T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

INTEGRATIVE ANALYSIS OF MULTI-CELLULAR GENOME-SCALE METABOLIC NETWORKS WITH CELL TYPE SPECIFIC TRANSCRIPTOME DATA PREDICTED BY DECONVOLUTION ALGORITHMS: APPLICATION TO PARKINSON'S DISEASE

KADİR KOCABAŞ

A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE

DEPARTMENT OF BIOENGINEERING
BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAMME

GEBZE
2021

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

INTEGRATIVE ANALYSIS OF MULTI-CELLULAR GENOME-SCALE METABOLIC NETWORKS WITH CELL TYPE SPECIFIC TRANSCRIPTOME DATA PREDICTED BY DECONVOLUTION ALGORITHMS: APPLICATION TO PARKINSON'S DISEASE

KADİR KOCABAŞ

A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF BIOENGINEERING
BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAMME

THESIS SUPERVISOR
ASSOC. PROF. DR. TUNAHAN ÇAKIR

GEBZE
2021

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DEKONVOLUSYON ALGORİTMALARIYLA TAHMİN EDİLEN HÜCRE TİPİNE ÖZGÜ PARKİNSON HASTALIĞI TRANSKRİPTOM VERİLERİNİN ÇOK HÜCRELİ GENOM ÖLÇEKLİ METABOLİK AĞA HARİTALANMASI

KADİR KOCABAŞ
YÜKSEK LİSANS TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI
BİYOİNFORMATİK VE SİSTEM BİYOLOJİSİ PROGRAMI

DANIŞMANI
DOÇ. DR. TUNAHAN ÇAKIR

GEBZE
2021

GEBZE TEKNİK ÜNİVERSİTESİ	YÜKSEK LİSANS JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu’nun 08/01/2021 tarih ve 2021/02 sayılı kararıyla oluşturulan jüri tarafından 13/01/2021 tarihinde tez savunma sınavı yapılan Kadir Kocabaş’ın tez çalışması Biyomühendislik Anabilim Dalı Biyoinformatik ve Sistem Biyolojisi programı YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE
(TEZ DANIŞMANI) : Doç. Dr. Tunahan ÇAKIR

ÜYE : Dr. Öğr. Üyesi Pınar PİR

ÜYE : Doç. Dr. Ali ÇAKMAK

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu’nun
…..…/…..…/..…..… tarih ve ………/……... sayılı kararı.

İMZA/MÜHÜR

SUMMARY

Transcriptome-incorporated Flux Balance Analysis (FBA) of genome scale metabolic network (GMN) models has been proven to be useful for the elucidation of metabolic alterations in diseases. However, these studies were performed using bulk transcriptomic data, which comprise various cell types. More specific approaches are necessary to identify underlying reasons for complex diseases since most of the effect of diseases on metabolism are observed at tissue even cellular level. Cell type specific (CTS) gene expression data and a brain-specific GMN model were used in GMN-based flux prediction in this study to obtain a better insight on molecular mechanisms of Parkinson's Disease (PD) from a metabolic perspective. Multi-measure individual deconvolution (MIND), cellR and CIBERSORTx algorithms were used to predict CTS gene expression data since there are no available such data for PD. The predicted CTS gene expression data were evaluated based on principal component analysis (PCA), differential gene expression analysis and gene ontology (GO) enrichment analysis techniques. Results showed that the predictions by cellR and CIBERSORTx algorithms were poor based on the evaluation criteria, and they were not further pursued in this study. Later, MIND-predicted CTS gene expression data was mapped on iBrain671, a brain-specific GMN model that comprises 994 astrocytic and neuronal reactions controlled by 671 genes, to predict metabolic fluxes using Flux balance analysis (FBA) framework through GIMME algorithm. The predicted flux distribution of the disease-specific GMN that was generated by mapping CTS gene expression data predicted by using MIND algorithm was shown to be compatible with literature information.

Keywords: Parkinson's Disease, Brain-Specific Metabolic Network, Deconvolution, Transcriptome Data Integration, Flux Balance Analysis.

ÖZET

Genom ölçekli metabolik ağların (GMA) akı denge analizi (ADA) tekniği ile analiz edilmesinin hastalıkların metabolizmasının analizinde oldukça faydalı olduğu kanıtlanmıştır. Hastalıkların altında yatan nedenleri belirlemek için çeşitli GMA tabanlı çalışmalar yapılmıştır. Fakat bu çalışmalarda hücre spesifik olmayan toplu gen ekspresyon verileri kullanılmıştır. Hastalıkların metabolizma üzerindeki etkilerinin çoğu dokusal hatta hücresel düzeyde görüldüğünden, karmaşık hastalıkların altında yatan nedenleri belirlemek için daha spesifik yaklaşımlar gereklidir. Bu çalışmada Parkinson hastalığının (PH) moleküler mekanizmalarını hücresel seviyede analiz etmek için hücre tipine özgü (HTÖ) gen ekspresyon verileri ve beyne özgü GMA modeli kullanılmıştır. PH'nın böyle bir verisi olmadığından, HTÖ gen ekspresyon verilerini elde etmek için Multi-measure individual deconvolution (MIND), cellR ve CIBERSORTx algoritmaları kullanılmıştır. Tahmin edilen HTÖ gen ekspresyon verileri, temel bileşenler analizi, diferansiyel gen ekspresyon analizi ve gen ontolojisi (GO) zenginleştirme analizi tekniklerine dayalı olarak değerlendirilmiştir. GO zenginleştirme analizinde, nöronlar ve astrositler için PH ve normal örnekler ayrı ayrı kullanılarak tahmin edilen ekspresyon verileri karşılaştırılmıştır. cellR ve CIBERSORTx algoritmalarıyla elde edilen sonuçlar değerlendirme kriterlerine göre başarısız bulunduğu için çalışmanın geri kalanında MIND algoritmasının sonuçları kullanılmıştır. Akı denge analizi (ADA) 671 gen tarafından kontrol edilen 994 reaksiyondan oluşan beyne özgü GMA modeli iBrain671 kullanılarak gerçekleştirilmiştir. iBrain671, astrosit ve nörondan oluştuğu için HTÖ gen ekspresyon verilerinin haritalanmasına olanak sağlar. GIMME algoritması kullanılarak HTÖ gen ekspresyon verilerinin iBrain671’e haritalanmasıyla koşula özgü GMA modelleri elde edilmiş ve PH tepkime hızlarının (akıların) tahmini gerçekleştirilmiştir. MIND algoritması ile tahmin edilen HTÖ gen ekspresyon verileriyle oluşturulan hastalığa özgü GMA modeli kullanılarak tahmin edilen tepkime hızı dağılımının literatür bilgileri ile uyumlu olduğu gözlemlenmiştir.

Anahtar Kelimeler: Genom Ölçekli Metabolik Ağ, Parkinson Hastalığı, Dekonvolüsyon, Transkriptom Veri Entegrasyon, Akı Dengesi Analizi.

ACKNOWLEDGEMENTS

I would like to express my utmost gratitude first and foremost to my supervisor Assoc. Prof. Tunahan Çakır for his constant guidance and encouragement, he was always there for me when I needed to enlighten my way. Studying under his support was a joyful journey that always pushed me to the next step. I would like to thank the other members of my thesis committee, Assist. Prof. Pınar Pir and Assoc. Prof. Ali Cakmak for expending my horizon with their comments and suggestions on my thesis. I would like to thank PHITECH Bioinformatics Inc. and Özgenur Akdaş for their preprocessing of raw RNA-seq data of GSE68719 dataset. TUBİTAK is gratefully acknowledged for the financial support for this thesis through a grant (Project Code: 315S302).

I want to express my gratitude to my colleagues from Computational Systems Biology Group at Gebze Technical University for cheering me up when I needed providing a fun environment to work at.

TABLE of CONTENTS

LIST of ABBREVIATIONS and ACRONYMS

Abbreviations and Acronyms		**Explanations**
ATP	:	Adenosine Triphosphate
BA9	:	Brodmann Area 9
CBM	:	Constraint-Based Modelling
CMR	:	Cerebral Metabolic Rate
CTS	:	Cell-Type Specific
ETC	:	Electron Transport Chain
FBA	:	Flux Balance Analysis
GABA	:	Gamma Aminobutyric Acid
GEO	:	Gene Expression Omnibus
GIMME	:	Gene Inactivity Moderated by Metabolism and Expression
GPR	:	Gene-Protein-Reaction
GTEx	:	Genotype-Tissue Expression
LB	:	Lewy Body
LDA	:	Linear Discriminant Analysis
LP	:	Linear Programing
MIND	:	Multi-measure Individual Deconvolution
MOMA	:	The Method of Minimization of Metabolic Adjustment
MPTP	:	1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine
PCA	:	Principal Component Analysis
PD	:	Parkinson's Disease
PMI	:	Post Mortem Interval
PT	:	Putamen
QP	:	Quadratic Programing
ROS	:	Reactive Oxygen Species
scRNA-seq	:	Sigle-cell RNA sequencing
S	:	Stoichiometric matrix
SN	:	Substantia Nigra
SNpc	:	Substantia Nigra pars compacta
TCA	:	Tricarboxylic Acid

LIST of FIGURES

LIST of TABLES

1. INTRODUCTION

As human lifespan expanded in recent decades, neurodegenerative diseases are becoming one of the most serious threats to human health since they affect mostly elderly population. Effective treatments are needed to cure neurodegenerative diseases since there is no effective treatment strategy yet [1]. Parkinson's Disease (PD) is one of the most common neurodegenerative diseases in the world. Since an effective treatment method has not been developed to treat PD yet, it is essential to study it thoroughly. Studying PD is challenging because it is a complex disease that is caused by more than one factors [2, 3]. It is important to elucidate the complexity of the disease to find an effective treatment, and systems biology approaches are promising since they use computational power to overcome complexity.

Biological entities interact with each other in the cell, and most of the bio-interactions are altered in diseases via loss/gain or weakening/strengthening. Genome-scale metabolic networks (GMN) are reconstructed to simulate these bio-interactions using the information of biological entities such as DNA, proteins and metabolites. Different systems biology approaches have been developed to simulate bio-molecule interactions by using GMNs. Analysis of various conditions by using GMN models has been possible by the development of methods for the integration of transcriptomic data on the GMN models [4]. Another advantage of transcriptome data integration methods is generating organ-specific GMNs, which is very useful considering the fact that most of the diseases affect particular organs. There is also organ-specific models that were reconstructed to simulate specific organs by considering active reactions and specific cells inside the organ of interest, such as brain-specific GMN model iMS570 that consists of astrocyte and neuron. iMS570 has been used to analyze neurodegenerative diseases [5]. Another important feature of iMS570 is enabling cell-type-specific (CTS) gene expression data integration. This feature has potential to increase the resolution of the integrative metabolic modelling approaches, and it might give better insight on molecular mechanisms of neurodegenerative diseases from a metabolic perspective.

Although the number of single-cell RNA sequencing (scRNA-seq) studies has increased in last decades, it has still some disadvantages that cannot be overcome yet, such as it is not suitable for large cohorts [6]. It has great potential but still in evolvement, and there is no publicly available scRNA-seq data for certain cases such as PD. Bulk gene expression data deconvolution algorithms have been developed for two main purposes; (i) overcoming the aforementioned issues, (ii) enabling the usage of bulk gene expression data as CTS gene expression.

This study aims to analyze PD metabolism by using a brain-specific and cell-type-specific GMN model and predicted CTS gene expression data. It consists of two major steps; (i) deconvolution of bulk gene expression data, (ii) analysis of condition-specific GMN models by mapping CTS gene expression data.

Next chapter gives literature information about PD, transcriptome, bulk gene expression deconvolution techniques, GMN models, constraint-based modelling approaches and Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm that was used in this study to map CTS gene expression data on the GMN model. Chapter 3 focuses on the deconvolution of bulk gene expression data and the results of deconvolution processes. Assessment of results of the deconvolution algorithm is given in Chapter 4. In Chapter 5, analysis of condition-specific GMN model produced by the integration of CTS gene expression data is explained. Chapter 6 and 7 are about the discussion and conclusions of the study, respectively. The flowchart of the thesis study is represented in Figure 1.1.

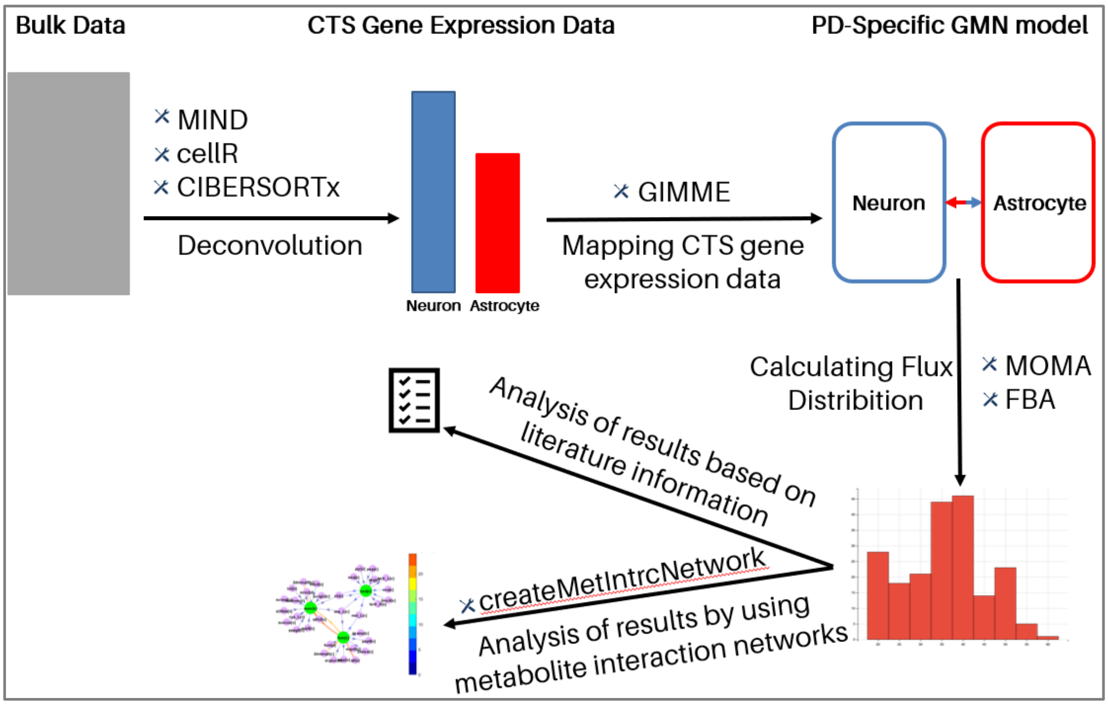

Figure 1.1: The flowchart of thesis study.

2. BACKGROUND INFORMATION

2.1. Parkinson's Disease

Parkinson's disease (PD) is one of the most common neurodegenerative disorders that mostly affect the elderly population. Although the ratio of PD among the population aged 40 years and over is about 0.1%, the ratio reaches 2-3% for the population older than 65 years old [2, 3]. PD is a multifactorial disease caused by both environmental and genetic factors. The clinical symptoms of PD include motor disorders, rigidity, rest tremor, suffering from slowness of movement, and disturbances in balance. Although some effective symptomatic therapies are available, a proven neuroprotective treatment is not available yet [2].

2.1.1. Neuropathology

The clinical hallmarks of PD were described by James Parkinson in 1817 for the first time. The loss of neurons in the substantia nigra pars compacta (SNpc) was identified as the main pathological characteristic of Parkinson's Disease about a century later. Hereafter, discovery of dopamine (DA) in the mammalian brain by Arvid Carlsson in 1958 has accelerated the pathological characterization of PD [7].

The nigrostriatal pathway that links SNpc with the dorsal striatum is involved in movement and cognition [8]. Neuronal loss in SNpc causes striatal DA insufficiency, which is the reason for the main clinical symptoms of PD. Although the replenishment of DA with the DA precursor, levodopa, to minimize the effect of DA deficiency is an effective treatment for clinical symptoms of PD, most patients start to suffer from uncontrolled movements termed "dyskinesia" after several years of levodopa treatment [7].

One of the main neuropathological characteristics of PD is the accumulation of α-synuclein in the cytoplasm of neurons in various brain regions [3]. α-synuclein protein, coded by SNCA gene in humans, is mainly located at the presynaptic terminals, and it is deeply related to the neuropathology of PD [9]. Deposits of α-synuclein proteins called Lewy bodies were discovered decades ago in PD patients.

The Lewy body (LB) accumulation takes place in various brain regions of PD patients based on the progression of the disease [3].

Braak and colleagues have described different stages of the progression of Lewy pathology [10]. Lewy pathology is found in the cholinergic and monoaminergic lower brainstem neurons in the early stage of PD patients (Braak stage I and II). It occurs in cholinergic and monoaminergic neurons of the midbrain and basal forebrain in the progressed PD patients (Braak stage III and IV). Pathology affects the limbic and neocortical regions in the last stages of PD patients (Braak stage V and stage VI). The major neuropathological features of PD are summarised in Figure 2.1 [3].

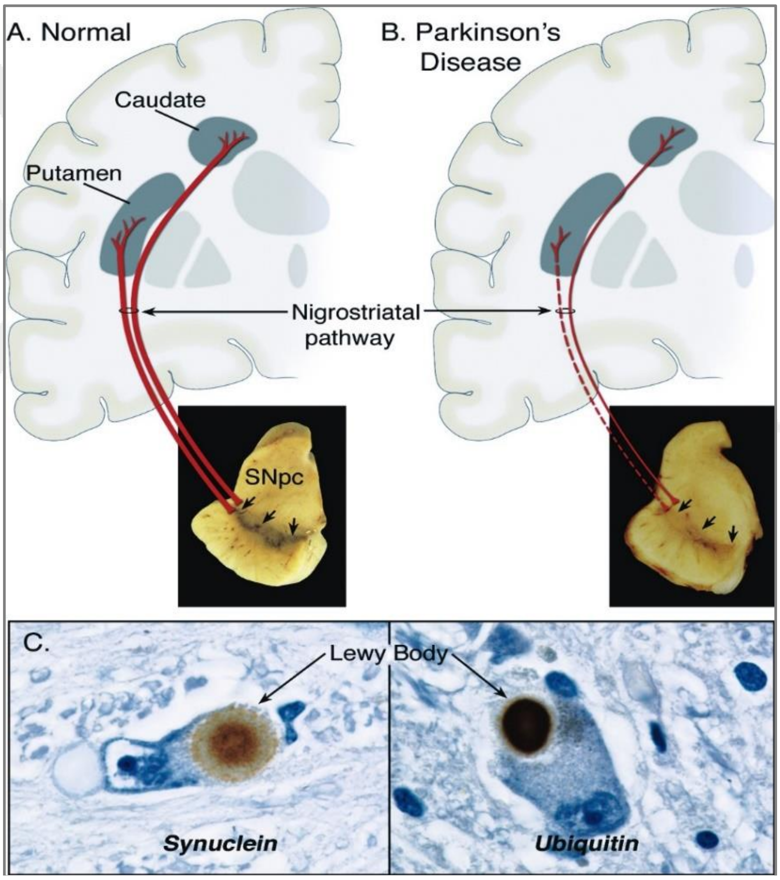

Figure 2.1:The main neuropathological characteristics of PD. A) The healthy nigrostriatal pathway (in red). B) The diseased nigrostriatal pathway (in red). The dashed line indicates degeneration of the nigrostriatal pathway.
C) Immunohistochemical labelling based images of LBs located in the SNpc dopaminergic neuron.

2.1.2. The Effect of Genetics and Environmental Factors on Parkinson's Disease

Although the cause of PD has not been fully discovered yet, it is accepted as a combination of genetic and environmental factors. Environmental factors cause PD as much as genetic factors. PD developed through genetic inheritance is called familial PD, while PD developed under the combination of environmental and genetic factors is named as sporadic PD. Sporadic PD comprises ~85–90% of all PD cases with the remainder classified as familial PD.

Some studies have verified that getting exposed to pesticides and heavy metals increases the risk of developing PD. Pesticides such as rotenone and paraquat cause oxidative stress by disrupting mitochondrial functions [11]. In the 1980s, some studies confirmed that 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) led to parkinsonism by causing damage in the substantia nigra. Researchers realized that 1-methyl-4-phenylpyridinium (MPP+), a product of MPTP metabolism in astrocytes, has a very similar composition with paraquat, revealing the role of pesticides in the PD pathogenesis [12].

It is known that some rare mutations lead to PD such as mutations in the alpha-synuclein gene (SNCA). In 1997, Polymeropoulos and colleagues showed that mutations in the SNCA gene are linked to PD, and thus revealed the role of α-synuclein proteins, produced from SNCA gene, in PD [13].

