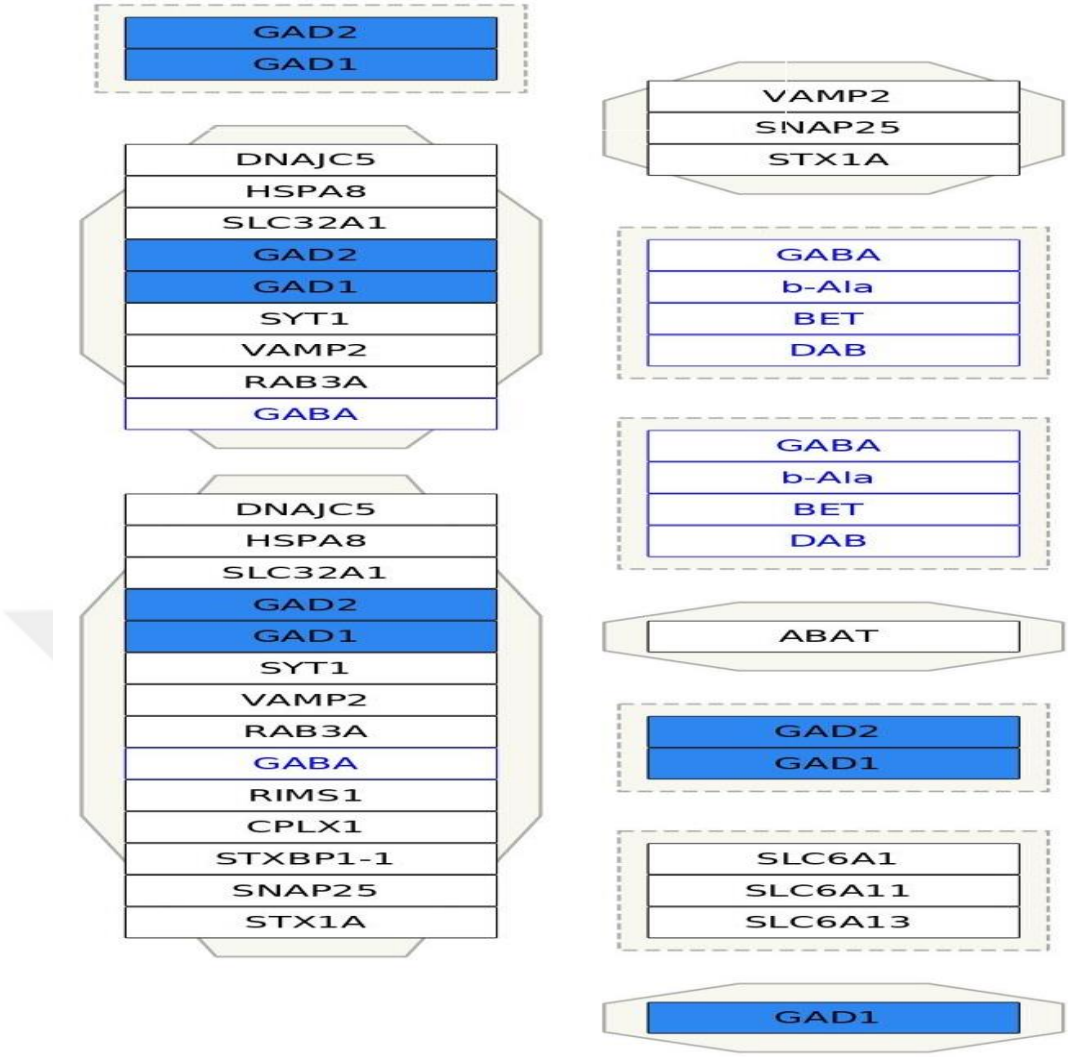

Figure 4.8. Protein complexes, dimers, and metabolites present in the pathways of GABA synthesis, release, reuptake and degradation.

Finally, the variants in all three diseases were compared and found no variants common to all three diseases. Hence in the particular case, instead of intersecting the three diseases' variants, the variant lists were merged among which 3500 variants ended up with the SIFT predictor is D (deleterious). I filtered 850 nonsynonymous, exonic variants whose Polyphen2 HVAR predictor is not B (benign) from this list. Then, the gene lists were combined against 200 DEGs. 36/3500 and 8/850 variant containing genes were also present in 200 genes, respectively. Interestingly, all detected variants of these genes had either highly low (< 0.01) or unknown frequencies overall considering 1000Genomes and ExAC. *GAD1*, *GAD2*, *HSP90AA1*, *HSPD1*, *HSPB1*, *BCL2L1*, *PPID* and *DNAJB6* are these 8 significantly differentially expressed

genes carrying mutations. These genes are primarily involved in cellular heat stress response and GABA synthesis pathways based on my earlier finding.

When we shift focus onto heat shock proteins, we encounter with other mutations on their encoding genes. 13 variants on Hsp70 cofactors from DNAJ subfamily (*DNAJB6* is a significant DEG. Others are: *DNAJA4, DNAJB2, DNAJB6, DNAJB9, DNAJC12, DNAJC13, DNAJC17, DNAJC2, DNAJC21, DNAJC4, DNAJC6*), 9 variants on Hsp70 core protein-encoding genes (*HSPD1* is significant DEG. Others: *HSPA14, HSPA1B, HSPA4, HSPA5, HSPA9*), 1 variant on a Hsp90 core protein-encoding gene (*HSP90AA1*, namely), 3 mutations on chaperonin genes (*CCT5, CCT6A, CCT8*), 1 mutation on Hsp90 gene localised in endoplasmic reticulum or Golgi (*HSP90B1*) and 1 mutation on a small HSP core protein-encoding gene (*HSPB21*) mutation on a Hsp70 cofactor (*BAG3*) were detected (Kampinga & Bergink, 2016).

5. DISCUSSION

In this thesis, four independent datasets belonging to three neurodegenerative disorder studies were analysed to explore DEGs and pathways shared among these disorders. Based on my meta-analysis and subsequent pathway analysis and variant analysis, it was revealed that the heat shock-related pathways are upregulated and GABA associated pathways are downregulated significantly in neurodegenerative diseases, supporting the previous researches.

Heat shock proteins (HSPs) are a large family of naturally heat durable proteins. Hsp70, Hsp90, chaperonins, co-chaperones and small HSPs are machinery complexes composed of various HSPs. They form complexes that provide proper folding and degradation of proteins. Therefore, they are crucial hub proteins in overall protein homeostasis (Kampinga & Bergink, 2016). In addition, they stabilize the cytoskeleton and prevent apoptosis (Alnemri, 2000; Schmitt et al., 2003). Regarding these and based on *in vivo* studies (Kaltenbach et al., 2007; Pemberton et al., 2011), the association of HSPs and neurodegenerative diseases seems obvious. This observation is supported by the correlation of a decline in the quantity of HSPs with increasing age, which can be given as evidence of HSPs effect on protein aggregations in NDDs (Balch, Morimoto, Dillin, & Kelly, 2008).

