2020

Gene–Environment Interaction In Parkinson Disease: The Gut Microbiome

Zachary D. Wallen
University of Alabama at Birmingham

GENE–ENVIRONMENT INTERACTION IN PARKINSON DISEASE: THE GUT MICROBIOME

by

ZACHARY D. WALLEN

DAVID G. STANDAERT, COMMITTEE CHAIR
ELLIOT J. LEFKOWITZ
HAYDEH PAYAMI, MENTOR
ERIK D. ROBERSON
HEMANT K. TIWARI

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2020

GENE–ENVIRONMENT INTERACTION IN PARKINSON DISEASE: THE GUT MICROBIOME

ZACHARY D. WALLEN

GENETICS, GENOMICS, AND BIOINFORMATICS

ABSTRACT

Parkinson disease (PD) is a progressive neurodegenerative disease with no cure. Majority of cases are idiopathic, and the cause is unknown. Studies have been conducted in human and animals to identify PD risk factors, resulting in a list of genetic and environmental factors that modestly increases risk of PD. Still, no individual risk factor fully explains the cause of PD, and neither has the combination of these factors. Additional avenues of research are being investigated to find potential triggers of PD, and factors that might modify the progression of PD. This includes research into the gut microbiome, as gut health perturbations are frequently documented and studies have shown a dysbiotic gut microbial community in PD. This also includes the search for genetic modifiers of PD onset, which might provide a mechanism to modify disease progression and prolong onset of PD. This dissertation focuses on both of the aforementioned areas of PD research.

We first sought to identify genetic modifiers for idiopathic PD by performing a genome-wide association study (GWAS) of age at PD diagnosis using 2,000 PD patients. Then, we performed a microbiome-wide association study (MWAS) in two additional cohorts to characterize gut microbial alterations observed in PD. We then investigated if opportunistic pathogens found enriched in PD gut might interact with genetic susceptibility of PD.

From our GWAS of age at PD diagnosis, we detected two potentially independent signals associated with an earlier PD diagnosis of ~6 years in a gene involved in neuronal plasticity and response to injury. Our MWAS of the PD gut microbiome revealed 15 bacterial genera significantly associated with PD, three of which were opportunistic pathogens, enriched in PD, that were part of a poly-microbial group of correlated genera, also enriched in PD. We detected potential interaction between these opportunistic pathogens and genetic susceptibility of PD conferred by genetic variants at the 3′ end of *SNCA*, the gene most highly associated with PD risk, which codes for the pathological hallmark of PD. Results provide potential leads for further research in humans and animals to see if findings have biological implications for PD disease progression.

Keywords: Parkinson disease, GWAS, gut microbiome, age at diagnosis, opportunistic pathogens, SNCA

DEDICATION

I dedicate this dissertation to my loved ones. To my wife, thank you for all that you have done to support me through this program. You have been my rock and the most amazing life partner one could ask for during this time. To my parents, thank you for always supporting and believing in me and pushing me to reach for the highest of goals. Most of all, thank you to my son (and any of his future siblings), you are my world and the biggest motivator in my life.

ACKNOWLEDGEMENT

I would like to thank my mentor, Dr. Haydeh Payami, for her guidance, support, insight, and teachings throughout my dissertation study. The time spent as her mentee has been crucial to my development as a strong, well-rounded scientist who is able to think critically and creatively to ask and answer important scientific questions. I would also like to thank my committee members, Drs. David Standaert, Elliot Lefkowitz, Erik Roberson, and Hemant Tiwari, for their support and advice during my dissertation. Lastly I would like to thank Payami lab members past and present, especially Erin Hill-Burns for allowing me to learn from her early on in my program, and Matthew Lewis for being a great lab mate, continued friend, and critical component to the research presented for this dissertation.

I would like to thank the Graduate Biomedical Sciences and my theme, Genetics, Genomics, and Bioinformatics, for their support and assistance during the program; the Neurology department, and specifically the Center for Neurodegeneration and Experimental Therapeutics, in the School of Medicine for their support and award of my T32 training grant; and the National Institutes of Health for their financial support through my T32 training grant.

TABLE OF CONTENTS

Page

APPENDIX

LIST OF TABLES

Table *Page*

PLASTICITY-RELATED GENE 3 (*LPPR1*) AND AGE AT DIAGNOSIS OF PARKINSON DISEASE

APPLICATION OF SIXTEEN DIFFERENTIAL ABUNDANCE METHODS TO TWO LARGE PARKINSON DISEASE GUT MICROBIOME DATASETS

CHARACTERIZING DYSBIOSIS OF GUT MICROBIOME IN PD: EVIDENCE FOR OVERABUNDANCE OF OPPORTUNISTIC PATHOGENS

EVIDENCE FOR INTERACTION BETWEEN GENETIC VARIATION IN THE *SNCA* LOCUS AND ABUNDANCE OF OPPORTUNISTIC PATHOGENS IN THE PARKINSON DISEASE GUT MICROBIOME

LIST OF FIGURES

Figure *Page*

PLASTICITY-RELATED GENE 3 (*LPPR1*) AND AGE AT DIAGNOSIS OF PARKINSON DISEASE

APPLICATION OF SIXTEEN DIFFERENTIAL ABUNDANCE METHODS TO TWO LARGE PARKINSON DISEASE GUT MICROBIOME DATASETS

CHARACTERIZING DYSBIOSIS OF GUT MICROBIOME IN PD: EVIDENCE FOR OVERABUNDANCE OF OPPORTUNISTIC PATHOGENS

EVIDENCE FOR INTERACTION BETWEEN GENETIC VARIATION IN THE *SNCA* LOCUS AND ABUNDANCE OF OPPORTUNISTIC PATHOGENS IN THE PARKINSON DISEASE GUT MICROBIOME

GENERAL INTRODUCTION

Parkinson disease

Overview

Parkinson disease (PD) is a common, progressive, and debilitating neurodegenerative disease affecting approximately 1% of the population over the age of 60 [de Lau & Breteler 2006]. The direct and indirect costs of PD including medical treatment, payments from social security, and lost income due to the inability for individuals to work is estimated to be around $14.4 billion dollars every year in the U.S. alone [Kowal et al. 2013]. PD is an age related disease with age being the largest risk factor for developing PD [Pange et al. 2019]. As the population ages, the economic burden of PD will continue to grow in the next 20 years. By 2037, PD prevalence has been projected to increase to 1.6 million in the United States with an economic burden greater than $79 billion [Yang et al. 2020].

Parkinson disease is primarily considered a movement disorder due to the core clinical features of PD involving deficits in motor function. The cardinal motor symptoms, used in the diagnostic criteria for PD, includes bradykinesia, resting tremor, and rigidity [Obeso et al. 2017]. Postural instability is also a common motor symptom of PD, but it is not used as part of the core diagnostic criteria as it usually occurs at later stages in the disease course [Obeso et al. 2017]. Motor symptoms of PD are usually caused by reduced levels of dopamine from loss of dopaminergic neurons in the substantia nigra pars compacta of the brain [Obeso et al. 2017]. The most widely used therapy to date for managing the motor symptoms of PD is levodopa, which is converted in the body to dopamine, and is used to replace the dopamine lost to neuronal cell loss. This therapy has its own drawbacks, however, as long term use of this drug causes its own motor dysfunctions and other non-motor issues [Obeso et al. 2017], and it does not modify the disease progression, only the motor symptoms experienced by a patient. The mechanisms causing dopaminergic neuronal cell loss is still unknown, therefore, a disease modifying therapy is still not available for use with PD. Neuronal cell death may be due to the dysfunction in a number of cellular pathways including misfolding of a key pathological protein in PD called α-synuclein [Michel, Hirsch & Hunot 2016]. The pathological hallmarks of idiopathic PD are inclusions of misfolded α-synuclein in neuronal cell bodies or neuronal processes, collectively referred to as Lewy pathology (LP) [Dickson et al. 2009]. The presence of LP in conjunction with dopaminergic neuronal cell loss in the substantia nigra is used at autopsy for post-mortem confirmation of a PD diagnosis [Kalia & Lang 2015]. Although LP and its main component α-synuclein are key pathological features of PD, the role they play in the initiation and/or progression of PD is still under investigation.

