T.C.

ISTANBUL MEDIPOL UNIVERSITY

INSTITUTE OF HEALTH SCIENCES

MASTER THESIS

METAPROTEOMIC ANALYSIS OF SALIVA SAMPLES FROM PARKINSON'S DISEASE PATIENTS WITH COGNITIVE IMPAIRMENTS

BALKIS KURBEH

DEPARTMENT OF MOLECULAR MEDICINE AND BIOTECHNOLOGY

SUPERVISOR

Prof. Dr. SÜLEYMAN YILDIRIM

ISTANBUL - 2022

THESIS APPROVAL FORM

Institute	: Istanbul Medipol University Institute of Health Sciences
Level of Program	: Master (X) Doctorate ( )
Department	: Molekular Medicine and Biotechnology
Student	: Balkıs KURBEH
Title	: METAPROTEOMIC ANALYSIS OF SALIVA SAMPLES FROM PARKINSON'S DISEASE PATIENTS WITH COGNITIVE IMPAIRMENTS
Place	: Istanbul Medipol University South Campus
Examination Date	: 27.07.2022

We have reviewed the present thesis in regard to content and quality and approved as a Master Thesis.

Supervisor	**Institution**	**Signature**
Prof.Dr. Süleyman YILDIRIM	Istanbul Medipol University	
Members of Examination Jury		
Dr.Öğr.Üyesi Kıvanç KÖK	Istanbul Medipol University	
Dr.Öğr.Üyesi Cüneyd PARLAYAN	Bahçeşehir University	

The above jury decision has been approved by the Institute Board of Directors with the decision number ..../ ..... , ... dated ..../ ..../ ....

Prof.Dr. Neslin EMEKLİ

Director of Institute of Health Sciences

DECLARATION OF CONFORMITY

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

Signature :

Name, Surname: BALKIS KURBEH

ACKNOWLEDGEMENT

I would like to thank my advisor Prof.Dr.Süleyman YILDIRIM as well as Assist.Prof.Dr.Muzaffer ARIKAN for their guidance and patience throughout this project. Also, I thank Assist.Prof.Dr.Elizabeth HEMOND for her help in applying plagiarism check for this written thesis report, and my friend Damla YAY for tolerating my questions about the thesis submission process. Additionally, I mention my appreciation for my cousins Dr.Kenana ALADEM, Lavdi AHMEDI, and Judy ALAZEM for their helpful comments and motivational words.

Balkis Kurbeh

August, 2022

TABLE OF CONTENTS

THESIS APPROVAL FORM i
DECLARATION OF CONFORMITY ii
ACKNOWLEDGEMENT iii
TABLE OF CONTENTS iv
LIST OF ABBREVIATIONS AND SYMBOLS v
LIST OF FIGURES viii
LIST OF TABLES ix
1. ABSTRACT 1
2. ÖZET 2
3. INTRODUCTION AND PURPOSE 3
3.1 Neurodegenerative Diseases (NDD) and Parkinson's Disease (PD): The call for non-invasive and accurate biomarkers for Cognitive Impairment (CI) 3
3.2 Proteomics and MS Methods for NDD and PD 4
3.3 Salivary MetaProteomics 6
4. LITERATURE REVIEW 7
5. MATERIALS AND METHODS 8
5.1 Study Cohort 9
5.2 Sampling: Saliva Samples Collection and Storage 9
5.3 Sample Preparation 9
5.3.1. Protein isolation “cell disruption” 9
5.3.2. Protein digestion: MS sample preparation 10
5.4 Mass Spectrometry Measurement 10
5.5 Data Analysis 11
5.5.1. Preprocessing 11
5.5.2. Data comparison 11
5.5.3. Post-identification and statistical analysis 11
6. RESULTS 13
6.1 Study Cohort 13
6.2 Customization of Protein Reference Database 13
6.3 Data Comparison: Proteins Abundance Table 14
6.4 Post-identification and Statistical Analysis 17
6.4.1 Human to bacterial protein groups comparison 17
6.4.2 Protein groups and features associations detection 20
6.4.3. Bacterial taxonomy demonstration 25
7. DISCUSSION 26
8. CONCLUSIONS 28
9. REFERENCES 29
10. ETHICS COMMITTEE APPROVAL 34
11. CURRICULUM VITAE 35

LIST OF ABBREVIATIONS AND SYMBOLS

ACN : Acetonitrile

α-syn : alpha-synclein

α : Alpha

BB : Bead-Beating

CDR : Clinical Dementia Rating

db :data base

DDA : Data Dependent Acquisition

DIA : Data Independent Acquisition

DIA-NN : Data-Independent Acquisition Neural Networks

eEF1A : eukaryotic Elongation Factor Tu

eEF2 : eukaryotic Elongation Factor G

eEF2K : eukaryotic Elongation Factor G Kinase

EF-G : Elongation Factor G

EF-Tu : bacterial Elongation Factor Tu

ESI : Electro Spray Ionization

FA : Formic Acid

FASP : Filter Aided Sample Preparation

FDR : False Discovery Rate

fmol : femto-mole

fMRI : functional Magnetic Resonance Imaging

GC : Gas Chromatography

GN :Gene Name

HC : Healthy Control

IMS : Ion Mobility Separation

IPA : Ingenuity Pathway Analysis software

LC : Liquid Chromatography

LC-MS/MS : Liquid Chromatography tandem Mass Spectrometry

MALDI : Matrix Assisted Laser Desorption Ionization

µ : Micro

MaAsLin 2 : Microbiome Multivariate Association with Linear Models

MMSE : Mini-Mental State Examination

MCI : Mild Cognitive Impairment stage

MRI : Magnetic Resonance Imaging

MCPs : ion trap Micro Channel Plates

MPA : MetaProteome Analyzer

Mud-PIT : Multidimensional Protein Identification Technologies

MS : Mass Spectrometry

m/z : mass to charge

ml : milli-liter

nLC-MS/MS :nano Liquid Chromatography tandem Mass Spectrometry

NCI : No Cognitive Impairment stage

NDD : NeuroDegenerative Diseases

ng : nano-gram

OS : Organism name

OX : Organism identifier

PCA : Principal Component Analysis

PD : Parkinson’s Disease

PD-MCI : Parkinson’s Disease with Mild Cognitive Impairment

PDD : Parkinson’s Disease with Dementia

PE : Protein Existence

PERMANOVA ;Permutational Multivariate Analysis of Variance

PET : Positron Emission Tomography

Prodigal : PROkaryotic DYnamic programming Gene-finding Algorithm,

ROC : Receiver Operating Characteristic plot

s : seconds

SCI : Subjective Cognitive Impairment stage

sp :UniProtKB/Swiss-Prot

sMRI : structural and functional Magnetic Resonance Imaging

std : standard deviation

SV : Sequence Version.

TOF : Time Of Flight

UPX : Universal Protein extraction reagent

V : Volts

WS : Working Solution

XML : Extensible Markup Language

yrs : years

LIST OF FIGURES

Figure 5-1 The study workflow outline from wet lab to dry lab. 8
Figure 6-1 Human to bacterial protein groups percentages across the collected saliva samples.. 14
Figure 6-2 Heat-map of human protein groups abundance among the three study groups. 15
Figure 6-3 PCA for the study groups distinction according to both human and bacterial protein groups.. 18
Figure 6-4 PCA for the study groups distinction according to human protein groups.. 18
Figure 6-5 PCA for the study groups distinction according to bacterial protein groups.. 19
Figure 6-6 MaAsLin2 box plot for PD-MCI association with bacterial EF-Tu protein group abundance (molecule/cell) 21
Figure 6-7 MaAsLin2 box plot for PDD association with bacterial EF-Tu protein group abundance (molecule/cell).. 22
Figure 6-8 MaAsLin2 plot for age feature association with bacterial EF-Tu protein group abundance (molecule/cell).. 23
Figure 6-9 MaAsLin2 box plot for sex (M; male, F; female) feature association with bacterial AP-4-A phoshorylase protein group abundance (molecule/cell).. 24
Figure 6-10 Common taxonomy tree and significant protein group and associated features 25

LIST OF TABLES

Table 5-1 Data analysis bioinformatics tools used.. 12
Table 6-1 Features of the study cohort.. 13
Table 6-2 Top 20 most abundant human protein groups... 16
Table 6-3 Bacterial Protein Groups Identified. 17
Table 6-4 MaAsLin 2 protein groups to metadata associations results..................... 20

1. ABSTRACT

METAPROTEOMIC ANALYSIS OF SALIVA SAMPLES FROM PARKINSON'S DISEASE PATIENTS WITH COGNITIVE IMPAIRMENTS

Parkinson's Disease (PD) is a NeuroDegenerative Disease (NDD) that includes impairment and death of dopaminergic neurons at the brain and alpha-synuclein (α-syn) protein accumulation (Lewy bodies). PD symptoms include both motor symptoms, such as tremor, and non-motor symptoms, such as abnormal saliva production and cognitive impairment (CI). CI can develop to dementia tragically affecting the lives of PD patients and their caregivers, reaching to life threating stages. Therefore, developing non-invasive and accurate biomarkers for detecting and monitoring CI and dementia progression in PD is urgent. Although, cerebrospinal fluid (CSF) is a useful source for CI-biomarkers in PD, its obtainment method is considered invasive. Thereby, saliva is a good replacement aiding the detection of both human and microbial proteins, allowing MetaProteomics analysis considering microbiome-host relationship with health/disease. Additionally, the development of Mass Spectrometry (MS) technologies and bioinformatics tools aid metaproteomics studies. Here, we analyzed salivary metaproteome content of three study groups; Healthy Control (HC: 28), Parkinson's patients with Mild Cognitive Impairment (PD-MCI: 40), and Parkinson's patients with Dementia (PDD: 40), using nano-Liquid Chromatography tandem Mass Spectrometry (nLC-MS/MS), DIA-NN pipeline, and R-program. Results showed that bacterial proteins differentiate the three study groups, and out the 435 protein groups abundances identified, the decrease of the protein group Elongation Factor Tu (EF-Tu) showed significant association with CI progression ($p<0.05$). Finally, salivary metaproteomics shows potentials to enrich non-invasive biomarkers discovery for CI/dementia detection and monitoring in PD and other NDD.

Keywords:

Cognitive impairemet/dementia, nLC-MS/MS, Metaproteomics biomarkers, Parkinson's disease, Saliva microbiome

2. ÖZET

KOGNİTİF BOZUKLUĞU OLAN PARKİNSON HASTALARINDAN ALINAN TÜKÜRÜK ÖRNEKLERİNİN METAPROTEOMİK ANALİZİ

Parkinson Hastalığı (PD), beyindeki dopaminerjik nöronların bozulmasını ve ölümünü ve alfa-sinüklin (α-syn) protein birikimini (Lewy cisimcikleri) içeren bir Nörodejeneratif Hastalıktır (NDD). PD semptomları, hem titreme gibi motor semptomları hem de anormal tükürük üretimi ve bilişsel bozulma (CI) gibi motor olmayan semptomları içerir. CI, PD hastalarının ve bakıcılarının yaşamlarını trajik bir şekilde etkileyen ve yaşamı tehdit eden aşamalara ulaşan demansa dönüşebilir. Bu nedenle, PH'de CI ve demans ilerlemesini saptamak ve izlemek için invaziv olmayan ve doğru biyobelirteçler geliştirmek acildir. Beyin omurilik sıvısı (BOS), PH'de CI biyobelirteçleri için yararlı bir kaynak olmasına rağmen, elde etme yöntemi invaziv olarak kabul edilir. Böylece tükürük, hem insan hem de mikrobiyal proteinlerin saptanmasına yardımcı olan iyi bir ikamedir ve sağlık/hastalık ile mikrobiyom-konak ilişkisini göz önünde bulundurarak MetaProteomik analizine izin verir. Ek olarak, Kütle Spektrometrisi (MS) teknolojilerinin ve biyoinformatik araçlarının geliştirilmesi, metaproteomik çalışmalara yardımcı olur. Burada, üç çalışma grubunun tükürük metaproteome içeriğini analiz ettik; nano-Sıvı Kromatografi tandem Kütle Spektrometresi (nLC-MS/MS), DIA-NN kullanılarak Sağlıklı Kontrol (HC: 28), Hafif Bilişsel Bozukluğu olan Parkinson hastaları (PD-MCI: 40) ve Demanslı Parkinson hastaları (PDD: 40) boru hattı ve R-programı. Sonuçlar, bakteriyel proteinlerin üç çalışma grubunu farklılaştırdığını ve tanımlanan 435 protein grubu bolluğu dışında, protein grubunun Uzama Faktörü Tu (EF-Tu) azalmasının CI ilerlemesi ile anlamlı bir ilişki gösterdiğini gösterdi ($p<0.05$). Son olarak, tükürük metaproteomikleri, PD ve diğer NDD'de CI/demans tespiti ve izlenmesi için invaziv olmayan biyobelirteç keşfini zenginleştirme potansiyeli gösterir..