On the contrary, there are two reasons to suspect this association. First, protein aggregations in NDD usually occur outside of the cell, whereas the majority of HSPs stay inside the cell. As evidence of this, HSP70 injections into a mouse model with α-synucleinopathy showed worse pathophysiology (Shimshek, Mueller, Wiessner, Schweizer, & Herman van der Putten, 2010). In addition, the mechanism for the exocytosis of HSPs is unclear (Kampinga & Bergink, 2016). Second, protein aggregations in NDD, such as Aβ aggregate and Lewy bodies, do not form due to hydrophobic interactions which are the targets of canonical chaperones. These two reasons imply that protein aggregations outside of the cell cause the diseases.

On the other hand, these reasons have weak points. As mentioned in the case of AD, the effect of intercellular aggregations in pathogenesis is being evaluated as causative less day-by-day (Kametani & Hasegawa, 2018). Instead, Hsp70 and Hsp90 can bond with hydrophobic regions of intracellular tau protein, which is naturally unfolded, and they have been shown to counteract tau aggregation due to this feature (Kampinga & Bergink, 2016). α-synuclein in PD (Pemberton et al., 2011), polyglutamine proteins in Huntington disease (Hageman et al., 2010) and *SOD1* in ALS (Kakkar, Meister-Brookema, Minoia, Carra, & Kamping, 2014) have been demonstrated as targets of HSPs, mainly HSP70 and HSP90, basically owing to their similar features.

Besides to HSPs, meta-analysis verified the importance of another set of proteins. My analysis identified two glutamic acid dehydrogenases *GAD1* and *GAD2*, responsible for the synthesis of the GABA for synaptic release, which were both downregulated in all three of the diseases. This observation was supported by the discovery of variants in both GAD genes, specific to patients.

Gamma-aminobutyric acid (GABA) is the primary inhibitory neurotransmitter in the cortex (Petroff, 2002). Decarboxylation of glutamate by glutamate decarboxylases, i.e. *GAD1* and *GAD2*, is the crucial step in GABA synthesis (Bu et al., 1992). Unsurprisingly, a decrease in levels of *GAD1* and *GAD2* reduces GABA synthesis drastically, which in turn prevents the inhibition of postsynaptic GABAergic neurons (having GABA receptors). It can be speculated that this leads to continuous excitation of the postsynaptic neuron, or glutamate excitotoxicity, in other words.

Association of GABA receptors and NDD has been studied for a while. Increasing evidence has linked changes in GABA levels and *GAD* activity to mutant *HTT*, neuroinflammation and excitotoxicity in Huntington disease (Hsu, Chang, & Chern, 2018). Another study linked loss of GABA function to AD pathology based on *in vitro* experiments they made on *Xenopus* oocytes (Limon, Reyes-Ruiz, & Miledi, 2012). Another study brings this further, concludes that tau pathology may impair hippocampal GABAergic function in tauopathy model mice (Levenga et al., 2013).

To summarize, the associations of HSPs and GABA receptors with NDDs have been studied extensively. HSPs have crucial functions for regulating protein homeostasis

and neuroprotective features, while GABA synthase directly affects signalling mechanism in neurons, which may devastate neuronal health in the end.

The upregulation of HSPs indicates an increase in aggregating misfolded or unfolded proteins. There is no evidence to speculate this upregulation as a cause of neurodegeneration. Therefore, the question is why the upregulation of these neuroprotective elements is not sufficient to protect cells from neurodegeneration. The reason may be related to the progressive nature of the diseases. Stressors, such as ageing and head trauma, may reduce the resistance of neurons and glial cells as a whole by time. Cellular health, especially in sensitive cells like neurons, depends on the proper operation of multiple mechanisms. ROS must be catalysed to H_2O and O_2, toxins must be detoxified, impaired organelles and cellular structures must be lysed and re-synthesized. In addition, intracellular aggregations must be degraded. Although neurons respond against these stress factors by neuroprotective elements, e.g. HSPs, after a threshold this may not prevent the destruction of homeostasis, which ends in apoptosis. Furthermore, the intake of intercellular aggregations may accelerate this progress.

The downregulation of *GAD1* and *GAD2* decreases GABA synthesis, may result in the suppression of GABA receptors. The relation between GABA levels and excitotoxicity has been mentioned above. Reduced inhibition of neuron due to a decrease in GABA receptor activation matches increased excitation of the neuron. As evidence of this, Jose Luis Nieto-Gonzalez and colleagues found that reduced inhibition of GABA receptors causes hyperexcitability in a mouse model of ALS (Nieto-Gonzalez, Moser, Lauritzen, Schmitt-John, & Jensen, 2011). As introduced previously, overexcitation or excitotoxicity may cascade apoptotic progresses; it may trigger or enhance failure in calcium homeostasis, mitochondrial dysfunction, production of ROS (Prentice, Modi, & Wu, 2015).

Therefore, *GAD1*, *GAD2* and HSPs are good candidates for the treatment of neurodegenerative diseases. Heat shock proteins are recently in the radar for a potential target of neurodegenerative diseases. For instance, Colchicine, originally a medication to treat gout attacks, was recently repurposed as a potential vaccine for its function of enhancing the HSPs on ALS patients (Crippa et al., 2016). In addition to enhancing HSP activity, disaggregation of pre-existing aggregates, e.g. using bacterial and yeast-

derived HSPs in collaboration with HSP70 system (Doyle & Wickner, 2009) and increasing chaperone-mediated autophagy (Schneider & Cuervo, 2013; Xu et al., 2017) may be developed as treatments targeting HSPs (Kampinga & Bergink, 2016). We have a broader range of drugs for GABA targeted treatments, cause the GABA receptor has been studied extensively revealing that it has five distinct sites (Figure 5.1.). Some GABA antagonists such as bicuculline, picrotoxin (Yoshiike et al., 2008) and propofol (a common anesthetic) (Shao et al., 2014) showed ameliorating effects on mice models.