Once thought to be solely a disease of the brain, it is now appreciated that PD is a systemic condition that affects multiple areas of the human body. Multiple non-motor symptoms have been documented regularly in the course of PD progression, some of which occur years before onset of motor symptoms, and therefore, are thought to be part of a prodromal phase of PD. Conditions and symptoms such as rapid eye movement (REM) sleep behavioral disorder, constipation, and hyposmia (decreased sense of smell) are some of the earliest manifestations of PD, having been documented to occur 2 to >15 years before onset of motor symptoms with an estimated relative risk for PD of 50, 2.5, and 5 respectively [Pont-Sunyer et al. 2015; Obeso et al. 2017]. The prevalence of these non-motor manifestations are common in prodromal PD with REM sleep behavior disorder, constipation, and hyposmia affecting approximately 40 – 60% of PD patients before onset of motor symptoms [Pont-Sunyer et al. 2015]. Other non-motor manifestations of PD that might occur earlier or later in the disease course include cognitive impairment such as mild cognitive impairment and dementia, anxiety, depression, apathy, hallucinations, and autonomic dysfunction, some of which may be caused by the actual dopamine replacement therapies currently used in PD [Obeso et al. 2017].

Although the mechanism behind the initiation and progression of idiopathic PD is still unknown, multiple factors have been identified through human and animal studies that associate with risk of PD. Through epidemiological and experimental animal studies, multiple environmental factors have been associated with increased risk of PD including exposure to pesticides, chlorinated solvents, head injury, and polychlorinated biphenyls as well as certain occupations, environments, and activities such as farming, rural living, and well water consumption [Tanner 2010; Obeso et al. 2017]. Through the same study mechanisms, environmental factors have also been associated with decreased risk of PD including lifestyle factors such as consumption of caffeinated coffee and/or tea, smoking, lower cholesterol levels, certain dietary patterns, physical activity, and use of non-steroidal anti-inflammatory drugs [Tanner 2010; Obeso et al. 2017], some of which have been shown to have a combinatory effect on PD risk [Powers et al. 2008]. Through large meta-analyses of genome-wide association studies (GWAS) of PD risk, upwards of 90 genetic susceptibility loci have been identified for PD, the most significant being located near the α-synuclein gene, *SNCA* [Chang et al. 2017; Nalls et al. 2019]. Performing gene-environment interaction studies in both human and animals have also revealed combinations of genetic variants and environmental exposures that influence the risk of PD [Cannon & Greenamyre 2013; Hamza et al. 2011; Hill-Burns et al. 2013; Biernacka et al. 2016]. Unfortunately, no identified environmental or genetic factor, individually or in combination, has fully explained the cause of PD, therefore, the investigation for a triggering event of PD is ongoing.

Genetics of risk

In respect to genetics, PD is usually referred to as monogenic or idiopathic. Monogenic PD refers to PD that is caused by highly penetrant, rare variants that are sufficient to cause disease on their own [Blauwendraat, Nalls & Singleton 2020]. This form of PD is rare, as all monogenic forms of PD combined only make up 30% of familial PD cases and 3% to 5% of sporadic PD cases [Kumar, Djarmati-Westenberger & Grunewald 2011]. Genetic variants linked to monogenic forms of PD with high confidence includes missense mutations in *SNCA*, *PRKN*, *PINK1*, *DJ-1*, *ATP13A2*, *FBXO7*, *PLA2G6*, and *VPS35*, which are inherited in either a autosomal dominant (*SNCA*, *VPS35*) or recessive (*PRKN*, *PINK1*, *DJ-1*, *ATP13A2*, *FBXO7*, *PLA2G6*) manner [Blauwendraat, Nalls & Singleton 2020]. Duplications and triplications of the *SNCA* gene are also causes of monogenic, autosomal dominant PD with an increasing phenotype severity with number of gene copies [Klein & Schlossmacher 2006]. All mutations referenced above most likely result in the loss of function of the encoded protein, except for mutations and multiplications in *SNCA,* which cause a gain of function or overexpression of α-synuclein [Blauwendraat, Nalls & Singleton 2020]. Mutations in the genes *LRRK2* and *GBA* have also been reported to cause an autosomal dominant form of PD, but these differ from the previously mentioned mutations as they are common variants with an incomplete penetrance [Hernandez, Reed & Singleton 2016; Blauwendraat, Nalls & Singleton 2020]. Mutations in *LRRK2* are currently considered the most common genetic causes of late-onset PD, found in approximately 10% of autosomal dominant familial PD cases and 4% of sporadic PD cases [Hernandez, Reed & Singleton 2016]. Majority of genes linked to monogenic forms of PD seem to revolve around similar biological pathways that play roles in vesicular trafficking, mitochondrial function and health, endosome-lysozome pathway, and cellular response to stress [Hernandez, Reed & Singleton 2016].

Idiopathic PD, also called sporadic PD, refers to cases of PD that have no known cause, which make up ~95% of PD cases and will be the focus of this dissertation. Although the cause of PD is still unknown, we now have a long list of genetic risk factors that may predispose someone to develop PD [Nalls et al. 2019]. Unlike the variants discovered for monogenic forms of PD, discovered through the study of families, common genetic risk factors for idiopathic PD have been identified through large GWASs and meta-analyses [Hernandez, Reed & Singleton 2016]. The largest meta-analysis of PD risk to date was performed in 2019, when Nalls and colleagues meta-analyzed 17 different datasets from previous PD GWASs totaling 37,688 PD cases, 18,618 UK Biobank "proxy-cases" (subjects who did not have PD, but had an afflicted first degree relative), and 1.4 million control subjects [Nalls et al. 2019]. This brought the total number of identified genetic risk factors to 90, which span the entire genome across 78 genomic regions and 305 genes leaving chromosome 22 as the only autosome without a PD genetic risk factor according to the European Bioinformatics Institute's (EBI) GWAS catalog (https://www.ebi.ac.uk/gwas/efotraits/EFO_0002508, accessed 8/20/2020) [Nalls et al. 2019]. Genomic regions containing *SNCA*, *LRRK2, MAPT*, and *TMEM175* were among the top most associated with PD risk, with *SNCA* being the highest signal. Other regions contained genes previously associated with other diseases such as *NOD2* previously associated with Crohn's disease [Nalls et al. 2019]. The identification of these susceptibility loci and their functionally relevant genes provides evidence for biological pathways involved in PD and leads for functional studies. The issue remains, however, that no identified genetic risk factor(s) fully explain the cause of PD as all have modest effect sizes individually (odds ratios = 1.1 – 2, excluding rarer *LRRK2* variants), and in combination (odds ratio = 3.7 – 6.3) [Nalls et al. 2019]. It has also been estimated that the currently identified genetic variants only explain 16 – 36% of the heritable risk of PD, and provide an area under the curve and balanced accuracy of 0.6 – 0.7 when used in a genetic predictive model of PD [Nalls et al. 2019]. This suggests that there is a large portion of PD risk that is not being accounted for, and requires further investigation into the genome, and elsewhere, to find that missing portion.

Genetics of age at onset

The vast majority of genetic studies in PD have focused on finding modifiers of PD risk, however, most of these genetic risk factors do not explain the highly variable nature of age at PD onset or diagnosis [Blauwendraat et al. 2019]. Age at onset and diagnosis of PD varies greatly from person to person ranging from the teens to within the 10th decade of life. Previous studies, some performed even before the advent of GWAS, have shown substantial evidence for the involvement of genetic factors in the age at onset of motor symptoms and age at diagnosis of PD, estimating the heritability of age at PD onset to be upwards of 98% [Zareparsi et al. 1998; Maher et al. 2002; McDonnell et al. 2006; Hamza et al. 2010; Nalls et al. 2015]. Even with such high heritability, GWASs of age at onset or diagnosis of PD have been given less attention than risk, with only ~10% of the current GWASs listed in the EBI GWAS catalog for “parkinson’s disease” focusing on age at onset or diagnosis (https://www.ebi.ac.uk/gwas/efotraits/EFO_ 0002508, accessed 8/20/2020). Even still, a handful of GWASs have provided evidence for putative genetic variants and genes that might play a role in the inter-individual variation in age at onset or diagnosis seen in the broader PD population [Hill-Burns et al. 2016; Wallen et al. 2018; Blauwendraat et al. 2019].

In one of the earliest and largest genome-wide age at onset studies, Hill-Burns et al. performed an age at onset GWAS on 431 familial (at least one affected first or second degree relative) and 1,544 non-familial PD cases recruited as part of the NeuroGenetics Research Consortium (NGRC) [Hill-Burns et al. 2016]. An additional 737 familial and 2,363 non-familial PD cases from seven additional cohorts were used for replication. In familial PD, they detected two genome-wide significant signals that replicated robustly in the replication dataset and were associated with earlier onset of PD motor symptoms by 9.3 – 15.3 years. Neither signal was associated with risk of PD, suggesting these variants might not play a role in risk of developing PD, but affect the progression of it once it has been triggered. These signals mapped to two genes, *LHFPL2* and *TPM1*, neither of which were associated with age at onset in non-familial PD, or when all PD subjects were combined. No genome-wide significant signals were detected for non-familial and all PD combined.