Parkin and PINK1 genes play a major role in the degradation of dysfunctional mitochondria by lysosome, a process named as "mitophagy". Mutations in Parkin and PINK1 genes lead to the accumulation of dysfunctional mitochondria, which causes oxidative stress and accumulation of proapoptotic proteins [14]. Parkin also plays an important role in the regulation of PGC-1alpha protein, which plays a central role in mitochondrial biogenesis and multiple antioxidant defenses [13, 15]. Consequently, mutations in Parkin and PINK1 genes lead to parkinsonism as a result of oxidative stress and damaged mitochondrial biogenesis pathway [16, 17].

DJ-1 plays a central role in the defense mechanism of the cell against oxidative stress by the regulation of NRF2 and stimulation of glutathione synthesis. Research has confirmed that mutations in the DJ-1 gene cause PD as a result of oxidative stress [13].

Mutation in the leucine-rich repeat kinase 2 (LRRK2) gene may contribute to the accumulation of α-synuclein, the main neuropathological characteristics of PD. It is reported that LRRK2 co-localizes with LB formation in lower brainstem. There are studies reporting that there is a high correlation between a mutation in the LRRK2 gene and sporadic PD development [18].

2.1.3. Metabolic Alterations in Parkinson's Disease

Although PD-linked metabolic alterations of the brain cells have not been fully elucidated, the dysfunction of redox homeostasis and bioenergetics are thought to be closely related to PD. Dysfunction of redox homeostasis and bioenergetics are associated with the pathways that are essential for cell survival, protein quality control and specific functions of nerve cells [19]. Accumulation of the reactive oxygen species (ROS) as a result of mitochondrial dysfunction may contribute to cell death [20].

The brain is a very complex organ that requires a considerably high amount of energy to carry its functions. Production of energy in the brain is nearly 20% of the total body energy production. Brain consumes 25% of the total glucose and 20% of the total oxygen. Energy metabolism of the brain has to be stable due to the requirement of continuous supply of high energy for maintenance of its functions. Glucose is the essential energy source of the brain. However, other energy sources can also be used, such as lactate and ketone bodies, in the necessary conditions that require high energy [19]. Neurons consume much more energy compared to glial cells for the maintenance of their functions demanding high energy such as delivering signals by using action potentials and post-synaptic potentials [21]. Mitochondrial dysfunction causes insufficient energy production, and this has a disruptive effect on neuronal functions such as the balance of ionic gradients and maintenance of neurotransmitter pathways. While increment of glycolysis capacity to avoid the effect of energy failure may have a potential to prevent neuronal cell death, accumulation of ROS is a serious risk factor for cell viability [19].

ROS are produced by electron transport chain (ETC) and the other reactions involved in the cellular homeostasis. The ROS are produced in the mitochondria via the reduction of oxygen with a combination of oxygen molecules and electrons. The source of electrons used in the production of mitochondrial ROS is the leakage of electrons in the ETC Complex I (Nicotinamide adenine dinucleotide [NADH] dehydrogenase) and Complex III (cytochrome bc1) [20]. Nicotinamide adenine dinucleotide phosphate (NADPH) oxidase (NOX) enzyme is another source of ROS production. NOX increases oxidative stress via the production of ROS as part of the defense mechanism of cells against infections [22]. Although production of ROS is a natural outcome of cells for carrying their functions, excessive accumulation of ROS is highly detrimental for cells. There are antioxidant defense mechanisms of cells for the avoidance of harmful effect of oxidative stress via the regulation of ROS. Superoxide dismutase (SOD) enzyme helps decreasing the level of superoxide anions as a result of generation of hydrogen peroxide using superoxide anions. The next step of the detoxification is carried out by the catalase (CAT) or glutathione peroxidase (GPX) enzymes via decomposition of hydrogen peroxide to water and molecular oxygen. Mitochondrial dysfunction and deterioration of regulation of the ROS mechanism may cause cell death as a consequence of the excessive accumulation of ROS [23].

Alteration of central carbon metabolism is another well-known metabolic phenomena of PD. Abnormal changes in lactate and pyruvate levels have been reported following a decrease in the activity of glucose metabolism. While lactate has protective effect against α-synuclein accumulation, inadequacy of glucose promotes formation of α-synuclein. Decrease in the activation of α-ketoglutarate dehydrogenase in the tricarboxylic acid cycle has also been observed as one of the metabolic phenomena of PD. [19].

2.2. Transcriptome Data

The genes carry information for the production of proteins. Although there is a large number of genes in an organism's DNA, only the ones providing information for the required proteins are active in a specific condition thanks to gene regulation mechanisms. For the purpose of producing proteins, active genes transmit genetic information to ribosome via messenger RNAs (mRNAs). There are some specific proteins, called enzymes, that catalyze reactions that are involved in the metabolism of cell. Analyzing the levels of mRNAs in a specific condition is a convenient way to enlighten the condition-associated metabolic activities inside a cell. Changing levels of mRNA may provide insights into metabolic alterations in cells under the influence of diseases.

Three widely used transcriptomics technologies have been developed for the purpose of measuring mRNA levels so far. Microarray technique, the oldest one, evolved at the end of 1990s [24]. Microarray technique uses hybridization probes that are placed in a particular way in the arrays. Each hybridization probe is specific to a known gene. The samples of interest that are tagged with fluorescent dye are loaded on microarrays. The signals are generated as a result of the hybridization of the transcripts on the probes, and the intensities of the signals are used to calculate relative levels of mRNAs. RNA-seq, the relatively new sequencing technology, was developed in the late 2000s [25]. The first step consists of producing complementary DNA (cDNA) from fragmented mRNAs and attachment of adapters to cDNAs. Subsequently, the reads are obtained by sequencing cDNAs in a sequencing machine. The machine-generated reads are then aligned to reference genome to get genome-scale mRNA levels [26]. The newest sequencing technology is single-cell RNA-sequencing (scRNA-seq) technology that allows measurement of cell specific mRNA levels. The first step of scRNA-seq is the isolation of viable cell from tissue. The next step is the production of cDNAs from mRNAs obtained from viable cells. cDNAs are tagged with unique identifiers for the prevention of losing information on cellular origin of mRNAs. Lastly, the tagged cDNAs are sequenced similar to RNA-seq [27].
A wide range of transcriptome data is publicly accessible through various databases such as Gene Expression Omnibus (GEO) [28] and ArrayExpress [29]. The records that are stored in GEO database consist of three main components as series, samples and platforms, and all of these components are related to each other. A Series record has a unique identifier as GSExxx, and it contains associated samples and descriptive information about a study. A Sample record contains descriptive and gene expression information of a sample that is part of the study. It has unique identifier as GSMxxx in the GEO database. A Platform record includes information about IDs such as probe ID, gene symbol and Entrez ID. It is stored with a unique identifier as GPLxxx in the database [30]. The number of total datasets stored in the GEO database (numbers obtained in Oct 2020) is given in Table 2.1.

Table 2.1: Number of datasets in the GEO database (as of Oct 2020).

	All Organisms	***Homo Sapiens***
Number of total datasets	137,638	58,364
Expression profiling by array	61,551	25,199
Expression profiling by high throughput sequencing	40,928	15,094

Publicly available gene expression datasets of PD are limited as it is hard to collect a sample of human brain. Post mortem tissues, tissues that are collected after the death of an individual, are used to generate gene expression data for neurodegenerative diseases. GSE20295 is one of the PD related microarray-based transcriptome datasets, which contains gene expression data taken from BA9, Putamen and substantia nigra regions of the brain. It was generated to investigate PD related metabolic pathways that are closely associated with bioenergetics and protein aggregation. In the associated study, the authors found strong indications of damage on electron transport and protein degradation systems caused by PD, and they identified novel genes that might be related to PD pathogenesis [31]. In another study, PD was studied by using the combination of proteomics and transcriptomics data that were taken from BA9 region of the brain, and they published their work on the GEO database with the associated ID GSE68719. They compared PD and control groups to identify functionally different pathways by using the significantly different proteins and genes separately. They found that functionally different pathways derived from the proteomics study are mostly related to mitochondrial functions. On the other hand, the results of transcriptomics study were mainly associated with protein folding pathways. Another PD related transcriptome study, GSE110716, is an RNA-seq study performed to analyze cingulate gyrus, which is one of the affected regions of the brain during PD. It consists of 16 samples, and equal number of samples were collected from PD and control groups [32]. GSE110716 is a subpart of GSE110720, which consists of 201 samples. They investigated the role of microRNAs and Piwi-interacting RNAs in PD. They identified that the expressions of microRNAs and Piwi-interacting RNAs are significantly changed during PD.

2.3. Deconvolution Methods of Bulk Transcriptome Data

Although scRNA-seq studies are emerging, they have some disadvantages such as noisy nature and being impractical for the analysis of large sample cohorts. Different bulk gene expression data deconvolution methods have been developed to eliminate the disadvantages of scRNA-seq by utilizing already available bulk gene expression data. Bulk gene expression data deconvolution algorithms are machine learning methods that infer cell-type-specific (CTS) gene expression profiles without physical cell isolation [6]. Although there is an error margin in such algorithms, they were shown to have a reasonable prediction power.

Multi-measure Individual Deconvolution (MIND) algorithm is one of the bulk gene expression data deconvolution algorithms that use empirical Bayes method [33]. Unlike a classical Bayesian method, the empirical Bayesian method uses existing data for the estimation of prior probabilities. After estimating prior probabilities, it applies classical Bayes rules [34]. The MIND algorithm works based on two main assumptions: (1) Percentage of cell types in bulk data can be predicted by deconvolution algorithms since bulk data consists of a combination of different CTS gene expression data. For the prediction, deconvolution algorithms use signature matrix that consists of marker genes, the genes that are specifically expressed in a certain cell type. (2) Bulk gene expression data from multiple brain regions of the same individual consists of common cell types over those regions. MIND algorithm requires bulk gene expression data from more than one brain region per subject [33]. For six cell types (astrocyte, oligodendrocyte, microglia, endothelial cell, inhibitory and excitatory neurons), MIND was used to predict CTS gene expression profiles of healthy brains [33], and the predicted expression profiles were validated by calculation of correlation between the predicted CTS gene expression data and single-nucleus RNA-seq data taken from the literature [35]. The high correlation has shown the potential of MIND method [33].

One of the most recently published deconvolution algorithms is cellR [36]. It uses reference scRNA-seq and population-wide expression profiles from GTEx as the prior information. Unlike MIND, cellR determines weights of marker genes using tissue-wise expression signatures. It predicts CTS gene expression by linear programming approach using a multivariate stochastic search algorithm by considering cellular correlations. cellR was used to predict CTS gene expression profiles of healthy and diseased brain tissue [36]. It was validated by using bulk transcriptome data of schizophrenia and pseudo-bulk RNA-seq samples data generated from real scRNA-seq data. The accuracy of predicted samples from pseudo-bulk RNA-seq samples was evaluated by using cosine similarity. For the evaluation of predicted samples from bulk transcriptome data of schizophrenia patients, differentially expressed (DE) gene analysis was conducted between predicted CTS gene expression profiles from healthy individuals and schizophrenia patients. The results were compared with DE genes reported in the literature [36, 37].

Another deconvolution tool that was developed for the prediction of CTS gene expression from bulk transcriptome data is CIBERSORTx [6]. It is an extended version of CIBERSORT, which was developed to predict cell fractions by ν-support vector regression method. CIBERSORTx uses linear equations to infer CTS gene expression data from bulk gene expression data, and it also allows to apply batch correction to decrease batch effect. The tool was used to predict CTS gene expression from various tumor types. Firstly, pseudo bulk data was generated from scRNA-seq data taken from melanoma tumors, and CTS gene expression from the bulk data was predicted. Secondly, the predicted CTS gene expression data were compared with the original scRNA-seq data to evaluate the performance of CIBERSORTx [6].

2.4. Genome-Scale Metabolic Networks

Metabolism is a complex network that consists of all enzymatic reactions that take place inside a cell. Thousands of metabolites are linked to each other via reactions that convert one metabolite to another, and thousands of enzymes, which are produced based on the instructions coming from mRNAs, catalyze these reactions. Development of systems biology techniques has made it possible to systematically analyze biological entities by computers. Systems biology approach is a systemic analysis of biological entities using biological data and computational techniques. Advancement in high-throughput biological data generation and in informatics technology has enabled systematic simulation of a cell by reconstructing Genome-Scale Metabolic Networks (GMNs). GMNs are mathematical representations of cell with the help of matrices called stoichiometry matrix, and it consists of reactions, metabolites and genes. GMNs are constructed based on gene-protein-reaction (GPR) rules that determine the link between reactions and genes through enzymes, and they enable simulation of metabolic phenotypes of various conditions by integrating transcriptome data [38]. The toy model of a GMN is shown in Figure 2.2 [4].

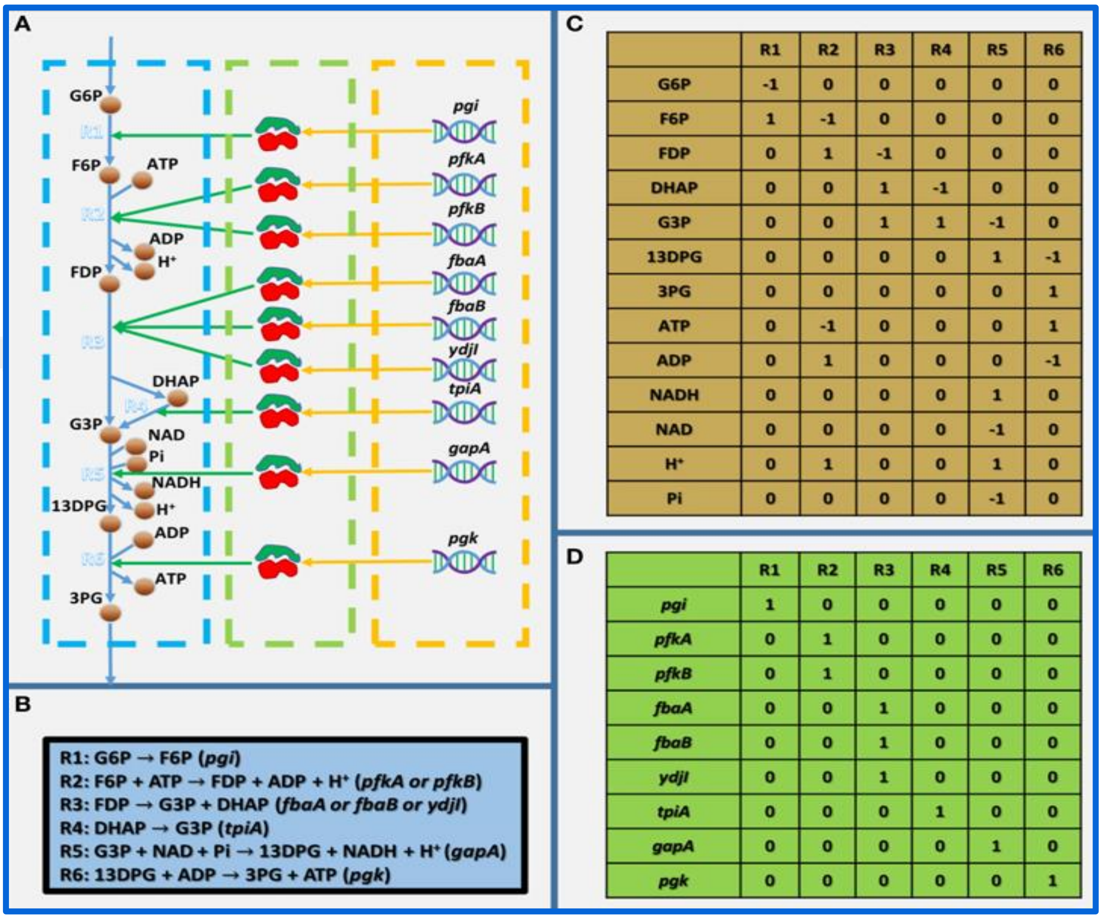

Figure 2.2. Toy model of GMN. A) The GPR rules that explain relationships between genes and reactions. B) Formulas of reactions. C) Mathematical representation of reactions in matrix format. D) Mathematical representation of GPR rules in matrix format.

The first GMN was reconstructed for *Haemophilus influenzae* in 1999 [39]. Publication of the first GMN has accelerated development of systems biology approaches on studying the metabolism of biological organisms in different conditions. Since then, GMNs were reconstructed for various organisms, and several different algorithms were developed to study these GMNs, such as mapping transcriptome data on GMNs [4]. First generic human GMN, called Recon 1, was published in 2007 [40]. This has enabled simulating the metabolism of human cells in different conditions in silico using systems biology approaches. Considering the complexity of metabolism of human cells, generic (i.e. organ/tissue non-specific) human GMNs have been constantly improved over the years to increase gene/pathway coverage and prediction quality [4]. Major generic human GMNs are listed in Table 2.2.

Table 2.2: Major generic human GMNs.

	Unique Metabolites	**Reactions**	**Genes**
Recon 1 [40]	1509	3743	1496
Recon 2 [41]	2626	7440	1789
Recon 3D [42]	4140	13543	3288
HMR [43]	3160	8100	3668
HMR 2.0 [44]	3160	8181	3765
HumanGEM [45]	8400	13101	3626

Mapping transcriptome data on generic human GMNs is widely used as an effective systems biology approach to elucidate disease associated metabolic alterations [46]. However, organ-specific GMNs, which are generated considering organ-specific metabolic pathways, are more useful for studying disease metabolism since most of the diseases mainly affect particular organs. Although every organ has its own unique metabolic functions, the brain has very special metabolic functions. There is complex networking between glial cells and neurons, which is essential for brain metabolism [47]. Several brain-specific GMNs were generated considering the complex relationship between astrocytes and neurons [5, 48]. One of the major brain-specific GMNs is iMS570, which was generated considering important pathways of brain metabolism such as ROS pathway, central carbon metabolism, neurotransmitter metabolism and lipid metabolism [5]. It was published in 2014, and it consists of 570 genes controlling 630 reactions via GPR rules to enable mapping transcriptome data. iMS570 is useful to study neurodegenerative diseases since (i) it was specifically reconstructed based on available information on brain metabolism, (ii) it separately includes neuron and astrocyte metabolism and it also includes interactions between these major brain cell types, and (iii) it was shown to predict healthy-state major brain metabolic fluxes via FBA simulations.

2.5. Constraint-Based Modeling Approaches

Constraint-based modelling (CBM) is a beneficial technique that helps to analyze complex systems of metabolism, and it was used for numerous studies so far [49]. CBM operates under physiochemical constraints with a steady-state assumption. Mass balances, thermodynamic reliability, substrate availability are main physiochemical constraints in CBM to simulate the metabolism of interest. The major constraint of CBM is the steady-state assumption, which assumes that the concentrations of intracellular metabolites do not change for a sufficiently long time [50]. Although determining a relevant objective function is challenging for higher organisms as most cell types have their own unique metabolic properties, CBM studies have commonly used various objective functions that are thought to be the main goal of the metabolism of interest.

Most widely used constraint-based computational approach for the analysis of GMNs is Flux Balance Analysis (FBA). FBA is a mathematical optimization technique that uses linear optimization to predict distribution of fluxes at steady state by employing the stoichiometric matrix (S matrix). FBA uses an objective function besides aforementioned constraints to select an optimum point from the flux solution space [51, 52]. Although ATP generation and biomass maximization are commonly used as objective functions, more specific ones are also used for a better representation of metabolism of organs such as maximization of glutamate/glutamine/GABA cycles between neurons and astrocytes for simulating brain metabolism [5]. The major components of FBA are S matrix, objective function and additional constraints. S matrix is generated using mass balance equations around intracellular metabolites as a function of reaction rates, and S_{mxn} consists of m metabolites and n reactions. Defining a proper objective function is crucial for meaningful results. Additional constraints play a critical role in narrowing solution space to meaningful results, and they can be determined based on wet-lab experiments. Accumulation of metabolites can be represented in mathematical form as dx/dt, where x is the vector of metabolite concentrations. The mathematical equation for the accumulation of metabolite concentrations can be written in terms of multiplication of S matrix with the vector of fluxes, which is represented as v in Equation (2.1). The assumption of the steady-state condition turns the complicated ordinary differential equation based problem into a much more straightforward linear system of equations, as shown in Equation 2.2.

$$\frac{dx}{dt} = S \times v \tag{2.1}$$

$$S \times v = 0 \tag{2.2}$$

The last step of FBA is performing optimization using necessary aforementioned constraints. All necessary steps for performing FBA are given in Figure 2.3 [51].