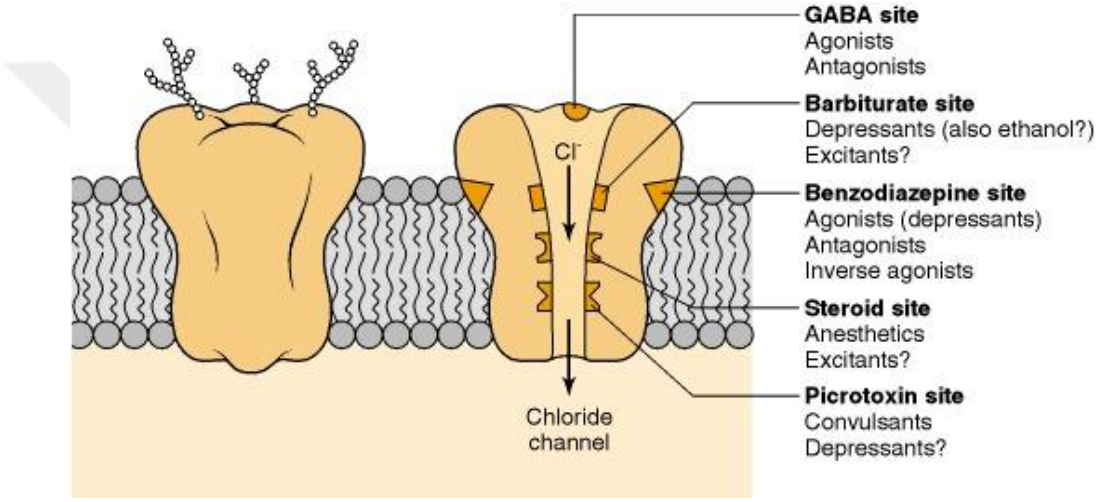

Figure 5.1. Drug binding sites on GABA receptors (Olsen & DeLorey, 1999)

Furthermore, an ongoing clinical trial is testing whether the administration of flumazenil improves the motor function for the Parkinson disease patients with low GABA receptor availability (https://clinicaltrials.gov/ct2/show/NCT03462641). Hence, this thesis further supports the significance of GABA synthesis pathway and related proteins in neurodegenerative diseases.

6. CONCLUSION

This study supported the association of heat shock proteins with neurodegenerative disorders and demonstrated the importance of GABA synthesizing proteins in them, or at least AD, PD and ALS. Several treatments targeting these proteins have been proposed, and others have possibility of re-purposed for the treatment of neurodegenerative diseases considering that they, i.e. *DNAJB1* and *GAD* genes, are novel findings. Focusing on these targets, specially GAD genes regarding that they place in upstream of GABA synthesis whose disruption may cascade neurodegeneration via excitotoxicity, may be more beneficial for our struggle against NDDs.

Furthermore, a recent study showed the possibility of detecting somatic mutations in normal tissue using RNA-Seq (Yizhak et al., n.d.). Considering the emergent importance of somatic mutations in neurological disorders (Poduri, Evrony, Cai, & Walsh, 2013), the set of mutations found in the thesis can be filtered further to reveal somatic mutations in a future work, and verify the significance of their mutational load.

REFERENCES

Alnemri, E. S. (2000). Negative regulation of the APAF-1 apoptosome by HSP70. *Article in Nature Cell Biology*.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Baba, M., Nakajo, S., Tu, P. H., Tomita, T., Nakaya, K., Lee, V. M., … Iwatsubo, T. (1998). Aggregation of alpha-synuclein in Lewy bodies of sporadic Parkinson's disease and dementia with Lewy bodies. *The American Journal of Pathology*, *152*(4), 879–884.

Badiola, N., de Oliveira, R. M., Herrera, F., Guardia-Laguarta, C., Gonçalves, S. A., Pera, M., … Lleó, A. (2011). Tau enhances α-synuclein aggregation and toxicity in cellular models of synucleinopathy. *PLoS ONE*, *6*(10), 1–9.

Balch, W. E., Morimoto, R. I., Dillin, A., & Kelly, J. W. (2008). Adapting Proteostasis for Disease Intervention. *Science*, *319*(916).

Barber, S. C., Mead, R. J., & Shaw, P. J. (2006). Oxidative stress in ALS: A mechanism of neurodegeneration and a therapeutic target. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1762*(11–12), 1051–1067.

Baruzzo, G., Hayer, K. E., Kim, E. J., Camillo, B. Di, Fitzgerald, G. A., & Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners HHS Public Access. *Nat Methods*, *14*(2), 135–139.

Bateman, R. J., Xiong, C., Benzinger, T. L. S., Fagan, A. M., Goate, A., Fox, N. C., Morris, J. C. (2012). Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease. *New England Journal of Medicine*, *367*(9), 795–804.

Beard, J. D., Engel, L. S., Richardson, D. B., Gammon, M. D., Baird, C., Umbach, D. M., Kamel, F. (2016). Military service, deployments, and exposures in relation to amyotrophic lateral sclerosis etiology. *Environment International*, *91*, 104–115.

Beyer, K., & Ariza, A. (2007). Protein aggregation mechanisms in synucleinopathies: Commonalities and differences. *Journal of Neuropathology and Experimental Neurology*, *66*(11), 965–974.

Binder, L. I., Guillozet-Bongaarts, A. L., Garcia-Sierra, F., & Berry, R. W. (2005). Tau, tangles, and Alzheimer's disease. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1739*(2), 216–223.

Bloom, G. S. (2014). Amyloid-β and Tau. *JAMA Neurology*, *71*(4), 505.

Borenstein, M., Hedges, L. V, Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*.

Borjabad, A., & Volsky, D. J. (2012). Common transcriptional signatures in brain tissue from patients with HIV-associated neurocognitive disorders, Alzheimer's disease, and Multiple Sclerosis. *Journal of Neuroimmune Pharmacology : The Official Journal of the Society on NeuroImmune Pharmacology*, *7*(4), 914–926.

Braak, H., Braak, E., Yilmazer, D., De Vos, R. A. I., Jansen, E. N. H., & Bohl, J. (1996). Pattern of brain destruction in Parkinson's and Alzheimer's diseases. *Journal of Neural Transmission*, *103*(4), 455–490.

Braak, Heiko, & Braak, E. (2008). Alzheimer's disease affects limbic nuclei of the thalamus. *Acta Neuropathologica*, *115*(2), 269–271.

Braak, Heiko, Braak, E., Yilmazer, D., de Vos, R. A. I., Jansen, E. N. H., Bohl, J., & Jellinger, K. (1994). Amygdala pathology in Parkinson's disease. *Acta Neuropathologica*, *88*(6), 493–500.

Braak, Heiko, Ghebremedhin, E., Rüb, U., Bratzke, H., & Del Tredici, K. (2004). Stages in the development of Parkinson's disease-related pathology. *Cell and Tissue Research*, *318*(1), 121–134.

Bu, D. F., Erlander, M. G., Hitz, B. C., Tillakaratne, N. J., Kaufman, D. L., Wagner-McPherson, C. B., … Tobin, A. J. (1992). Two human glutamate decarboxylases, 65-kDa GAD and 67-kDa GAD, are each encoded by a single gene. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(6), 2115–2119.

Bushnell, B. (2015). BBMap (version 37.75) [Software].

Castellano, J. M., Kim, J., Stewart, F. R., Jiang, H., DeMattos, R. B., Patterson, B. W., Holtzman, D. M. (2011). Human apoE isoforms differentially regulate brain amyloid-β peptide clearance. *Science Translational Medicine*, *3*(89).

Chang, L.-C., Lin, H.-M., Sibille, E., & Tseng, G. C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*, *14*(1), 368.