A follow-up study to Hill-Burns et al. was performed by Wallen et al., who performed an age at diagnosis GWAS on 1,950 PD cases from NGRC, the same cohort analyzed by Hill-Burns et al [Wallen et al. 2018]. An additional 726 PD cases from the Parkinson's, Genes and Environment (PAGE) study [Chen et al. 2010] were used for replication. This study is detailed in the first chapter of this dissertation under the title "PLASTICITY-RELATED GENE 3 (*LPPR1*) AND AGE AT DIAGNOSIS OF PARKINSON DISEASE". Briefly, GWAS revealed two association signals that tagged two, seemingly independent linkage disequilibrium (LD) blocks of variants inside the *LPPR1* gene. Only one of these blocks had a genome-wide significant hit, but gene-based analysis confirmed *LPPR1* as it reached multiple testing corrected significance. One LD block replicated robustly, while the other only replicated in a subset of the replication PD cases. Both signals were associated with earlier diagnosis of PD by ~6 years, and contained functionally relevant variants that potentially act to destabilize the LPPR1 protein, and alter the expression of another gene *GRIN3A*.

The most recent and largest GWAS of age at onset of PD to date was performed by Blauwendraat et al. where they performed a hybrid age at onset and diagnosis GWAS followed by meta-analysis on 28,568 PD cases from 17 cohorts in the International Parkinson's Disease Genomics Consortium (IPDGC) and one cohort from 23andMe [Blauwendraat et al. 2019]. Age at onset was used for the outcome of IPDGC GWASs unless not available, then age at diagnosis was used. Age at diagnosis was used for all subjects in the 23andMe GWAS. Meta-analysis resulted in two genome-wide significant associations that mapped to the 3′ end of *SNCA*, which is the most highly associated region for PD risk, and a coding variant in exon 11 of *TMEM175*. Both signals were associated with earlier onset, or diagnosis, by ~0.6 years. Nominal signals were detected for other loci associated with PD risk (*BST1*, *INPP5F/BAG3*, *FAM47E/SCARB2*, *MCCC1*), but to the author's surprise no significance was detected for well-established PD risk loci including *RAB7L1/NUCKS1* (PARK16), *GCH1*, and *MAPT*. Regardless, the authors state associations detected in this study provide evidence for a dual role of a subset of PD risk loci in both increasing PD risk and modulating PD progression.

Gut microbiome

Overview

The human body is home to trillions of microbial cells including bacteria, archaea, viruses, and eukaryotic microbes such as fungi, whose collective genomes are referred to as the human microbiome. These microbes have co-evolved with humans to form complex niches on and within our bodies, resulting in numerous body-area-specific ecosystems that can adapt to changing host physiology and exposures. For the most part, it is a symbiotic relationship between host and microbial needs as humans provide commensal microbes with the environment and nutrients they need to survive while commensals perform critical roles in keeping our bodies healthy. However, a disbalance in the composition of the microbiome has been associated with numerous diseases ranging from diseases of the gastrointestinal tract to neurological conditions [Lloyd-Price, Abu-Ali & Huttenhower 2016]. Due to its importance in human health and disease, and the rapid rate at which methods for studying the microbiome have become available, microbiome research has exploded into many areas of biomedical research where the most studied microbial community of the human body has been that of the gut microbiome.

The gastrointestinal tract houses the largest and most diverse collection of microbes on, or in, the human body. Microbes in the gut provide assistance with many essential activities needed for a healthy body including digestion and nutrient uptake, defense against infection from foreign microbes, detoxification of ingested compounds, development and priming of the immune system, and mediation of diseases [Liang et al. 2018]. Unfortunately in disease, if the lining of the gut is breached, usually harmless commensal microbes can become a source of inflammation to surrounding tissues and can lead to immune system perturbations [Segata et al. 2012], or potentially more systemic damage. Because of this dual nature of the gut microbiome, it is important to understand and characterize it in both healthy and diseased states. Our understanding of what organisms reside in the gut, their functional characteristics, and how they influence human health and disease has been growing at a rapid pace, largely facilitated by an exponential increase in computational tools and sequencing technologies that allows access to the gut microbiome, which is largely unculturable. These advanced tools and technologies have been, and are currently being used to establish and build important facets of gut microbiome research including (1) compilation of reference data for both taxonomic identities and gene content of gut microbiota, (2) associations between abundances of microbes and microbial genes with host factors, disease and environmental exposures, (3) functional characterization of microbial associations with host or external factors through a wider breadth of methodologies such as multi-omics or experimental manipulations, and (4) translation of gained knowledge into biomedical applications such as gut microbiome modulation for disease therapies [Schmidt, Raes & Bork 2018]. Studies of the gut microbiome outlined in this dissertation will pull from, or contribute to, 1-3 of the above list.

Methods for analyzing the gut microbiome

Two forms of sequencing technology are commonly used to assess the presence and abundance of microbes in the gut: shotgun metagenomic sequencing and amplicon sequencing of marker gene fragments (usually 16S rRNA gene fragments for bacteria/archaea and 18S rRNA/ITS gene fragments for fungi). Both methods have their utilities, pros, and cons. Although referred to as shotgun metagenomics, this technique is actually standard whole genome sequencing that is applied to a sample of DNA from a mixed community of microorganisms [Venter et al. 2004]. As with whole genome sequencing of a singular organism, DNA from the mixed community of microorganisms is sheared into tiny fragments and sequenced individually resulting in DNA sequence reads that can be aligned to reference genomes to determine what microbes are present in the sample and the relative abundance of each [Sharpton 2014]. Untargeted sequencing of DNA that can span across the genomes of microorganisms has key advantages over marker gene amplicon sequencing including the ability to detect all organisms present in a sample across multiple domains (Bacteria, Archaea, Eukaryota, viruses) at high resolutions (down to strain level) and gain functional insight by sequencing different gene coding portions of the genomes [Sharpton 2014]. Even with these benefits, shotgun metagenomic sequencing has drawbacks that keep it from becoming the main sequencing technique for surveying the gut microbiome. These include the increased cost of shotgun metagenomics when compared to amplicon sequencing as it can be 10x more expensive to perform than amplicon sequencing, the increased computational resources and expertise needed to process the massive amount of complex data resulting from shotgun metagenomics, and the lack of a comprehensive list of reference genomes to apply to sequenced microbial DNA [Sharpton 2014], although this is continually being improved upon by a number of large collaborative studies such as the MetaHIT consortium [Qin et al. 2010; Li et al. 2014] and Human Microbiome Project [Nelson et al. 2010; Human Microbiome Project Consortium 2012]. On the other hand, amplicon sequencing of marker gene fragments require much less funds to perform, have a much smaller data storage footprint than shotgun metagenomics, require less computational resources (although analysis of many samples will still require a fair amount), and have more comprehensive and specialized reference databases, of which the largest and most widely used is the SILVA database [Quast et al. 2013; Yilmaz et al. 2014; Balvočiūtė & Huson 2017].

The most popular choice of marker gene is the bacterial/archaeal 16S rRNA gene as it has a good balance between conservation among bacteria and archaea, but contains hypervariable regions that are more prone to genetic variation, which makes it an informative marker for taxonomical and phylogenetic differentiation [Pace et al. 1986; Hugenholtz & Pace 1996]. In amplicon sequencing of the 16S rRNA gene, extracted microbial DNA is submitted to a polymerase chain reaction (PCR) using primers targeting a specific hypervariable region (commonly used regions include hypervariable region 4 (V4) or hypervariable region 3 and 4 (V3-4)), then amplicons of the hypervariable region are sequenced. These sequences are then bioinformatically processed to infer what microorganisms are present in a sample and at what relative abundance [Sharpton 2014]. Two main methods are used for the detection of unique organisms, or unique clusters of closely related organisms, in a sequenced sample: clustering unique sequences into operational taxonomic units (OTUs) based on a particular sequence similarity threshold (97% commonly used) or inference of high resolution, exact sequence variants (referred to as amplicon sequence variants (ASVs)) representing unique microorganisms that can differ by as little as one nucleotide [Callahan, McMurdie & Holmes 2017]. Both methods are still used in the literature, however, leaders in the field have suggested replacing the use of OTUs with ASVs as ASVs provide a higher resolution for microorganism detection and are more biologically relevant and reproducible between studies as their identity corresponds to the exact microorganism sequence detected within a sample instead of a synthetic grouping of similar sequences based on an arbitrary similarity threshold with an arbitrary ID [Callahan, McMurdie & Holmes 2017; Boleyn et al. 2019]. After detection of OTUs/ASVs, taxonomic identities are assigned via a taxonomic classifier, which usually assigns identities from kingdom to genus taxonomic levels. Species level taxonomic assignments might be achievable for a subset of detected OTUs/ASVs, but classifications at the species level usually cannot be confidently made with 16S rRNA amplicon sequencing [Johnson et al. 2019], or produce ties between multiple species as the gene fragment sequenced does not contain enough genetic variation to distinguish between two or more species. An optional step in the bioinformatic processing of 16S rRNA amplicon data is the creation of a phylogenetic tree of OTUs/ASVs, but this is usually done only if a phylogenetic study is being performed, or if statistical methods will be implemented downstream that use phylogenetic relatedness in their analyses (e.g. the widely used UniFrac distances for measuring inter-individual variation in microbiome compositions) [Lozupone et al. 2005; Chen et al. 2012]. At this point a sample by feature (OTUs/ASVs or higher taxonomic levels) table with per sample feature abundances has been produced with accompanying taxonomic assignments and potentially phylogenetic tree, and is the main input, along with sample metadata (host factors such as disease status, age, and sex or technical variables such as collection method and sample storage), for further downstream statistical analyses. Before statistical analyses are conducted, however, sample by feature abundance tables are usually, and should be, processed further to account for inter-sample variation in sequencing depth. A number of methods have been developed to account for unequal sequencing depth between samples ranging from simple data transformations as the widely used relative abundance (dividing each feature count by the total sample count, also referred to as total sum scaling) to more intricate methodologies that calculate size factors for scaling feature counts such as those used in RNA-seq methods DESeq2 and edgeR [Robinson & Smyth 2008; Love, Huber & Anders 2014].