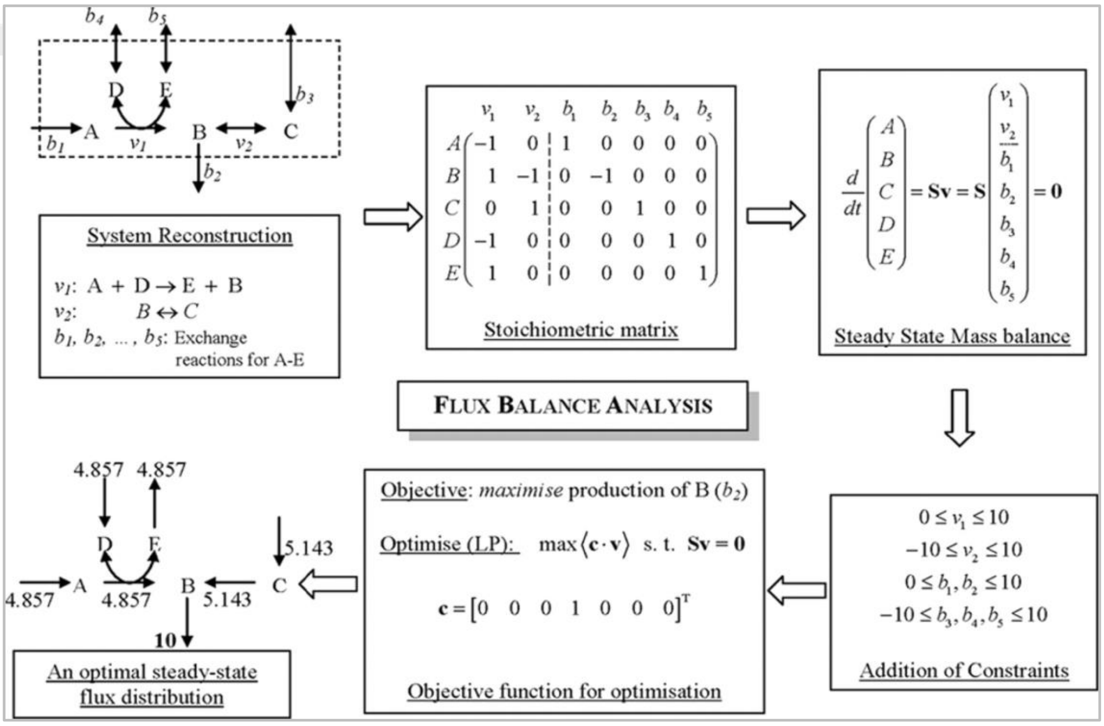

Figure 2.3. Representation of FBA. The system consists of 5 metabolites (A, B, C, D and E), 2 internal reactions (v_1 and v_2) and 5 exchange reactions (b_1, b_2, b_3, b_4, b_5).

One of the commonly used constraint-based approaches is Minimization of Metabolic Adjustment (MOMA), which was developed for the prediction of flux changes in perturbed metabolism [53]. The working principle of MOMA is based on the hypothesis that minimal redistribution will be observed in the flux distribution of perturbed metabolism with respect to the wild-type/healthy-state flux distribution. Although the working principles of MOMA are highly similar to FBA, the objective function is predicting a flux distribution of the perturbed metabolism such that it is closest to the wild-type/healthy-state metabolism. This is mathematically represented as the minimization of Euclidean distance between the two flux vectors. The mathematical formula of the objective function of MOMA is given by Equation 2.3, where v_{mt} represents the flux distribution of perturbed metabolism, and v_{wt} is the flux distribution of wild type/healthy state. v_{wt} can be obtained either from an FBA-based simulation or from flux-measurement experiments [53].

$$min \mid\mid v_{wt} - v_{mt} \mid\mid^{2} \tag{2.3}$$

2.6. Gene Inactivity Moderated by Metabolism and Expression (GIMME)

GMNs are generic metabolic networks that include all possible reactions in the corresponding organism without any separation of organ-specific reactions. There are two main limitations of the generic GMN models (i) every tissue has its own unique metabolism, (ii) most of the diseases are tissue/organ specific. Development of transcriptome data integration methods has enabled to overcome previously defined limitations of generic GMN models. Using biological data, organ-specific GMN models can be generated by removing reactions that are specific for other organs from the generic GMN model. For example, to study cardiovascular diseases, the scientists constructed human heart-specific GMN model by using transcriptome and proteome data. The human heart-specific GMN model was reconstructed by removing reactions from the generic human GMN model by the help of omic data [54]. Another useful application of transcriptome data integration methods is creating condition-specific GMNs, whereby effective studies can be performed for the discovery of disease associated metabolic alterations. Özcan and Çakır examined glioblastoma multiforme (GBM) by generating GBM-specific GMN models by mapping transcriptome data on brain-specific GMN model iMS570, and they successfully predicted metabolic phenomena of GBM [55].

One of the commonly used transcriptome data integration methods to generate condition/organ specific metabolic networks is GIMME [56]. The GIMME algorithm determines active reactions in the GMN based on inconsistency scores and mRNA levels belonging to a particular condition. It generates condition-specific GMN that can perform predefined metabolic functions called Required Metabolic Functionality

(RMF). GIMME requires four inputs: GMN model, threshold, reaction scores and objective fraction. The threshold parameter is used to determine active and inactive reactions, and it can be determined from transcriptome data. Reaction scores are calculated based on expression levels of the genes controlling the reactions, considering GPR rules. GPR rules are defined based on boolean logic, considering the structure and function of enzymes. There are three main GPR rules, (i) reaction is controlled by a single gene product (ii) reaction is controlled by enzyme complex that consists of multiple subunits (iii) reaction is controlled by isoenzymes (Figure 2.2) [57]. To put it simply, a reaction score is an indication of available enzymes or enzyme complexes that catalyze the reaction. The objective fraction parameter is used to ensure generating condition-specific GMN models with RMF. It is written as a fraction, and defines maximum allowable reduction in the value of objective function. For example, it should be chosen as 0.2 to allow maximum 80% reduction in the objective function value of organism-specific generic GMN model.

Firstly, a reduced GMN is generated by removing the reactions with scores below the predefined threshold. Reaction scores are calculated based on GPR rules and transcriptome data [56]. Secondly, the removed reactions that are needed for the achievement of RMF are reinserted to the reduced model based on the level of inconsistency score (IS). IS defines deviation between the transcriptome data and the generated condition-specific model, and it is calculated as the multiplication of flux with the difference between threshold and data for each reinserted reaction (Figure 2.4) [56]. The fluxes that are used to calculate IS are predicted via FBA. Reactions that are reinserted to the reduced model are chosen such that the smallest IS is guaranteed. GIMME generates a condition-specific GMN that can perform RMF with the smallest possible IS (Figure 2.4).

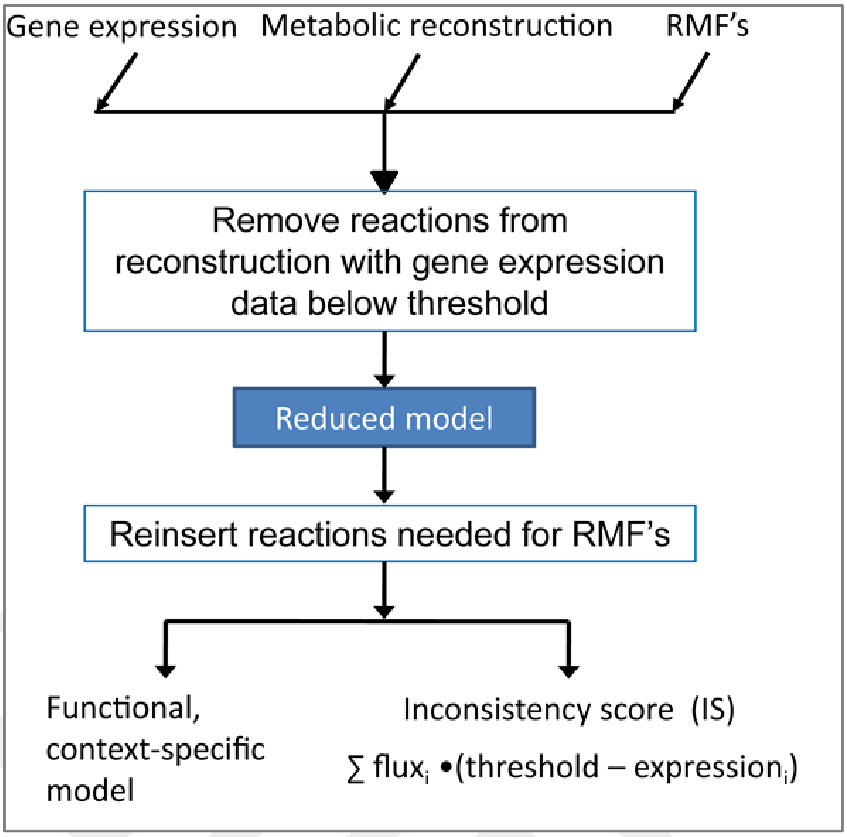

Figure 2.4: Flow chart of GIMME algorithm. GIMME requires transcriptome data, GMN and RMF. Firstly, a reduced model is created. Secondly, removed reactions with smallest IS are reinserted to produce a functional context/tissue specific model. Fluxes are calculated using FBA.

3. DECONVOLUTION OF BULK GENE EXPRESSION DATA

3.1. Datasets

Three different bulk gene expression datasets were used in this study for generating cell type-specific gene expression data. Datasets are publicly available on GEO database with unique IDs of GSE20295, GSE68719 and GSE110716 [31, 32, 58]. GSE20295 comprises two different groups as control and PD. The dataset consists of 93 samples taken from three separate brain regions, with 29 BA9 samples (15 control and 14 PD), 35 putamen samples (20 control and 15 PD) and 29 substantia nigra (SN) samples (18 control and 11 PD). Samples were chosen based on age, sex, post mortem interval (PMI) and brain pH values for ensuring the generation of consistent data since these parameters are critical for generating comparable data [31]. Another significant feature of GSE20295 is the fact that it contains gene expression data from different brain regions belonging to the same individuals, and there are 14 controls and 11 PD patients in the dataset with data from multiple brain regions. This transcriptome study was performed using microarray technology with [HG-U133A] Affymetrix Human Genome U133A Array. GSE68719 consists of 73 samples taken from BA9 region of the brain as 29 PD and 44 control groups [58]. The information of age, sex, RNA integrity number and PMI for each donor were provided with the sample ID (GSMxxx). This transcriptome study was carried out with RNA-Seq technology by using Illumina HiSeq 2000. GSE110716 consists of 20 samples taken from cingulate gyrus region of brain, and it comprises equal number of samples taken from PD and healthy individuals [32]. Unlike other datasets, the only provided information was the condition of individuals. This transcriptome study was performed with RNA-Seq technology by using Illumina HiSeq 1000.

3.2. Deconvolution of Bulk Data by MIND Algorithm

3.2.1. Prediction of Cell-Type Fractions

One of the newest approaches for the generation of cell-specific gene expression data from bulk transcriptome data is MIND algorithm, which was coded in R programing language. MIND requires bulk gene expression data taken from different brain regions of the same subject. Therefore, it can be applied to GSE20295 dataset. As PD and control groups, two bulk data matrices were generated to predict cell type fractions. With the gene expression data from 3 different brain regions (BA9, SN and putamen), PD group consists of 11 patients while control group has 14 individuals. MIND algorithm requires fractions for the cell types in the brain regions of each subject separately to predict cell type specific (CTS) gene expressions. Cell type fractions were estimated using est_frac() function of MIND R package. This function estimates cell proportions based on Non-negative matrix factorization (NMF) approach [59]. NMF is a numerical technique that factorizes the main matrix to submatrices such that they do not contain any negative values. It estimates cell fractions by establishing a linear relationship between signature matrix (C_{NxT}), bulk data (B_{NxS}) and cell type fractions (W_{TxS}) as represented in Equation 3.1, where N is number of marker genes, S is number of samples and T is number of cell types. The objective function of the NMF approach is represented in Equation 3.2.

$$B_{NxS} = C_{NxT} \times \mathrm{W}_{\mathrm{TxS}} \tag{3.1}$$

$$\min \left(|| \; B - C \times W \; ||^2 \right) \tag{3.2}$$

Cell-type fractions, which are represented as W_{TXS} in Equation 3.1, were predicted using bulk gene expression data and signature matrix. Bulk gene expression data GSE20295, which is represented as B_{NxS} in Equation 3.1, were used in log2 transformed format in accordance with the article of MIND algorithm. A signature matrix, which is represented as C_{NXT} in Equation 3.1, were taken from literature and were used as a template [33]. It contains 754 genes for six cell-types as excitatory neuron, inhibitory neuron, astrocyte, oligodendrocyte, endothelial and microglia. The signature matrix should contain marker genes that are highly expressed in a particular cell-type compared to other cell-types. To put it simply, marker genes should represent a particular cell-type in the sense of gene expression. The template signature matrix was updated in this study to establish a signature matrix with the aforementioned properties. Firstly, excitatory neuron and inhibitory neuron were united since iBrain671 model does not distinguish between neuron types. Secondly, to choose a marker, the threshold was set to be ten times greater than the sum of all expression levels in other cell types, compared to a particular chosen cell type. A signature matrix that consists of 248 marker genes was obtained as a result of performing the aforementioned elimination process. For both PD and control groups, cell fractions of astrocyte, neuron, microglia, endothelial and oligodendrocyte were estimated for each sample separately using est_frac() function of MIND. For the three brain regions, the means of predicted cell fractions and standard deviations are given in Table 3.1.

Table 3.1: Mean of the predicted cell fractions and their standard deviations (sd) considering brain regions and conditions. SN represents the substantia nigra region of brain.

Groups	Regions	Values	Neuron	Astrocyte	Oligo dendrocyte	Endothelial	Microglia
CONTROL	BA9	Mean	0.54	0.20	0.04	0.17	0.03
		Sd	0.21	0.12	0.02	0.08	0.01
	Putamen	Mean	0.23	0.22	0.22	0.26	0.05
		Sd	0.12	0.12	0.08	0.09	0.03
	SN	Mean	0.21	0.17	0.24	0.30	0.07
		Sd	0.15	0.07	0.05	0.09	0.03
PD	BA9	Mean	0.39	0.23	0.08	0.25	0.03
		Sd	0.22	0.12	0.06	0.11	0.01
	Putamen	Mean	0.17	0.21	0.24	0.30	0.06
		Sd	0.13	0.06	0.11	0.08	0.02
	SN	Mean	0.11	0.14	0.28	0.37	0.07
		Sd	0.06	0.06	0.07	0.11	0.03

The cell fraction prediction results were compared with literature information for validation. Predicted cell fractions for BA9 were compared with frontal cortex since there is no available information about cell fractions of BA9 in the literature, and BA9 is part of the frontal cortex. For PD and control groups, the means of predicted cell fractions for each sample were represented in Figure 3.1 [33].

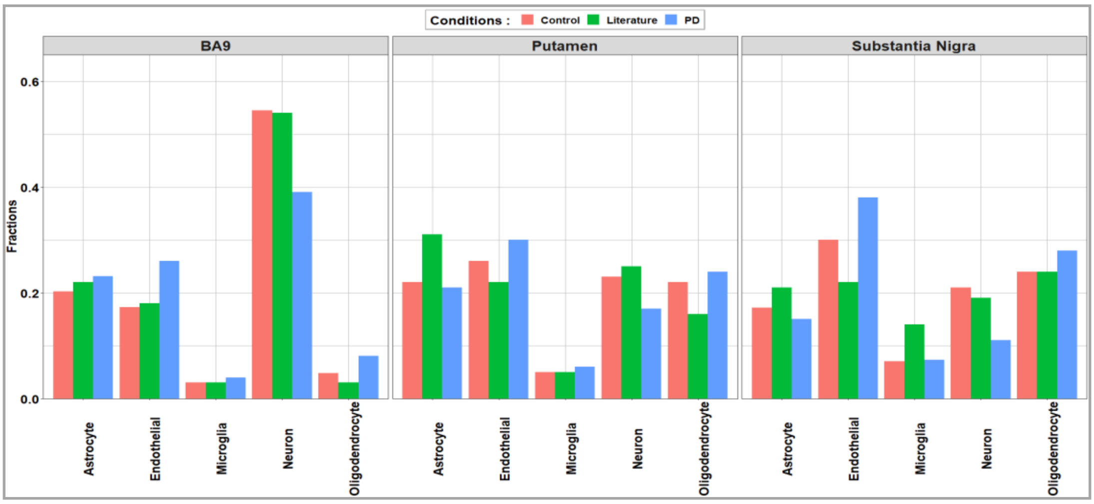

Figure 3.1: Bar plots of cell fraction predictions for each brain region. Red, blue and green bars represent the control group, PD group and literature information for the control group, respectively.

For the control group, the correlations between predicted and reported cell fractions were calculated to evaluate the results of cell-fraction predictions statistically. Pearson correlation was used to calculate magnitudes of correlations (R) and significance of correlations (p-value). The results are given in Table 3.2.

Table 3.2: Correlation between predicted and reported cell fractions for the control group.

	BA9	**Putamen**	**SN**
R	0.99	0.8	0.84
pvalue	0.000003	0.09	0.07

3.2.2. Cell Type Specific Gene Expression Predictions

For the prediction of CTS gene expression values, MIND algorithm requires cell fractions of K cell types and bulk data taken from n subject. For the K cell types, it models bulk gene expression matrix (B) using a linear equation that comprises cell fraction matrix (W), CTS gene expression matrix (A) and error margin (Equation 3.3). Each sample of B, which is a 3-dimensional data, consists of gene expression of different regions from the same individual. The algorithm needs the combination of two main measurement types; (i) measurements from different individuals for the same tissue, (ii) measurements from different regions from the same individual. T_i in Equation 3.3 represents the regions for i[th] subject (Figure 3.2 b and c) [33]. K represents cell types while j holds gene information (Figure 3.2 a). The predicted CTS gene expression, which is represented as A_{ij} in Equation 3.3, is sample-specific but does not change depending on regions. To put it simply, MIND predicts gene expression of each cell type for each individual without region specificity.

$$\underset{(T_i \, x\, 1)}{B_{ij}} = \underset{(T_i \, x \, K)}{W_i} \times \underset{(K \, x \, 1)}{A_{ij}} + \underset{(T_i \, x \, 1)}{e_{ij}} \tag{3.3}$$

In Equation 3.3, B_{ij} holds the information on gene expression of i[th] subject for gene j as a vector of quantified measurements T_i, which is equal to the product of CTS gene expressions A_{ij} and W_i, representing sample-specific and region-specific cell type fractions. e_{ij} is added to the given equation to represent the noise. A_{ij} is sample-specific but does not change depending on regions, and it is presumed to be distributed randomly as $A_{ij} \sim N(\alpha_j, \sum_c)$, with mean a_j (Kx1) and covariance matrix $\sum_c$ (KxK), where K represents cell types.

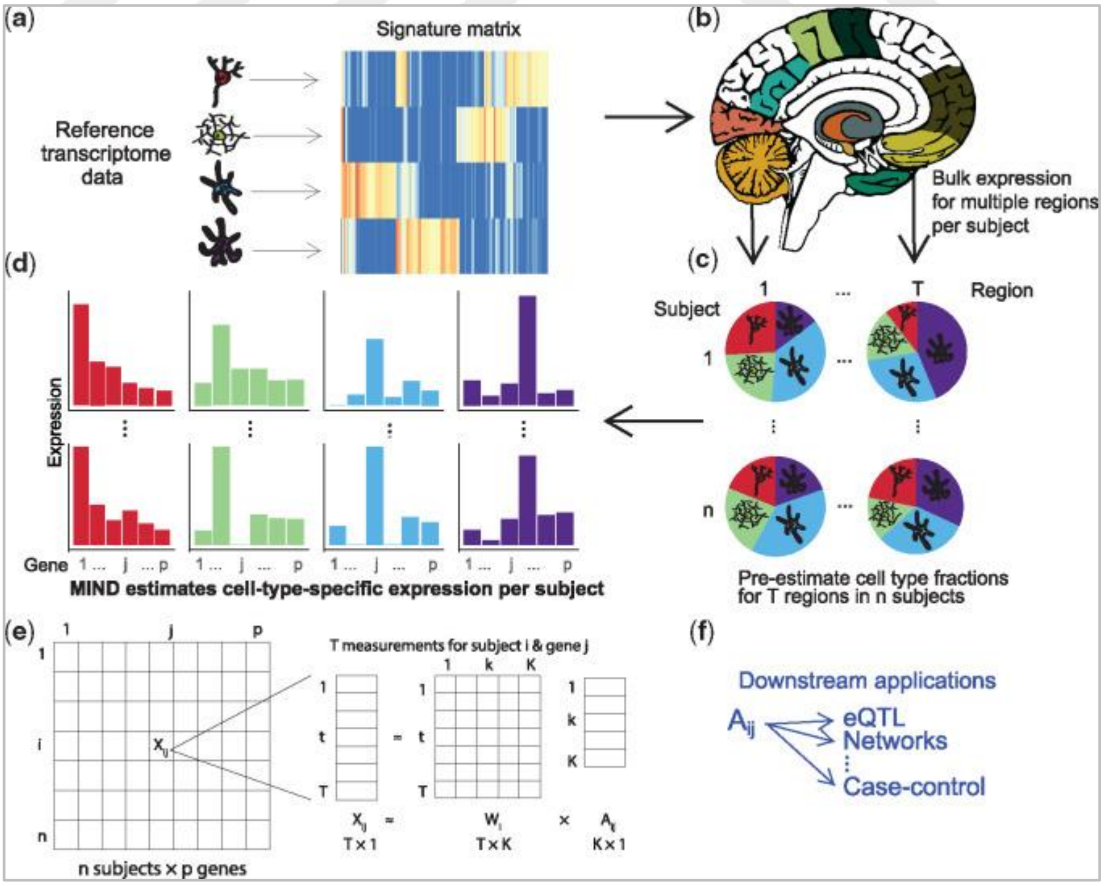

Figure 3.2: The flow chart of MIND. a) Representation of signature matrix, which consists of marker genes that are highly expressed in only a particular cell-type. b) description of bulk gene expression data. It consists of samples which contain measurements from different brain regions of the same individual. c) Representation of cell type fractions, which are subject and region specific. d) The predicted CTS gene expression, which is subject specific but is not region specific. e) Matrix representation of Equation 3.3. Bulk data X comprises T measurements of p gene from n subject. Cell fraction matrix W consists of K cell types from T measurements while A contains CTS gene expression from K cell types. f) Example application fields of CTS gene expression data.