Anahtar Kelimeler:

Bilişsel bozukluk/demans, nLC-MS/MS, Metaproteomik biyobelirteçler, Parkinson hastalığı, Tükürük mikrobiyom

3. INTRODUCTION AND PURPOSE

3.1 Neurodegenerative Diseases (NDD) and Parkinson's Disease (PD): The call for non-invasive and accurate biomarkers for Cognitive Impairment (CI)

Neurodegenerative Diseases (NDD) can be described as disorders which affect the central nervous system (CNS), and are related with loss of neural tissues in a progressive manner. The crucial point of this progressive loss is that after the neural damage occurs the neurons are unable to regenerate and compensate the occurred damage and loss [1]. There is still no cure for NDD, but some medications can be used to reduce symptoms of the disease and maintain quality of life [2]. From the most common NDD are Parkinson’s Disease (PD), Alzheimer’s Disease (AD), Amyotrophic Lateral Sclerosis (ALS), and Huntington’s Disease (HD) [1].

When talking about PD, it is a type of NDD that includes the impairment and death of dopaminergic neurons at Substantia nigra area at the brain. Also, the pathology of PD includes alpha-synuclein (α-syn) protein accumulation, these abnormal protein clumps are called Lewy bodies [3]. PD is mostly known with its motor symptoms, such as tremor, rigidity of limbs, and slowness of movement (Bradykinesia) [4]. Additionally, PD also includes non-motor symptoms, such as constipation, too much or too few saliva production, and cognitive impairment (CI) that can develop into dementia [3]. From the previously mentioned non-motor symptoms, CI and dementia development symptoms are interest of study as they can tragically affect the lives of PD patients along with the lives of their caregivers, and can even develop into life threating stages [5].
Briefly, CI stages of severity can be described as four main stages: *No CI (NCI)*; individual with normal cognition abilities, *Subjective CI (SCI)*; some functional or cognitive decline would be observed and the individual is aware of it and this decline does not interfere with his normal daily activities, *Mild CI (MCI)*; cognitive decline in abilities such as perception, language, reasoning of memory, and learning new skills would be observed, but the individual can still perform basic skills that were well learned such as driving, and then the most developed stage of CI is *dementia*; the decline in cognition that interferes with the individual’s ability of performing his normal daily skills even the well learned ones such as cooking, driving, dressing, bathing, and it can even develop into stages where it affects swallowing, control of speech and facial expressions [6].

Furthermore, the need for early diagnosis and monitoring for CI in NDD such as PD is urgent in order to avoid the progression of CI into dementia. From the current methods used for NDD diagnosis are: immunochemical, molecular and molecular imaging techniques. The basic aims of these methods is to identify biomarkers which can help in the diagnosis of a disease and its severity [7].

Biomarkers can be described as biological indicators of the disorder presence or severity. Diagnostic biomarkers can be genetic or biochemical biomarkers [1].

Also, biomarkers can be classified into fluid-based (such that in molecular diagnosis of body fluids) and image-based (from molecular imaging techniques results) biomarkers. From the most common sample sources for fluid-based biomarkers are cerebrospinal fluid (CSF), saliva and blood. For image-based biomarkers, structural and functional Magnetic Resonance Imaging (sMRI)(fMRI), and Positron emission tomography (PET) molecular imaging techniques are used [7]. Although these methods are commonly used for NDD detection, some are considered both invasive and expensive. To overcome these problems, some groups worked on developing digital technological-detecting framework which can identify early stages of dementia related-NDD based on behavioral and physiological measures [8]. But from the possible problems with behavioral and physiological-based identification methods is it is challenging to accurately identify the disorder type, and as the symptoms started to show on behavioral level this means that the disease had possibly started reaching an advanced level.

So as another possible solution to develop non-invasive and cost-effective early diagnostic and CI monitoring methods for NDD is the utilization of *omics-based technologies* applied on non-invasively obtained samples source such as saliva.

When talking of omics technologies, we should first explain the term omics as it is used to describe sciences end up with the suffix –omics, such as genomics, proteomics, and trarnsceriptomics. The suffix *-ome-* originate from intersection of various Greek terms in -ωμα (*-ōma*), and used in molecular biology to refer to a totality [9]. Also, this field considers life sciences which focus on large-scale data information in order to understand life summed up in “omes” and “omics” [10]. Moreover, omics can be defined as the study of a group of molecules, such as DNA, RNA, proteins, and metabolites in a global or comprehensive manner. Distinctively, the use of –ome refers to the type of molecules observed, such as saying genome is the total DNA of an organism, where on the other hand –omics refers to the study field of these groups, such as genomics is the study of the genome sequence and function of an organism, and proteomics studies complete sets of proteins, proteome, and their interactions. [11][12][13].

Interest shift toward proteomics in many NDD related studies is promising especially for studies at functional level as proteins play important roles in living organisms, such as regulating molecular processes, and pathologies progression, such as in PD. Therefore, different analyzing techniques were developed to study proteins.

3.2 Proteomics and MS Methods for NDD and PD

Techniques developed for studying proteins can be broadly classified into three main categories*: protein separation*, *western blotting*, and *identification techniques*. Underneath each category, different methods were developed. Protein separation include multiple electrophoresis methods, most commonly SDS-PAGE is used. For western blotting, immunoblotting is wildly known. Where for protein identification methods, Edman degradation and protein Mass Spectrometry (MS) are good examples. In interest of identifying large numbers of proteins, such that in proteomics studies, protein MS is a proper method to go with [14].

In general, two main approaches are followed in proteomics studies: *bottom-up* and *top-down proteomics* approaches. The main differences between these two approaches are: bottom-up proteomics includes protein digestion and it is more commonly used, while top-down proteomics approach does not include protein digestion and it is less commonly used as its data analysis process is considered more complex. Briefly, the concept of bottom-up proteomics, which is considered peptide-oriented, starts with protein digestion to peptides smaller in size via chemical or enzymatic reaction which are then delivered to a mass spectrometer machine, and afterward peptides identification step is done by using patterns of fragmentation and intact masses by referring to reference database of protein sequences [15] .
When talking about MS method for proteins identification, it depends on mass-charge ratio (m/z) measurement of charged particles [14]. MS system main components include: an *inlet system* (such as Liquid Chromatography (LC) and Gas Chromatography (GC)), the *ion source* (such as Electro Spray Ionization (ESI) and Matrix Assisted Laser Desorption Ionization (MALDI)), *mass analyzer* (such as Time Of Flight (TOF) and Ion Trap), and finally a *detector (*such as Micro Channel Plates (MCPs) and Electron Multiplier) [16].

Here, it is important to mention that MS resulted proteomics data analysis is challenged at both *biological level* (high variability of proteomes and their characteristics), and *technical level* (data analysis processes). When talking about the challenges over the biological level; proteome is highly variable, and proteins vary in their characteristics (such as polarity) and also their sizes (5-100kDa). Also, the detection of low abundant proteins, described as dark proteins, is another challenge. Fortunately, with the development of new and more advanced proteomics methods, such as Multidimensional Protein Identification Technologies (Mud-PIT), which replaces gel separation of proteins with 2-Dimensional LC and ionization methods such as MALDI with tandem mass spectrometry, the analysis process is becoming more efficient [17]. When looking at the MS-proteomics challenges over the technical level; the generation of different formats of data-files from different mass spectrometers restricts results sharing among different research groups and interpretation using bioinformatics tools. Therefore, different data representation formats, such as mzML, which represents MS data in an opened eXtensible Markup Language (XML) format and converting tools were developed [18].

From the approaches used for MS-proteomics data analysis are Data Dependent Acquisition (DDA) and Data Independent Acquisition (DIA) approaches. DDA is more commonly used, and the precursor ions found most abundant are selected by the mass spectrometer. But due the complexity of proteome, especially when working with high throughput experiments such as microbiome, large amount of data generated and increased rates of stochasticity for peptides detection, decreases the depth of the proteomics sampling challenges the analysis process. Therefore, DIA approaches, in which mass spectrometer refers to multiple predefined precursors, are developed to enhance reproducibility and proteomic depth, and decrease stochastic peptides detection. Although, DIA data processing is considered challenging, different analysis tools were developed to aid the analysis process [19]. Examples of programs and pipelines developed to aid MS-proteomics data analysis include: Galaxy-p [20], MetaProteome Analyzer (MPA)[21], DIA-NN pipeline [19] and Progenesis-QI (Waters) program [22].

In order to get useful information from protein MS regarding NDD such as PD, proper analytic fluid should be chosen.

3.3 Salivary MetaProteomics

When talking about analytic fluid sources used for PD proteomics biomarkers, CSF is used in different studies [23]. Despite the useful information obtained from CSF, its obtainment method, via lumbar puncture with a needle inserted into space surrounding spinal cord filled with CSF, is considered invasive, adds additional stress on patients, and can be associated with possible risks such as infection, bleeding in spinal cord, or even nerve damage [24]. Therefore, this calls a need for a replacement analytic fluid which can provide useful information and obtained in a non-invasive manner.

Here comes the choice for saliva which is obtained in a non-invasive manner, and can be possibly related with abnormal changes of saliva secretion levels such as hypersialorrhea, excessive saliva production [25], or dry mouth[26] commonly seen in PD patients. In addition, saliva is considered a good analytic fluid not only for human proteins but also for microbial proteins, thereby adding another dimension for biomarkers detection, the dimension of microbiome studies [27].

As the terms microbiome and microbiota are usually used interchangeably, microbiota is the collection of all microorganisms occupying a certain organ or region [28], where microbiome is the total genomes of these microorganisms [29].

Different studies had showed evidence of connection between microbiota and NDD development, and it is a promising source of biomarkers for many diseases due to microbiota-health/disease relationship [30][31].

Omics studies interested in microbiota deals with large data size, as the large numbers of microbial communities generates larger amounts of data, therefore, the prefix *"Meta-"*, which means: beyond, higher level, or more comprehensive [32], is added before the omic data type, such as saying metagenomics and metaproteomics.

Omics technologies help in studying microbiomes to identifying possible microbial types or products in which may contribute with NDD development, therefore, opening the door for microbiome-based biomarkers that can be developed for early diagnosis of NDD and monitoring CI.

Hereby, when we want to look for CI and dementia monitoring biomarkers from a non-invasively and easily collected sample source, with considering both functional level of proteins and microbiome-disease association, and using high throughput technologies such as MS; *salivary metaproteomics* is the choice to go with.

Salivary metaproteomics studies proteins from multiple specie sources, such as studying and characterizing proteins from both human and bacteria simultaneously, specifically from saliva sample source [33]. Saliva sample collection can be done by both unstimulated or stimulated saliva collection methods [34]. But some studies show that using stimulated saliva collection methods, via paraffin gum chewing, might change bacterial content [35], therefore using unstimulated saliva collection method would more preferred [36].

Additionally, different studies in literature considered salivary metaproteomics biomarkers for CI in NDD such as in AD and PD.

4. LITERATURE REVIEW

Different studies consider salivary metaproteomics biomarkers for NDD-CI such as in early detection of dementia in both AD and PD, as CI mechanisms and research methods can overlap in some points.