Statistical analyses implemented in study of the gut microbiome differ based on study goals, but one of the most commonly used analyses in gut microbiome research is differential abundance analysis [Thorsen et al. 2016]. Differential abundance testing has been especially useful when studying the gut microbiome and disease as a number of diseases have been associated with alterations of individual taxa in the gut microbiome using differential abundance testing [Schmidt, Raes & Bork 2018]. Differential abundance testing involves the use of serial univariate statistical tests to determine if certain taxa are significantly different between groups [Thorsen et al. 2016]. Numerous differential abundance testing methods exist and include classical statistical tests (e.g. Kruskal-Wallis rank sum test), methods developed to detect differential expression of gene transcripts in RNA-Seq data (e.g. DESeq2, edgeR), methods specifically designed for detecting differentially abundant taxa in microbiome data (e.g. metagenomeSeq), and methods designed to detect differentially abundant features in compositional data (e.g. ALDEx2). Choice of differential abundance method can greatly influence what, and how many, differentially abundant taxa are detected in a disease state, and most, if not all, methods will respond differently to microbiome data due to differences in their underlying characteristics. Multiple studies have previously assessed and compared the performance of popularly used differential abundance testing methods, measuring their false positive rates, false discovery rates, sensitivities, and/or specificities from simulated data [McMurdie & Holmes 2014; Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019], with only one of these studies testing different methods on real data [Weiss et al. 2017]. The literature lacks examples of the use of different differential abundance testing methods on real gut microbiome datasets, therefore, a study detailed later in this dissertation (chapter "APPLICATION OF SIXTEEN DIFFERENTIAL ABUNDANCE METHODS TO TWO LARGE PARKINSON DISEASE GUT MICROBIOME DATASETS") applied 16 differential abundance testing methods to two large PD-gut microbiome datasets to compare results between these methods when performed on real, complex disease datasets.

Other statistical analyses commonly used in gut microbiome research include visualization and/or testing of inter-individual variations in overall gut microbiome compositions and co-occurrence network analysis to detect groups of correlated taxa.

Inter-individual variation in microbiome composition (also referred to as β-diversity) can be visualized and tested in a variety of ways. Methods such as principal component analysis, principal coordinate analysis, and non-metric multidimensional scaling are commonly used ordination tools to observe patterns of sample clustering based on within and between group differences in microbiome composition [Buttigieg & Ramette 2014]. Obvious clustering between groups of samples might indicate a strong effect size of the grouping variable on composition of the microbiome. Differences between groups can then be tested using permutational multivariate analysis of variance (PERMANOVA), which tests if between group differences in microbiome composition are significantly larger than within group differences signifying that a separation between groups is occurring based on microbiome compositional differences [Anderson 2001].

Co-occurrence network analysis provides a way to infer biological interactions *in silico* through construction of a correlation network between taxa of interest [Friedman & Alm 2012]. A number of methods exist to construct co-occurrence networks, but the main workflow includes calculation of pairwise correlations between taxa abundances, and then plotting correlations in a network framework followed by clustering of network nodes (taxa) using some form of clustering algorithm to visually observe different communities of correlated taxa. Community detection algorithms can be used to test if what is being observed visually are actually distinct correlated communities [Blondel et al. 2008]. Similar to how LD is used in GWAS for associated genetic variants, using this technique can be informative for observing whether or not taxa associations detected via differential abundance testing are in fact independent, or part of a correlated group of taxa. It may also provide leads for additional candidate taxa to study that may not have been detected via differential abundance testing.

Complications of gut microbiome analysis

Even with an expansive list of tools, analysis of the gut microbiome still has its difficulties ranging from the initial collection of data and samples to the end statistical analyses. Composition of the gut microbiome is influenced by a myriad of factors from the host, environment, and gut microbiome itself [Schmidt, Raes & Bork 2018]. Through large population studies of 2,000 subjects combined, gut microbiome composition was shown to be influenced by approximately 130 different host intrinsic and life-style factors including, but not limited to, stool consistency, body mass index, age, metabolite levels, gender, diet, presence of gastrointestinal diseases, and medication use [Falony et al. 2016; Zhernakova et al. 2016]. Even still, one study estimated that the identified host factors only explained 18.7% of the variation in inter-individual differences in microbiome composition, suggesting that the majority of factors affecting gut microbiome composition have yet to be identified [Zhernakova et al. 2016]. The large range of host factors that influence the gut microbiome makes subject data collection difficult at the onset of a gut microbiome study, and also complicates later statistical analyses when trying to detect potential confounding variables. In addition to host factors, numerous technical factors also influence detected gut microbiome compositions including choice of sample storage method [Choo, Leong & Rogers 2015], DNA extraction method

[Mackenzie, Waite & Taylor 2015], and 16S rRNA gene region [Yu et al. 2008; Klindworth et al. 2013; Yang, Wang & Qian 2016]. Even after molecular processing, biases of what gut microbes are detected from a sample can be introduced through choice of bioinformatic pipeline [Prodan et al. 2020], and taxonomic reference database [Balvočiūtė & Huson 2017].

Performing statistical analyses on gut microbiome data is, in itself, difficult due to multiple underlying characteristics of the data. The majority of microbiome data is heavily skewed toward zero (80 – 95% of the counts equal to zero) and over-dispersed (variances are larger than their means) [Thorsen et al. 2016], usually following a negative-binomial distribution. This makes microbiome data difficult to analyze using standard statistical methods that expect the data to be normally distributed with no over-dispersion, requiring the use of specialized methods, or transformations to try and bring the data closer to a normal distribution. As mentioned earlier, usually there is large variation in sequencing depth, or total sample count, between samples, which must be accounted for by normalizing individual taxa counts by the total sample count, or by including total sample count in the statistical model. Microbiome sequencing data is also compositional in nature, meaning that all sequence counts given to individual taxa within a sample adds up to a particular constant, whether that constant is 1 when data is transformed to relative abundances, or 100,000 if data is raw abundances [Gloor et al. 2016]. This translates to all taxa within a sample being, at least, somewhat correlated with one another just from technical artifact. In order to break this compositionality, ratios of one taxon to others can be used to capture the relationships between individual taxa without the compositionality constraint [Gloor et al. 2017]. Taking the log of these ratios, hence why this transformation is referred to as log-ratio, makes the data more symmetrical and analyzable by standard statistical methods [Gloor et al. 2017]. Log-ratios are used in the differential abundance method ANCOM (Analysis of composition of microbiomes) [Mandal et al. 2015], while another differential abundance method, ALDEx2, uses the centered-log ratio transformation (log-ratio transformation with the denominator being the geometric mean of the sample) [Fernandes et al. 2014].