Chang, X. L., Mao, X. Y., Li, H. H., Zhang, J. H., Li, N. N., Burgunder, J. M., Tan, E. K. (2011). Association of GWAS loci with PD in China. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, *156*(3), 334–339.

Chen, S., Sayana, P., Zhang, X., & Le, W. (2013). Genetics of amyotrophic lateral sclerosis: An update. *Molecular Neurodegeneration*, *8*(1), 1.

Chio, A., Benzi, G., Dossena, M., Mutani, R., & Mora, G. (2005). Severely increased risk of amyotrophic lateral sclerosis among Italian professional football players. *Brain*, *128*(3), 472–476.

Chung, T., Park, J. S., Kim, S., Montes, N., Walston, J., & Höke, A. (2017). Evidence for dying-back axonal degeneration in age-associated skeletal muscle decline. *Muscle & Nerve*, *55*(6), 894–901.

Ciervo, Y., Ning, K., Jun, X., Shaw, P. J., & Mead, R. J. (2017). Advances, challenges and future directions for stem cell therapy in amyotrophic lateral sclerosis. *Molecular Neurodegeneration*, *12*(1), 85.

Cleveland, D. W., Hwo, S. Y., & Kirschner, M. W. (1977). Purification of tau, a microtubule-associated protein that induces assembly of microtubules from purified tubulin. *Journal of Molecular Biology*.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771.

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, *12*(12), e0190152.

Couchie, D., Mavilia, C., Georgieff, I. S., Liem, R. K., Shelanski, M. L., & Nunez, J. (2006). Primary structure of high molecular weight tau present in the peripheral nervous system. *Proceedings of the National Academy of Sciences*.

Crippa, V., D'Agostino, V. G., Cristofani, R., Rusmini, P., Cicardi, M. E., Messi, E., Poletti, A. (2016). Transcriptional induction of the heat shock protein B8 mediates the clearance of misfolded proteins responsible for motor neuron diseases. *Scientific Reports*, *6*(February), 1–16.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, *39*(Database), D691–D697.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*.

Daves, M. H., Hilsenbeck, S. G., Lau, C. C., & Man, T. K. (2011). Meta-analysis of multiple microarray datasets reveals a common gene signature of metastasis in solid tumors. *BMC Medical Genomics*, *4*(1), 56.

De Calignon, A., Polydoro, M., Suárez-Calvet, M., William, C., Adamowicz, D. H., Kopeikina, K. J., Hyman, B. T. (2012). Propagation of Tau Pathology in a Model of Early Alzheimer's Disease. *Neuron*, *73*(4), 685–697.

Del Tredici, K., Rüb, U., De Vos, R. A. I., Bohl, J. R. E., & Braak, H. (2002). Where does Parkinson disease pathology begin in the brain? *Journal of Neuropathology and Experimental Neurology*, *61*(5), 413–426.

Difiglia, M., Sapp, E., Chase, K. O., Davies, S. W., Gillian, P., Vonsattel, J. P., Aronin, N. (2016). Aggregation of Huntingtin in Neuronal Intranuclear Inclusions and Dystrophic Neurites in Brain Published by : American Association for the Advancement of Science Stable, *277*(5334), 1990–1993.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21.

Doecke, J. D., Laws, S. M., Faux, N. G., Wilson, W., Burnham, S. C., Lam, C. P., Martins, R. N. (2012). Blood-based protein biomarkers for diagnosis of Alzheimer disease. *Archives of Neurology*, *69*(10), 1318–1325.

Doody, R. S., Raman, R., Farlow, M., Iwatsubo, T., Vellas, B., Joffe, S., Mohs, R. (2013). A Phase 3 Trial of Semagacestat for Treatment of Alzheimer's Disease. *New England Journal of Medicine*.

Doyle, S. M., & Wickner, S. (2009). Hsp104 and ClpB: protein disaggregating machines. *Trends in Biochemical Sciences*, *29*.

Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, L., Blennow, K., Cummings, L. (2014). Position Paper Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Www.Thelancet.Com/Neurology*, *13*(June).

Dumitriu, A., Golji, J., Labadorf, A. T., Gao, B., Beach, T. G., Myers, R. H., Latourelle, J. C. (2016). Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease. *BMC Medical Genomics*, *9*(1), 1–17.

Ebrahimi-Fakhari, D., Wahlster, L., & McLean, P. J. (2011). Molecular chaperones in Parkinson's disease - Present and future. *Journal of Parkinson's Disease*, *1*(4), 299–320.

Edwards, T. L., Scott, W. K., Almonte, C., Burt, A., Powell, E. H., Beecham, G. W., Foundation, T. M. (2010). Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease, *74*(2), 97–109.

Elobeid, A., Libard, S., Leino, M., Popova, S. N., & Alafuzoff, I. (2016). Altered proteins in the aging brain. *Journal of Neuropathology and Experimental Neurology*, *75*(4), 316–325.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *46*(D1), D649–D655.

Fargo, K., & Bleiler, L. (2014). Alzheimer's Association Report: 2014 Alzheimers disease facts and figures. *Alzheimer's and Dementia*, *10*(2), e47–e92.

Feigin, V. L., Abajobir, A. A., Abate, K. H., Abd-Allah, F., Abdulle, A. M., Abera, S. F., Vos, T. (2017). Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Neurology*, *16*(11), 877–897.

Ferrer, I. (2011). Neuropathology and Neurochemistry of Nonmotor Symptoms in Parkinson's Disease. *Parkinson's Disease*, *2011*, 1–13.

Friedman, B. A., Srinivasan, K., Ayalon, G., Meilandt, W. J., Lin, H., Huntley, M. A., Hansen, D. V. (2018). Diverse Brain Myeloid Expression Profiles Reveal Distinct Microglial Activation States and Aspects of Alzheimer's Disease Not Evident in Mouse Models. *Cell Reports*.

Gardiner, P., Dai, Y., & Heckman, C. J. (2006). Effects of exercise training on α-motoneurons. *Journal of Applied Physiology*, *101*(4), 1228–1236.

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, *32*(90001), 258D – 261.

Goedert, M., Clavaguera, F., & Tolnay, M. (2010). The propagation of prion-like protein inclusions in neurodegenerative diseases. *Trends in Neurosciences*, *33*(7), 317–325.

Goedert, M., Spillantini, M. G., Del Tredici, K., & Braak, H. (2013). 100 years of Lewy pathology. *Nature Reviews Neurology*, *9*(1), 13–24.

Gonzales, J. C., Gordts, P. L. S. M., Foley, E. M., & Esko, J. D. (2013). Apolipoproteins e and AV mediate lipoprotein clearance by hepatic proteoglycans. *Journal of Clinical Investigation*, *123*(6), 2742–2751.

Hageman, J., Rujano, M. A., van Waarde, M. A. W. H., Kakkar, V., Dirks, R. P., Govorukhina, N., Kampinga, H. H. (2010). A DNAJB Chaperone Subfamily with HDAC-Dependent Activities Suppresses Toxic Protein Aggregation. *Molecular Cell*, *37*(3), 355–369.