When considering AD, in a study by François et al 2021 [37] they compared salivary metabolomics and proteomics of 80 participants (healthy control: 40, MCI group: 20, AD group: 20) in aim to specify altered cellular pathways with AD progression to aid early CI detection and biomarkers identification. LC-MS was applied for proteomics, and they used *SIMCA* software and *MetaboAnalyst 4.0* for multivariate data analysis for both metabolomics and proteomics. In conclusion, they found several significant cellular pathways alterations and their multi-omic analysis method using SIMCA software showed distinctions between the three study groups, therefore their method can be possibly applied to similar studies.

For PD protein biomarkers, in addition to α-syn, several studies reported DJ-1 multifunctional protein level elevation in PD patients, nominating it as a possible indicator of PD progression [38][39].

Also, in a study by done by Figura et al.2021[36], they mentioned the presence of lewy bodies, pathological hallmark of PD, was detected in submandibular saliva glands in almost 75% of PD patients, so they hypothesized that saliva can be a suitable biomarkers source for PD. Therefore, in their study they obtained saliva samples using *RNA-Pro-Sal kit* from 24 patients with PD and 15 health controls and used *Label-free LC-MS/MS* (nanoAcquity UPLC-Waters coupled to a QExactive MS) to characterize the samples' proteome composition that were then preprocessed using *Mascot Distiller software* and identified using *Mascot Server 2.5* over Swiss-Prot Homo sapiens protein database with size of 20,490 protein sequences. For the identified peptide lists they merged them into one list and then overlayed it to LC-MS spectra 2-D heatmaps to determine the peptides abundances. For statistical analysis they used *Diffport software,* and *ROCit 2.1.1 software* for generation of Receiver Operating Characteristic (ROC) plot, and for biological pathways identification along with related upstream regulators they used the software *Ingenuity Pathway Analysis (IPA)*. For their results about the salivary proteome composition, they identified 530 proteins from corresponding 1328 peptides obtained, found that concentration of the proteins VPS4B, ARP2/3, and S100-A16 were lower in PD group compared to HC. Where for results related with upstream regulators and biological protein networks they found that pathways the proteins concentrations related with inflammation, formation of adipose tissue and exosome processes do differ between PD and HC. Therefore, they conclude that the proteomics content of saliva can aid PD diagnosis and further understand its pathology.

5. MATERIALS AND METHODS

This study was approved by Medipol University Non-Interventional Clinical Trials Ethics Committee, with authorization number 10840098-604.01.01-E.3958, and the approval form in both English and Turkish are shown in appendix 1. All experiments took place at Medipol University: microbiology lab at REMER for wet lab experiments, and proteomics lab at the Genetic Center for mass spectrometry protocol. The whole study process from wet lab to dry lab is shown in figure1.

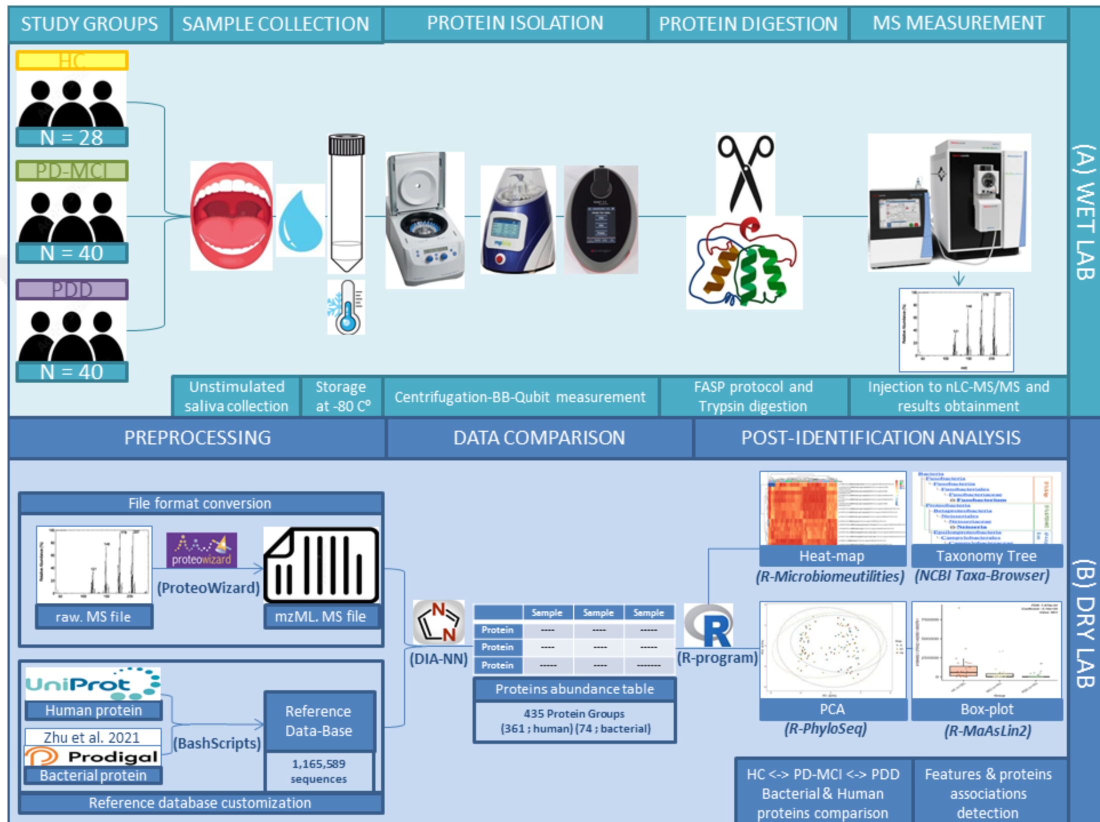

Figure 5-1 The study workflow outline from wet lab to dry lab. *A) Wet Lab:* Unstimulated saliva samples were collected from the three study groups: Healthy Control (HC: 28), Parkinson's patients with Mild Cognitive Impairment (PD-MCI: 40), and Parkinson's patients with Dementia (PDD: 40), then stored at -80 ^{0}C until proceeded with Protein Isolation including cell disruption using centrifugation, Bead-Beating (BB) method, and concentration measurement using Qubit method, followed by Protein Digestion by following FASP protocol, and then injecting the protein extract to nLC-MS/MS for MS measurement receiving MS spectra raw file. *B) Dry Lab:* Preprocessing included file format conversion of the received raw MS spectra file to mzML file format using ProteoWizard software, and reference protein data-base customization by merging bacterial proteins identified by protein annotation using Prodigal program to genome bins over species-level from study by Zhu et al. 2021 for 20 most abundant bacterial genera determined by 16s rRNA amplicon sequencing, with human saliva proteins from UniProt database using different bash scripts. For data comparison the proteins abundance table was generated using DIA-NN pipeline. For Post-Identification Process the packages: PhyloSeq for PCA, MaAsLin2 for association test, and Microbiomeutilities for heat-map generation were applied in R-program to compare bacterial and human proteins among the three study groups features. Also, NCBI taxonomy browser was used for common taxonomy tree generation.

5.1 Study Cohort

Saliva samples were collected from 108 participants; 28 HC, 40 PD-MCI, and 40 PDD, from two health centers: Medipol training hospital, and Bakirkoy Research and Training Hospital for Psychiatric and Neurological Diseases, both centers are in Istanbul-Turkey.

5.2 Sampling: Saliva Samples Collection and Storage

Unstimulated saliva samples were collected from participants, divided into 3 aliquots (1ml for each aliquot), and stored at –80 C° until used for down-stream analysis.

5.3 Sample Preparation

Bottom-up proteomics workflow, including protein digestion, was applied.

5.3.1. Protein isolation "cell disruption"

Protein extraction:

250 µL saliva samples from -80 ⁰C storage were kept to thaw at room temperature, then briefly got homogenized by vortexing, centrifuged at 10 000 x g for 8 minutes, and supernatants were discarded. The pellet in the tubes were suspended using 250 µL of UPX extraction buffer, transferred to BB-BeadBug™ tubes with capacity of 2 ml and 1.0 mm Zirconium beads, and then 2.5 µL of protein inhibitor cocktail and 2.5 µL of EDTA were added to each tube.

Bead-Beating step:

To homogenize samples was applied using *Next Advance Bullet Blender machine* in the order: Bead-Beating at level-4 for 30 seconds: Samples kept on ice for 30 seconds: Bead-Beating at level-4 for 30 seconds. Afterward, the samples were placed on shaker at 100 ⁰C at speed of 600 rpm for 10 minutes, and then kept at +4 ⁰C for 1 hour. The samples were centrifuged at 16 000 x g (15 000 X g) for 10 minutes and the supernatants were transferred to LoBind (Qubit) clean microcentrifuge tubes with 1.5 ml capacity.

Qubit measurement

Qubit^R 2.0 Fluorometer was used to measure concentration of the samples to unify their concentration to 50 µg of proteins in 30 µL and that by using the kit: *"QubitTM Protein Assay kit, 500 assays, Invitrogen by ThermoFisher Scientific (LOT: 1948667)"*. This was done by the following steps: first, qubit Working Solution (WS) was prepared, 200 µL per sample, by mixing 199 µL Qubit Buffer and 1 µL qubit dye/reagent, and then briefly vortexed. Afterward, for the 3 standards and samples preparation for measurements: each of the 3 standards solutions were prepared by adding 190 µL WS and 10 µL of the standard, where for samples; 198 µL WS and 2 µL of the sample were added, then all were incubated for 15 minutes in the dark. For the measurement, the 3 standards were measured at a time using qubit

device and set to be read as standards, then the samples were measured and the qubit results of the samples' concentrations (µg/ mL) were recorded. The samples' protein concentrations were adjusted to 50 µg / 30 µL to be used in downstream analyses.

5.3.2. Protein digestion: MS sample preparation

Filter Aided Sample Preparation (FASP) protocol was applied for protein digestion for mass spectrometry step preparation by using the kit *"ab270519 FASP Protein Digestion Kit, from abcam company (LOT: GR3418360-1)"* by the following steps: first of all, the required reagent were freshly prepared (urea sample solution; 1000 µL Tris hydrochloride + urea powder, 10x Iodoacetamide solution; 100 µL of urea sample + iodoacetamide, and digestion solution; 4 µg of trypsin + 75 µL of 50Mm Ammonium Bicarbonate), and then 30 µL of the protein extract and 200 µL of urea sample solution were added to spin filter tube labeled with sample name and then centrifuged at 14 000 x g for 15 minutes. 200 µL of urea sample solution was added again to the spin filter and then centrifuged at 14 000 x g for 15 minutes; the flow-through of the collection tube was discarded. Next, 10 µL of 10x Iodoacetamide solution and 90 µL of urea sample solution were added to sample tubes, vortexed for 1 minute and then incubated (without mixing) for 20 minutes in the dark. After incubation, tubes were centrifuged at 14 000 x g for 10 minutes, and then 100 µL of urea sample solution was added and centrifuged at 14 000 x g for 10 minutes, with repeating this step twice. Afterward, the flow-through of the collection tube was discarded and 100 µL of 50mM Ammonium Bicarbonate solution was added and centrifuged at 14 000 x g for 10 minutes, with repeating this step twice. 75 µL of digestion solution was added, vortexed for 1 minute, and the tops of tubes were wrapped with stretch-film (to minimize evaporation), then incubated at 37 ^{0}C for 4-28 hours (overnight). After incubation was done, the spin filters from each tube was transferred to a new labeled collection tube, and then 40 µL of mM of Ammonium Bicarbonate solution was added then tubes were centrifuged at 14 000 x g for 10 minutes, with repeating this step once. Finally, 50 µL of 0.5 M Sodium Chloride was added then centrifuged at 14 000 x g for 10 minutes, spin filters were discarded and the sample tubes were closed and placed at concentrator device at 30 ^{0}C for 6 hours. The samples were then stored at -80 ^{0}C until later use.