For the CTS gene expression prediction, PD and control groups were deconvolved separately using mind() function. Inputs of mind() function are cell type fractions and bulk gene expression data. The output of the function is CTS gene expression of each subject without region information. CTS gene expressions of astrocyte, neuron, microglia, endothelial and oligodendrocyte cells were predicted for both PD and control groups separately. Consistency of the predicted CTS gene expressions was analyzed using principal component analysis (PCA). This technique is a mathematical transformation technique that enables large-scale multivariate data sets to be reduced to a smaller size with minimal information loss. It is an unsupervised machine learning method that can be used to investigate patterns in data [60]. PCA can be used for various purposes, such as grouping variables and data compression. The similar variables will be grouping together as a result of PCA, and this can give an idea about patterns in data. In this study, PCA was used to assess predicted CTS gene expression data in which same cell types are expected to be grouping together considering conditions. The results of PCA are given in Figure 3.3.

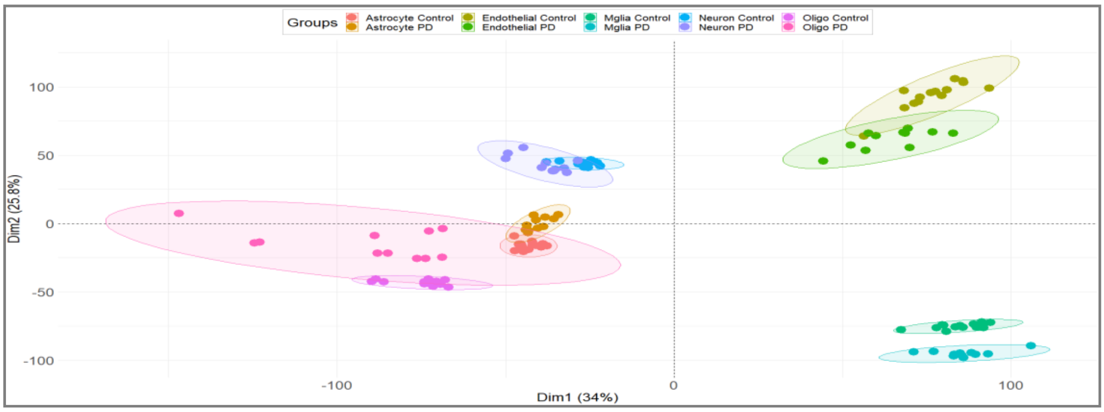

Figure 3.3: PCA results of CTS gene expression values. For PD and control groups, each cell type is represented by a different colour.

3.3. Deconvolution of Bulk Data by cellR

One of the newest bulk gene expression data deconvolution algorithms is cellR, which was published in 2020, and it is coded in R programing language [36]. Unlike MIND algorithm, it predicts cell fractions using its own algorithm. It uses reference scRNA-seq data and population-wide expression profiles to predict cell fractions and CTS gene expression data. cellR is compatible with bulk RNA-seq count data unlike MIND algorithm.

3.3.1. Prediction of Cell-Type Fractions

The first step of cellR algorithm is predicting cell fractions by using reference scRNA-seq and expression signatures taken from Genotype-Tissue Expression (GTEx) project. GTEx project was started to identify genetic variations across tissues by examining tissue-specific expression levels along with genotype-associated variants [61]. cellR uses the information on tissue-specific expressions from the GTEx project, and it adjusts the weights of cell-specific gene markers based on this information. cellR estimates cell-fractions based on an optimization problem (Equation 3.4).

$$\min f = |G_{IxT} - ((P_{CxI})^T \times B + \lambda \times M)| \tag{3.4}$$

$$s \times t \times P \geq 0 \tag{3.5}$$

$$\forall i \in I, \sum_k P_{ki} = 1 \tag{3.6}$$

In the equations, G_{IxT} represents the bulk gene expression data consisting I samples and T marker genes. Bulk data is provided by the user, and cellR identifies marker genes by using scRNA-seq data. P_{Cx1} represents proportions of C cell-types, and B indicates signature matrix that was generated based on reference scRNA-seq data. P_{Cx1} and scRNA-seq data should contain same types of cells. The terms for calculating error term, $\boldsymbol{\lambda}$ and M, represent complexity and penalty factor respectively.

For the calculation of cell-fractions, cellR uses GTEx data to calculate weights for marker cells (Equation 3.7).

$$w_i = 1 + \frac{1}{S_i X \; \sum_{t=1} \frac{1}{S_t + \varepsilon}}, w_i = 1 \; if \; s_i = 0 \tag{3.7}$$

In Equation 3.7, w_i represents the weight for marker gene i, s_i indicates standard deviation calculated from GTEx data and ε were added for preventing a value of infinity. With all of these aforementioned equations, cellR calculates cell-fractions from bulk data by using reference.

Two different bulk gene expression datasets were used to predict CTS gene expression data by using cellR algorithm since adequate performance could not be achieved for the first one. GSE68719 and GSE110716, which were collected from BA9 and Cingulate gyrus regions of the brain respectively, were separately used for predicting CTS gene expression by using cellR algorithm. GSE110716 is a suitable dataset since it was obtained with RNA-seq technology and the count values were published. The count values of the other dataset, GSE68719, were obtained by preprocessing raw fastq files since count values were not available. Reference scRNA-seq data taken from frontal cortex was obtained from the literature [62]. Genome wide expression signatures were obtained from GTEx [61]. The same reference scRNA-seq and GTEx data were used for both GSE68719 and GSE110716. Prediction of cell-fractions were performed for both GSE68719 and GSE110716 separately. For both PD and control groups, cell fractions of neuron, endothelial, pericyte, astrocyte, oligodendrocyte, oligodendrocyte progenitor cells (OPC) and microglia were estimated for each sample separately using Deconvolution() function of cellR. For brain regions, means of predicted cell fractions and standard deviations for GSE68719 and GSE110716 are given in Table 3.3.

Table 3.3: Mean of the predicted cell fractions and their standard deviations (sd) considering brain regions and conditions. The predicted cell fractions obtained by both datasets were given. OPC represents oligodendrocyte progenitor cell.

Cell-Types	**BA9 - GSE68719**				**Cingulate Gyrus- GSE110716**			
	Control		**Parkinson**		**Control**		**Parkinson**	
	Mean	**Sd**	**Mean**	**Sd**	**Mean**	**Sd**	**Mean**	**Sd**
Neuron	65.24	6.25	59.82	4.88	33.47	17.92	42.56	17.76
Endothelial	0.54	0.34	0.79	0.47	4.89	4.60	3.22	4.45
Pericytes	0.05	0.12	0.06	0.13	8.06	3.78	4.42	4.79
Astrocyte	14.83	3.10	16.06	2.63	45.61	19.45	41.12	14.30
Oligodendrocyte	16.50	5.88	19.33	6.13	3.98	5.07	3.96	5.74
OPC	2.17	0.95	2.75	0.81	2.89	4.96	4.73	5.26
Microglia	0.67	0.39	1.18	0.77	1.10	2.91	0.00	0.00

The cell fraction prediction results obtained by cellR algorithm were compared with literature information for validation. There is no reported information of cell fractions of BA9 and cingulate gyrus in the literature. Predicted cell fractions were compared with reported cell fractions of the frontal cortex, which is the closest region of the brain to BA9 and cingulate gyrus. The means of predicted cell fractions for each sample were represented in Figure 3.4.

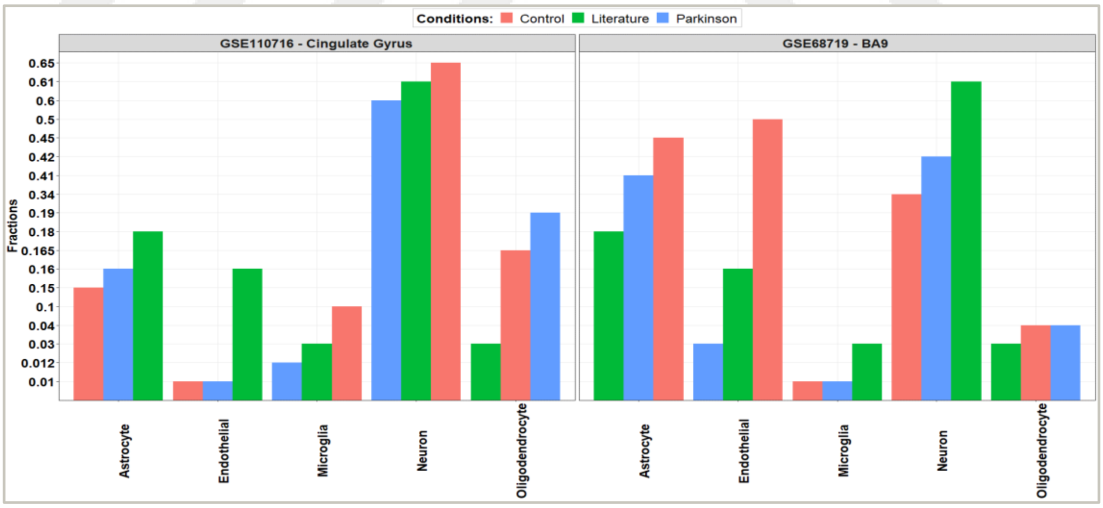

Figure 3.4: Bar plots of cell fraction predictions by cellR algorithm. Red, blue and green bars represent the control group, PD group and literature information for the frontal cortex for health people, respectively.

The cell-fraction prediction accuracy was analyzed by calculating correlation between predicted and reported cell-fractions. Pearson correlation was used to calculate correlation coefficients (R) and p-values, and the results are given in Table 3.4.

Table 3.4: The results of correlation analysis between predicted and reported cell fractions for BA9 and Cingulate gyrus region of the brain.

	BA9 - GSE68719	Cingulate Gyrus- GSE110716
R	0.91	0.61
p-value	0.03	0.26

3.3.2. Cell Type Specific Gene Expression Predictions

For the prediction of CTS gene expression data, cellR requires bulk RNA-seq data, reference scRNA-seq data and cell fractions. cellR estimates CTS gene expression by predicting one gene at a time. It models the expression of one gene for each sample for every step (Equation 3.8).

$$g_{ij} = \sum_{c=1}^{C} (p_{cj}\ X\ e_{icj}) \tag{3.8}$$

In Equation 3.8, g_{ij} represents bulk expressions of i^{th} gene in j^{th} sample, and C represents cell types. P_{cj} indicates cell fraction of c^{th} cell in j^{th} sample. $e_{icj,}$ represents CTS gene expression of i^{th} gene for C^{th} cell in j^{th} sample, and the distribution of e_{icj} is accepted to show features of the negative binomial distribution NB(r_{ic}, d_{ic}), where r is mean and d is dispersion parameter.

For the prediction of CTS gene expression of a gene, cellR estimates parameters of the distribution for each cell type separately. At the beginning of the parameter estimation process, cellR assigns random values to parameters considering cell-types. For each iteration, algorithm estimates $\hat{g}_{ij}$, which is the estimated bulk gene expression for i^{th} gene and j^{th} sample, by using estimated parameters and Equation 3.9.

$$\hat{g}_{ij} = \sum_{c=1}^{C} (p_{cj}\ \times\ \breve{e}_{icj}) \tag{3.9}$$

In Equation 3.9, $\hat{g}_{ij}$ represents the estimated bulk gene expression of gene i in sample j, and $\breve{e}_{icj}$ represents CTS gene expression of gene i in sample j for c^{th} cell type.

For the start, $\breve{e}_{icj}$ is estimated after the estimation of $\hat{g}_{ij}$, which is predicted by using initially assigned parameters, and they are iteratively estimated in every step with updated parameters. For the NB(r_{ic}, d_{ic}), the parameter updating process is performed using the perturbation algorithm that randomly chooses one of the following two update techniques: (i) the parameters are updated by calculating the following pseudo equation: (current parameters) + (current parameters times random values that are generated between -0.5 and 0.5), (ii) for the r_{ic}, the parameters are updated by calculating the following pseudo equation: (new mean parameter(r_{ic})) ± (standard deviation of gene i across all samples) and d_{ic} is generated randomly between 0 and 1. For each iteration, cellR compares known bulk gene expression with predicted one that is calculated by using updated parameters. cellR calculates RMSE for each iteration and it repeats iterations as RMSE decreases (Equation 3.10). In Equation 3.10, N represents the number of samples.

$$RMSE_i = \sum_{j=1}^{N} (\mathrm{g}_{\mathrm{ij}} - \hat{g}_{ij})^2 \tag{3.10}$$

To put it simply, cellR estimates CTS gene expression of gene i for a particular cell type for each sample through the following steps: (i) assign random values to the parameters, (ii) create negative binomial distribution with mean parameter (r) and disperse parameter (d), (iii) generate random values for gene i from the generated negative binomial distribution, (iv) calculate root mean square error, (v) update parameters with one of the two methods mentioned above and return to step (ii). The parameter update method is chosen randomly. The prediction of CTS gene expression by using cellR requires long computational time since it needs too much repetition with considerably expensive calculation. For each cell type, the prediction of CTS gene expression of one gene takes approximately 90 seconds (Computer System: Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz (16 CPUs), ~2.1GHz processor and 128GB RAM). For neuron and astrocyte, CTS gene expression was performed only for 671 genes that are included in iBrain671 to overcome time issue.

For PD and control groups, CTS gene expression for astrocyte and neuron were separately predicted by using GSE110716. The results were evaluated based on PCA, LDA and correlation matrix. PCA and linear discriminant analysis (LDA) were performed to examine the grouping of cells considering conditions. LDA, which is a supervised method unlike PCA, can be used for grouping variables like PCA. LDA is a machine learning technique that aims to find a vector that maximizes the distance between priori known classes while minimizing standard deviation within classes [63]. The results of PCA are given in Figure 3.5.

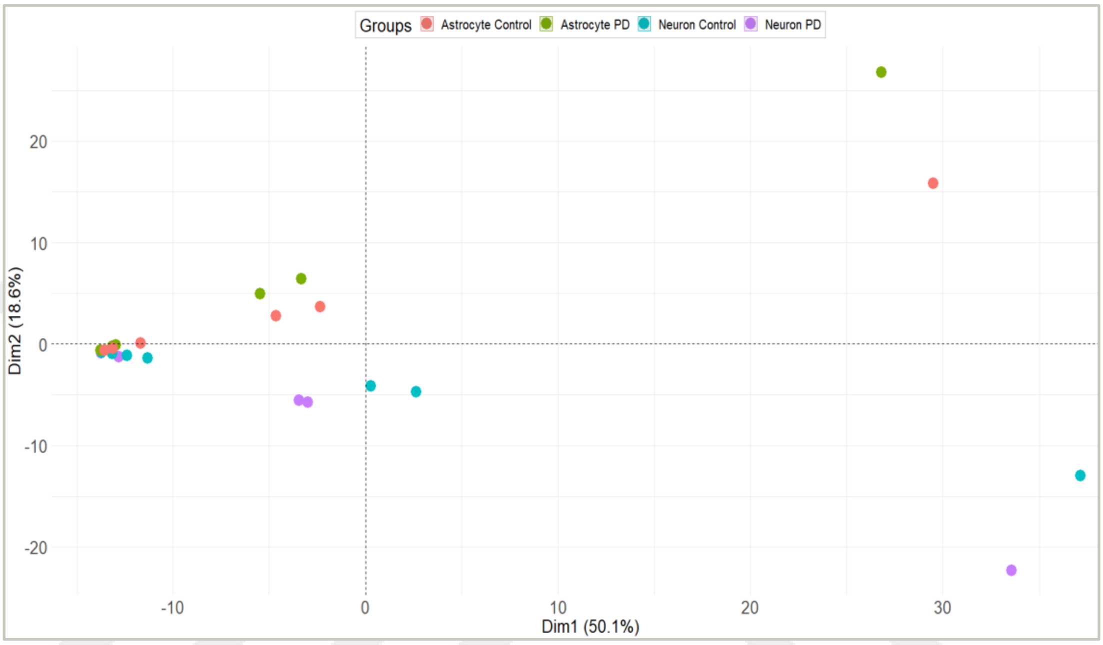

Figure 3.5: PCA results of CTS gene expression values. For PD and control groups, each cell type is represented by a different colour.

The cell types are distributed with a specific pattern, which is not expected, in the PCA results. Each cell types spread through a line with a different direction, and most of the variables are grouping together even if they are different cell types in different conditions. Although the total number of samples in the dataset is 64, only few points are seen in the PCA graph since most of the sample points overlap, which implies a poor prediction of CTS gene expressions by CellR. Examining data from a different aspect might give a deeper understanding of the problem. Therefore, LDA analysis was performed to examine the grouping of cell types with a different machine learning algorithm. The LDA graph is given in Figure 3.6.

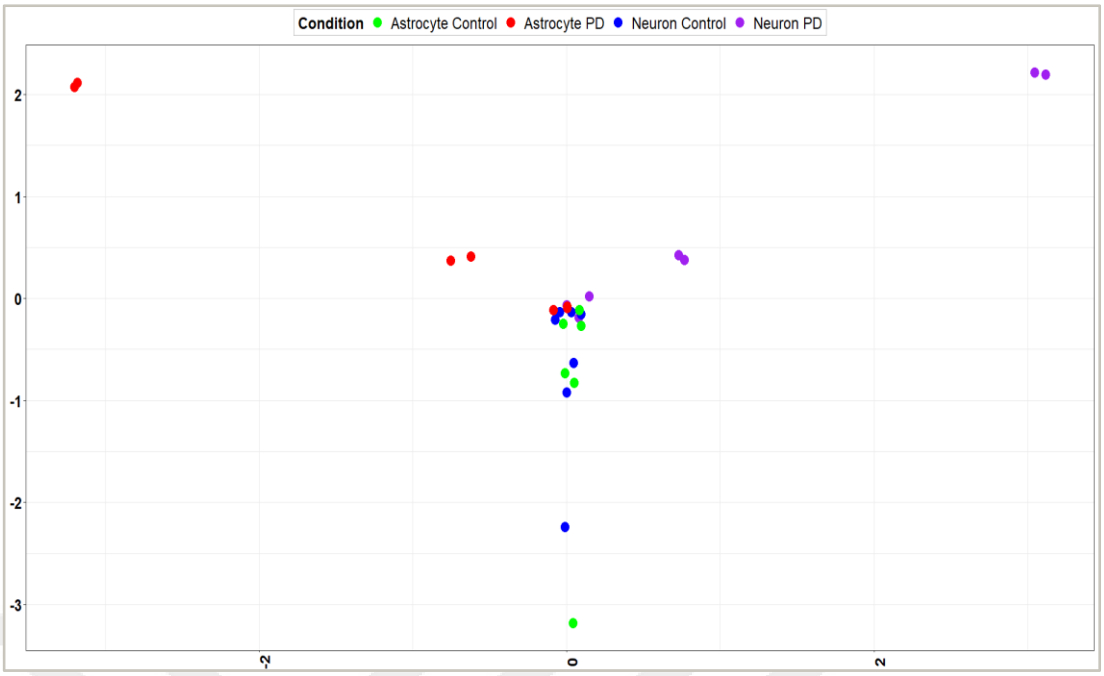

Figure 3.6: LDA results of CTS gene expression values. For PD and control groups, each cell type is represented by a different colour.