Parkinson disease and gut connection

Multiple lines of evidence from previous studies point to the involvement of the gut and gut microbiome in PD. As stated previously, some of the first non-motor symptoms of PD, occurring well in advance of motor symptom onset, are gastrointestinal disturbances, including constipation [Cersosimo et al. 2013; Chen et al. 2015]. Presence of α-synuclein has been shown in the gastrointestinal tract of persons with early PD [Shannon et al. 2012], Lewy body disease [Breen, Halliday & Lang 2019], and rapid eye movement disorder [Knudsen et al. 2018], which has a high conversion rate to PD. Presence of α-synuclein in the gut was also found in conjunction with increased intestinal permeability in early stage PD [Forsyth et al. 2011]. Large epidemiological studies have suggested a reduction in PD risk for those who have undergone truncal vagotomy years before PD onset [Svensson et al. 2015; Liu et al. 2017], and a study in mouse saw that truncal vagotomy and endogenous α-synuclein deficiency prevented gut to brain spread of injected preformed α-synuclein fibrils and development of PD-like neurodegeneration and behavioral deficits [Kim et al. 2019]. Studies in human have shown a role of α-synuclein in pathogen response where infection of the gut or olfactory system triggered α-synuclein expression, which in turn mobilized the immune system to respond to the infection [Stolzenberg et al. 2017; Tomlinson et al. 2017]. Experimentally, it has been shown in a *Pink1* knockout mouse model of PD that intestinal infection may act as a trigger for dopaminergic cell loss and motor impairment through activation of T cells in the periphery [Matheoud et al. 2019]. A hypothesis that has gained popularity in recent years, termed "Braak's hypothesis", states that non-inherited forms of PD may be caused by a yet to be identified pathogen that invades the gastrointestinal tract and, through the enteric nervous system, makes its way to the brain [Braak et al. 2003; Braak et al. 2003]. This hypothesis has been further modified to state that it may not be an actual pathogen making its way to the brain, but pathogenic species of α-synuclein initiated in the gut by a pathogen, or altered microbial state, and traveling to the brain. Multiple studies in human have associated a dysbiotic gut microbiome with PD, all finding individual microorganisms significantly enriched or depleted in PD, albeit with varying results [Gerhardt & Mohajeri 2018; Boertien 2019]. One experimental study showed enhanced neuro-inflammation and motor symptoms in germ free α-synuclein overexpressing when colonized with gut microbiota derived from PD patients compared to mice colonized with control microbiota [Sampson et al. 2016].

Even with an overwhelming amount of evidence pointing to gut and gut microbiome involvement in PD, it is still under investigation whether or not the gut is a site involved in the initiation of PD, or gut and gut microbiome perturbations are just a byproduct of disease, an unhealthy gut, or weakened immune system.

Aims, rationale, and brief description of dissertation

It is clear that a large portion of age at onset and diagnosis of PD is heritable, therefore, our overarching hypothesis is that there is a genetic component to the variability in age at onset and diagnosis of PD. Genome-wide association studies have mostly focused on PD risk while only a handful of genome-wide studies have been performed for age at onset of PD. This has resulted in 90 genetic risk loci being detected for PD, while only ~5 genetic loci have been nominated through genome-wide methods as potential age at onset or diagnosis modifiers in PD. More genetic studies targeting age at onset and diagnosis are needed to continue characterizing the genetic component of age at onset and diagnosis. To this end, the aim of the first chapter of this dissertation was to continue identifying genetic modifiers of age at diagnosis of PD. To do this, we performed a GWAS for age at diagnosis of PD using 1,950 PD patients. Through both SNP and gene-based GWAS, we identified an additional putative age at diagnosis modifier gene, which added to the relatively short list of current genes. Functional annotation of SNPs underlying the detected GWAS signal revealed an additional 9 brain expressed genes that might be candidates for further functional studies to determine if they play a role in modifying the progression of PD.

The remaining parts of this dissertation focus around PD and the gut microbiome: first testing and comparing the behavior of different differential abundance testing methods on real microbiome data, then characterizing the PD gut microbiome in two large cohorts, and finally performing a candidate taxa, candidate gene study to see if abundances of certain taxa found enriched in PD and genetic variants in the *SNCA* locus influenced one another's associations with PD.

As methods for microbiome analysis are constantly evolving and updating, and no example is available in the literature for the comparison of different differential abundance methods on gut microbiome datasets of real, complex disease, we took it upon ourselves to perform 16 differential abundance testing methods on two large PD-gut microbiome datasets and compare their results. This study makes up the second chapter of the dissertation, which aimed to compare results between 16 differential abundance methods when performed on two large, real gut microbiome datasets that were created for study of a complex disease. We hypothesized inter-method variation in results would be evident as has been previously shown [McMurdie & Holmes 2014; Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019], but the degree to which method results would differ from one another was unknown to us. Indeed, a wide range of inter-method variation in results was observed, but some methods (mainly those previously shown to have low false positive and discovery rate) resulted in higher overall concordances than others. Despite differences observed between methods, a group of PD-microbe associations were found to be agreed upon by the majority of methods in both datasets. This study gave us a better understanding of how different differential abundance methods behaved with real, complex disease gut microbiome data and provided a unique method comparison study currently missing from the literature.

With increased understanding on how different differential abundance testing methods behave on our datasets, we then moved to perform a study characterizing the gut microbiome of PD in the largest PD-gut microbiome datasets to date, which is detailed in the third chapter of this dissertation. Similarly to previous PD-gut microbiome studies, we hypothesized that there is enrichment and/or depletion of certain microbes in the gut of persons with PD, but unlike previous studies we did not assume associations of gut microbes with PD would be independent, rather, might be correlated with one another as a whole or in part. For this chapter, we aimed to identify reproducible signals of association between PD and gut microbiota by using larger sample sizes, adjustment for potential confounding variables, robust statistical methods, stringent statistical criteria, and a replication paradigm, which has been lacking in previous PD-gut microbiome studies. Performing a hypothesis-free microbiome-wide association study at the genus level resulted in detection of robust associations between PD and 15 bacterial genera. Detected associations were not independent, but represented three clusters of co-occurring microorganisms, which, per literature search, included a group of opportunistic pathogens, short-chain fatty-acid producing bacteria, and supposed probiotic bacteria. Results from this study confirmed previous evidence for alterations of short-chain fatty-acid producing and probiotic bacteria in PD, while detecting novel associations with potential opportunistic pathogens that might be of interest for further functional studies and were the focus of the final chapter of this dissertation.

As mentioned earlier in the introduction, previous literature points to immune involvement of α-synuclein, and pathological α-synuclein has been observed in the gut of PD patients [Stolzenberg et al. 2017; Tomlinson et al. 2017; Shannon et al. 2012]. With this, and Braak's hypothesis in mind (also mentioned earlier in the introduction) [Braak et al. 2003; Braak et al. 2003], we posited that the opportunistic pathogens associating with PD in chapter three of this dissertation might be connected to aberrant α-synuclein presence in the gut as both were found to be enriched in PD previously. To begin investigating whether a potential relationship exists between overabundance of opportunistic pathogens in the PD gut and α-synuclein, we performed a genetic study (detailed in chapter 4 of this dissertation) to test if a potential interaction existed between opportunistic pathogens reported in chapter three and genetic variation in and around the *SNCA* gene. Not only is the *SNCA* region the highest peak in GWAS of PD risk [Nalls et al. 2019], but genetic variants in this region have been previously associated with increased expression of *SNCA* [GTEx Consortium 2015; Soldner et al. 2016; Emelyanov et al. 2016]. As both presence of pathogens in the gut and genetic variants in and around *SNCA* have been previously shown to increase *SNCA* expression, and increasing dosages of *SNCA* is important in PD pathogenesis as seen in *SNCA* duplication and triplication cases [Devine, Gwinn, Singleton & Hardy 2011], we hypothesized that the combination of both opportunistic pathogens in the gut and genetic variation in the *SNCA* region might increase the risk of PD. In order to establish a putative connection between previously detected opportunistic pathogens in PD and genetic variation in the *SNCA* region, the aim of the fourth chapter of this dissertation was to investigate whether or not genetic variation in the *SNCA* region moderated the associations between PD and opportunistic pathogens reported in chapter 3, and then, if presence of opportunistic pathogens enhanced the detected genetic variants' associations with PD. Interaction analyses resulted in detection of two variants that showed an obvious genotype effect on PD and opportunistic pathogen associations. Then, associations between detected genetic variants and PD were found to be enhanced when testing in subjects who were positive for opportunistic pathogens of chapter 3. While results from this study only provides a suggestive connection between opportunistic pathogens in the PD gut and genetic variation in the *SNCA* region, it is the first study to investigate the interaction between host genetics and gut microbiome in relation to PD, and provides further leads for testing the involvement of these gut microbes in functional studies of PD including animal models that overexpress α-synuclein.