5.4 Mass Spectrometry Measurement

By following a previously published protocol by Beker et al. [40] the nLC-MS/MS analysis was performed including the following steps: before injecting to the nLC-MS/MS system, the peptides were dissolved with formic acid, 0.1 percent, and diluted to 100ng/µl and then loaded into *ACQUITY UPLC M-Class system* coupled with *SYNAPT G2-Si high definition mass spectrometer (Waters)*[41] and by using 97% mobile phase; the columns were equilibrate at temperature set to 55 ^{0}C. For peptides separation from trap column to the analytic column; gradient elution for 90 minutes at flow rate of 0.400 µl/minute with 4%-40% acetonitrile (ACN) gradient including Formic Acid (FA) of 0.1 % was applied. Also, sequential MS ad MS/MS

scans with cycle time of 0.7 seconds with positive ion modes were performed. By using Ion Mobility Separation (IMS) with velocity of 1000 m/s-55 m/s, 500 µs release time trapping mobility at 15 V and 1000 µs wave delay, ions were separated. All ions were fragmented within the range 50–1900 m/z without any ion precursor pre-selection. In addition, lockmass reference Glu-1-fibrinopeptide B, 100 fmol/µl, was infused with interval of 60 s.

5.5 Data Analysis

For salivary metaproteomics MS data analysis, DIA method was applied.

5.5.1. Preprocessing

Customization of protein reference database

To customize the reference protein database used for proteins identification step, both bacterial and human salivary proteins were merged using different bash scripts. The bacterial salivary proteins were annotated using PROkaryotic DYnamic programming Gene-finding ALgorithm (Prodigal) program [42] to genomes of the bacterial genera determined by 16s rRNA amplicon sequencing analysis applied to the saliva samples as part of a larger project. The genome bins over the species-level were obtained from a study on human oral microbiome done by Zhu et al. 2021[43]. On the other hand, human saliva proteins were obtained from *UniProt* database official website [44].

File conversion

In order to analyze the file obtained from nLC-MS/MS system in raw DIA closed-Waters format the file format was converted into mzML open Extensible Markup Language (XML)-based format by using Proteowizard software [45].

5.5.2. Data comparison

For proteins identification and quantification, the MS file in mzML format was compared to the customized protein database using DIA-NN pipeline [46] with library-free mode.

5.5.3. Post-identification and statistical analysis

Features of the study cohort significance test

For standard deviation (std) and mean calculation for numerical data (age, education, MMESE and CDR scores) Microsoft Excel functions (average and STDEV) were used, and for the study cohort p-value significance test ANNOVA-one way test was used for numerical data (age, education, MMSE, and CDR), and Chi-Square test for categorical data (sex).

Human to bacterial proteins comparison

For human proteins demonstration, a heat-map for the top 20 most abundant human protein groups was generated using Microbiomeutilities R package [47] with hierarchical clustering method. Principal Component Analysis (PCA) was applied using

PhyloSeq package [48] in R-program (R 4.1.2) to show whether bacterial and or human proteins can significantly differentiate between the three studied groups. For PCAs p-value significance calculation, from the Vegan package[49] the Permutational Multivariate Analysis of Variance (PERMANOVA) test was applied in R-program (R 4.1.2).

***Protein groups* and features associations detection**

For association test among study groups features, Microbiome Multivariate Association with Linear Models (MaAsLin2) package [50] was used in R-program (R 4.1.2) using default mode with association significance $p < 0.05$ and $q < 0.25$.

Bacterial taxonomy demonstration

Common taxonomy tree of the bacteria found related with the most significantly associated protein groups was constructed using NCBI Taxonomy Browser [51] .

Brief description of the data analysis bioinformatics tools used and their applications in this project are shown in table 5-1.

Table 5-1 Data analysis bioinformatics tools used

Name	Type	Description	Application in this project	Ref
Proteowizard	Software	An open source software that provides tools and libraries for proteomics data analysis	Conversion of MS raw. file format to mzML open XML file format	[45]
Prodigal	Software	An open source software tool for gene prediction	Annotation of bacterial proteins from genome bins	[42]
DIA-NN	Software/ Pipeline	An open source software that aids DIA analysis of proteomics data by using deep neural networks along with correction inference strategies	Generation of proteins abundance table	[19] [46]
Phyloseq	R-package	An open source R-package that aid microbiome data analysis and graphical demonstrations	Applying PCA for human and bacterial proteins	[48]
Microbiome - utilities	R package	An open source R-package provides set of functions aid analyzing sequencing data of gene marker amplicon in microbial research	Heat map generation for human protein groups	[47]
MaAsLin2	R-package	An open source R-package developed to perform association analysis between metadata-features and multi-omics microbiome data using both linear and mixed models	Association analysis between metadata features and protein groups	[50]

Prodigal; PROkaryotic DYnamic programming Gene-finding Algorithm, DIA-NN; Data-Independent Acquisition Neural Networks, PCA; Principal Component Analysis , MaAsLin2; Microbiome Multivariate Association with Linear Models, Ref; Reference

6. RESULTS

6.1 Study Cohort

The detailed features including age, sex, education, and Mini-Mental State Examination (MMSE), and Clinical Dementia Rating (CDR) scores for each study group are demonstrated in table 6-1.

Table 6-1 Features of the study cohort

Characteristics	HC	PD-MCI	PDD	P-Value
Number (n)	28	40	40	-
Age (yrs: mean ± std)	(59.6 ± 8)	(68.5 ±8.9)	(71.2 ± 7.3)	<0.00001*
Sex (n: female, male)	(14,14)	(17,23)	(21,19)	0.653
Education (yrs: mean ± std)	(10.6 ± 5.1)	(6.9 ± 4.7)	(4 ± 4.1)	<0.00001*
MMSE (score: mean ± std)	(27.9 ± 1.8)	(23.7 ± 2.6)	(18.8 ±3.1)	<0.00001*
CDR (score: mean ± std)	(0 ± 0)	(0.04 ±0.13)	(1.2 ± 0.5)	<0.00001*

Number of participants (n), age (mean and standard deviation of years), sex (female, male), education (mean and standard deviation of years), Mini-Mental State Examination; MMSE (mean and standard deviation of scores), and Clinical Dementia Rating; CDR (mean and standard deviation of scores) for each of the study groups: HC; Healthy Control, PD-MCI; Parkinson's Disease with Mild Cognitive Impairment, and PDD; Parkinson's Disease with Dementia. Significance test P-value <0.05, significant;*

6.2 Customization of Protein Reference Database

For the bacterial proteins customization, the16s rRNA amplicon sequencing analysis applied to the saliva samples as part of a larger project identified 20 most abundant bacterial genera including: *Actinomyces, Aggregatibacter, Alloprevotella, Campylobacter, Capnocytophaga, Fusobacterium, Gemella, Granulicatella, Haemophilus, Lactobacillus, Leptotrichia, Neisseria, Porphyromonas, Prevotella, Prevotella_6, Prevotella_7, Rothia, Selenomonas_3, Streptococcus*, and *Veillonella*. After merging both bacterial and human proteins sequences, the finally produced protein database included 1,165,589 protein sequences.

6.3 Data Comparison: Proteins Abundance Table

The comparison between the mzML MS file to the customized protein reference database using DIA-NN pipeline resulted with proteins abundance table including 435 protein groups, 361 from human origin and 74 are bacterial proteins, demonstrated in pie chart figure 6-1. The top 20 most abundant human protein groups are demonstrated in the heat-map in figure 6-2 showing no clear distinction between the study groups, and the human protein groups names from figure 6-2 are shown in table 6-2 with the same order as in the figure. The protein groups identified from bacterial source and the number of times they were assigned are demonstrated in table 6-3.

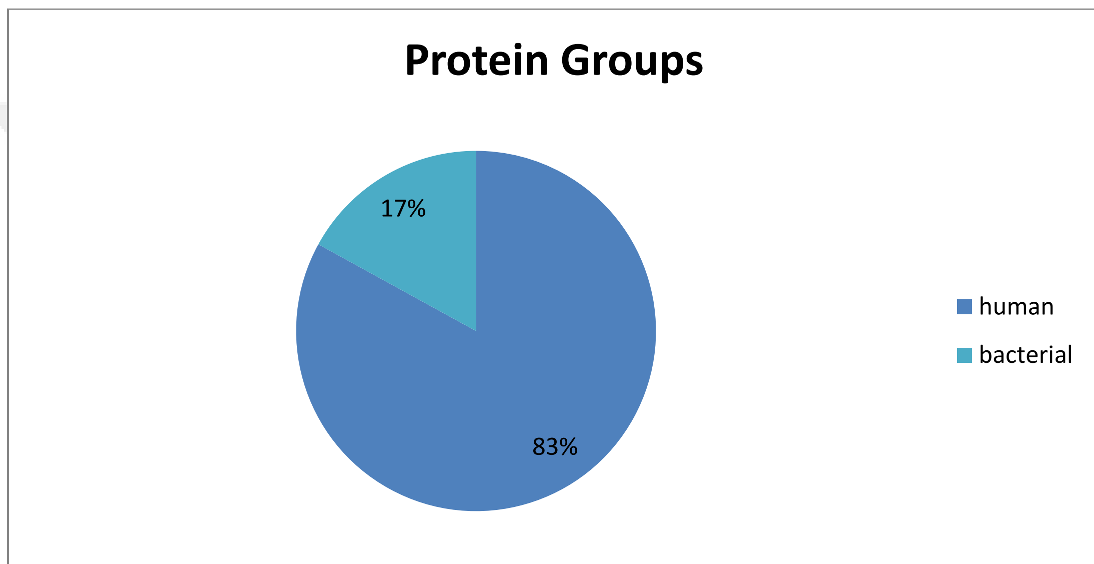

Figure 6-1 Human to bacterial protein groups percentages across the collected saliva samples. Out of the 435 protein groups identified, 361 were from human origin (83%) and 74 were bacterial protein groups (17%), showing human protein groups dominance over bacterial protein groups in terms of protein groups count.

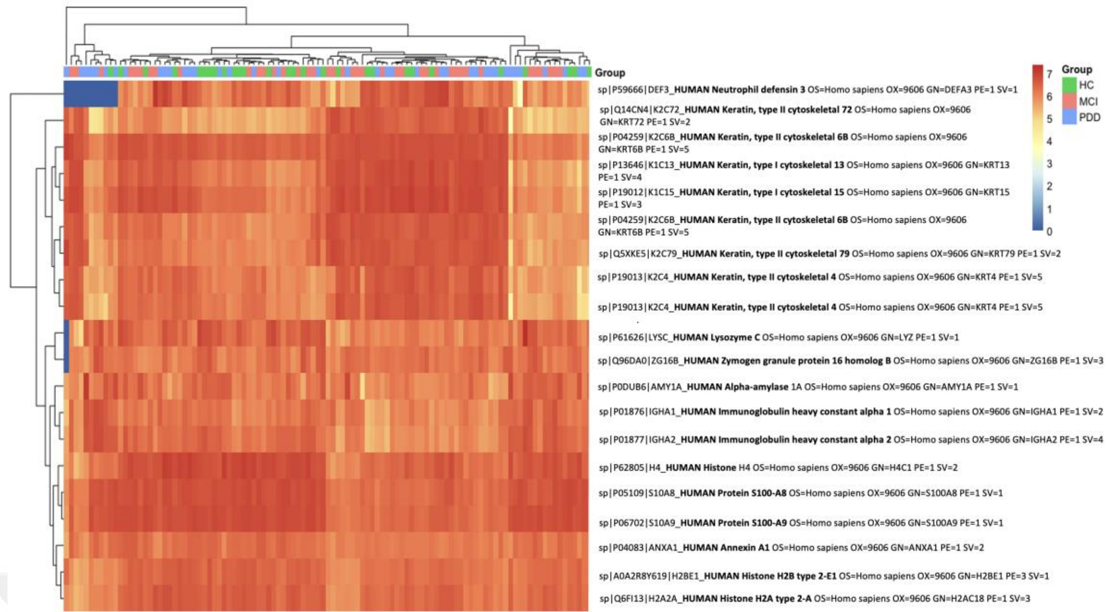

Figure 6-2 Heat-map of human protein groups abundance among the three study groups. The abundance of the top 20 most abundant human protein groups (right side) demonstrated through the three study groups (on top side): HC; Healthy Control (green), MCI; Parkinson's Disease with Mild Cognitive Impairment (pink), and PDD; Parkinson's Disease with Dementia (blue) in this heat-map generated using hierarchical clustering method showing no clear distinction between the study groups based on human protein groups abundances.