LDA results have the same problem as PCA results. There is a specific pattern in the distribution of variables, It might be an evidence of the bias in the prediction process. In nature, variables should be distributed randomly and the same types of cells should group together.

The correlation matrix of cellR-predicted CTS samples was calculated to examine variables more deeply. The correlation matrix shows relationships between variables. The correlation matrix consists of R values, which can be between 1 and -1, which is an indication of a strong or weak relationship between two vectors. 1 and -1 of R value are an indication of strong positive and strong negative relationship respectively. An R value of zero is an indication of no relationship between vectors. The correlation matrix for astrocytes and neurons are shown in Table 3.5 and 3.6 respectively.

Table 3.5: Correlation matrix for astrocyte samples predicted by cellR algorithm by using GSE110716. S indicates samples (S1 means sample 1). PD and Control indicates condition.

	Samples (S)	PARKINSON							CONTROL					
		S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**
PARKINSON	**S1**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	0.82	0.82	0.82	0.82	0.82
	S2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	0.82	0.82	0.82	0.82	0.82
	S3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	0.82	0.82	0.82	0.82	0.82
	S4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	0.82	0.82	0.82	0.82	0.82
	S5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	0.82	0.82	0.82	0.82	0.82
	S 6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	0.82	0.82	0.82	0.82	0.82
	S7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	0.82	0.82	0.82	0.82	0.82
CONTROL	**S8**	0.82	0.82	0.82	0.82	0.82	0.82	0.82	1.00	1.00	1.00	1.00	1.00	1.00
	S9	0.82	0.82	0.82	0.82	0.82	0.82	0.82	1.00	1.00	1.00	1.00	1.00	1.00
	S10	0.82	0.82	0.82	0.82	0.82	0.82	0.82	1.00	1.00	1.00	1.00	1.00	1.00
	S11	0.82	0.82	0.82	0.82	0.82	0.82	0.82	1.00	1.00	1.00	1.00	1.00	1.00
	S12	0.82	0.82	0.82	0.82	0.82	0.82	0.82	1.00	1.00	1.00	1.00	1.00	1.00
	S13	0.82	0.82	0.82	0.82	0.82	0.82	0.82	1.00	1.00	1.00	1.00	1.00	1.00

Table 3.6: Correlation matrix for neuron samples predicted by cellR algorithm by using GSE110716. S indicates samples. Parkinson and Control indicates condition.

	Samples (S)	PARKINSON							CONTROL					
		S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**
PARKINSON	**S1**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	0.25	0.25	0.25	0.25	0.25
	S2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	0.25	0.25	0.25	0.25	0.25
	S3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	0.25	0.25	0.25	0.25	0.25
	S4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	0.25	0.25	0.25	0.25	0.25
	S5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	0.25	0.25	0.25	0.25	0.25
	S6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	0.25	0.25	0.25	0.25	0.25
	S7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	0.25	0.25	0.25	0.25	0.25
CONTROL	**S8**	0.25	0.25	0.25	0.25	0.25	0.25	0.25	1.00	1.00	1.00	1.00	1.00	1.00
	S9	0.25	0.25	0.25	0.25	0.25	0.25	0.25	1.00	1.00	1.00	1.00	1.00	1.00
	S10	0.25	0.25	0.25	0.25	0.25	0.25	0.25	1.00	1.00	1.00	1.00	1.00	1.00
	S11	0.25	0.25	0.25	0.25	0.25	0.25	0.25	1.00	1.00	1.00	1.00	1.00	1.00
	S12	0.25	0.25	0.25	0.25	0.25	0.25	0.25	1.00	1.00	1.00	1.00	1.00	1.00
	S13	0.25	0.25	0.25	0.25	0.25	0.25	0.25	1.00	1.00	1.00	1.00	1.00	1.00

The correlation analysis shows that the relationship between cells in the same condition is very strong. Although the strong relationship was expected between cells in the same condition, there must be some perturbations in relationships since samples were obtained by different experiments and brain cells must not be identical since they have very complex metabolism. The relationship between cells in different conditions are the same for all cells although one would expect a random relationship between cells since biological cells have complex metabolism. There is no randomness in the results of PCA, LDA and correlation matrices. Therefore, the predicted CTS gene expression data by using GSE110716 were not seen adequate for further analysis.

cellR-based prediction of CTS gene expression values were also repeated for the other dataset, GSE68719. For astrocyte and neuron, CTS gene expression of 671 genes, which were included in iBrain671, were predicted considering PD and control groups. The results were evaluated based on PCA, LDA and correlation matrix in the same way as the results of GSE110716. The results of PCA are given in Figure 3.7.

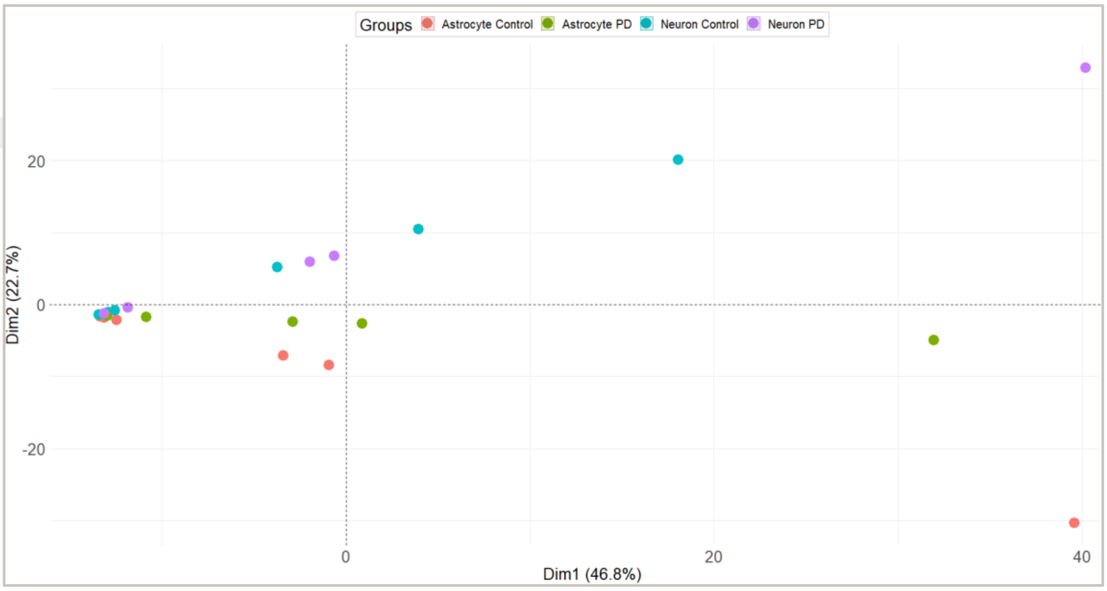

Figure 3.7: PCA results of CTS gene expression values. For PD and control groups, each cell type is represented by a different colour.

For the cellR algorithm, PCA results for CTS gene expression of cells that were predicted by using GSE68719 has the same problems as the first study that was performed with GSE110716. Distribution of variables is in a very specific pattern without randomness. Although there are 224 samples in the PCA graph, only few number of samples are seen in the graph since most of the samples overlap. Most of the cells are grouped together irrespective of cell-type and condition. LDA analysis was performed for deeper investigation of distribution of cells. LDA results are given in Figure 3.8.

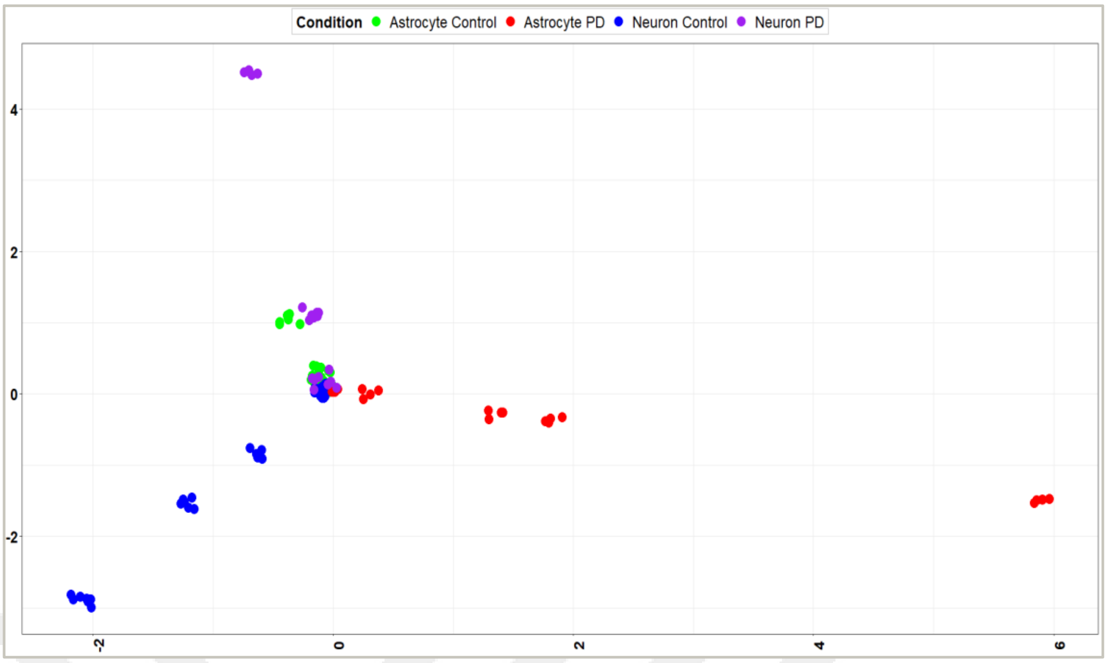

Figure 3.8: LDA results of CTS gene expression values. For PD and control groups, each cell type is represented by a different colour.

The LDA result has the same problem as the PCA result. There is a specific pattern in the distribution of variables without randomness. The same problems were encountered in the results of LDA and PCA analyses of CTS gene expressions obtained by using GSE110716 and GSE68719. The correlation matrix of CTS gene expressions obtained by using GSE68719 was calculated to examine the relationships between samples. The results show same properties as the result of first study that were performed by using cellR algorithm. All of the R values for the cells in the same conditions are 1 for the predicted astrocyte and neuron data separately. All of the R values across astrocyte samples in the different conditions is 0.92, and those across neuron samples in the different conditions are 0.8. CTS gene expressions predicted by using GSE68719 were inadequate to perform further analysis as a result of the evaluation with PCA, LDA and correlation analyses.

3.4. Deconvolution of Bulk Data by CIBERSORTx

CIBERSORTx is a web based deconvolution algorithm published in 2019 [6]. It predicts cell fractions by its own algorithm. For the estimation of CTS gene expression data, it requires a signature matrix that can be given by the user or calculated by

CIBERSORTx by using a reference scRNA-seq data. There are three analysis steps available on CIBERSORT website: create signature matrix, impute cell fractions and impute cell expressions. Creating signature matrix analysis option is used to create signature matrix by using reference scRNA-seq data. Imputing cell fractions option is used to predict cell fractions by using signature matrix and bulk gene expression data. Signature matrix can be given by user or calculated by using create-signature-matrix option. Impute-cell-expressions options is used to predict CTS gene expressions. There are two modes in this option as group mode and high resolution mode. Group mode calculates CTS gene expression of the data without sample differences. To put it simply, it calculates one CTS gene expression value for each cell type by using bulk gene expression data consisting of multiple samples. It requires signature matrix and bulk gene expression data as mandatory inputs. RNA-seq or microarray bulk gene expression data can be used. In the group mode, there are two optional inputs as grand truth and merged class files. Grand truth file, which consists of CTS gene expressions of each cell-type, is used to compare predicted and known CTS gene expressions for each cell-type. Merged class file, which consists of the names of cell-types, is used to merge some cells, such as merging excitatory and inhibitory neurons. High resolution mode is used to predict CTS gene expression for each sample. It requires bulk gene expression data and signature matrix and a subset of genes as mandatory inputs. For the high-resolution mode, a subset of genes is required since CIBERSORTx permits to predict CTS gene expression of 1000 genes at maximum. This mode also has merge class and ground truth files as optional inputs, the same as the group mode. In this thesis, the updated signature matrix used in the MIND algorithm simulations for the prediction of cell-fractions was used also in CIBERSORTx based simulations. The details of the update process is available in Section 3.2.1. GSE20295 was used to predict cell-fractions and CTS gene expression data by using CIBERSORTx.

3.4.1. Prediction of Cell-Type Fractions

CIBERSORTx predicts cell type fractions with the help of a linear set of equations. It finds factors of bulk gene expression data based on Equation 3.11 [6].

$$B = C \times S \tag{3.11}$$

In Equation 3.11, B, C and S represent bulk gene expression data, cell fractions and signature matrix respectively. C is predicted by using known parameters, and B and S are given by user. Cell-fractions were predicted for each sample in a given condition by using impute cell fraction option on CIBERSORTx web page. For brain regions, means of predicted cell fractions and standard deviations are given in Table 3.7.

Table 3.7: Means of the predicted cell fractions and their standard deviations (sd) considering brain regions and conditions.

Groups	Regions	Values	Neuron	Astrocyte	Oligo dendrocyte	Endothelial	Microglia
CONTROL	**BA9**	Mean	0.60	0.22	0.04	0.14	0.00
		Sd	0.20	0.13	0.02	0.07	0.00
	Substantia	Mean	0.31	0.08	0.24	0.34	0.02
		Sd	0.19	0.08	0.10	0.17	0.03
	Putamen	Mean	0.29	0.21	0.25	0.25	0.01
		Sd	0.16	0.17	0.11	0.12	0.01
PD	**BA9**	Mean	0.52	0.19	0.10	0.18	0.01
		Sd	0.32	0.16	0.06	0.17	0.01
	Substantia	Mean	0.19	0.03	0.34	0.41	0.02
		Sd	0.08	0.05	0.09	0.10	0.02
	Putamen	Mean	0.19	0.16	0.27	0.38	0.00
		Sd	0.16	0.08	0.11	0.16	0.00

Predicted cell-fractions by using CIBERSORTx were compared with literature information for validation. Reported cell fractions of frontal cortex were compared with BA9 region since there is no available information about cell-fraction of BA9 in the literature and frontal cortex is a very close region to BA9 (Figure 3.9).

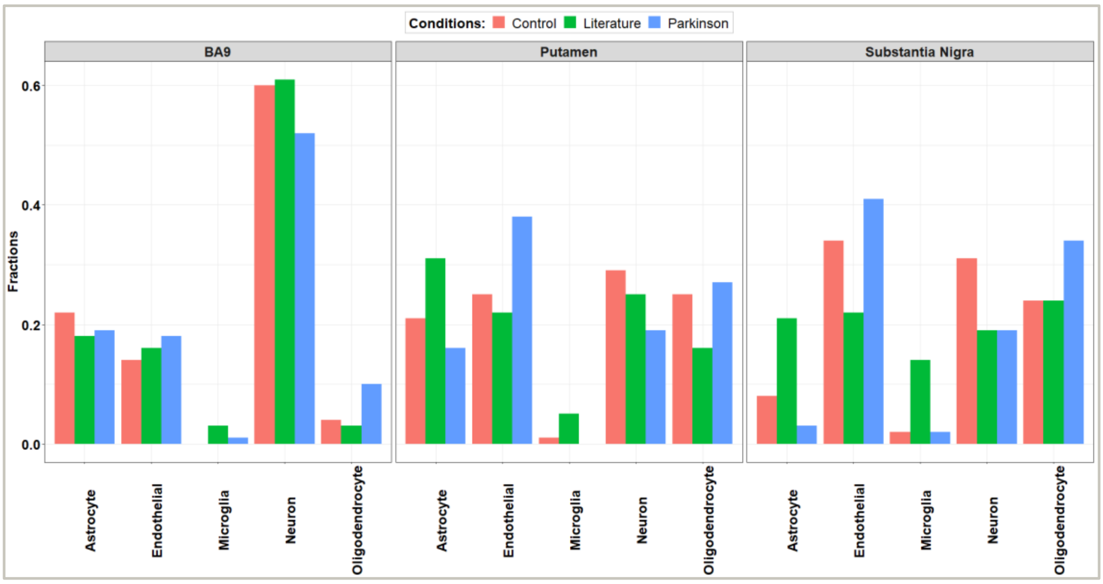

Figure 3.9: For each brain region, bar plots of cell fractions predicted by using CIBERSORTx. Red, blue and green bars represent the control group, PD group and literature information for the control group, respectively

The correlation between predicted and reported cell-fractions for each region were calculated. Pearson correlation technique was used to calculate r and p values. The correlation test results are given in Table 3.8.

Table 3.8: The results of correlation between predicted and reported cell fractions for CIBESORTx method.

	BA9	**Putamen**	**Substantia Nigra**
R	0.99	0.75	0.59
p-value	0.0004	0.14	0.28

3.4.2. Cell Type Specific Gene Expression Predictions by Using CIBERSORTx

For the prediction of CTS gene expression, the high resolution mode of CIBERSORTx was used. Details of the high resolution mode were explained in the previous section. CIBERSORTx infers cell-fractions by modelling bulk data (B), cell-fractions (C) and CTS gene expressions (G) (Equation 3.12).

$$diag\left(G_{n,k,c} \times C_{c,k}\right) = B_{n,k} \tag{3.12}$$

In Equation 3.12, CTS gene expression, which is denoted as G, consists of n genes in c cell types for k samples. In other words, it is a three dimensional matrix. C represents cell-fractions consisting of c cell types in k samples. B denotes bulk gene expression data consisting of n genes in k samples. For the prediction of CTS gene expression, CIBERSORTx uses the heuristic approach defined below. CIBERSORTx calculates CTS gene expressions of genes one by one. Firstly, it predicts CTS gene expression of gene i by using Equation 3.13, and checks whether it is significantly expressed in any cell types, and gives NA as the result if the gene is not significantly expressed in any cell-type.

$$G_{i,c} \, X \, C_{c,k} = B_{i,k} \tag{3.13}$$

In Equation 3.13, G represents CTS gene expression of gene i in cell-type c. C denotes cell-fractions of c cell types in k samples. B represents the bulk gene expression of gene i in k samples. Predicted CTS gene expressions by using CIBERSORTx algorithm contained too much NA values. All predicted samples nearly contained 30% NA values. Therefore, further analysis was not performed with the CTS gene expressions predicted by using CIBERSORTx algorithm.

3.5. Assessment of CTS Astrocyte and Neuron Gene Expression

Assessment of the results of MIND algorithm was performed via differential gene expression analysis and gene ontology (GO) analysis. The analyses were performed using only predicted CTS gene expression profiles of astrocyte and neuron cells, by excluding other cell types, since the metabolic model used for the analysis of PD consists of astrocyte and neuron cells.

3.5.1. Differential Gene Expression Analysis

The differential gene expression analysis is performed to determine genes whose expression levels are significantly changed between PD and control groups. Significance is determined based on the chosen cut off for p-value, which is the probability of an observed value extremity while accepting the null hypothesis. Mostly, it is performed by Student's t-test (t-test), which is a parametric statistical method. There are different R packages that were developed for differential gene expression analysis such as Linear Models for Microarray Data (Limma) [64]. Each of these packages is compatible with different types of gene expression data, and Limma package was used in this study since it is suited with microarray data. For PD and control groups, differential gene expression analysis was separately performed for astrocytes and neurons. Benjamini-Hochberg correction method was used to decrease false positives that are incorrectly identified as significantly changed genes between conditions. To put it simply, this method decreases the number of genes that are incorrectly identified as significantly changed. It calculates adjusted p-values called q-values by using p-values obtained by classical statistical tests. For the identification of significantly changed genes, the cut off value for q-value was chosen as 0.01. For neuron and astrocyte cells, 481 and 2473 significantly changed genes were determined respectively.

3.5.2. Gene Ontology Enrichment Analysis

Gene Ontology (GO) enrichment analysis, which compares given genes with the reference gene set by statistical techniques, is performed to obtain the profiles of over-represented genes in the given set. The analysis can be performed to retrieve a profile of molecular functions, cellular components or biological processes for genes of interest, and it gives results as a vocabulary of terms called GO terms. GO terms are given in both general and more specific terms in a structured way. For example, cellular response to chemical stimulus is a general biological GO term while cellular response to zinc ion is a more specific one. It is beneficial to examine profiles of significantly changed genes in different conditions.

For PD and control groups, GO enrichment analysis was performed to obtain affected biological processes using significantly changed genes in astrocytes and neurons separately. For neuron and astrocyte, 268 and 971 significantly affected biological process GO terms were determined respectively, and most significantly over-represented 25 GO biological process terms are given in Table 3.9

Table 3.9: Most significantly over-represented 30 GO biological process terms for astrocyte and neuron for control-PD comparison.