PLASTICITY-RELATED GENE 3 (*LPPR1*) AND AGE AT DIAGNOSIS OF PARKINSON DISEASE

by

ZACHARY D. WALLEN, HONGLEI CHEN, ERIN M. HILL-BURNS, STEWART A. FACTOR, CYRUS P. ZABETIAN, AND HAYDEH PAYAMI

Neurology Genetics 2018;4:e271. doi:10.1212/NXG.0000000000000271
PMID: 30338293   PMCID: PMC6186025

Format adapted for dissertation

ABSTRACT

The objective of this study was to identify modifiers of age at diagnosis of Parkinson disease (PD). We performed a genome-wide association study (GWAS) that included 1,950 individuals with PD from the NeuroGenetics Research Consortium (NGRC) study. Replication was conducted in the Parkinson's, Genes and Environment study, including 209 prevalent (PAGE$_{\text{P}}$) and 517 incident (PAGE$_{\text{I}}$) PD cases. Cox regression was used to test association with age at diagnosis. Individuals without neurologic disease were used to rule out confounding. Gene-level analysis and functional annotation were conducted using Functional Mapping and Annotation of GWAS platform (FUMA). GWAS revealed 2 linked, but seemingly independent association signals that mapped to *LPPR1* on chromosome 9. *LPPR1* was significant in gene-based analysis (*P* = 1E-8). The top signal (rs17763929, hazard ratio [HR] = 1.88, *P* = 5E-8) replicated in PAGE$_{\text{P}}$ (HR = 1.87, p = 0.01), but not in PAGE$_{\text{I}}$. The second signal (rs73656147) was robust with no evidence of heterogeneity (HR = 1.95, *P* = 3E-6 in NGRC; HR = 2.14, *P* = 1E-3 in PAGE$_{\text{P}}$ + PAGE$_{\text{I}}$, and HR = 2.00, *P* = 9E-9 in meta-analysis of NGRC + PAGE$_{\text{P}}$ + PAGE$_{\text{I}}$). The associations were with age at diagnosis, not confounded by age in patients or in the general population. The PD-associated regions included variants with Combined Annotation Dependent Depletion (CADD) scores = 10–19 (top 1%–10% most deleterious mutations in the genome), a missense with predicted destabilizing effect on LPPR1, an expression quantitative trait locus (eQTL) for *GRIN3A* (false discovery rate [FDR] = 4E-4), and variants that overlap with enhancers in *LPPR1* and interact with promoters of *LPPR1* and 9 other brain-expressed genes (Hi-C FDR < 1E-6). Through association with age at diagnosis, we uncovered *LPPR1* as a modifier gene for PD. *LPPR1* expression

promotes neuronal regeneration after injury in animal models. Present data provide a strong foundation for mechanistic studies to test *LPPR1* as a driver of response to damage and a therapeutic target for enhancing neuro-regeneration and slowing disease progression.

INTRODUCTION

The underlying neurodegenerative process that causes Parkinson disease (PD) begins decades before the disease is diagnosed.[1] The current view is that following an initial insult (e.g., toxicity, trauma, or genetic), the disease starts with an asymptomatic phase of unknown duration, followed by development of prodromal nonmotor symptoms such as constipation, anosmia, and sleep disorders. Years later, cardinal motor signs appear, at which point a diagnosis of PD is made. Age at onset of motor signs, and therefore the age at diagnosis of PD, is highly variable, ranging from teen ages to the 10th decade of life. The reason for this variation is unknown, and understanding it will likely shed light on factors that affect the rate of disease progression.

There is substantial evidence that genetic factors play a major role in age at onset of motor signs and age at diagnosis of PD.[2-6] Genome-wide studies have identified numerous loci that associate with the risk of developing PD,[7] but the risk factors do not explain the variation in age at onset.[8-10] Three loci have been nominated as modifiers of age at onset in familial PD.[11,12] The present study was aimed at identifying genetic modifiers for common idiopathic PD. We hypothesized that identification of the genetic basis to interindividual variability in age at diagnosis will provide insights into the intrinsic mechanisms that determine the rate of deterioration during preclinical disease.

METHODS

This study was a case-control GWAS, followed by replication and functional annotation.

Standard protocol approvals, registrations, and patient consents

The study was approved by the institutional review boards at all participating institutions. Written informed consent was obtained from all patients and controls for participation in the study.

Participants

The study included 2 data sets. The NeuroGenetics Research Consortium (NGRC) data set[13] was used for the discovery GWAS, gene-based test, and functional annotations. The Parkinson's, Genes and Environment (PAGE) study[14] was used for replication. Participants' characteristics are shown in table 1 and figure e-1 (links.lww.com/NXG/A66).

NGRC is a case-control study of genetically unrelated participants, including 2000 PD cases and 1986 controls.[13] Patients were enrolled sequentially from movement disorder clinics in Portland (OR), Seattle (WA), Albany (NY), and Atlanta (GA). Controls were spouses of patients or community volunteers, self-reported as being free of neurologic disease. The eligibility criterion for cases was diagnosis of PD by a movement disorder specialist according to the UK Brain Bank criteria.[15] The eligibility criteria for controls were no neurologic disease and genetically unrelated to patients. Age was defined as age at study entry. Age at diagnosis was extracted from medical records or ascertained by self-report. Age at onset of the first motor sign was obtained using a self-administered questionnaire. Age at onset and age at diagnosis were highly correlated in the NGRC ($r^2 = 0.91$, $P < 2E\text{-}16$). All participants were whites of European descent.[13]

Table 1. Data sets and participants' characteristics. Data on the NGRC participants were collected at enrollment: patients already had the diagnosis of PD and controls were free of neurologic disease. NGRC participants were enrolled at four sites: Oregon, Washington, New York and Georgia. Age at onset mean ± SD were Oregon=56.6 ± 12.8, Washington=58.7 ± 11.8, New York=59.4 ± 11.5, Georgia=58.7 ± 11.1. Age at diagnosis mean ± SD were Oregon=59.6 ± 11.7, Washington=60.7 ± 11.6, New York=60.9 ± 11.1, Georgia=60.3 ± 10.6. PAGE participants were originally enrolled in the longitudinal NIH-AARP diet study in 1995-1997. Their PD status was investigated in 2004-2006. Participants who had the diagnosis of PD before 1998 were classified as prevalent PD ($PAGE_P$), participants who were diagnosed with PD during follow-up (between 1998 and 2006) were classified as incident PD ($PAGE_I$), and participants who did not have PD were designated as controls. Since PAGE participants were of similar age at entry, the method of classifying the participants into prevalent vs. incident cases inevitably assigned earlier ages-at-diagnosis to the prevalent group and later diagnoses to the incident group. Abbreviations: NA = not available; NGRC = NeuroGenetics Research Consortium; NR = not relevant; PAGE = Parkinson's, Genes, and Environment. Participants were non-Hispanic whites and genetically unrelated.

	Discovery (NGRC)		Replication (PAGE)		
	PD	Control	$PAGE_P$	$PAGE_I$	Controls
N	2,000	1,986	209	517	1,549
Male / Female	1,346 / 654	769 / 1,217	164 / 45	396 / 121	1,213 / 336
Age at enrollment mean ± SD	67.3 ± 10.7	70.3 ± 14.1	62.6 ± 4.9	63.2 ± 4.9	63.4 ± 4.9
Age at follow-up mean ± SD	NR	NR	73.9 ± 4.9	74.5± 4.9	74.0 ± 4.9
N with age at onset data	1,999	NR	0	0	NR
Age at onset mean ± SD	58.3 ± 11.9	NR	NA	NA	NR
N with age at diagnosis data	1,950	NR	209	517	NR
Age at diagnosis range	25 - 90	NR	42 - 72	53 - 81	NR
Age at diagnosis mean ± SD	60.4 ± 11.4	NR	59.9 ± 6.6	69.4 ± 5.4	NR

PAGE is a cross-sectional study nested in the longitudinal NIH-American Association of Retired Persons Diet and Health Study.[14] Participants were enrolled in 1995–1997 (irrespective of PD) via a food frequency questionnaire mailing[16] and in the 2004–2006 follow-up visit were asked if they had been diagnosed with a major chronic disease including PD. Participants who had been diagnosed with PD before enrollment (before 1998) were designated as prevalent PD (PAGE$_P$, N = 209), participants who were diagnosed during follow-up (1998–2006) were designated as incident PD (PAGE$_I$, N = 517), and participants who did not have PD were designated as controls (N = 1,549). All participants in this study were non-Hispanic whites.

Genotyping

NGRC participants were genotyped on Illumina HumanOmni1-Quad v1-0 B array and Immunochip array. Genotypes and samples were filtered by call rate, minor allele frequency (MAF) < 0.01, Hardy-Weinberg, and cryptic relatedness, as described before.[13] Imputation was performed using IMPUTE v2.3.0,[17] with the 1000G Phase3 integrated variant set (October 2014) as reference. Imputed single nucleotide polymorphisms (SNPs) with info score < 0.9 or MAF < 0.01 were excluded. A total of 8.5 million SNPs (900,000 genotyped and 7.6 million imputed) were used in the analysis.