Table 6-2 Top 20 most abundant human protein groups

	db	UniqueIdentifier	EntryName	ProteinName	OS	OX	GN	PE	SV
1	sp	P59666	DEF3	Neutrophil defensin 3	Homo sapiens	9606	DEFA3	1	1
2	sp	Q14CN4	K2C72	Keratin, type II cytoskeletal 72	Homo sapiens	9606	KRT72	1	2
3	sp	P04259	K2C6B	Keratin, type II cytoskeletal 6B	Homo sapiens	9606	KRT6B	1	5
4	sp	P13646	K1C13	Keratin, type I cytoskeletal 13	Homo sapiens	9606	KRT13	1	4
5	sp	P19012	K1C15	Keratin, type I cytoskeletal 15	Homo sapiens	9606	KRT15	1	3
6	sp	P04259	K2C6B	Keratin, type II cytoskeletal 6B	Homo sapiens	9606	KRT79	1	5
7	sp	Q5XKE5	K2C79	Keratin, type II cytoskeletal 79	Homo sapiens	9606	KRT6B	1	2
8	sp	P19013	K2C4	Keratin, type II cytoskeletal 4	Homo sapiens	9606	KRT4	1	5
9	sp	P19013	K2C4	Keratin, type II cytoskeletal 4	Homo sapiens	9606	KRT4	1	5
10	sp	P61626	LYSC	Lysozyme C	Homo sapiens	9606	LYZ	1	1
11	sp	Q96DA0	ZG16B	Zymogen granule protein 16 homolog B	Homo sapiens	9606	ZG16B	1	3
12	sp	P0DUB6	AMY1A	Alpha-amylase 1A	Homo sapiens	9606	AMY1A	1	1
13	sp	P01876	IGHA1	Immunoglobulin heavy constant alpha 1	Homo sapiens	9606	IGHA1	1	2
14	sp	P01877	IGHA2	Immunoglobulin heavy constant alpha 2	Homo sapiens	9606	IGHA2	1	4
15	sp	P62805	H4	Histone H4	Homo sapiens	9606	H4C1	1	2
16	sp	P05109	S10A8	Protein S100-A8	Homo sapiens	9606	S100A8	1	1
17	sp	P06702	S10A9	Protein S100-A9	Homo sapiens	9606	S100A9	1	1
18	sp	P04083	ANXA1	Annexin A1	Homo sapiens	9606	ANXA1	1	2
19	sp	A0A2R8Y619	H2BE1	Histone H2B type 2-E1	Homo sapiens	9606	H2BE1	3	1
20	sp	Q6FI13	H2A2A	Histone H2A type 2-A	Homo sapiens	9606	H2AC18	1	3

db;data base, sp; UniProtKB/Swiss-Prot, OS; Organism name, OX; Organism identifier, GN; Gene Name, PE; Protein Existence, SV; Sequence Version.

Table 6-3 Bacterial Protein Groups Identified. The description of the protein group from bacterial source and the number of times it was assigned

Bacterial Protein Groups	
Protein group classification	**Number of assignments**
Bacterial Elongation Factor Tu	24
Bacterial Glyceraldehyde-3-phosphate dehydrogenase	1
Bacterial ATP synthase subunit beta	4
Bacterial hypothetical protein	13
Bacterial Fructose-bisphosphate aldolase	1
Bacterial Alkyl hydroperoxide reductase C	3
Bacterial Riboflavin biosynthesis protein RibBA	1
Bacterial BACT 30S ribosomal protein S19	1
BACT 50S ribosomal protein L7/L12	8
Methylmalonyl-CoA mutase large subunit	1
DNA-binding protein HU	2
DNA-binding protein HU 1	1
Methylmalonyl-CoA mutase	2
Glyceraldehyde#-3-phosphate dehydrogenase	2
Outer membrane protein 41	1
Methionine-binding lipoprotein MetQ	1
putative oxidoreductase YjmC	3
Enolase	3
AP-4-A phosphorylase	1
Putative dipeptidase	1
Ferritin BfrB	1

6.4 Post-identification and Statistical Analysis

6.4.1 Human to bacterial protein groups comparison

When comparing the PCA plots generated: all proteins including both human and bacterial (figure 6-3), human proteins (figure 6-4), and bacterial proteins (figure 6-5), it showed that only bacterial proteins do significantly differentiate the three study groups HC, PD-MCI and PDD ($p<0.05$).

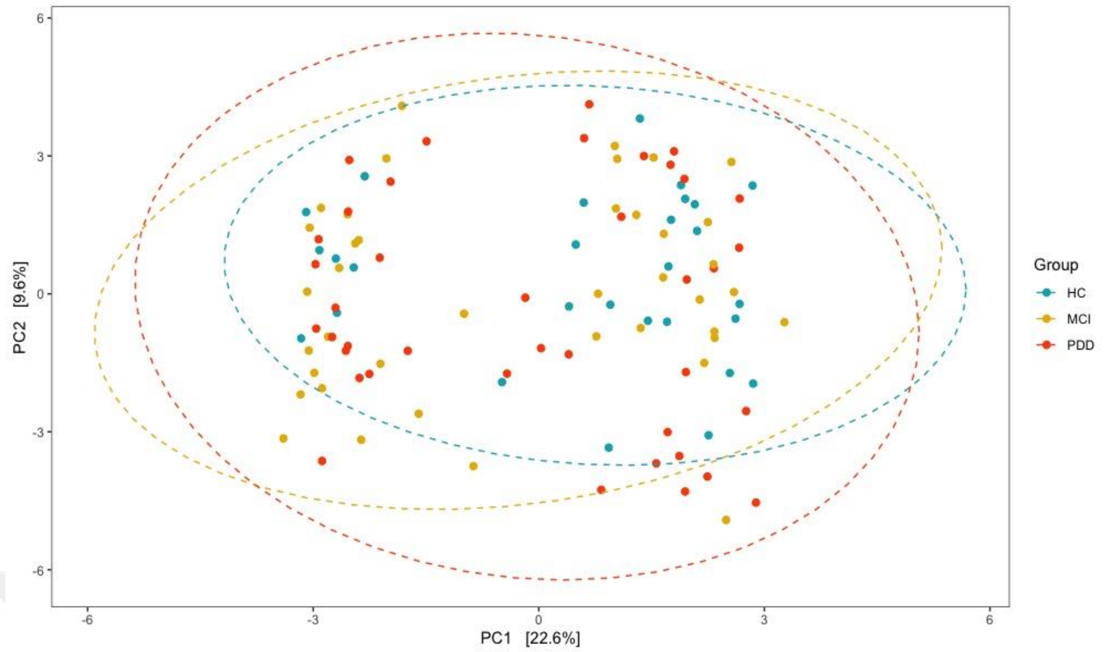

Figure 6-3 PCA for the study groups distinction according to both human and bacterial protein groups. When applied PCA including both human and bacterial protein groups, no significant distinction was found between the three study groups: HC; Healthy Control (blue), MCI; Parkinson's Disease with Mild Cognitive Impairment (yellow), and PDD; Parkinson's Disease with Dementia (red). (p=0.612).

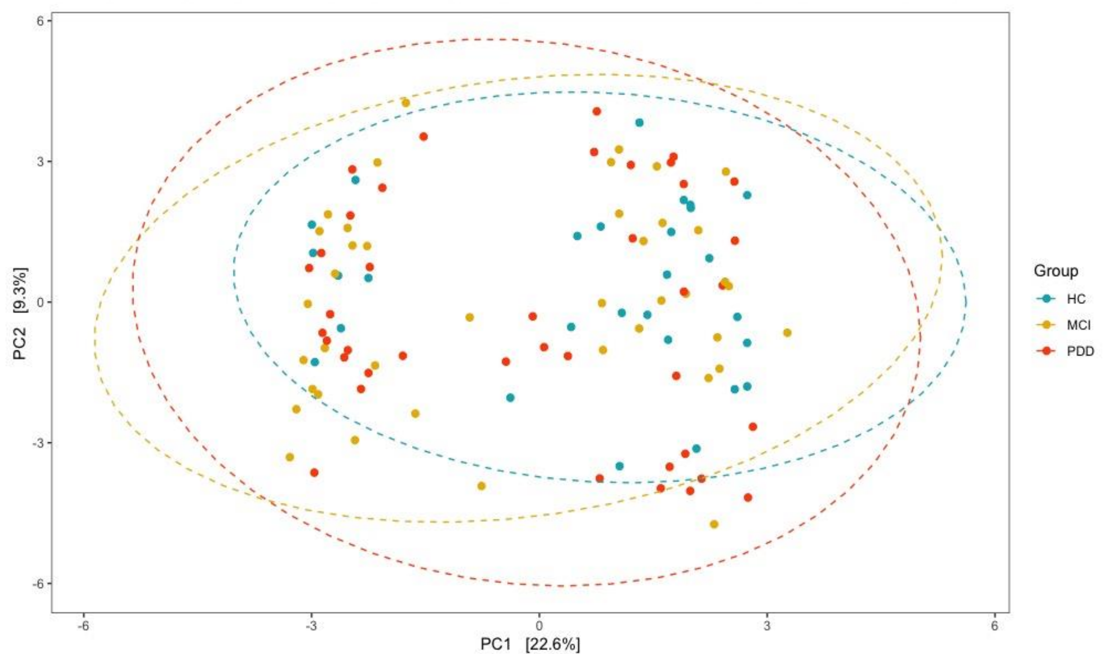

Figure 6-4 PCA for the study groups distinction according to human protein groups. When applied PCA including only human protein groups, no significant distinction was found between the three study groups: HC; Healthy Control (blue), MCI; Parkinson's Disease with

Mild Cognitive Impairment (yellow), and PDD; Parkinson's Disease with Dementia (red). (p=0.76).

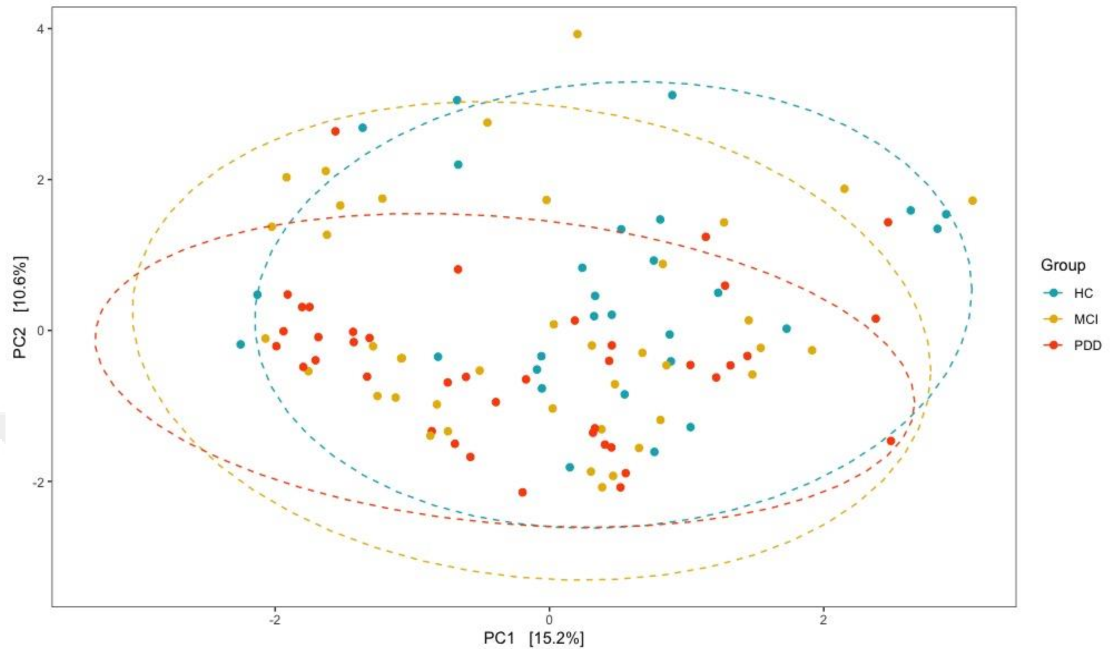

Figure 6-5 PCA for the study groups distinction according to bacterial protein groups. When applied PCA including only bacterial protein groups, significant distinction was found ($p<0.05$) between the three study groups: HC; Healthy Control (blue), MCI; Parkinson's Disease with Mild Cognitive Impairment (yellow), and PDD; Parkinson's Disease with Dementia (red). (p=0.046).