Hardiman, O., Benatar, M., Brooks, B. R., Chio, A., de Carvalho, M., Ince, P. G., Martin R Turner, Orla Hardiman, Michael Benatar, Benjamin R Brooks, Adriano Chio, Mamede de Carvalho, Paul G Ince, Cindy Lin, Robert G Miller, Hiroshi Mitsumoto. (2013). Controversies and priorities in amyotrophic lateral sclerosis. *The Lancet. Neurology*, *12*(3), 310–322.

Harrison, P. W., Alako, B., Amid, C., Cerdeño-Tárraga, A., Cleland, I., Holt, S., Cochrane, G. (2019). The European Nucleotide Archive in 2018. *Nucleic Acids Research*, *47*(D1), D84–D88.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Hubbard, T. J. (n.d.). GENCODE: The reference human genome annotation for The ENCODE Project.

Hong, F., & Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, *24*(3), 374–382.

Hongshan Jiang, Rong Lei, Shou-Wei Ding, & Shuifang Zhu. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, *15*, 1–12.

Hsu, Y.-T., Chang, Y.-G., & Chern, Y. (2018). Insights into GABAAergic system alteration in Huntington's disease. *Open Biology*, *8*(12).

Hung, A. S. M., Liang, Y., Chow, T. C. H., Tang, H. C., Wu, S. L. Y., Wai, M. S. M., & Yew, D. T. (2016). Mutated tau, amyloid and neuroinflammation in Alzheimer disease—A brief review. *Progress in Histochemistry and Cytochemistry*, *51*(1), 1–8.

Illumina. (2011). *Quality Scores for Next-Generation Sequencing*.

Jankovic, J. (2008). Parkinson's disease: Clinical features and diagnosis. *Journal of Neurology, Neurosurgery and Psychiatry*, *79*(4), 368–376.

Javed, K., & Daly, D. T. (2019). *Neuroanatomy, Lower Motor Neuron Lesion*. *StatPearls*. StatPearls Publishing. Retrieved from

Kakkar, V., Meister-Brookema, M., Minoia, M., Carra, S., & Kamping, H. H. (2014). Barcoding heat shock proteins to human diseases: looking beyond the heat shock response.

Kalia, S. K., Lee, S., Smith, P. D., Liu, L., Crocker, S. J., Thorarinsdottir, T. E., Lozano, A. M. (2004). BAG5 inhibits parkin and enhances dopaminergic neuron degeneration. *Neuron*, *44*(6), 931–945.

Kaltenbach, L. S., Romero, E., Becklin, R. R., Chettier, R., Bell, R., Phansalkar, A., Hughes, R. E. (2007). Huntingtin Interacting Proteins Are Genetic Modifiers of Neurodegeneration. *PLoS Genetics*, *3*(5), e82.

Kametani, F., & Hasegawa, M. (2018). Reconsideration of Amyloid Hypothesis and Tau Hypothesis in Alzheimer's Disease. *Frontiers in Neuroscience*, *12*, 25.

Kampinga, H. H., & Bergink, S. (2016). Heat shock proteins as potential targets for protective strategies in neurodegeneration. *The Lancet Neurology*, *15*(7), 748–759.

Karran, E., & Hardy, J. (2014). A critique of the drug discovery and phase 3 clinical programs targeting the amyloid hypothesis for Alzheimer disease. *Annals of Neurology*, *76*(2), 185–205.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, *8*(2).

Kiernan, M. C., Vucic, S., Cheah, B. C., Turner, M. R., Eisen, A., Hardiman, O., Zoing, M. C. (2011). Amyotrophic lateral sclerosis. *The Lancet*, *377*(9769), 942–955.

Kim, J., Chakrabarty, P., Hanna, A., March, A., Dickson, D. W., Borchelt, D. R., Janus, C. (2013). Normal cognition in transgenic BRI2-Aβ mice. *Molecular Neurodegeneration*, *8*(1), 15.

Kolarova, M., García-Sierra, F., Bartos, A., Ricny, J., & Ripova, D. (2012). Structure and Pathology of Tau Protein in Alzheimer Disease. *International Journal of Alzheimer's Disease*, *2012*, 1–13.

Kramer, M. L., & Schulz-Schaeffer, W. J. (2007). Presynaptic -Synuclein Aggregates, Not Lewy Bodies, Cause Neurodegeneration in Dementia with Lewy Bodies. *Journal of Neuroscience*, *27*(6), 1405–1410.

Krishnaswamy, S., Verdile, G., Groth, D., Kanyenda, L., & Martins, R. N. (2009). The structure and function of Alzheimer's gamma secretase enzyme complex. *Critical Reviews in Clinical Laboratory Sciences*.

Kritsilis, M., Rizou, S. V., Koutsoudaki, P. N., Evangelou, K., Gorgoulis, V. G., &

Papadopoulos, D. (2018). Ageing, cellular senescence and neurodegenerative disease. *International Journal of Molecular Sciences*, *19*(10).

Kutmon, M., Lotia, S., Evelo, C. T., & Pico, A. R. (2014). WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization. *F1000Research*.

Kuusisto, E., Parkkinen, L., & Alafuzoff, I. (2003). Morphogenesis of Lewy Bodies: Dissimilar Incorporation of α-Synuclein, Ubiquitin, and p62. *Journal of Neuropathology and Experimental Neurology*, *62*(12), 1241–1253.

Laferla, F. M. (2010). The Biology of Tau and its Role in Tauopathies Pathways linking Aβ and tau pathologies. *Biochemical Society Transactions*.

Landles, C., & Bates, G. P. (2004). Huntingtin and the molecular pathogenesis of Huntington's disease. Fourth in molecular medicine review series. *EMBO Reports*, *5*(10), 958–963.

Langston, J. W., Ballard, P., Tetrud, J. W., & Irwin, I. (1983). Chronic Parkinsonism in humans due to a product of meperidine-analog synthesis. *Science (New York, N.Y.)*, *219*(4587), 979–980.

Lazarov, O., & Hollands, C. (2016). Hippocampal neurogenesis: learning to remember. *Prog Neurobiol*, *138140*, 1–18.

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(SUPPL. 1), 2010–2012.

Levenga, J., Krishnamurthy, P., Rajamohamedsait, H., Wong, H., Franke, T. F., Cain, P., Hoeffer, C. A. (2013). Tau pathology induces loss of GABAergic interneurons leading to altered synaptic plasticity and behavioral impairments. *Acta Neuropathologica Communications*, *1*(1), 34.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, M. D., Burns, T. C., Morgan, A. A., & Khatri, P. (2014). Integrated multi-cohort transcriptional meta-analysis of neurodegenerative diseases. *Acta Neuropathologica Communications*, *2*(1), 93.

Li, S., & Li, X. J. (2006). Multiple pathways contribute to the pathogenesis of Huntington disease. *Molecular Neurodegeneration*, *1*(1), 1–10.