PAGE participants were genotyped for rs73656147 (block 1) and rs17763929 (block 2). SNPs were chosen based on statistical significance and availability of predesigned validated TaqMan assay from Thermo Fisher (rs73656147 assay number = C__97534229_10; rs17763929 assay number = C__34297681_10).

Population structure

Principal component (PC) analysis[18] is used to infer population-specific genetic differences, which arise from ancestry differences in allele frequencies and can obscure genetic association studies if not accounted for. NGRC PC analysis was conducted using a pruned subset of 100K SNPs from the GWAS as previously described.[13] The top 3 PCs (effect sizes PC1 = 0.2%, PC2 = 0.06%, and PC3 = 0.06%) were included in the GWAS and adjusted for in all downstream analyses involving the NGRC. The PAGE data sets used for replication did not have ancestry informative markers (AIMs); however, a subset of the participants (396 of 726 PD cases) was previously genotyped with the Immunochip array. We conducted PC analysis using a pruned set of 20K SNPs from the Immunochip array, using PLINK. Tests were conducted once using the full PAGE data set, with no PC adjustment, and again with a PAGE subset, adjusting for PC1-3 (effect sizes PC1 = 0.48%, PC2 = 0.20%, and PC3 = 0.17%). NGRC and PAGE cluster with Europeans in the 1000G_Phase_3 global data set (figure e-2, links.lww.com/NXG/A67).

Statistics

For discovery, GWAS was conducted using PD cases only (1,950 NGRC participants with known age at diagnosis). Association between 8.5M SNPs and age at diagnosis was tested using Cox regression in ProbABEL v0.5.0.,[19] specifying an additive genetic model, treating age at diagnosis as a quantitative trait, and adjusting for PC1-3. The statistical outcome of Cox regression was hazard ratios (HRs) and corresponding *P* values. Statistical significance was set at $P < 5E\text{-}8$. Manhattan plots and quantile-quantile (QQ) plots were generated using FUMA v1.3.0.[20] Genomic inflation factor (λ) was calculated using the estlambda function in GenABEL v1.8 in R.[21] LocusZoom[22] was used to visualize the chr9:103,865,000–104,055,000 region (GWAS peak). Haploview v4.2[23] was used to generate linkage disequilibrium (LD) plots of D′ and r^2 for SNPs in the chr9:103,865,000–104,055,000 region with GWAS $P < 1E\text{-}4$. LD between 2 SNPs was calculated using 1000G Phase3 v5 in LDlink.[24] Linear regression was used to estimate and test differences in mean age at diagnosis (β). Conditional analysis was performed using coxph function in the survival v2.41 R package. Moving average plots (MAPs) were generated using the freqMAP v0.2 R package.[25] Gene-based analysis was conducted using summary statistics from the GWAS and LD from the 1000G Phase3 EUR to map the GWAS SNPs to 18,985 protein-coding genes (hg19 build) and to calculate gene-based *P* values, using MAGMA v1.06,[26] as implemented in FUMA v1.3.0.[20] Statistical significance was set at Bonferroni-corrected $P < 2.6E-6$ (0.05/18,985).

For replication, Cox regression (coxph function in the survival v2.41 R package) was used to replicate the association of 2 SNPs with age at diagnosis. We used the same model as the NGRC (additive genetic model, treating age at diagnosis as a quantitative trait). Because of the availability of PCs only in a subset of PAGE, analyses were conducted twice: using the full PAGE data set without PC adjustment and using the subset that had AIMs and adjusting for PC1-3. $PAGE_I$ and $PAGE_P$ were treated separately and were combined using meta-analysis after testing for heterogeneity. If *P* of heterogeneity was <0.1, the random-effect model was used. Meta-analysis was performed using the metagen function in the meta v4.8 R package.

Functional annotation

Functional annotation was conducted in FUMA v1.3.0,[20] using SNPs with GWAS $P < 1E-6$ and all variants in $r^2 \geq 0.6$ with them, and included CADD analysis,[27] eQTL mapping,[28] 3D chromatin interaction mapping (Hi-C),[29] annotation of enhancers,[30] tissue-specific expression of genes identified via Hi-C and eQTL mapping,[28] and their age-specific expression in the brain (BrainSpan.org). The false discovery rate (FDR) was used to correct for multiple testing. STRUM was used to predict the effect of a missense on the structural stability of a protein.[31]

Data availability

NGRC genotype and phenotype data are available at dbGaP (ncbi.nlm.nih.gov/gap; accession number phs000196.v3.p1).

RESULTS

GWAS

In SNP-based GWAS, the most significant signal for association, at $P = 5\text{E-}8$, mapped to *LPPR1* on chromosome 9q31.1 (figure 1, A and B). In the gene-based test, *LPPR1* achieved $P = 1\text{E-}8$, surpassing the genome-wide statistical significance threshold of $P < 2.6\text{E-}6$ (figure 1, C and D). The *P* values were not inflated ($\lambda = 1.007$ SNP based, $\lambda = 1.04$ gene based). Analysis of LD in the region revealed 2 haplotype blocks with seemingly independent signals for association (figure 1, E and F). There was strong LD among SNPs in each block, but weak LD between the blocks ($r^2 \leq 0.2$) because of a recombination hot spot between them (figure 1F). The 2 blocks were in a ~200 Kb region inside *LPPR1*. Block 1 consisted of 51 SNPs with MAF~0.01, which yielded HR = 2.02–1.88, with $P = 9\text{E-}7$ to $2\text{E-}5$ for association with age at diagnosis. Block 2 consisted of 39 SNPs with MAF~0.02, which yielded HR = 1.88–1.85, with $P = 5\text{E-}8$ to $7\text{E-}7$. We chose 1 SNP to represent each block for replication: rs73656147 for block 1 (MAF = 0.01, HR

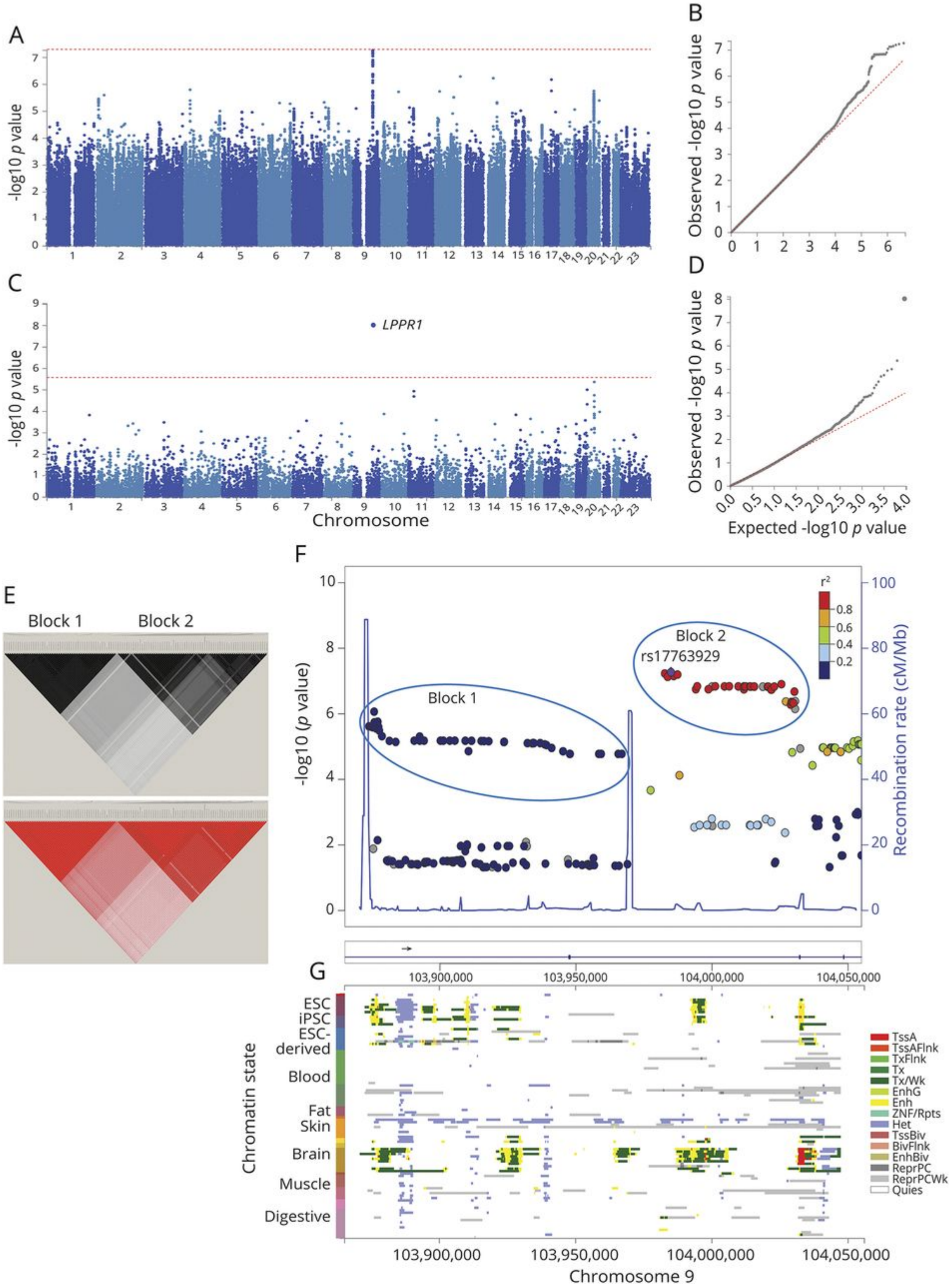


Figure 1: Results of genome-wide association study for age at diagnosis of PD.