6.4.2 Protein groups and features associations detection

The results of MaAsLin2 association test between protein groups (human and bacterial, human, bacterial) and features of the study groups (HC, PD-MCI, PDD, age, sex, MMSE, and CDR) association were only found between bacterial proteins and the features PD-MCI, PDD, age, and sex. MaAsLin 2 protein groups to metadata associations with significant results, $p<0.05$ and $q<0.25$, are shown in (Table 6-4).

For MaAsLin2 proteins groups association to PD-MCI (figure 6-6) and PDD (figure6-7) showed that proteins group described as bacterial Elongation factor Tu (EF-Tu) have continuously decrease from HC to MCI-PD and to PDD groups, with $p<0.05$.

For age (figure 6-8), results showed slight increase of EF-Tu protein group by age.

Whereas for sex (figure 6-9), results showed association with the bacterial AP-4-A phoshorylase protein and increase in females compared to males.

Table 6-4 MaAsLin 2 protein groups to metadata associations results

Protein Group	metadata	value	coef	std	N	N.not.0	p-val	q-val	Fig
EF-Tu	Group	PDD	-6.16055	1.25996	108	59	3.78E-06	0.00085	6-7
EF-Tu	Group	MCI	-4.28282	1.12144	108	59	0.00023	0.025907	6-6
AP-4-A Phosphorylase	Sex	M	-2.10273	0.71650	108	100	0.00412	0.17372	6-9
EF-Tu	Age	Age	0.67659	0.25007	108	9	0.00799	0.19982	6-8

Coef; coefficient, N; number of data points, N.not.0; number of nonzero data points, p-val; p-value, q-val; q-value. Fig; the number of the figure demonstrating the result. EF-Tu; Bacterial Elongation Factor Tu, MCI; Parkinson's Disease with Mild Cognitive Impairment , PDD; Parkinson's Disease with Dementia (red)

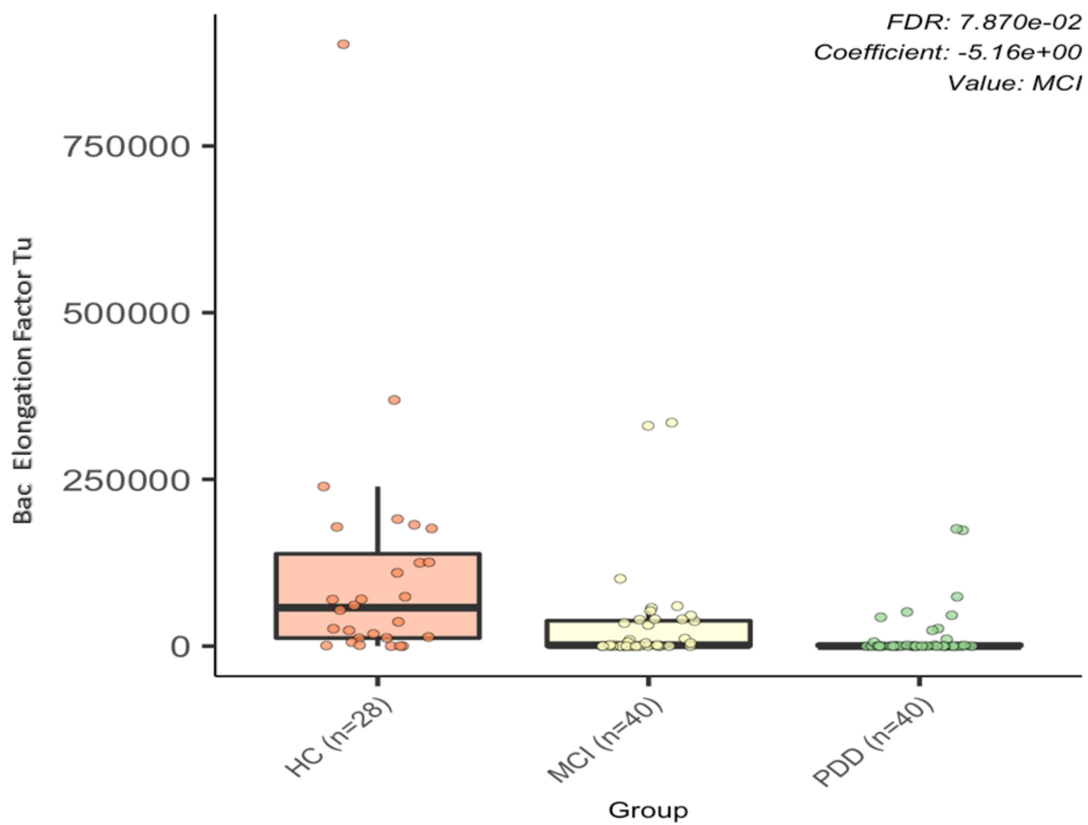

Figure 6-6 MaAsLin2 box plot for PD-MCI association with bacterial EF-Tu protein group abundance (molecule/cell). Using MaAsLin2, significant association was found between Mild Cognitive Impairment (MCI) and bacterial Elongation Factor-Tu (EF-Tu) protein group ($p<0.05$) and observed its gradual decrease in its abundance among the three study groups from HC; Healthy Control (red), to MCI; Parkinson's Disease with Mild Cognitive Impairment (yellow), and to PDD; Parkinson's Disease with Dementia (green).

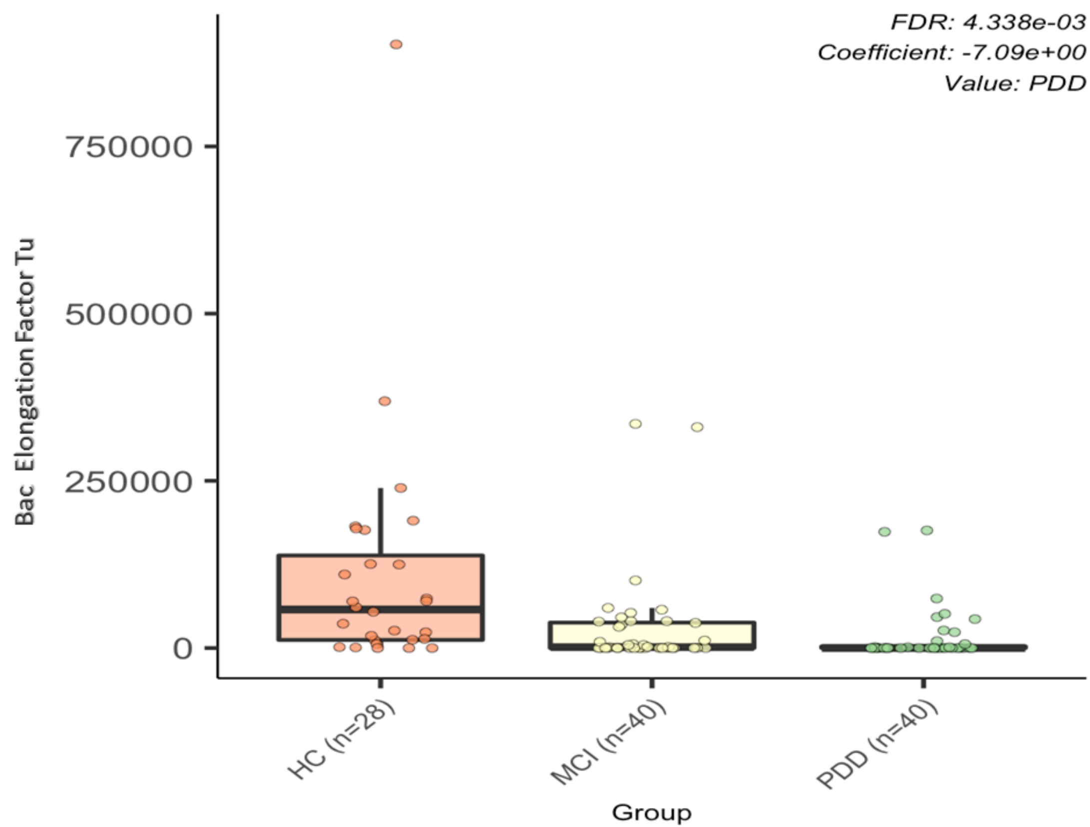

Figure 6-7 MaAsLin2 box plot for PDD association with bacterial EF-Tu protein group abundance (molecule/cell). Using MaAsLin2, significant association was found between PDD and bacterial Elongation Factor-Tu (EF-Tu) protein group ($p<0.05$) and observed its gradual decrease in its abundance among the three study groups from HC; Healthy Control (red), to MCI; Parkinson's Disease with Mild Cognitive Impairment (yellow), and to PDD; Parkinson's Disease with Dementia (green).

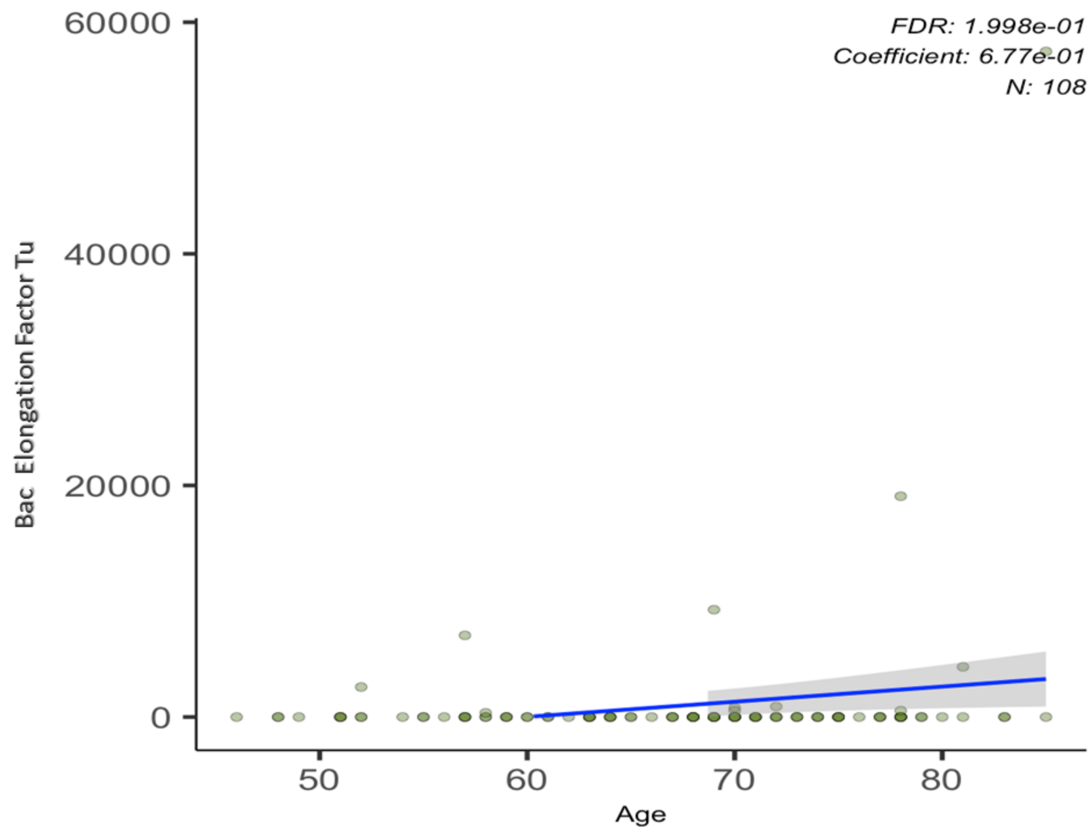

Figure 6-8 MaAsLin2 plot for age feature association with bacterial EF-Tu protein group abundance (molecule/cell).Using MaAsLin2, significant association was found between age and bacterial Elongation Factor-Tu (EF-Tu) protein group ($p<0.05$) and observed its increase by age.