Li, W. (2016). No Title.

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930.

Limon, A., Reyes-Ruiz, J. M., & Miledi, R. (2012). Loss of functional GABAA receptors in the Alzheimer diseased brain. *Proceedings of the National Academy of Sciences*, *109*(25), 10071–10076.

Liu, T., Perry, G., Chan, H. W., Verdile, G., Martins, R. N., Smith, M. A., & Atwood, C. S. (2004). Amyloid-b-induced toxicity of primary neurons is dependent upon differentiation-associated increases in tau and cyclin-dependent kinase 5 expression. *J. Neurochem*, *88*, 554–563.

Liu, Y. F. (1998). Expression of Polyglutamine-expanded Huntingtin Activates the SEK1-JNK Pathway and Induces Apoptosis in a Expression of Polyglutamine-expanded Huntingtin Activates the SEK1-JNK Pathway and Induces Apoptosis in a Hippocampal N. *Cell Biology and Metabolism*, *273*(44), 28873-.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*.

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*.

McCrate, M. E., & Kaspar, B. K. (2008). Physical Activity and Neuroprotection in Amyotrophic Lateral Sclerosis. *NeuroMolecular Medicine*, *10*(2), 108–117.

Nieto-Gonzalez, J. L., Moser, J., Lauritzen, M., Schmitt-John, T., & Jensen, K. (2011). Reduced GABAergic Inhibition Explains Cortical Hyperexcitability in the Wobbler Mouse Model of ALS. *Cerebral Cortex*, *21*(3), 625–635.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745.

Olsen, R. W., & DeLorey, T. M. (1999). GABA Receptor Physiology and Pharmacology.

Opazo, F., Krenz, A., Heermann, S., Schulz, J. B., & Falkenburger, B. H. (2008). Accumulation and clearance of α-synuclein aggregates demonstrated by time-lapse imaging. *Journal of Neurochemistry*, *106*(2), 529–540.

Organization, W. H. (2014). WHO | Noncommunicable diseases country profiles 2014. *Who*.

Ostrowitzki, S., Deptula, D., Thurfjell, L., Barkhof, F., Bohrmann, B., Brooks, D. J., Santarelli, L. (2012). Mechanism of Amyloid Removal in Patients With Alzheimer Disease Treated With Gantenerumab. *Archives of Neurology*, *69*(2), 198.

Parkinson, J. (2002). An essay on the shaking palsy. *Journal of Neuropsychiatry and Clinical Neurosciences*, *14*(2), 899–900.

Patel, H., Dobson, R. J. B., Newhouse, S. J., Dobson, P. R. J. B., Informatics, H., & London, S. (2018). A Meta-Analysis of Alzheimer ' s Disease Brain Transcriptomic Data.

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, *85*(8), 2444–

2448.

Pemberton, S., Madiona, K., Pieri, L., Kabani, M., Bousset, L., & Melki, R. (2011). Hsc70 protein interaction with soluble and fibrillar α-synuclein. *Journal of Biological Chemistry*, *286*(40), 34690–34699.

Peng, L., Bian, X. W., Li, D. K., Xu, C., Wang, G. M., Xia, Q. Y., & Xiong, Q. (2015). Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. *Scientific Reports*, *5*(August), 1–18.

Petroff, O. A. (2002). GABA and Glutamate in the Human Brain. *The Neuroscientist*.

Phan, J. H., Young, A. N., & Wang, M. D. (2012). Robust Microarray Meta-Analysis Identifies Differentially Expressed Genes for Clinical Prediction. *The Scientific World Journal*, *2012*, 1–9.

Pico, A., Hanspers, K., Willighagen, E., Fehrhart, & Kutmon, M. (n.d.). Parkinsons Disease Pathway (Homo sapiens) - WikiPathways. Retrieved May 31, 2019, from

Poduri, A., Evrony, G. D., Cai, X., & Walsh, C. A. (2013). Somatic Mutation, Genomic Variation, and Neurological Disease. *Science (New York, N.Y.)*, *341*(6141), 1237758.

Prentice, H., Modi, J. P., & Wu, J.-Y. (2015). Mechanisms of Neuronal Protection against Excitotoxicity, Endoplasmic Reticulum Stress, and Mitochondrial Dysfunction in Stroke and Neurodegenerative Diseases. *Oxidative Medicine and Cellular Longevity*, *2015*, 1–7.

Prudencio, M., Belzil, V. V., Batra, R., Ross, C. A., Gendron, T. F., Pregent, L. J., … Petrucelli, L. (2015). Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. *Nature Neuroscience*.

Rau, A., Marot, G., & Jaffrézic, F. (2014). Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics*, *15*(1), 1–10.

Reactome Team, Bohler, A., & Kutmon, M. (n.d.). GABA synthesis, release, reuptake and degradation (Homo sapiens) - WikiPathways.

Reactome Team, & Kutmon, M. (n.d.). Cellular response to heat stress (Homo sapiens) - WikiPathways. Retrieved May 31, 2019,

Recasens, A., Dehay, B., Bové, J., Carballo-Carbajal, I., Dovero, S., Pérez-Villalba, A., Vila, M. (2014). Lewy body extracts from Parkinson disease brains trigger α-synuclein pathology and neurodegeneration in mice and monkeys. *Annals of Neurology*, *75*(3), 351–362.

Reeve, A., Simcox, E., & Turnbull, D. (2014). Ageing and Parkinson's disease: Why is advancing age the biggest risk factor? *Ageing Research Reviews*, *14*(1), 19–30.

Refolo, L. M., Pappolla, M. A., Malester, B., LaFrancois, J., Bryant-Thomas, T., Wang, R., Duff, K. (2000). Hypercholesterolemia accelerates the Alzheimer's amyloid pathology in a transgenic mouse model. *Neurobiology of Disease*, *7*(4),

321–331.

Reiman, E. M., Quiroz, Y. T., Fleisher, A. S., Chen, K., Velez-Pardo, C., Jimenez-Del-Rio, M., Lopera, F. (2012). Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: A case-control study. *The Lancet Neurology*, *11*(12), 1048–1056.

Rhodes, S. L., Sinsheimer, J. S., Bordelon, Y., Bronstein, J. M., & Ritz, B. (2011). Replication of GWAS Associations for GAK and MAPT in Parkinson's Disease. *Annals of Human Genetics*, *75*(2), 195–200.

Riutta, A., Hanspers, K., Willighagen, E., Pico, A., Kutmon, M., Peryea, T., … Coort, S. (n.d.). Amyotrophic lateral sclerosis (ALS) (Homo sapiens) - WikiPathways. Retrieved May 31, 2019,

Rose, J. M., Novoselov, S. S., Robinson, P. A., & Cheetham, M. E. (2011). Molecular chaperone-mediated rescue of mitophagy by a Parkin RING1 domain mutant. *Human Molecular Genetics*, *20*(1), 16–27.