GO Biological Process Neuron	**GO Biological Process Astrocyte**
Regulation of dendritic spine development	SRP-dependent cotranslational protein targeting to membrane
Regulation of dendrite morphogenesis	Cytoplasmic translation
Regulation of dendritic spine morphogenesis	Antigen processing and presentation of exogenous peptide antigen via MHC class I
Positive regulation of dendrite development	Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
Positive regulation of dendrite morphogenesis	Regulation of cellular amine metabolic process
Stress response to metal ion	Regulation of cellular amino acid metabolic process
Protein refolding	Interleukin-1-mediated signaling pathway
negative regulation of inclusion body assembly	Regulation of transcription from RNA polymerase II promoter in response to hypoxia
Positive regulation of dendritic spine morphogenesis	Ribosome assembly
Positive regulation of dendritic spine development	Mitochondrial electron transport, NADH to ubiquinone
Regulation of cellular response to heat	Chaperone-mediated protein folding
Positive regulation of neuron death	Wnt signaling pathway, planar cell polarity pathway
Cellular response to cadmium ion	Response to axon injury
Stress response to copper ion	S phase
Detoxification of copper ion	Mitotic S phase
Interferon-gamma-mediated signaling pathway	Neuron projection regeneration

Table 3.9 Continued.

Synapse pruning	positive regulation of neuron death
Response to cocaine	Negative regulation of G2/M transition of mitotic cell cycle
Cell junction disassembly	Regulation of alternative mRNA splicing, via spliceosome
Positive regulation of proteasomal ubiquitin-dependent protein catabolic process	Regulation of phospholipid metabolic process
Androgen receptor signaling pathway	NIK/NF-kappaB signaling
Zinc ion homeostasis	Axon regeneration
Cellular response to zinc ion	Mitochondrial ATP synthesis coupled electron transport
Actin filament bundle assembly	Purine-containing compound catabolic process

Although GO enrichment analysis that was performed by using predicted gene expressions is promising, another analysis was performed by using bulk gene expression data of SN from GSE20295 in order to compare results. SN region was chosen to perform GO enrichment analysis since it is the most affected part of the brain during PD. GO enrichment analysis was performed by using significantly changed genes between control and PD conditions. The result of GO enrichment analysis is given in Table 3.10.

Table 3.10: GO enrichment analysis results for SN region of brain.

GO biological process complete Substantia Nigra
positive regulation of phosphatidylinositol 3-kinase activity
positive regulation of lipid kinase activity
positive regulation of lipid metabolic process
regulation of lipid metabolic process
regulation of body fluid levels
chemical synaptic transmission
anterograde trans-synaptic signaling
cellular response to growth factor stimulus
trans-synaptic signaling
response to growth factor

4. ANALYSIS OF GMN MODEL

4.1. Brain Specific GMN Model iBrain671

iBrain671, which consists of astrocyte and neuron cells, is a more comprehensive version of iMS570 brain-specific GMN model that was published in 2014 [5, 65]. It consists of 812 metabolites, 994 reactions and 671 genes, and the reactions are controlled by genes via GPR rules that explain relationships between reactions and genes. Defining GPR rules in the model enables mapping transcriptome data to the model. iBrain671 contains key pathways of brain metabolism such as glycolysis, TCA cycle, metabolism of lipids, mitochondrial biogenesis, pentose phosphate pathway and metabolism of amino acids. For the analysis of the metabolism of PD, iBrain671 is very suitable since it consists of astrocyte and neuron that contain main biological pathways affected by neurodegenerative diseases. In the model, there are 448 and 413 internal reactions occurring within cytoplasm and mitochondria of astrocyte and neuron respectively. The model includes 22 intercellular reactions that enable metabolite exchanges between astrocyte and neuron, such as glutamate-glutamine cycle. Astrocyte and neuron respectively contain 46 and 44 exchange reactions, which define the interactions between the model and the extracellular environment. Astrocyte and neuron can uptake or secrete metabolites via these exchange reactions. In constraint-based modelling approaches, exchange reactions are used to constrain the GMN model in order to predict a flux distribution. The bounds of the exchange reactions are mostly determined based on experimental measurements in order to predict a meaningful condition-specific flux distribution. Lower and upper bounds in an FBA framework define minimum and maximum flux values that reactions can take, and they help obtaining flux distributions via constraining GMN model. To introduce a measured value to a GMN, the lower and upper bounds of the corresponding exchange reaction is set to the measured value.

In iBrain671, for the simulation of resting-state fluxes, the exchange limits of the cerebral glucose demand rate ($CMR_{Glucose}$) was set to 0.32 μmol/g/min, and it was assumed that astrocyte and neuron share glucose equally based on literature information [66–69]. CMR_{O2} was set to 1.76 μmol/g/min in total, and 70% of total oxygen was allocated to neuron. Most of the oxygen is used by neuron to meet high energy demand [66, 69, 70]. In the human brain, the ratio between oxygen uptake and carbon dioxide release (O2/CO2) was reported to be between 0.91 to 1 in the literature [71]. For astrocyte and neuron, carbon dioxide release rates (CMR_{CO2}) were set to 0.481 μmol/g/min and 1.232 μmol/g/min respectively considering the study of Çakır et al. in 2007 [72]. By considering literature information, the exchange bounds of amino acid uptake rates were defined based on the brain metabolism of elderly population since transcriptome data used in this study were obtained from old people [73]. The exchange limits of the GMN model compiled from literature and used for the simulation of resting-state fluxes are given in Table 4.1.

Table 4.1: The limits of the exchange reactions of iBrain 671. The lb and ub represent lower and upper limits that the flux values of the reactions can take, respectively. The units of the values are μmol/g/min.

Rate	lb	ub	Rate	lb	ub
$CMR_{Glucose}^{A}$	0.16	0.16	**Tyrosine**	-0.0023	0.001
$CMR_{Glucose}^{N}$	0.16	0.16	**Tryptophan**	-0.0012	0.0044
CMR_{O2}^{A}	0.53	0.53	**Lysine**	-0.014	0.013
CMR_{O2}^{N}	1.23	1.23	**Phenylalanine**	-0.0016	0.0059
CMR_{CO2}^{A}	0.48	0.53	**Histidine**	-0.0049	0.0012
CMR_{CO2}^{N}	1.12	1.23	**Ammonia**	0.0035	0.0035
$Cysteine^{A}$	-0.0049	0.0063	**Ornithine**	-0.011	0.0031
$Leucine^{A}$	0.0016	0.01	**Methionine**	-0.0008	0.0029
$Isoleucine^{A}$	0	0.0055	**Threonine**	-0.0008	0.011
$Valine^{A}$	-0.01	0.0078	**$Biomass^{A}$**	1.623×10^{-6}	1000
$Arginine^{A}$	0.0016	0.0051	**$Biomass^{N}$**	1.727×10^{-6}	1000

4.2. Generating Disease-Specific Metabolic Model

Context-specific GMNs are generated to simulate altered metabolisms under certain conditions such as metabolism of PD. They are produced from a generic model by mapping transcriptome data. Generic metabolic models are organism-specific, and they are reconstructed by adding as much reactions as possible for the given organism or tissue. For the generation of context-specific models, inactive reactions are removed from the generic models by considering gene activity information derived from transcriptome data.

There are various algorithms developed to generate context-specific GMN models, and one of the most common ones, the GIMME algorithm, was used in this study to create a disease-specific GMN. GIMME algorithm decides on active and inactive reactions based on two parameters: threshold and objective fraction parameters. Objective fraction parameter is adjusted to generate context-specific GMN models with desired functionality. GIMME algorithm treats reactions with a transcriptome-based score under the given threshold as inactive, unless removal of those reactions violates the desired functionality. Threshold value for gene activity was chosen as 30th percentile of entire transcriptome data and the bottom 30th percentile of the entire transcriptome data was assumed as inactive genes. The objective fraction parameter, which represents maximum allowed reduction in the objective function value after the removal of inactive reactions, was set to 0.2. Glutamate/glutamine/GABA cycles were chosen as the objective function of the model since this objective function was shown to predict resting state brain metabolic fluxes accurately [5]. A small value was chosen for the objective function parameter since the activity of glutamate/glutamine/GABA cycles between neuron and astrocyte is known to be hampered during PD. For a better simulation of PD metabolism, upper and lower bounds of exchange reactions were chosen in a wide range. Objective fraction parameter ensures that the generated context specific GMN model has active flux through its respective objective function. Despite that, defining a high objective fraction value along with a highly constrained GMN model might lead to a reduction of effectiveness of transcriptome data by preventing removal of lowly scored reactions to achieve minimum objective function value which is defined directly by this objective fraction parameter. The metabolism of the brain changes during PD, and thus leads to shifting of metabolite exchange rates. Therefore, determination of exchange limits of metabolites in a wide range before generating the context-specific GMN model is important. Upper and lower bounds of the exchange reactions of amino acid uptakes were chosen as 0.05 and -0.05 respectively since amino acid metabolite uptake rates of healthy elderly people roughly oscillate between 0.05 and -0.05. Before generating context-specific GMN model for simulating PD metabolism, setting wider ranges for glucose and oxygen uptake rates is important to account for alterations in those rates in PD metabolism since they are known to change. To introduce a wide range, the bounds of the oxygen and glucose exchange fluxes in iBrain671 were multiplied with 2 and 0.5 of the resting state values (Table 4.2) respectively since there is no available wet-lab experiments information of exchange rates of oxygen and glucose in PD condition. The exchange limits of CO2 were also multiplied with 2 and 0.5 since it is a natural product of oxygen metabolism. For astrocyte and neuron in the model, the reaction scores, which is an indication of available enzymes or enzyme complexes, were calculated separately by using CTS gene expression data of astrocyte and neuron predicted by MIND algorithm. With all these parameters and reaction scores, a PD-specific model with 906 reactions was generated by using GIMME algorithm.

4.3. Constraint-Based Analysis

MOMA was used in this study to obtain flux distribution of PD-specific GMN model. MOMA predicts flux values by minimizing Euclidean distance between reference and context-specific flux distribution. Reference (healthy resting state) flux distribution was obtained by standard FBA. Standard FBA requires an objective function to obtain flux distribution, and maximization of the rates through Glutamate/glutamine/GABA cycles was chosen as the primary objective function. The constraints introduced in Table 4.2 were used in the simulations.

The constraint-based modelling approach is a combination of mathematical techniques that optimize hundreds of mass balance derived equations written based on reactions inside organism of interest. In constraint-based metabolic models, alternate optima can be an issue in interpreting the results of FBA since there might be multiple flux distributions that result in the same value for the objective function (Figure 4.1). There are multiple approaches to overcome this issue, such as sampling-based methods or minimization of the sum of all flux values [74, 75]. Sampling-based methods focus on sampling the solution space to obtain a high number of (eg. 1000) flux distributions and subsequent comparison of the distribution of flux values for each reaction. In this thesis study, minimization of the sum of all flux values was chosen to prevent alternate optima. The principle of this method was proposed based on an evolutionary process that can be defined as the accomplishment of the goal with minimal resource expenditure since the flux values are an indication of the amount of depletion of resources [75], and it was successfully applied to brain metabolic networks before [5, 76].

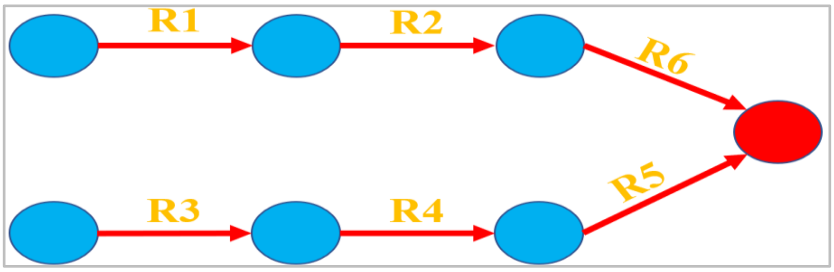

Figure 4.1: Representation of flux variability on a toy model. The red lines with arrow and R1,R2,R3,R4,R5 and R6 are indication of irreversible reactions and reaction names respectively. Circles represent metabolites. The flux variability will be an issue if maximization of the red metabolite were chosen as the objective function since the red metabolite can be produced through R1, R2 and R6 as well as R3, R4 and R5 reaction chains.

Constraint-based analysis was performed with all aforementioned objective functions and constraints to predict flux distributions of healthy and PD-specific model. Predicted flux values for some key metabolic events are given in Table 4.2.

Table 4.2: Predicted flux values of healthy-state and PD-specific GMN model, and experimental flux values for healthy human subjects. The unit of the values are μmol/g/min.

Percentage of Flux Ratio	**Healthy model**	**PD-specific model**	**Experimental values for healthy people**
Lactate production flux with respect to $CMR_{Glucose}$	11.47	15.01	3-9 [5]
Glutamate-glutamine cycle flux with respect to $CMR_{Glucose}$	74.67	66.25	40-80 [5]
Relative oxidative metabolism of astrocyte to total oxidative metabolism	28.01	39.30	30 [5]
Total PPP flux with respect to $CMR_{Glucose}$	5.5	0.0033	3-6 [5]
Pyruvate carboxylase flux with respect to $CMR_{Glucose}$	9.38	12.65	10 [5]

The activities of some potentially PD related pathways were additionally examined by comparing the total flux activity in the healthy and PD-specific conditions. Total flux activity was calculated by summing the absolute values of fluxes that are related to the metabolic pathways of interest. Metabolic pathways and their flux activities in control and PD conditions are given in Table 5.3.

Table 4.3: Metabolic activities of pathways in PD and control conditions. PPP, TCA cycle and ROS indicate pentose phosphate pathway, citric acid cycle and reactive oxygen species respectively.

	Helthy Condition		**PD Condition**	
	Astrocyte	**Neuron**	**Astrocyte**	**Neuron**
Glycolysis	2.55	2.81	2.3	2.44
PPP	0.048	0.04	0	0
TCA Cycle	0.91	3.69	1.03	3.56
Oxidative Phosphorylation	3.96	7.6	3.56	7.49
Alanine Pathway	0.015	0.017	0	0.003
Glycine Pathway	0.36	0.002	0.04	0.01
Lysine Pathway	0	0.13	0.39	0
Arginine Pathway	0.076	0.43	0.042	0.18
ROS pathway	0.079	0.88	0.16	0.96

4.4. Metabolite-Metabolite Interactions

Alterations in the metabolic activities reflect the effect of diseases on metabolism. Metabolites connect to each other since they are converted to one another during metabolic processes. Perturbations in the activity of one metabolite have a potential to affect all metabolic activities. In another perspective, it can be said that constraint-based modelling is the investigation of metabolite interactions. In this thesis study, a new graph-based tool, createMetIntrcNetwork(), was developed to examine metabolite interactions in a better perspective. The tool creates a directed or undirected network of metabolite-metabolite interactions based on given inputs. It has two mandatory inputs as GMN model and metabolite list, and twelve optional inputs as fluxes, Graphtitle, excludedMets, nodeColour, Hnodecolour, scaleMin, scaleMax, nodeSize, HnodeSize, arrowSize, threshold and excNodesWithDeg. If flux values of a GMN model are given, it can generate directed metabolite-metabolite interaction network with adjusted thickness and colour of edges based on given fluxes. Desired metabolites can be excluded from the graph by excludedMets parameter via metabolite names and by excNodesWithDeg via the degrees of metabolites. scaleMin and scaleMax parameters were added to generate comparable networks by scaling colorbar and colour of edges. The threshold parameter was added to exclude edges (fluxes) based on a threshold. Graphtitle, arrowSize, nodeColour, Hnodecolour, nodeSize and HnodeSize were added to customize generated figures. This tool was written in MATLAB language.

Pyruvate, which is the end product of glycolysis and the main fuel of the TCA cycle, is a key metabolite in bioenergetics [77]. Generating pyruvate centered metabolite-metabolite interaction network can give an idea about metabolic characteristics of PD since disruption of bioenergetics is well-known characteristics of PD, especially in neurons. Therefore, metabolite interactions of cytosolic pyruvate were examined by using createMetIntrcNetwork() function. The network was generated by using flux information from the previous section, and colorbar that represents flux values was scaled between 0-0.32 to generate a comparable graph between PD and control conditions since the biggest flux value inside the network was nearly 0.32. Thus, flux values of edges can be represented by a colour that has the same meaning in both conditions. For neuron, pyruvate centered metabolite-metabolite interaction network is given in Figure 4.2.

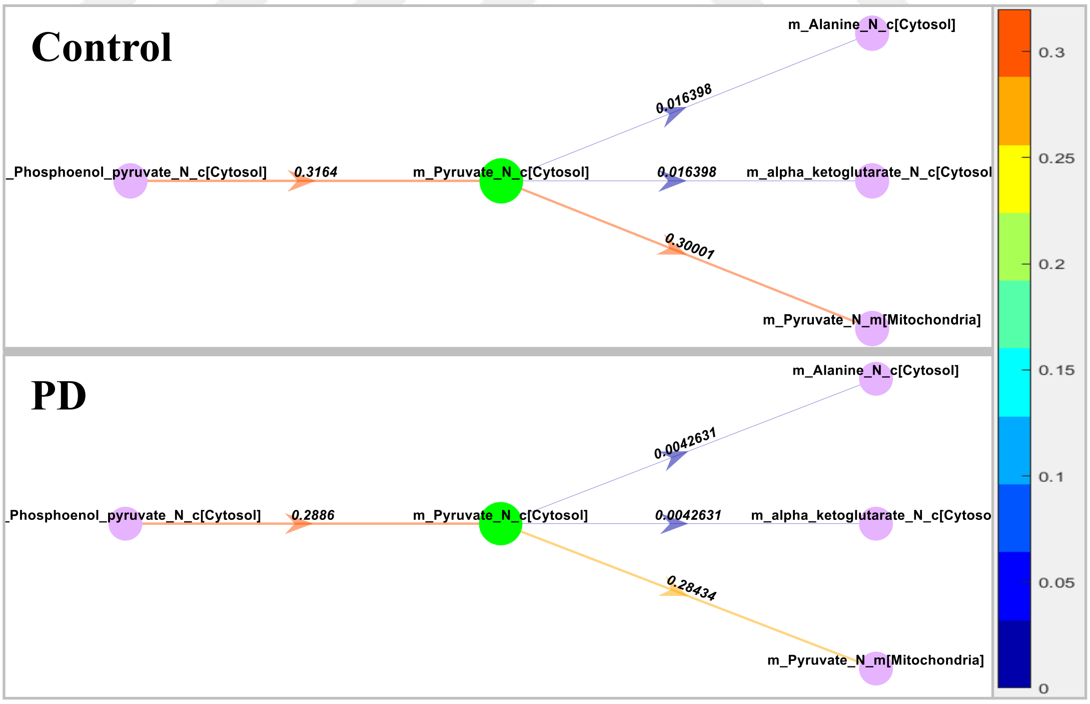

Figure 4.2: For PD and control conditions, pyruvate centered metabolite-metabolite interaction network for neuron. The colour of edges represents flux values that are defined in colorbar. The values on the edges represent flux values. The direction of arrows indicates the direction of reactions. The values on the colorbar describe which colour represents which flux value.

Tyrosine is a precursor metabolite of L-dopa, which is the precursor metabolite of the dopamine, and dopaquinone which is the precursor metabolite of melanin. Dopamine and melanin levels are changed during PD, and there is a competition between the production of dopamine and melanin in neurons. Production of dopamine is decreased during PD while the production of melanin increases [78]. Therefore, tyrosine centered metabolite-metabolite interaction network was generated to examine this phenomenon. Tyrosine centered metabolite-metabolite interaction network is given in Figure 4.3. L-dopa production is not seen in the PD section of the graph since edges with zero flux were excluded.