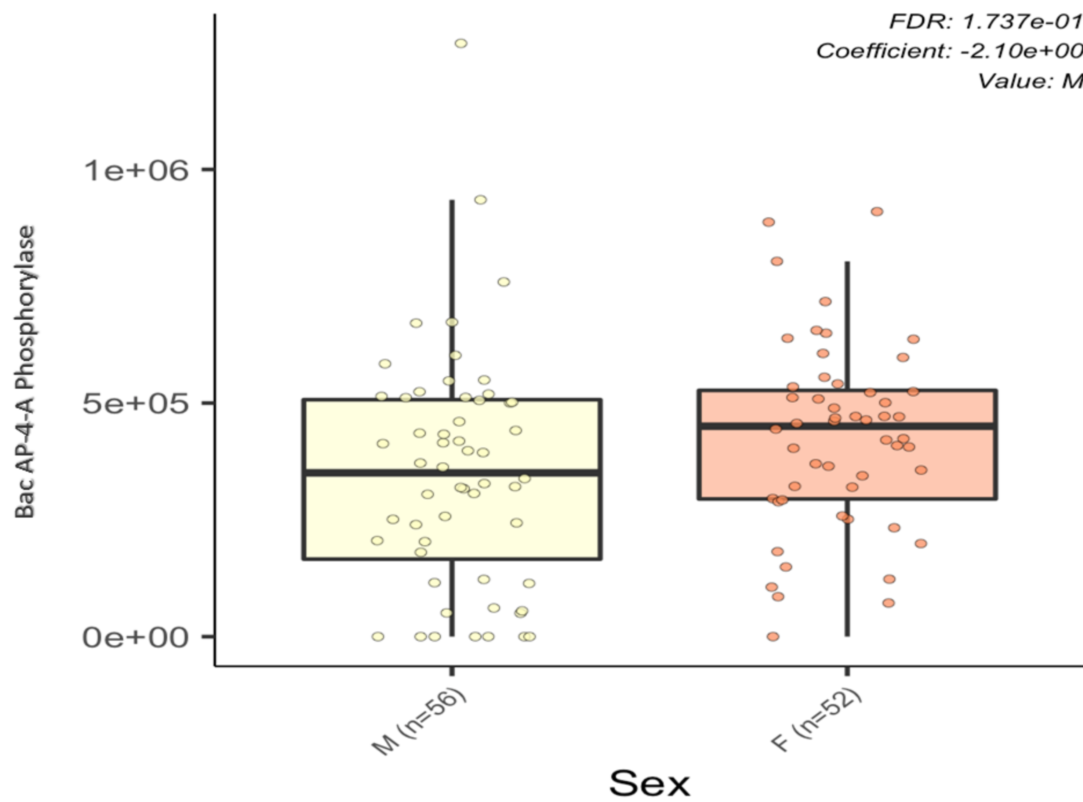

Figure 6-9 MaAsLin2 box plot for sex (M; male, F; female) feature association with bacterial AP-4-A phoshorylase protein group abundance (molecule/cell).Using MaAsLin2, significant association was found between sex feature and bacterial AP-4-A phoshorylase protein group ($p<0.05$) and observed its increase in F; females (red) compared to M; males (yellow).

6.4.3. Bacterial taxonomy demonstration

The bacterial common taxonomy tree for the most significant bacterial proteins with their associated feature are demonstrated in figure 6-10.

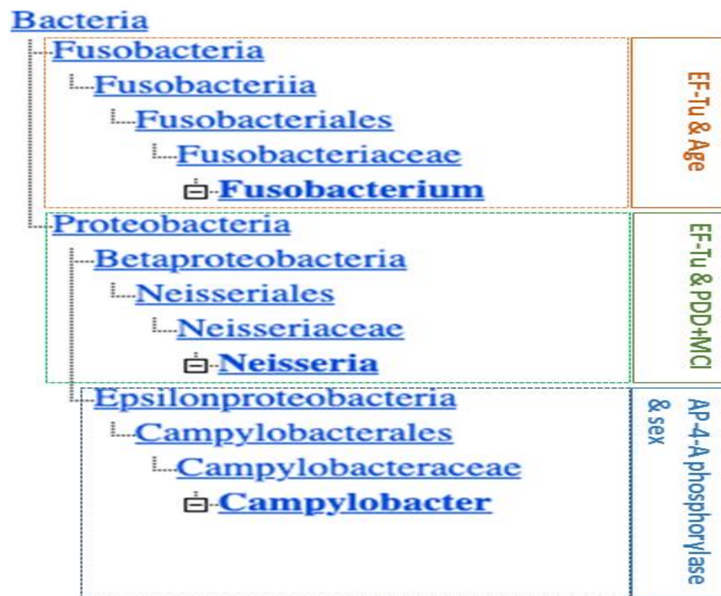

Figure 6-10 Bacterial common taxonomy tree with significant protein group and associated feature. *EF-Tu & Age;* Elongation Factor Tu protein group associated with age feature (orange), *EF-Tu & PDD+MCI*; Elongation Factor Tu protein group associated with the study groups and PDD; Parkinson's Disease with Dementia and MCI; Parkinson's Disease with Mild Cognitive Impairment (green), *AP-4-A phosphorylase & sex*; AP-4-A phosphorylase protein group associated with sex feature (blue).

7. DISCUSSION

Unstimulated saliva collection method was preferred over using stimulated method in order to avoid any possible subsequent changes in bacterial proteins due to use of paraffin gum in stimulated saliva collection method [35]. For sample preparation and protein isolation, Bead-Beating (BB) method was used as it provides effective disruption of cells and aid homogenization of samples [52]. Whereas for MS sample preparation, bottom-up proteomics work flow including protein digestion was followed as it is more commonly used and aided with more analytic tools compared to top-down proteomics, and as it is a peptide-oriented method so it suites the aims of this study [15]. For protein digestion, Filter-aided sample preparation (FASP) was chosen as it is considered an efficient and versatile method for protein extract processing [53]. For mass spectrometry, nano-Liquid Chromatography tandem Mass Spectrometry (nLC-MS/MS (Waters)) was chosen as it provides high sensitivity and efficiency [17][54].

The data analysis process was challenged with the large size of reference protein database including 1,165,589 protein sequences, along with the received MS raw data file in closed Waters raw format. In order to overcome these challenges the MS raw file format was converted to mzML open XML format using ProteoWizard [45] and then analyzed by using DIA-NN pipeline. DIA-NN pipeline was chosen as it can be freely downloaded and it provides data-independent acquisition (DIA) analysis of proteomics data, reliable statistical analysis, robust and flexible data modeling, user-friendly interphase, and high analysis speed [46]. DIA approach was chosen over DDA traditional method as some research show that DIA enhance reproducibility and proteomic depth, and decrease stochastic peptides detection [55], and the challenges associated with DIA data processing can be overcomed by using DIA-NN pipeline [19]. During peptides and proteins classification into protein groups, some protein groups had similar classifications, such that in human proteins number 8 and 9 in table 6-2, or bacterial proteins assignments in table 6-3, and this can be due to the classification method applied by DIA-NN or having protein orthologs.

As PCA results showed, despite the overlaps between some of the human and bacterial proteins among the three study groups, bacterial proteins PCA showed more clear distinctions between the three study groups compared to human proteins and both human/bacterial proteins PCA plots.

Also, as shown by the MaAsLin2 box plots, the protein group related with bacterial EF-Tu significantly showed decreasing between the three study groups, HC, MCI-PD and to PDD. Although EF-Tu seems to decrease with CI development, MaAsLin2 results for age showed that EF-Tu increased by age, demonstrating negative CI-age proportion in term of EF-Tu abundance. This result was an unexpected, where CI-aging relation was expected to be positive proportion; CI increases by age, thereby, further investigation is needed.

Briefly, EF-Tu is a G protein which aids aminoacyl-tRNA delivery to ribosome A-site and catalyzes its binding to allow proper mRNA to protein translational process [56]. EF-Tu along with the Elongation Factor G (EF-G), which facilitates tRNA-mRNA complex translocation on the ribosome, are considered the main players in the protein elongation stage during protein translation, and the elongation rate in overall is affected by other elongation factors, like the EF-Ts [57]. Even though different studies show that key regulator of translational process and its relation to health and disease is the protein elongation stage [57], no studies were reported on bacterial EF-Tu relation to PD. However, some studies were reported on the eukaryotic ortholog of the bacterial EF-Tu (eEF1A) [58] [59], the eukaryotic ortholog of the bacterial EF-G (eEF2) [58] [59], and eEF2 kinase (eEF2K) [60] and their possible relations with NDD such as AD and PD. Furthermore, several studies showed decrease of eEF1A and eEF2 expression in frontal lobe of AD and PD patients [58], and one of these studies was done by Garcia-Esparcia et al. 2015 [59], where they observed decrease of eEF1A and eEF2 proteins expression in PD patients' frontal cortex. Also, a study by Jan etl. 2018 [60] showed increase in the expression of eEF2K.

When looking at MaAsLin2 results for sex feature, although association was observed with bacterial AP-4-A phoshorylase protein, no meaningful conclusion in terms of CI monitoring was drawn.

8. CONCLUSIONS

In conclusion, it can be said that, bacterial proteins from the saliva samples differentiate the three study groups (HC,PD-MCI, and PDD), and out of the 435 protein groups obtained from the proteins abundance table the decrease in the abundance of the bacterial protein group described as *Elongation factor Tu* was found to be significantly associated with the progression of CI ($p<0.05$). Therefore, these results draw attentions to study bacterial proteins and look toward the protein Elongation Factor Tu as a possible biomarker for CI detection and monitoring in PD. Also, this addresses saliva metaproteomics as a promising non-invasive biomarkers source for CI detection and monitoring in PD and other NDD that need to be further investigated.

9. REFERENCES

[1] A. B. Megha Agrawal, “Molecular diagnostics of neurodegenerative disorders,” *Front. Mol. Biosci.*, 2015.

[2] L. Reed-Guy, “Brain disorders,” 2017. [Online]. Available: https://www.healthline.com/health/brain-disorders.

[3] P. Foundation, “What is Parkinson’s.” [Online]. Available: https://www.parkinson.org/understanding-parkinsons/what-is-parkinsons.

[4] M. H. A Berardelli, J C Rothwell, P D Thompson, “Pathophysiology of bradykinesia in Parkinson’s disease,” *oxord Univ. Press*, 2001.

[5] R.-M. 2015 Chin-HsienLin, “Review: Biomarkers of cognitive decline in Parkinson’s disease,” *Elsevier*, 2015.

[6] Presbyterian Hoag Memorial Hospital, “Cognitive Severity Stages (Normal Aging - Dementia),” 2005. [Online]. Available: https://www.hoag.org/specialties-services/neurosciences/programs/memory-cognitive-disorders/types-of-memory-cognitive-disorders/cognitive-severity-stages/.

[7] Khatun Ayesha and M. S. , Ross W. Paterson, “University College London/University of Gothenburg PhD course ‘Biomarkers in neurodegenerative diseases’ 2019—course organisation,” 2020.

[8] M. A. Anna-Lena Frey, Michael Karran, Rafael Jimenez, James Baxter, C. Dennis Chan, John Crawford, Paul Dagum, Richard Everson, Chris Hinds, C. Holmes, Zoe Kourtzi, David J. Llewellyn, Ann-Marie Mallon, Richard Milne, C. Mummery, Mike Oldham, Sarah Slight, Kirstie Whitaker, Zuzana Walker, and Routledge, “Harnessing the Potential of Digital Technologies for the Early Detection of Neurodegenerative Diseases,” 2019.

[9] “Appendix:Suffixes -ome and -omics.” [Online]. Available: https://en.wiktionary.org/wiki/Appendix:Suffixes_-ome_and_-omics.

[10] Satya P. Yadav, “The Wholeness in Suffix -omics, -omes, and the Word Om,” *J. Biomol. Tech.*, 2007.

[11] A. S. C.-P. Maria E. Diaz-Ortiz, “Omics in neurodegenerative disease: Hope or hype?,” 2019.

[12] H. F. W. Geoffrey S. Ginsburg, *Essentials of Genomic and Personalized Medicine*. 2010.

[13] NIH, “definition of proteome.” [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/proteome.