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerrière, A., Vital, A., Campion, D. (2006). APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genetics*, *38*(1), 24–26.

Roy, B., & Jackson, G. R. (2014). Interactions between tau and α-synuclein augment neurotoxicity in a Drosophila model of parkinson's disease. *Human Molecular Genetics*, *23*(11), 3008–3023.

Sai, Y., Zou, Z., Peng, K., & Dong, Z. (2012). The Parkinson's disease-related genes act in mitochondrial homeostasis. *Neuroscience & Biobehavioral Reviews*, *36*(9), 2034–2043.

Salomonis, N., Hanspers, K., Willighagen, E., Fehrhart, Thakur, L., Pico, A., Coort, S. (n.d.). Alzheimers Disease (Homo sapiens) - WikiPathways. Retrieved May 31, 2019,

Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Toda, T. (2009). Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nature Genetics*, *41*(12), 1303–1307.

Schmidt, S., Kwee, L. C., Allen, K. D., & Oddone, E. Z. (2010). Association of ALS with head injury, cigarette smoking and APOE genotypes. *Journal of the Neurological Sciences*, *291*(1–2), 22–29.

Schmitt, E., Daugas, E., Gurbuxani, S., Cande, C., Parcellier, A., Hammann, A., Garrido, C. (2003). Heat shock protein 70 binding inhibits the nuclear import of apoptosis-inducing factor Immuno-Oncology View project Lipotox View project Heat shock protein 70 binding inhibits the nuclear import of apoptosis-inducing factor. *Oncogene*, *22*, 6669–6678.

Schneeberger, A., Mandler, M., Otava, O., Zauner, W., Mattner, F., & Schmidt, W. (2009). Second Generation Ad Vaccines Development of Affitope Vaccines for

Alzheimer ' S Disease (Ad) – From Concept To Clinical Testing. *Health (San Francisco)*, *13*(3), 264–267.

Schneider, J. L., & Cuervo, A. M. (2013). Chaperone-mediated autophagy: dedicated saviour and unfortunate victim in the neurodegeneration arena. *Biochemical Society Transactions*, *41*(6), 1483–1488.

Schulz-Schaeffer, W. J. (2010). The synaptic pathology of α-synuclein aggregation in dementia with Lewy bodies, Parkinson's disease and Parkinson's disease dementia. *Acta Neuropathologica*, *120*(2), 131–143.

Schulz-Schaeffer, W. J. (2015). Is cell death primary or secondary in the pathophysiology of idiopathic Parkinson's disease? *Biomolecules*, *5*(3), 1467–1479.

Schurch, N. J., Schofield, P., Gierlin´ski, M., Gierlin´ski, G., Cole, C., Sherstnev, A., Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

Seals, R. M., Kioumourtzoglou, M.-A., Gredal, O., Hansen, J., & Weisskopf, M. G. (2016). ALS and the Military: A Population-Based Study in the Danish Registries. *Epidemiology (Cambridge, Mass.)*, *27*(2), 188.

Selkoe, D. J., Hardy, J., Selkoe, D., & Hardy, J. (2016). The amyloid hypothesis of Alzheimer ' s disease at 25 years. *EMBO Molecular Medicine*.

Seward, M. E., Swanson, E., Norambuena, A., Reimann, A., Nicholas Cochran, J., Li, R., Bloom, G. S. (2012). Amyloid-b signals through tau to drive ectopic neuronal cell cycle re-entry in Alzheimer's disease. *Journal of Cell Science*, *126*, 1278–1286.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., 大森敏行. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(22), 6.

Shao, H., Zhang, Y., Dong, Y., Yu, B., Xia, W., & Xie, Z. (2014). Chronic Treatment with Anesthetic Propofol Improves Cognitive Function and Attenuates Caspase Activation in Both Aged and Alzheimer's Disease Transgenic Mice. *Journal of Alzheimer's Disease : JAD*, *41*(2), 499.

Shimshek, D. R., Mueller, M., Wiessner, C., Schweizer, T., & Herman van der Putten, P. (2010). The HSP70 molecular chaperone is not beneficial in a mouse model of α-synucleinopathy. *PLoS ONE*.

Simón-Sánchez, J., Schulte, C., Bras, J. J. M., Sharma, M., Gibbs, J. R., Berg, D., Bras, J. J. M. (2009). Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Genetics*, *41*(12), 1308–1312.

SRA-Tools - NCBI. (n.d.). Retrieved May 28, 2019,

Steinberg, S., Stefansson, H., Jonsson, T., Johannsdottir, H., Ingason, A., Helgason, H., Stefansson, K. (2015). Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nature Genetics*, *47*(5), 445–447.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-vance, M., Enghild, J., Salvesen, G. S., Bryan Alzheimer, K. (1993). *Apolipoprotein E: High-avidity binding to ,B-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc. Natl. Acad. Sci. USA* (Vol. 90). Retrieved from

Sweeney, T. E., Haynes, W. A., Vallania, F., Ioannidis, J. P., & Khatri, P. (2017). Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Research*, *45*(1), 1–14.

Takahashi, M., Miyata, H., Kametani, F., Nonaka, T., Akiyama, H., Hisanaga, S., & Hasegawa, M. (2015). Extracellular association of APP and tau fibrils induces intracellular aggregate formation of tau. *Acta Neuropathologica*, *129*(6), 895–907.

Teng, M., Love, M. I., Davis, C. A., Djebali, S., Dobin, A., Graveley, B. R., Irizarry, R. A. (2016). A benchmark for RNA-seq quantification pipelines.

Tseng, G. C., Ghosh, D., & Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, *40*(9), 3785–3799.

Wakabayashi, K., Matsumoto, K., Takayama, K., Yoshimoto, M., & Takahashi, H. (1997). NACP, a presynaptic protein, immunoreactivity in lewy bodies in Parkinson's disease. *Neuroscience Letters*, *239*(1), 45–48.

Wakabayashi, K., Tanji, K., Mori, F., & Takahashi, H. (2007). The Lewy body in Parkinson's disease: Molecules implicated in the formation and degradation of α-synuclein aggregates. *Neuropathology*, *27*(5), 494–506.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*.

Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*.

Wood, L. B., Winslow, A. R., Strasser, S. D., Wood, L., & Israel, B. (2015). Systems Biology of Neurodegenerative Diseases. *Integr Biol (Camb Integr Biol (Camb)*, *7*(7), 758–775.

Wu, G., & Haw, R. (2017). Functional Interaction Network Construction and Analysis for Disease Discovery (pp. 235–253). Humana Press, New York, NY.

Wu, T., & Hallett, M. (2013). The cerebellum in Parkinson's disease. *Brain*, *136*(3), 696–709.

Xu, C.-Y., Kang, W.-Y., Chen, Y.-M., Jiang, T.-F., Zhang, J., Zhang, L.-N., Chen, S.-D. (2017). DJ-1 Inhibits α-Synuclein Aggregation by Regulating Chaperone-Mediated Autophagy. *Frontiers in Aging Neuroscience*, *9*, 308.