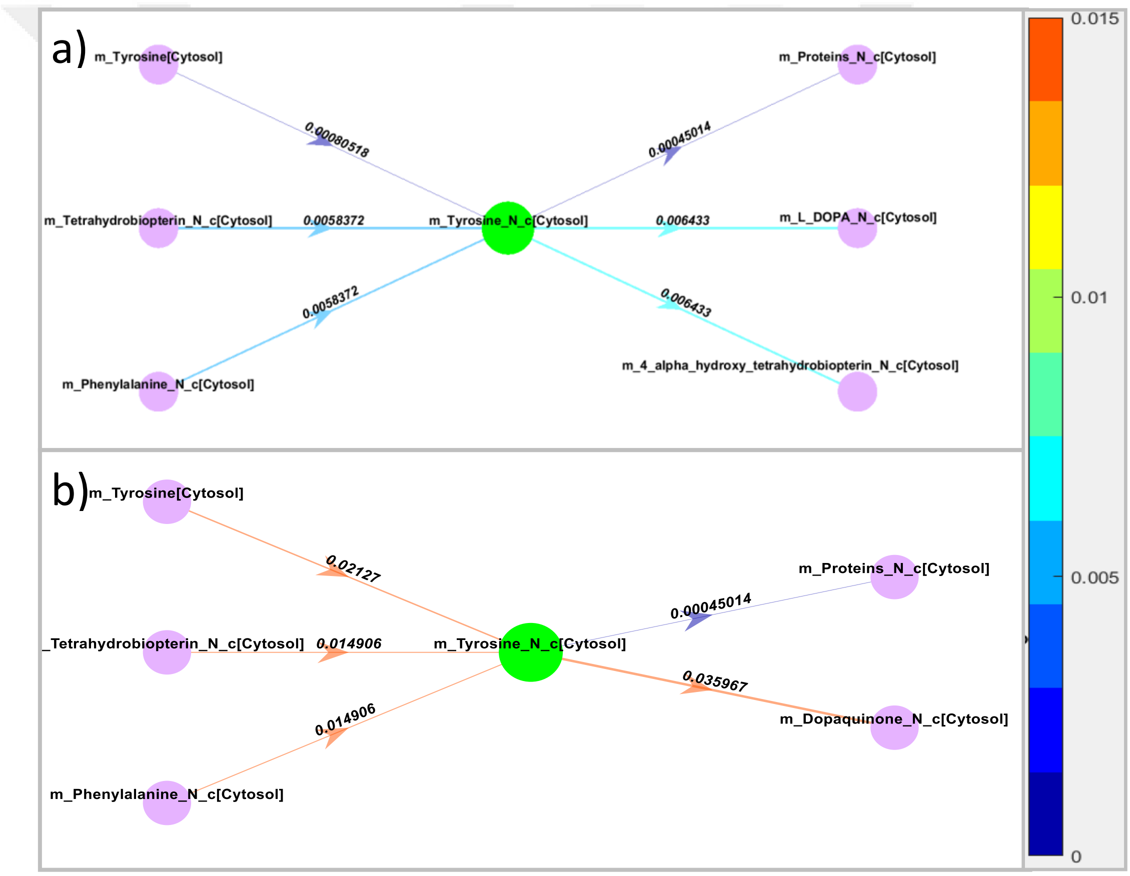

Figure 4.3: For PD and control conditions, tyrosine centered metabolite-metabolite interaction network for neuron. a and b represents control and PD respectively. The colour of edges represents flux values that are defined in colorbar. The values on the edges represent flux values. The direction of arrows indicate the direction of reactions. The values on the colorbar describe which colour represents which flux value.

5. DISCUSSION

PD is a complex disease caused by genetic and environmental factors, and it strongly affects the metabolism of brain cells, primarily dopaminergic neurons, by shifting the activities of metabolic pathways. Systems biology approaches might have the potential to elucidate the effect of PD on the metabolism of brain cells [79]. Integrative analysis of GMN models, which is one of the most widely used systems biology approach for the investigation of metabolic alterations in diseases, can bring affected metabolic pathways to light since GMN models are reconstructed based on the active metabolic pathways in the organism of interest [4]. Tissue-specific GMN models, which were created based on active metabolic pathways of the tissue of interest, have an advantage in elucidating disease metabolism since most of the diseases mainly affect a particular tissue.

The metabolic alterations of astrocyte and neuron are different during PD since the effects of PD are seen at the cellular level. Therefore, the analysis of PD metabolism at the cellular level by taking into account CTS expression data might help to determine the mechanism of PD more precisely. Even if deconvolution of bulk gene expression data methods and scRNA-seq technique are at the beginning of their development process, they offer a great advantage in the observation of metabolic alterations at the cellular level. Analysis of GMN model consisting of different cell-types by using CTS gene expression data might help to discover metabolic perturbations at the cellular level, unlike the analysis by using bulk gene expression data that cannot provide cell-type specific resolution.

In this study, condition-specific GMN models were created by integration of predicted CTS gene expression data with the brain-specific GMN model iBrain671 to analyze PD metabolism. This is the first study in literature that uses predicted CTS gene expression data in GMN-based flux prediction. iBrain671 enables mapping of CTS gene expression data since it consists of two cell types, astrocyte and neuron, and their metabolic interactions. Usage of more specific data and GMN models increase the resolution of the integrative metabolic modelling approach and potentially give better insight on molecular mechanisms of PD from a metabolic perspective. CTS gene expression data were predicted by using deconvolution algorithms since there was no publicly available CTS gene expression data of PD.

Three different deconvolution algorithms, MIND, cellR and CIBERSORTx, were used in this study. cellR requires count RNAseq data unlike MIND and CIBERSORTx algorithms. The mandatory inputs of MIND and cellR algorithms are signature matrix and bulk gene expression data, while cellR requires GTEx file and reference scRNA-seq data besides bulk gene expression data. The MIND algorithm requires very special gene expression data such that gene expression values must be taken from more than one tissue per subject. MIND and cellR algorithms can be used by using R platform but CIBERSORTx is a web-based tool that requires registration. cellR predicts CTS gene expression data very slowly compared to MIND and CIBERSORTx.

Assessment of predicted CTS gene expression data was conducted via GO enrichment analysis since there were no publicly available CTS gene expression data of PD for the comparison. For PD and control groups, GO enrichment analysis was performed for astrocyte and neuron separately by using significantly changed genes. Predicted CTS gene expression data by MIND algorithm were accepted as meaningful since GO enrichment analysis gives PD related ontology terms.

5.1. Prediction of CTS gene expression data

Prediction of the cell-type fraction is the first step of the prediction of CTS gene expression data. Most of the bulk data deconvolution algorithms require cell-type fractions to predict CTS gene expression data. Prediction of cell-type fractions is crucial since bulk gene expression data deconvolution algorithms are statistical techniques that are not robust against cell-type fractions. A signature matrix is used for the prediction of cell-type fractions, and the choice of signature matrix highly affects the results. MIND algorithm expects a user-defined signature matrix while cellR extracts it from reference scRNA-seq data by using its algorithm. In this study, the signature matrix was taken from literature to predict CTS gene expression data by using MIND and CIBERSORTx algorithms. The genes that do not have enough discrimination power between cell types were removed from the signature matrix, and it was seen that fraction of cell-types became more accurate compared to the literature. For cellR, the signature matrix was generated by using Deconvolution() function. It produces the signature matrix from based on reference scRNA-seq that were retrieved from literature [62].

5.1.1. MIND Algorithm

CTS gene expressions were predicted by using MIND algorithm and GSE20295. The results of predicted cell-type fractions were compared with the information from MIND algorithm's paper since it is about predicting CTS gene expression of the healthy brain considering the difference in brain regions, and most of the results are consistent with literature (Figure 3.1). It was seen that the fraction of neurons decreases in PD condition as expected since death of neurons is the main pathological characteristic of PD [7]. The biggest decrease of fraction of neuron was observed in the SN region of the brain. Therefore, this is another promising result since the mainly affected region of the brain is SN in PD, where about 70% of neurons are already lost at the time of diagnosis.

CTS gene expression data were obtained by using predicted cell-type fractions and bulk gene expression data that were retrieved from publicly available database GEO. The predicted CTS gene expression data were evaluated in two steps, (i) PCA approach to see if the predicted samples from different groups are grouping together, (ii) GO enrichment analysis of significantly changed genes between control and PD groups, considering astrocyte and neuron separately.

5.1.1.1. Assessment of Predicted CTS Gene Expressions

PCA is a powerful tool to observe distribution of patterns in data. It was used to check condition-specific grouping of each cell-type. The result of PCA is promising since each cell-type is grouped together (Figure 3.3). Therefore, GO enrichment analysis was performed to evaluate whether predicted CTS gene expression data shows the metabolic properties of PD.

GO enrichment analysis is a powerful tool to observe mostly affected metabolic pathways between different conditions. The aim of performing GO enrichment analysis is to investigate differences in the patterns of predicted CTS gene expression in control and PD conditions. The differences in the patterns of gene expression data can indicate affected metabolic pathways in PD condition compared to the control group since gene expression is directly related to metabolic pathways. The results of GO enrichment analysis are promising since most of the GO terms were related to PD (Table 3.9). The first five GO terms for neuron are about dendrites, and it was reported that there is pathological and morphological changes in dendrites during PD [80]. One of the main pathological features of PD is the accumulation of protein deposits called LB as a result of dysfunction in protein refolding mechanisms [9]. Protein refolding, negative regulation of inclusion body assembly, regulation of inclusion body assembly and cell junction disassembly terms, all of which were obtained as a result of GO enrichment analysis, are directly related to LB pathology. Although most of the PD studies are related to neurons, GO enrichment analysis for astrocyte were able to predict known dysfunctions in some metabolic processes that were reported for astrocyte in the literature such as antigen processing [81]. As a result, GO enrichment analysis shows that MIND-based predicted CTS gene expression values have potential to reflect pathology of PD at cellular level.

GO enrichment analysis was performed by using bulk gene expression data and the results were compared with the results of GO enrichment analysis that were obtained by using predicted CTS gene expression data (Table 3.9 and 3.10). It was seen that CTS gene expression data better reflect the metabolic alterations of the brain during the PD at the cellular level compared to bulk gene expression data. For example, CTS specific gene expression data of neuron reflects the main pathological feature of PD, which is protein refolding process, but this kind of precise results could not be seen in the results of GO enrichment analysis from bulk gene expression data.

5.1.2. cellR Algorithm

Prediction of CTS gene expression data of gyrus and BA9 regions of brain was performed by using cellR algorithm using two transcriptomic datasets, GSE68719 and GSE110716. GSE68719 and GSE110716 are RNA-seq data from BA9 and gyrus region of the brain respectively. Firstly, cell-type fractions were predicted (Table 3.2). The results were not good compared to MIND algorithm. Secondly, CTS gene expression data were predicted, and the results were evaluated with PCA, LDA and correlation matrix. Firstly, PCA was performed to evaluate predicted CTS gene expressions (Figure 3.3 and 3.5). PCA results were very poor since the distribution of the samples in the graph shows a specific pattern without randomness for both transcriptome data. Brain cells are complex biological systems, and they should not be distributed in a specific pattern such as being distributed through a line. LDA was performed for a deeper analysis of the distribution of cells (Figure 3.4 and 3.6). It is a supervised machine learning algorithm that can identify the distribution patterns of variables. The results of LDA show the same problematic behavior as PCA results for both datasets. Lastly, a correlation matrix was calculated to observe relationships of samples based on the predicted CTS expressions. Results show that all of the R values that were calculated by comparing samples of same cell-types in the same condition are 1, and the calculated R values across the samples of different conditions are the same for all samples (Table 3.5 and 3.6). The randomness were not seen in the correlation matrix. After the evaluation process, CTS gene expression values that were predicted by using cellR algorithm were not seen to be adequate for further analysis.

5.1.3. CIBERSORTx

CTS gene expression data were predicted by using GSE20295, which consists of BA9, PT and SN regions of the brain, and CIBERSORTx algorithm. It is a web-based tool that requires signature matrix and bulk gene expression data to predict cell-type expressions. The same signature matrix was used for both MIND and CIBERSORTx. Firstly, predicted cell-type fractions were evaluated (Table 3.6 and Figure 3.7). CIBERSORTx showed a considerably poorer performance about predicting cell-type fractions compared to MIND algorithm. Secondly, the predicted CTS gene expressions were evaluated. It was seen that the predicted CTS gene expressions contain too much NA values. Therefore, CTS gene expression data predicted by using CIBERSORTx were not used for further analysis.

5.2. Constraint-based Analysis

MOMA and FBA methods are powerful tools for the analysis of metabolism. Such constraint-based methods are becoming very promising as technology develops. Simulating disease metabolism in silico has some advantages, such as being cost effective and the usage of computational power. Comparison of fluxes predicted by PD-specific and healthy GMN models shows alterations in metabolic pathways. One such alteration predicted by the constraint-based analysis in this thesis study is the increase in the production of lactate, which is compatible with literature (Table 4.2). Raquel and her coworkers generated a PD model with mouse embryonic cortical neurons, and they detected increase in lactate production caused by shifting in glucose metabolism [82]. Neurons and glial cells are in a tight relationship in brain metabolism in order to carry out their functions [83]. A decrease was predicted in glutamate/glutamine/GABA cycle flux rate by the constraint-based analysis (Table 4.2). Decreased glutamate/glutamine/GABA cycle flux rate is one of the expected results since it is one of the main indicators of the tight relationship between astrocyte and neuron [5]. Alteration in the metabolism of brain cells, especially neurons, might disrupt the relationship between astrocyte and neuron in PD. The predictions showed a decrease in the proportion of oxidative metabolism of astrocyte to all oxidative metabolism (Table 4.2). One of the well-known phenomena in PD metabolism is mitochondrial dysfunction in neurons [20]. The shift in the bioenergetic metabolism of neurons leads to decrease in their percentage in total oxidative metabolism. Another reported phenomena of PD is the dysfunction of PPP metabolism, and it was detected as a result of the constraint-based analysis in this study (Table 4.2) [84].

The alterations in the activities of important metabolic pathways were analyzed by using the results of constraint-based analysis (Table 4.3). The activities of the pathways were calculated by summing the absolute values of the flux of reactions in the pathway of interest. For astrocyte and neuron, the metabolic activities of pathways in healthy and PD conditions were compared. It was seen that activities of glycolysis, PPP, oxidative phosphorylation, alanine and arginine pathway were decreased, while the activity of ROS pathway increased, for both neuron and astrocyte. On the other hand, activities of TCA cycle, glycine and lysine pathways were decreased for neuron while increased for astrocyte. Two well-known pathological features of PD were successfully captured as a result of metabolic pathway activity analysis. One feature is the decrease in the activities of the oxidative pathway during PD, which was previously reported in the literature [85]. The second feature is increase in the metabolic activities of ROS during PD, which was also previously reported [17].

5.2.1. Metabolite-Metabolite Interaction Analysis

Analysis of the interactions between metabolites may have the potential to elucidate the origin of perturbations of metabolic activities. Generating metabolite-metabolite interaction networks from flux distributions is a convenient way to analyse metabolic interactions specific to the studied condition since all interactions can be visually tracked. There was no available practical tool for generating metabolite-metabolite interaction networks from flux distribution data in literature to our knowledge. Therefore, a practical tool, createMetIntrcNetwork(), which can generate metabolite-metabolite interaction network very easy and fast, was generated in this study. The tool has many practical features that can help users modify a metabolic network and generate functional networks. Some of the key metabolites of PD metabolism were investigated by using createMetIntrcNetwork ().

Pyruvate, which is one of the key metabolites in bioenergetics, were investigated by metabolite-metabolite interaction network created by createMetIntrcNetwork() tool (Figure 4.2). Perturbations of bioenergetics of neuron were reported in the literature [19]. It was seen that the production of pyruvate in the neuron is decreased in PD compared to the control condition. This may lead to a decrease in energy production of neurons during PD. Tyrosine, which is the precursor metabolite of melanin and L-dopa, is one of the key metabolites in brain metabolism. There is a phenomenon called melanin steal, which can be described as the increase in the production of melanin from tyrosine in PD compared to L-dopa production. Decrease in the production of L-Dopa leads to dopamine deficiency in neurons [78]. Therefore, tyrosine centered metabolite-metabolite interaction was generated to observe this phenomenon, and it was seen that there is no production of L-Dopa from tyrosine in Neuron in the PD condition based on the predicted fluxes (Figure 4.3).

6. CONCLUSIONS

Constraint-based analysis of GMN models is known to be powerful in mimicking cell metabolism in silico. Transcriptome integrated constraint based analysis of GMN models enable the analysis of various conditions including disease metabolism. Predicted CTS gene expression data were used in this study as a novel approach. Predicted CTS gene expression values were evaluated with PCA and GO enrichment analysis. It was concluded that predicted CTS gene expressions by using MIND algorithm were meaningful. All MIND-predicted cell type samples showed grouping in the PCA result in accordance with the particular conditions and their types. Results of GO enrichment analysis is consistent with shifting in brain metabolism during PD. MIND algorithm has shown its potential to predict CTS gene expression data. CIBERSORTx and cellR were shown to exhibit poor performance, unlike MIND algorithm.

Constraint-based analysis was performed with PD-specific GMN models that were generated by mapping CTS gene expression data that were predicted by using MIND algorithm. The main alterations in brain metabolism during PD were analyzed, and promising results were identified. Shifting in glucose metabolism and decrease in oxidative metabolism in neurons were the mechanisms captured by the comparison of predicted fluxes of healthy and PD conditions. Increase in lactate production in the PD condition indicates shifting in glucose metabolism. Decrease in the percentage of neurons in total oxidative metabolism indicates bioenergetics disruption.

Systems biology approach is a relatively new field that is developing rapidly. Analysis of diseases by using GMN models is a promising approach, and more specific GMN models were reconstructed in recent years such as iBrain671, which is a brain-specific GMN model consisting of two different cell-types. This thesis study aimed proof of concept that a condition-specific GMN model can be obtained by mapping CTS gene expression data predicted by deconvolution algorithms on a generic GMN model.

All of the deconvolution algorithms require cell-fractions, which is predicted by using a signature matrix, to predict CTS gene expression data. The signature matrix should be chosen carefully since the prediction of cell-fractions is highly dependent on the signature matrix. In this thesis study, to choose a marker gene, the threshold for the fold change was set to be ten times greater than the sum of all expression levels in other cell types, compared to a particular chosen cell type, and it was seen that the determination of high threshold for the fold change gives better results. A signature matrix can be obtained by using scRNA-seq data or literature information. If the necessary information is not available in the literature, usage of more than one scRNA-seq data to determine marker cells might help to eliminate false positives since the scRNA-seq technique is still evolving and there can be some errors in the expression values. Genes that are expressed much higher in a particular cell-type compared to the other cell types should be chosen as marker genes. Usage of fold change besides statistical testing might help to obtain a signature matrix that represents all cell-types well enough. The fold change can be used to eliminate falsely identified marker genes. Context-specific GMN model generation process is very important for the analysis of disease metabolism by using constraint-based analysis. The threshold value and objective function fraction parameter can be optimized based on transcriptome data and GMN model if the GIMME method is used to produce context-specific GMN. Individual-based analysis can be performed by producing a context-specific GMN model for every subject and statistical analysis can be performed to determine significantly changed metabolic activities. Another important parameter for the FBA method is the objective function. Different kind of objective functions that might explain PD condition can be chosen to increase the performance of the constraint-based analysis.

REFERENCES

[1] Gitler A. D., Dhillon P., Shorter J., (2017), "Neurodegenerative disease: Models, mechanisms, and a new hope". DMM Dis. Model. Mech. 10:499–502.

[2] Dawson T. M., Dawson V. L., (2003), "Molecular Pathways of Neurodegeneration in Parkinson’s Disease". Science (80-.). 302:819–822.

[3] Poewe W., Seppi K., Tanner C. M., Halliday G. M., Brundin P., Volkmann J., Schrag A. E., Lang A. E., (2017), "Parkinson disease", Nat Rev Dis Prim, 3, 1–21.

[4] Zhang C., Hua Q., (2016), "Applications of genome-scale metabolic models in biotechnology and systems medicine". Front. Physiol. 6:413.

[5] Sertbaş M., Ülgen K., Çakır T., (2014), "Systematic analysis of transcription-level effects of neurodegenerative diseases on human brain metabolism by a newly reconstructed brain-specific metabolic network", FEBS Open Bio, 4, 542–553.

[6] Newman A. M., Steen C. B., Liu C. L., Gentles A. J., Chaudhuri A. A., Scherer F., Khodadoust M. S., Esfahani M. S., Luca B. A., Steiner D., Diehn M., Alizadeh A. A., (2019), "Determining cell type abundance and expression from bulk tissues with digital cytometry", Nat Biotechnol, 37(7), 773–782.

[7] Dauer W., Przedborski S., (2003), "Parkinson’s disease: Mechanisms and models". Neuron 39:889–909.

[8] Ledonne A., Mercuri N. B., (2017), "Current concepts on the physiopathological relevance of dopaminergic receptors", Front Cell Neurosci, 11, 27.

[9] Stefanis L., (2012), "α-Synuclein in Parkinson’s disease", Cold Spring Harb Perspect Med, 2(2).

[10] Burke R. E., Dauer W. T., Vonsattel J. P. G., (2008), "A critical evaluation of the Braak staging scheme for Parkinson’s disease". Ann. Neurol. 64:485–491.

[11] Zhang P. L., Chen Y., Zhang C. H., Wang Y. X., Fernandez-Funez P., (2018), "Genetics of Parkinson’s disease and related disorders". J. Med. Genet. 55:73–80.