[14] “PROTEIN ANALYSIS TECHNIQUES EXPLAINED.” [Online]. Available: https://www.atascientific.com.au/3-protein-analysis-techniques/.

[15] G. B. D. Rafael D. Melani, Fabio C. S. Nogueira, “It is time for top-down venomics,” *J. Venom. Anim. Toxins Incl. Trop. Dis.*, 2017.

[16] K. Harrata, “Mass Spectrometry Tutorial.” [Online]. Available: https://www.cif.iastate.edu/mass-spec/ms-tutorial. [Accessed: 24-Jun-2022].

[17] L. G. Eric C Schirmer, John R Yates-Iii, “MudPIT: A Powerful Proteomics Tool for Discovery,” *Discov. Med.*, 2009.

[18] B. R. Patrick G A Pedrioli, Jimmy K Eng, Robert Hubley, Mathijs Vogelzang, Eric W Deutsch, C. C. Brian Pratt, Erik Nilsson, Ruth H Angeletti, Rolf Apweiler, Kei Cheung, S. G. O. Henning Hermjakob, Sequin Huang6, Randall K Julian, Jr, Eugene Kapp, Mark E McComb, W. Z. Gilbert Omenn, Norman W Paton, Richard Simpson, Richard Smith, Chris F Taylor, and & R. Aebersold, “A common open representation of mass spectrometry data and its application to proteomics research,” *Nature*, 2004.

[19] M. R. Vadim Demichev, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, “DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput,” *Nat. Methods*, 2020.

[20] “Galaxy-p.” [Online]. Available: http://129.114.16.192/.

[21] “MetaProteome Analyzer (MPA).” [Online]. Available: www.mpa.ovgu.de.

[22] “Progenesis-QI (Waters) software.” [Online]. Available: https://www.nonlinear.com/progenesis/qi-for-proteomics/.

[23] M. B. S. Lana M. Chahine, “Parkinson’s Disease Biomarkers: Where Are We and Where Do We Go Next?,” *Int. Park. Mov. Disord. Soc.*, 2017.

[24] D. Weatherspoon, “Cerebrospinal Fluid Culture,” 2016. .

[25] M. P. M. Melisa Proulx BA, François P. De Courval MSc, Michael A. Wiseman DDS, FASGD, “Salivary production in Parkinson’s disease,” *Mov. Disord.*, 2004.

[26] E. E. B. Maria G Cersosimo, Gabriela B Raina, Cristian R Calandra, Alejandro Pellene, Cristiam Gutiérrez, Federico E Micheli, “Dry mouth: an overlooked autonomic symptom of Parkinson’s disease,” *J. Parkinsons. Dis.*, 2011.

[27] Z. Khurshid, S. Zohaib, S. Najeeb, M. S. Zafar, P. D. Slowey, and K. Almas, “Human saliva collection devices for proteomics: An update,” *Int. J. Mol. Sci.*, vol. 17, no. 6, 2016.

[28] K. R. Amato, “An introduction to microbiome analysis for human biology applications,” *Am. J. Hum. Biol.*, 2016.

[29] P. Amon and I. Sanderson, “What is the microbiome ?,” pp. 258–261, 2017.

[30] S. K. M. Gil Sharon, Timothy R. Sampson, Daniel H. Geschwind, “The Central Nervous System and the Gut Microbiome,” 2016.

[31] G. A. O. Fabrice Jean-Pierre1, Michael A. Henson, “Metabolic Modeling to Interrogate Microbial Disease: A Tale for Experimentalists,” *Frotiers Mol. Biosci.*, 2021.

[32] "Meaning of meta-." [Online]. Available: https://www.dictionary.com/browse/meta-.

[33] D. J. G. Robert H Mills, Yoshiki Vázquez-Baeza , Qiyun Zhu, Lingjing Jiang, James Gaffney, Greg Humphrey, Larry Smarr, Rob Knight, "Evaluating Metagenomic Prediction of the Metaproteome in a 4.5-Year Study of a Patient With Crohn's Disease," *mSystems*, 2019.

[34] A. M. Sonia Gomar-Vercher, Aurea Simón-Soro, "Stimulated and unstimulated saliva samples have significantly different bacterial profiles," *PLoS One*, 2018.

[35] C. D. W. Taichi Inui, Robert J. Palmer Jr., Nehal Shah, Wei Li, John O. Cisar, "Effect of mechanically stimulated saliva on initial human dental biofilm formation," *Nature*, 2019.

[36] A. F. Monika Figura, Ewa Sitkiewicz, Bianka ´ Swiderska, Łukasz Milanowski, Stanisław Szlufik, Dariusz Koziorowski, "Proteomic Profile of Saliva in Parkinson ' s Disease Patients : A Proof of Concept Study," *brain Sci.*, 2021.

[37] W. R. L. Maxime François, Avinash Karpe, Jian-Wei Liu, David Beale, Maryam Hor, Jane Hecker, Jeff Faunt, John Maddison, Sally Johns, James Doecke, Stephen Rose, "Salivaomics as a Potential Tool for Predicting Alzheimer's Disease During the Early Stages of Neurodegeneration," *J. Alzheimer's Dis.*, 2021.

[38] G. B. P. Joseph M.Mastersab, Alastair J.Noyceac, Thomas T.Warnerc, Gavin Giovannonia, "Elevated salivary protein in Parkinson's disease and salivary DJ-1 as a potential marker of disease severity," *Parkinsonism Relat. Disord.*, 2015.

[39] X.-Y. W. Wen-Yan Kang, Qiong Yang, Xu-Feng Jiang, Wei Chen1, Lin-Yuan Zhang, J. L. Li-Na Zhang, Thomas J. Quinn, and S.-D. Chen, "Salivary DJ-1 could be an indicator of Parkinson's disease progression," *Front. Aging Neurosci.*, 2014.

[40] R. J. R. Mustafa C. Beker, Berrak Caglayan, Ahmet B. Caglayan, Taha Kelestemur, Esra Yalcin, Aysun Caglayan, Ulkan Kilic, Ahmet T. Baykal and E. Kilic, "Interaction of melatonin and Bmal1 in the regulation of PI3K/AKTpathway components and cellular survival," *Nature*, 2019.

[41] "Proteomics | Core Facilities | Research Institute for Health Sciences and Technologies |Istanbul Medipol University." [Online]. Available: https://www.medipol.edu.tr/en/research/research-centers/health-science-and-technologies-research-institute/core-facilities/proteomics.

[42] L. J. H. Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, 2010.

[43] Z. J. J. Zhu, L. Tian, P. Chen, M. Han, L. Song, X. Tong, X. Sun, F. Yang, Z. Lin, X. Liu, C. Liu, X. Wang, Y. Lin, K. Cai, Y. Hou, X. Xu, H. Yang, J. Wang, K. Kristiansen, L. Xiao, T. Zhang, H. Jia, "Over 50,000 Metagenomically Assembled Draft Genomes for the Human Oral Microbiome Reveal New Taxa," *Genomics, Proteomics Bioinforma.*, 2021.

[44] "UniProt official website." [Online]. Available: https://www.uniprot.org/.

[45] "ProteoWizard." [Online]. Available: https://proteowizard.sourceforge.io/.

[46] "DIA-NN pipeline." [Online]. Available: https://github.com/vdemichev/DiaNN.

[47] L. L. Sudarshan A. Shetty, "microbiomeutilities: An R package for utilities to guide in-depth marker gene amplicon data analysis." Zenodo, 2018.

[48] S. H. Paul J. McMurdie, "phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data," *PLoS One*, 2013.

[49] J. W. [aut] Jari Oksanen [aut, cre], Gavin L. Simpson [aut], F. Guillaume Blanchet [aut], Roeland Kindt [aut], Pierre Legendre [aut], Peter R. Minchin [aut], R.B. O'Hara [aut], Peter Solymos [aut], M. Henry H. Stevens [aut], Eduard Szoecs [aut], Helene Wagner [aut], , "vegan: Community Ecology Package," 2022. [Online]. Available: cran.r-project.org/package=vegan.

[50] H. C. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, Tickle TL, Weingart G, Ren B, Schwager EH, Chatterjee S, Thompson KN, Wilkinson JE, Subramanian A, Lu Y, Waldron L, Paulson JN, Franzosa EA, Bravo HC, "Multivariable association discovery in population-scale meta-omics studies," *PLoS Comput. Biol.*, 2021.

[51] National Library of Medicine (NIH), "Taxonomy Tools." [Online]. Available: https://www.ncbi.nlm.nih.gov/taxonomy.

[52] SPEX SamplePrep, "BEAD BEATING INTRODUCTION & GUIDELINES." [Online]. Available: https://www.spexsampleprep.com/knowledge-base/resources/application_notes/0118-162623-Bead_beating_Intro_guide_3.pdf.

[53] J.R.Wiśniewski, "Filter-Aided Sample Preparation: The Versatile and Efficient Method for Proteomic Analysis," in *Methods in Enzymology*, 2016.

[54] G. C. Marco Gaspari, "Nano LC–MS/MS: A Robust Setup for Proteomic Analysis," in *Nanoproteomics*, 2011.

[55] L. R. Roland Bruderer, Oliver M.Bernhardt, Tejas Gandhi, Yue Xuan, Julia Sondermann, Manuela Schmidt, David Gomez-Varela, "Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results," *Mol. Cell. Proteomics*, 2017.

[56] S. P. D. Kate L. Harvey , Veronica M. Jarocki, Ian G. Charles, "The Diverse Functional Roles of Elongation Factor Tu (EF-Tu) in Microbial Pathogenesis," *Front. Mol. Biosci.*, 2019.

[57] G. S. Benjin Xu, Ling Liu, "Functions and Regulation of Translation Elongation Factors," *Frotiers Mol. Biosci.*, 2022.

[58] S. L. A. Mridu Kapur, Caitlin E. Monaghan, "Regulation of mRNA Translation in Neurons—A Matter of Life and Death," *Neuron*, 2017.

[59] I. F. Paula Garcia-Esparcia, Karina Hernández-Ortega, Anusha Koneti, Laura Gil, Raul Delgado-Morales, Ester Castaño, Margarita Carmona, "Altered machinery of protein synthesis is region- and stage-dependent and is associated with α-synuclein oligomers in Parkinson's disease," 2015.

[60] P. H. S. Asad Jan , Brandon Jansonius, Alberto Delaidelli, Forum Bhanshali , Yi Andy An , Nelson Ferreira, Lisa M. Smits, Gian Luca Negri, Jens C. Schwamborn , Poul H. Jensen, Ian R. Mackenzie, Stefan Taubert, "Activity of translation regulator eukaryotic elongation factor-2 kinase is increased in Parkinson disease brain and its inhibition reduces alpha synuclein toxicity," *Acta Neuropathol. Commun.*, 2018.

10.ETHICS COMMITTEE APPROVAL

İSTANBUL MEDİPOL ÜNİVERSİTESİ GİRİŞİMSEL OLMAYAN KLİNİK ARAŞTIRMALAR ETİK KURULU KARAR FORMU

E-10840098-772.02-4353 01/08/2022

APPLICATION INFORMATION					
	THE FULL NAME OF THE RESEARCH	METAPROTEOMIC ANALYSIS OF SALIVA SAMPLES FROM PARKINSON'S DISEASE PATIENTS WITH COGNITIVE IMPAIRMENTS			
	COORDINATOR / RESPONSIBLE INVESTIGATOR TITLE / NAME / SURNAME	Prof. Dr. Süleyman YILDIRIM			
	EXPERTISE OF THE COORDINATOR / RESPONSIBLE RESEARCHER	Medical Microbiology			
	COORDINATOR / RESPONSIBLE RESEARCH CENTER	Istanbul Medipol University International School of Medicine			
	SUPPORTING	-			
	CENTERS PARTICIPATING IN RESEARCH	ONE CENTER ☐	MULTI-CENTERED ☒	NATIONAL ☒	INTERNATIONAL ☐

Bu belge, güvenli elektronik imza ile imzalanmıştır.
Evrağınızı https://turkiye.gov.tr/istanbul-medipol-universitesi-ebys linkinden 32B2E64BXE kodu ile doğrulayabilirsiniz.

Sa