Yaron, A., & Schuldiner, O. (2016). Common and Divergent Mechanisms in Developmental Neuronal Remodeling and Dying Back Neurodegeneration. *Current Biology*, *26*(13), R628–R639.

Yizhak, K., Aguet, F., Kim, J., Hess, J. M., Kübler, K., Grimsby, J., Getz, G. (n.d.). RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues.

Yoshiike, Y., Kimura, T., Yamashita, S., Furudate, H., Mizoroki, T., Murayama, M., & Takashima, A. (2008). GABAA Receptor-Mediated Acceleration of Aging-Associated Memory Decline in APP/PS1 Mice and Its Pharmacological Treatment by Picrotoxin. *PLoS ONE*, *3*(8), e3029.

CURRICULUM VITAE

ABDULAHAD BAYRAKTAR

GSM: +905058745693
E-mail: abdulahadbayraktar@gmail.com

EDUCATION

Master of Science September 2016 – June 2019
Muğla Sıtkı Koçman University
Department of Bioinformatics
CGPA: 3.81/4.00
Thesis on Integrative Analysis of Neurodegenerative Disease Omics Data

Bachelor of Science September 2010 – 20.06.2016
Middle East Technical University
Department of Molecular Biology and Genetics

SKILLS

Bioinformatics Skills

Programming Languages:	R, Bash, Python, Perl, MySQL, C++, MATLAB
Variant Discovery:	GATK, ANNOVAR, BWA, Picard, FastQC, rnaseqmut
RNA-Seq Analysis:	DESeq2, STAR, BWA, StringTie, Cufflinks2, TopHat2, Flexbar
Metagenomics:	R packages (DADA2, DESeq2, vegan, phyloseq etc.), QIIME2, Metaphlan2
Network Visualization:	Cytoscape
Modelling of Molecular Structure:	Pymol, Chimera, MOE
Sequence/Alignment Search Tool:	BLAST
Version Management:	Bitbucket, GitHub
Statistics Tool:	R, Minitab, SPSS
Operating Systems:	Windows, Linux Ubuntu
Microsoft Office:	Word, Excel, PowerPoint, Visio
Text Editing:	Notepad++, LaTex

Laboratory Skills

Electrophoresis, PCR, Microbiological techniques, Western blotting, Light Microscopy

EXPERIENCE

Master's Thesis

Supervisor: Omur Baysal, PhD, Professor
Co-advisor: Tugba Onal Suzek, PhD, Assistant Professor

Doing differential expression analyses on publicly available RNA-seq data – belonging to Alzheimer's disease, Parkinson's disease and ALS studies – and meta analyses on gene lists in order to find out significant gene expression patterns among and specific to these diseases versus healthy status. Doing variant analyses in order to understand how mutations affect gene expressions and whether they support previous results. Doing pathway analyses in order to find out affected biological processes based on verified DEGs. I am using STAR, rnaseqmut, R (packages: DESeq2, vegan) and Cytoscape and write Bash scripts for analyses.

Bioinformatician/Biologist February 2018 – Present day

Regenerative and Restorative Medicine Research Center, Istanbul Medipol University, Istanbul, Turkey

Director: Gurkan Ozturk, MD, PhD, Professor
Supervisor: Baris E. Suzek, PhD, Assistant Professor

Built a fully automated pipeline for variant analysis on WES data, which is being run on IBM high performance computing cluster, LSF suite. I used GATK, ANNOVAR, BWA and Picard programs, Bash and Python scripts, which I wrote, and Perl one-line commands.

Supervisor: Suleyman Yildirim, PhD, Professor

Performed analyses on 16S rRNA sequencing data (MiSeq) and shotgun sequencing data (ONT) for projects granted by Turkish government. I used R (packages: DADA2, DESeq2, edgeR, vegan, phyloseq, ggplot, ape etc), BWA, Metaphlan2 and online tools/platforms for taxonomical profiling and statistical analyses, e.g. PCoA, nMDS, PERMANOVA, t-SNE. We expect to publish in near future.

Research Assistant June 2017 – Februrary 2018

Suzek Bioinformatics Lab, Mugla Sitki Kocman University, Mugla/Turkey

Supervisor: Baris E. Suzek, PhD, Assistant Professor

Developed and tested variant analysis pipeline on WES data for Istanbul Medipol University, using GATK, ANNOVAR, BWA and Picard programs on computing server in Mugla Sitki Kocman University

Student Assistant September 2014 - June 2015

Molecular Endocrinology Lab, Middle East Technical University. Ankara, Turkey

Supervisor: Tulin Yanik, PhD, Associate Professor

Attended weekly lab meetings, presentations and assisted ELISA and Western blotting experiments with focus on obesity and satiety.

Voluntary Intern July 2014 - September 2014

Ecological Research Society (EKAD), Belek, Antalya/Turkey

Supervisor: Ali Fuat Canbolat, PhD

Observed nests and hatchlings of turtles -mostly *Caretta Caretta*- during morning shifts, informed tourists about turtles to take protective measures.

PUBLICATIONS

Oral Presentations

Sinan Findik, Abdulahad Bayraktar, Zafer Yüksel and Barış E. Süzek. Usage of the Human Phenotype Ontology in Request Platforms and Analysis of Clinical Genetic Sequencing Tests in Turkey. 11th Congress of Medical Informatics. November 16-17 2018. Ankara, Turkey

Poster Presentation

Bayraktar, A. A Study of Common Pathways of Neurodegenerative Diseases. 10th International Symposium on Health Informatics and Informatics (HIBIT). June, 28-30, 2017. Güzelyurt, Northern Cyprus.

CONGRESS-CONFERENCE PARTICIPATION

11th Congress of Medical Informatics. November 16-17 2018. Hacettepe University. Ankara, Turkey.

Bioinformatics Forum, 2018. May 11, 2018. Gebze Technical University. Kocaeli, Turkey.

10th International Symposium on Health Informatics and Informatics (HIBIT 2017). June, 28-30, 2017. Güzelyurt, Northern Cyprus

NanoDay 2016. May 26, 2016. National Nanotechnology Research Center, Bilkent University, Ankara, Turkey.

International Workshop on Machine Learning for Understanding the Brain. June 28-29, 2016. Middle East Technical University, Ankara, Turkey.

Deep Learning Tutorial. March 11 2015. Neuroscience and Neurotechnology, Middle East Technical University, Ankara, Turkey.

8th Aykut Kence Evolution Conference. April 26-27, 2014. Middle East Technical University, Ankara, Turkey.

LANGUAGE

Turkish: Native

English: IELTS Academic – Paper based = 7/9 – overall 12.01.2019

- Listening : 7.5
- Reading : 8.5
- Writing : 5.5
- Speaking : 6.0

METU EPE = 89.5/100