[12] Sonninen T. M., Hämäläinen R. H., Koskuvi M., Oksanen M., Shakirzyanova A., Wojciechowski S., Puttonen K., Naumenko N., Goldsteins G., Laham-Karam N., Lehtonen M., Tavi P., Koistinaho J., Lehtonen Š., (2020), "Metabolic alterations in Parkinson’s disease astrocytes", Sci Rep, 10(1).

[13] Simon D. K., Tanner C. M., Brundin P., (2020), "Parkinson Disease Epidemiology, Pathology, Genetics, and Pathophysiology", Clin Geriatr Med, 36(1), 1–12.

[14] Teismann P., Mounsey R. B., (2011), "Mitochondrial dysfunction in Parkinson's disease: Pathogenesis and neuroprotection". Parkinsons. Dis.

[15] Cheng A., Wan R., Yang J. L., Kamimura N., Son T. G., Ouyang X., Luo Y., Okun E., Mattson M. P., (2012), "Involvement of PGC-1α in the formation and maintenance of neuronal dendritic spines", Nat Commun, 3, 1250.

[16] Pickrell A. M., Youle R. J., (2015), "The roles of PINK1, Parkin, and mitochondrial fidelity in parkinson's disease". Neuron 85:257–273.

[17] Dias V., Junn E., Mouradian M. M., (2013), "The role of oxidative stress in parkinson's disease". J. Parkinsons. Dis. 3:461–491.

[18] Taymans J.-M., Greggio E., (2016), "LRRK2 Kinase Inhibition as a Therapeutic Strategy for Parkinson's Disease, Where Do We Stand?", Curr Neuropharmacol, 14(3), 214–225.

[19] Anandhan A., Jacome M. S., Lei S., Hernandez-Franco P., Pappa A., Panayiotidis M. I., Powers R., Franco R., (2017), "Metabolic Disorder Dysfunction in Parkinson's Disease: Bioenergetics, Redox Homeostasis and Central Carbon Metabolism", Brain Res Bull, 133, 12–30.

[20] Puspita L., Chung S. Y., Shim J. W., (2017), "Oxidative stress and cellular pathologies in Parkinson's disease". Mol. Brain 10.

[21] Vergara R. C., Jaramillo-Riveri S., Luarte A., Moënne-Loccoz C., Fuentes R., Couve A., Maldonado P. E., (2019), "The Energy Homeostasis Principle: Neuronal Energy Regulation Drives Local Network Dynamics Generating Behavior", Front Comput Neurosci, 13, 49.

[22] Panday A., Sahoo M. K., Osorio D., Batra S., (2015), "NADPH oxidases: An overview from structure to innate immunity-associated pathologies". Cell. Mol. Immunol. 12:5–23.

[23] Ighodaro O. M., Akinloye O. A., (2018), "First line defence antioxidants-superoxide dismutase (SOD), catalase (CAT) and glutathione peroxidase (GPX): Their fundamental role in the entire antioxidant defence grid", Alexandria J Med, 54(4), 287–293.

[24] Nelson N. J., (2001), "Microarrays have arrived: Gene expression tool matures". J. Natl. Cancer Inst. 93:492–493.

[25] Lowe R., Shirley N., Bleackley M., Dolan S., Shafee T., (2017), "Transcriptomics technologies", PLoS Comput Biol, 13(5), e1005457.

[26] Wang Z., Gerstein M., Snyder M., (2009), "RNA-Seq: A revolutionary tool for transcriptomics". Nat. Rev. Genet. 10:57–63.

[27] Haque A., Engel J., Teichmann S. A., Lönnberg T., (2017), "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications". Genome Med. 9:1–12.

[28] Barrett T., Edgar R., (2006), "[19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis". Methods Enzymol. 411:352–369.

[29] Athar A., Füllgrabe A., George N., Iqbal H., Huerta L., Ali A., Snow C., Fonseca N. A., Petryszak R., Papatheodorou I., Sarkans U., Brazma A., (2019), "ArrayExpress update - From bulk to single-cell expression data", Nucleic Acids Res, 47(D1), D711–D715.

[30] Edgar R., Domrachev M., Lash A. E., (2002), "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", Nucleic Acids Res, 30(1), 207–210.

[31] Zhang Y., James M., Middleton F. A., Davis R. L., (2005), "Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms", Am J Med Genet - Neuropsychiatr Genet, 137 B(1), 5–16.

[32] Schulze M., Sommer A., Plötz S., Farrell M., Winner B., Grosch J., Winkler J., Riemenschneider M. J., (2018), "Sporadic Parkinson's disease derived neuronal cells show disease-specific mRNA and small RNA signatures with abundant deregulation of piRNAs", Acta Neuropathol Commun, 6(1), 58.

[33] Wang J., Devlin B., Roeder K., Kelso J., (2020), "Using multiple measurements of tissue to estimate subject- And cell-type-specific gene expression", Bioinformatics, 36(3), 782–788.

[34] Klebanov I., Sikorski A., Schütte C., Röblitz S., (2016), "Objective Priors in the Empirical Bayes Framework". Scand J Stat.

[35] Habib N., Avraham-Davidi I., Basu A., Burks T., Shekhar K., Hofree M., Choudhury S. R., Aguet F., Gelfand E., Ardlie K., Weitz D. A., Rozenblatt-Rosen O., Zhang F., Regev A., (2017), "Massively parallel single-nucleus RNA-seq with DroNc-seq", Nat Methods, 14(10), 955–958.

[36] Torshizi A. D., Duan J., Wang K., (2020), "A computational method for direct imputation of cell type-specific expression profiles and cellular compositions from bulk-tissue RNA-Seq in brain disorders", bioRxiv, 2020.05.28.121483.

[37] Fromer M., Roussos P., Sieberts S. K., Johnson J. S., Kavanagh D. H., Perumal T. M., Ruderfer D. M., Oh E. C., Topol A., Shah H. R., Klei L. L., Kramer R., Pinto D., Gümüş Z. H., Cicek A. E., Dang K. K., Browne A., Lu C., Xie L., Readhead B., Stahl E. A., Xiao J., Parvizi M., Hamamsy T., Fullard J. F., Wang Y. C., Mahajan M. C., Derry J. M. J., Dudley J. T., Hemby S. E., Logsdon B. A., Talbot K., Raj T., Bennett D. A., De Jager P. L., Zhu J., Zhang B., Sullivan P. F., Chess A., Purcell S. M., Shinobu L. A., Mangravite L. M., Toyoshiba H., Gur R. E., Hahn C. G., Lewis D. A., Haroutunian V., Peters M. A., Lipska B. K., Buxbaum J. D., Schadt E. E., Hirai K., Roeder K., Brennand K. J., Katsanis N., Domenici E., Devlin B., Sklar P., (2016), "Gene expression elucidates functional impact of polygenic risk for schizophrenia". Nat. Neurosci. 19:1442–1453.

[38] Mo M. L., Jamshidi N., Palsson B., (2007), "A genome-scale, constraint-based approach to systems biology of human metabolism", Mol Biosyst, 3(9), 598–603.

[39] Edwards J. S., Palsson B. O., (1999), "Systems properties of the Haemophilus influenzae Rd metabolic genotype", J Biol Chem, 274(25), 17410–17416.

[40] Duarte N. C., Becker S. A., Jamshidi N., Thiele I., Mo M. L., Vo T. D., Srivas R., Palsson B., (2007), "Global reconstruction of the human metabolic network based on genomic and bibliomic data", Proc Natl Acad Sci U S A, 104(6), 1777–1782.

[41] Thiele I., Swainston N., Fleming R. M. T., Hoppe A., Sahoo S., Aurich M. K., Haraldsdottir H., Mo M. L., Rolfsson O., Stobbe M. D., Thorleifsson S. G., Agren R., Bölling C., Bordel S., Chavali A. K., Dobson P., Dunn W. B., Endler L., Hala D., Hucka M., Hull D., Jameson D., Jamshidi N., Jonsson J. J., Juty N., Keating S., Nookaew I., Le Novère N., Malys N., Mazein A., Papin J. A., Price N. D., Selkov E., Sigurdsson M. I., Simeonidis E., Sonnenschein N., Smallbone K., Sorokin A., Van Beek J. H. G. M., Weichart D., Goryanin I., Nielsen J., Westerhoff H. V., Kell D. B., Mendes P., Palsson B. O., (2013), "A community-driven global reconstruction of human metabolism", Nat Biotechnol, 31(5), 419–425.

[42] Brunk E., Sahoo S., Zielinski D. C., Altunkaya A., Dräger A., Mih N., Gatto F., Nilsson A., Preciat Gonzalez G. A., Aurich M. K., Prlic A., Sastry A., Danielsdottir A. D., Heinken A., Noronha A., Rose P. W., Burley S. K., Fleming R. M. T., Nielsen J., Thiele I., Palsson B. O., (2018), "Recon3D enables a three-dimensional view of gene variation in human metabolism", Nat Biotechnol, 36(3), 272–281.

[43] Mardinoglu A., Agren R., Kampf C., Asplund A., Nookaew I., Jacobson P., Walley A. J., Froguel P., Carlsson L. M., Uhlen M., Nielsen J., (2013), "Integration of clinical data with a genome-scale metabolic model of the human adipocyte", Mol Syst Biol, 9(1).

[44] Mardinoglu A., Agren R., Kampf C., Asplund A., Uhlen M., Nielsen J., (2014), "Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease", Nat Commun, 5(1), 1–11.

[45] Robinson J. L., Kocabaş P., Wang H., Cholley P. E., Cook D., Nilsson A., Anton M., Ferreira R., Domenzain I., Billa V., Limeta A., Hedin A., Gustafsson J., Kerkhoven E. J., Svensson L. T., Palsson B. O., Mardinoglu A., Hansson L., Uhlén M., Nielsen J., (2020), "An atlas of human metabolism", Sci Signal, 13(624).

[46] Kim M. K., Lun D. S., (2014), "Methods for integration of transcriptomic data in genome-scale metabolic models". Comput. Struct. Biotechnol. J. 11:59–65.

[47] Hertz L., Zielke H. R., (2004), "Astrocytic control of glutamatergic activity: Astrocytes as stars of the show". Trends Neurosci. 27:735–743.

[48] Lewis N. E., Schramm G., Bordbar A., Schellenberger J., Andersen M. P., Cheng

J. K., Patel N., Yee A., Lewis R. A., Eils R., König R., Palsson B. O., (2010), "Large-scale in silico modeling of metabolic interactions between cell types in the human brain", Nat Biotechnol, 28(12), 1279–1285.

[49] Bordbar A., Monk J. M., King Z. A., Palsson B. O., (2014), "Constraint-based models predict metabolic and associated cellular functions". Nat. Rev. Genet. 15:107–120.

[50] Martins Conde P. do R., Sauter T., Pfau T., (2016), "Constraint Based Modeling Going Multicellular", Front Mol Biosci, 3(FEB), 3.

[51] Raman K., Chandra N., (2009), "Flux balance analysis of biological systems: Applications and challenges". Brief. Bioinform. 10:435–449.

[52] Orth J. D., Thiele I., Palsson B. O., (2010), "What is flux balance analysis?". Nat. Biotechnol. 28:245–248.

[53] Segrè D., Vitkup D., Church G. M., (2002), "Analysis of optimality in natural and perturbed metabolic networks", Proc Natl Acad Sci U S A, 99(23), 15112–15117.

[54] Zhao Y., Huang J., (2011), "Reconstruction and analysis of human heart-specific metabolic network based on transcriptome and proteome data", Biochem Biophys Res Commun, 415(3), 450–454.

[55] Özcan E., Çakır T., (2016), "Reconstructed Metabolic Network Models Predict Flux-Level Metabolic Reprogramming in Glioblastoma", Front Neurosci, 10(APR), 156.

[56] Becker S. A., Palsson B. O., (2008), "Context-Specific Metabolic Networks Are Consistent with Experiments", PLoS Comput Biol, 4(5), e1000082.

[57] Jensen P. A., Lutz K. A., Papin J. A., (2011), "TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks", BMC Syst Biol, 5, 147.

[58] Dumitriu A., Golji J., Labadorf A. T., Gao B., Beach T. G., Myers R. H., Longo K. A., Latourelle J. C., (2015), "Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease", BMC Med Genomics, 9(1), 5.

[59] Wang D., Liu S., Warrell J., Won H., Shi X., Navarro F. C. P., Clarke D., Gu M., Emani P., Yang Y. T., Min X., Gandal M. J., Lou S., Zhang J., Park J. J., Yan C., KyongRhie S., Manakongtreecheep K., Peters M., Mattei E., E., Kalayci S., Gümüş Z. H., Crawford G. E., Roussos P., Akbarian S., Jaffe A. E., White K. P., Weng Z., Sestan N., Geschwind D. H., Knowles J. A., Gerstein M. B., (2018), "Comprehensive functional genomic resource and integrative model for the human brain", Science (80-), 362(6420), 8464.

[60] Jolliffe I. T., Cadima J., (2016), "Principal component analysis: a review and recent developments", Philos Trans R Soc A Math Phys Eng Sci, 374(2065), 20150202.

[61] Ardlie K. G., DeLuca D. S., Segrè A. V., Sullivan T. J., Young T. R., Gelfand E. T., Trowbridge C. A., Maller J. B., Tukiainen T., Lek M., Ward L. D., Kheradpour P., Iriarte B., Meng Y., Palmer C. D., Esko T., Winckler W., Hirschhorn J. N., Kellis M., MacArthur D. G., Getz G., Shabalin A. A., Li G., Zhou Y. H., Nobel A. B., Rusyn I., Wright F. A., Lappalainen T., Ferreira P. G., Ongen H., Rivas M. A., Battle A., Mostafavi S., Monlong J., Sammeth M., Melé M., Reverter F., Goldmann J. M., Koller D., Guigó R., McCarthy M. I., Dermitzakis E. T., Gamazon E. R., Im H. K., Konkashbaev A., Nicolae D. L., Cox N. J., Flutre T., Wen X., Stephens M., Pritchard J. K., Tu Z., Zhang B., Huang T., Long Q., Lin L., Yang J., Zhu J., Liu J., Brown A., Mestichelli B., Tidwell D., Lo E., Salvatore M., Shad S., Thomas J. A., Lonsdale J. T., Moser M. T., Gillard B. M., Karasik E., Mosavel M., Siminoff L. A., Valley D. R., Rohrer D. C., Jewell S. D., Branton P. A., Sobin L. H., Barcus M., Qi L., McLean J., Hariharan P., Um K. S., Wu S., Tabor D., Shive C., Smith A. M., Buia S. A., Undale A. H., Robinson K. L., Roche N., Valentino K. M., Britton A., Burges R., Bradbury D., Hambright K. W., Seleski J., Korzeniewski G. E., Erickson K., Marcus Y., Tejada J., Taherian M., Lu C., Basile M., Mash D. C., Volpi S., Struewing J. P., Temple G. F., Boyer J., Colantuoni D., Little R., Koester S., Carithers L. J., Moore H. M., Guan P., Compton C., Sawyer S. J., Demchok J. P., Vaught J. B., Rabiner C. A., Lockhart, (2015), "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans", Science (80-), 348(6235), 648–660.

[62] Lake B. B., Chen S., Sos B. C., Fan J., Kaeser G. E., Yung Y. C., Duong T. E., Gao D., Chun J., Kharchenko P. V., Zhang K., (2018), "Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain", Nat Biotechnol, 36(1), 70.

[63] Martinez A. M., Kak A. C., (2001), "PCA versus LDA", IEEE Trans Pattern Anal Mach Intell, 23(2), 228–233.

[64] Smyth G. K., (2005), "limma: Linear Models for Microarray Data". In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer-Verlag, pp 397–420.

[65] Kaynar A., (2019), "Integrative analysis of transcriptome data and genome-scale metabolic networks to identify candidate drug targets and drugs for parkinson’s disease", Master Thesis, Gebze Technical University.

[66] Dienel G. A., Cruz N. F., (2004), "Nutrition during brain activation: Does cell-to-cell lactate shuttling contribute significantly to sweet and sour food for thought?", Neurochem Int, 45(2–3), 321–351.

[67] Gruetter R., Seaquist E. R., Ugurbil K., (2001), "A mathematical model of compartmentalized neurotransmitter metabolism in the human brain", Am J Physiol - Endocrinol Metab, 281(1 44-1).

[68] Nehlig A., Wittendorp-Rechenmann E., Lam C. D., (2004), "Selective uptake of [14c]2-deoxyglucose by neurons and astrocytes: High-resolution microautoradiographic imaging by cellular 14C-trajectography combined with immunohistochemistry", J Cereb Blood Flow Metab, 24(9), 1004–1014.

[69] Hertz L., Peng L., Dienel G. A., (2007), "Energy metabolism in astrocytes: High rate of oxidative metabolism and spatiotemporal dependence on glycolysis/glycogenolysis". J. Cereb. Blood Flow Metab. 27:219–249.

[70] Lebon V., Petersen K. F., Cline G. W., Shen J., Mason G. F., Dufour S., Behar K. L., Shulman G. I., Rothman D. L., (2002), "Astroglial contribution to brain energy metabolism in humans revealed by 13C nuclear magnetic resonance spectroscopy: Elucidation of the dominant pathway for neurotransmitter glutamate repletion and measurement of astrocytic oxidative metabolism", J Neurosci, 22(5), 1523–1531.

[71] Wahren J., Ekberg K., Fernqvist-Forbes E., Nair S., (1999), "Brain substrate utilisation during acute hypoglycaemia", Diabetologia, 42(7), 812–818.

[72] Çakir T., Alsan S., Saybaşili H., Akin A., Ülgen K. Ö., (2007), "Reconstruction and flux analysis of coupling between metabolic pathways of astrocytes and neurons: Application to cerebral hypoxia", Theor Biol Med Model, 4, 48.

[73] Lying-Tunell U., Lindblad B. S., Malmlund H. O., Persson B., (1980), "Cerebral blood flow and metabolic rate of oxygen, glucose, lactate, pyruvate, ketone bodies and amino acids: I. YOUNG AND OLD NORMAL SUBJECTS", Acta Neurol Scand, 62(5), 265–275.

[74] Herrmann H. A., Dyson B. C., Vass L., Johnson G. N., Schwartz J. M., (2019), "Flux sampling is a powerful tool to study metabolism under changing environmental conditions", npj Syst Biol Appl, 5(1).

[75] Holzhütter H. G., (2004), "The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks", Eur J Biochem, 271(14), 2905–2922.

[76] Çakir T., Kirdar B., Önsan Z. I., Ülgen K. Ö., Nielsen J., (2007), "Effect of carbon source perturbations on transcriptional regulation of metabolic fluxes in Saccharomyces cerevisiae", BMC Syst Biol, 1.

[77] Gray L. R., Tompkins S. C., Taylor E. B., (2014), "Regulation of pyruvate metabolism and human disease". Cell. Mol. Life Sci. 71:2577–2604.

[78] Hinz M., Stein A., Cole T., (2014), "Parkinson's disease-associated melanin steal", Neuropsychiatr Dis Treat, 10, 2331–2337.

[79] Glaab E., (2018), "Computational systems biology approaches for Parkinson's disease". Cell Tissue Res. 373:91–109.

[80] Patt S., Gertz H. J., Gerhard L., Cervos-Navarro J., (1991), "Pathological changes in dendrites of substantia nigra neurons in Parkinson's disease: A Golgi study", Histol Histopathol, 6(3), 373–380.

[81] Rostami J., Fotaki G., Sirois J., Mzezewa R., Bergström J., Essand M., Healy L., Erlandsson A., (2020), "Astrocytes have the capacity to act as antigen-presenting cells in the Parkinson's disease brain", J Neuroinflammation, 17(1).

[82] Requejo-Aguilar R., Lopez-Fabuel I., Fernandez E., Martins L. M., Almeida A., Bolaños J. P., (2014), "Pink1 deficiency sustains cell proliferation by reprogramming glucose metabolism through hif1", Nat Commun, 5(1), 1–9.

[83] Jha M. K., Morrison B. M., (2018), "Glia-neuron energy metabolism in health and diseases: New insights into the role of nervous system metabolic transporters". Exp. Neurol. 309:23–31.

[84] Dunn L., Fairfield V., Daham S., Bolaños J. P., Heales S. J., (2014), "Pentose-phosphate pathway disruption in the pathogenesis of Parkinson's disease". Transl. Neurosci. 5:179–184.

[85] Finsterer J., Jarius C., Baumgartner M., (2001), "Parkinson's disease associated with impaired oxidative phosphorylation", Neuroradiology, 43(11), 997–1000.

APPENDIX

Appendix A: Publications Done In the Scope of This Thesis Study

Kocabaş K., Çakır T., (2020), “Integrative analysis of multi-cellular genome-scale metabolic networks with cell type specific transcriptome data predicted by deconvolution algorithms: application to Parkinson’s Disease” International Eurasian Conference on Biotechnology and Biochemistry, 832-833, Ankara, TURKEY, 16-18 December.