Genome-wide association was tested between 8.5 million SNPs and age at diagnosis in 1,950 PD cases from the NGRC, using the Cox hazard ratio regression method and adjusting for principal components (PC1-3). (A) Manhattan plot of SNP-based GWAS. Tallest peak, at *P* = 5E-8, was on chromosome 9q31.1. (B) QQ plot of SNP-based GWAS. The observed *P* values were not inflated (λ = 1.007). (C) Manhattan plot of gene-based GWAS. *LPPR1* was at *P* = 1E-8. Statistical significance threshold was *P* < 2.6E-6, which is Bonferroni corrected for the 18,985 protein-coding genes tested. (D) QQ plot of gene-based GWAS. The observed p values were not inflated (λ = 1.04). (E) r^2 (top panel) and D' (bottom panel). Linkage disequilibrium (LD) across the SNPs that gave *P* < 1E-4 for association with age at diagnosis reveals 2 blocks represented by rs73656147 (left triangle) and rs17763929 (right triangle). (F) Magnified map of the associated region (chr9:103,865,000–104,055,000), showing that PD-associated SNPs map to *LPPR1* and form 2 haplotype blocks separated by recombination hot spots (blue spikes). (G) Chromatin state of *LPPR1* (Roadmap 111 Epigenomes), showing that active enhancers (yellow), transcription start site (red), and transcripts (green) of *LPPR1* are seen only in stem cells and the brain and that the GWAS SNPs align with regulatory elements. ESC = embryonic stem cell; iPSC = induced pluripotent stem cell; TssA = active transcription start site (TSS); TssAFlnk = flanking active TSS; TxFlnk = transcription at gene 5′ and 3′; Tx = strong transcription; TxWk = weak transcription; EnhG = genic enhancers; Enh = enhancers; ZNF/Rpts = zinc-finger genes and repeats; Het = heterochromatin; TssBiv = bivalent/poised TSS; BivFlnk = flanking bivalent TSS/enhancer; EnhBiv = bivalent enhancer; ReprPC = repressed polycomb; ReprPCWk = weak repressed polyComb; Quies = quiescent.

= 1.95, *P* = 3E-6) and rs17763929 for block 2 (MAF = 0.02, HR = 1.88, *P* = 5E-8), both in Hardy-Weinberg (*P* > 0.3), with little correlation between them (r^2 = 0.2). Conditional analysis conducted to determine whether the 2 blocks were tagging the same or different disease-associated variants was inconclusive because although the signals were weakened when adjusted for each other, neither was abolished when conditioned on the other (table e-1, links.lww.com/NXG/A69).

There are 2 caveats in interpreting statistical evidence for association with age at diagnosis. First, age at diagnosis is correlated with age (r^2 = 0.74, *P* < 2E-16), which can result in spurious conclusions if the driving force responsible for the association is not identified. Second, tests of age at diagnosis are conducted using patients only without the benefit of controls. For example, an SNP that appears to be associated with earlier PD diagnosis may in fact be associated with an age-related event unrelated to PD. To interpret the statistical evidence for association with age at diagnosis, we examined whether and how allele frequencies vary by age in cases or in controls. Allele frequencies were plotted in a moving average window as a function of age (figure e-3, links.lww.com/NXG/A68). Starting at age 45 years, allele frequencies were the same in cases and controls. In controls, allele frequencies remained the same across the age spectrum, whereas in cases, they decreased sharply and significantly by age and by age at diagnosis. The effect was therefore in cases and not in controls. Next, conditional analysis was conducted to tease age from age at diagnosis (table 2). The minor alleles of rs73656147 and rs17763929 were associated with age, as was expected, given their association with age at diagnosis. However, the association with age at diagnosis persisted when adjusted for age, but the association with age was abolished when adjusted for age at diagnosis. Hence, age at diagnosis was the driving force, and association with age was a by-product of the correlation.

To gauge robustness of the association signals with age at diagnosis and to test for heterogeneity, we stratified the data by 8 PD-relevant variables, tested the association of each SNP with age at diagnosis within each stratum, and compared the results across strata for evidence of heterogeneity (table e-2, links.lww.com/NXG/A70). The 8 categories of stratification were family history, sex, cigarette smoking, caffeine intake, nonsteroidal anti-inflammatory drugs use, recruitment site, Jewish heritage, and the European country of ancestral origin. The association signal for rs73656147 (block 1) was robust across all strata. rs17763929 (block 2) showed evidence of heterogeneity as a

Table 2. Association of *LPPR1* variants with age and age at diagnosis is driven by age at diagnosis. The associations were tested in the NGRC dataset using Cox regression, and the effect sizes were estimated using linear regression (LR). HR= hazard ratio, is the age-for-age increase in the odds of event per copy of the minor allele, as estimated using Cox regression. ß is the difference in years in age at diagnosis between carriers of one minor allele vs. no minor allele, as estimated using linear regression. Age at diagnosis was the primary outcome of the study. Minor alleles of rs73656147 and rs17763929 were associated with higher HR and younger age at diagnosis (Ia). The association was not influenced by sex (Ib), which was expected because, unlike PD risk which is significantly associated with sex (OR=3.26, *P*<2E-16), age at diagnosis is not associated with sex (HR=0.99, *P*=0.83). Minor alleles were also associated with younger ages in cases (II), but not in controls (III). Since age and age at diagnosis were correlated (r^2= 0.74, *P*<2E-16), an association with one will show as an association with both. In conditional analysis, the association with age at diagnosis persisted when adjusted for age (IV), but the association with age was abolished when adjusted for age-at-diagnosis (V), suggesting age at diagnosis was the driving force and association with age was a by-product of the correlation**.** Abbreviations: CI = confidence interval; HR = hazard ratio; LR = linear regression; ß = effect size on age at diagnosis (in years) per copy of minor allele.

	N	Block 1 rs73656147			Block 2 rs17763929		
		Cox		LR	Cox		LR
		HR	*P*	ß [95% CI]	HR	*P*	ß [95% CI]
Ia. Association with age at diagnosis in cases	1,950	1.95	3E-6	-6.00 [-9.18 to -2.83]	1.88	5E-8	-5.65 [-8.20 to -3.11]
Ib. Association with age at diagnosis in cases adjusted for sex	1,950	1.95	3E-6	-5.98 [-9.16 to -2.81]	1.88	6E-8	-5.61 [-8.16 to -3.07]
II. Association with age in cases	2,000	1.48	5E-3	-4.19 [-7.1 to -1.3]	1.53	2E-4	-3.56 [-5.9 to -1.2]
III. Association with age in controls	1,986	0.83	0.08	2.34 [-0.6 to 5.2]	0.84	0.07	2.37 [-0.3 to 5.1]
IV. Association with age at diagnosis in cases, adjusted for age	1,950	1.45	0.01	-2.30 [-3.9 to -0.7]	1.26	0.05	-2.11 [-3.4 to -0.8]
V. Association with age in cases, adjusted for age at diagnosis	1,950	0.92	0.56	0.78 [-0.8 to 2.3]	0.99	0.96	0.68 [-0.6 to 1.9]

function of recruitment site and the European country of ancestral origin. Given these results, we tested the association of the 2 SNPs with PCs. rs17763929 was associated with PC1 (*P* = 7E-6) and PC3 (*P* = 8E-3), and rs73656147 was not (*P* > 0.05 for PC1-3), indicating the presence of population structure in block 2, but not in block 1.

Replication

In comparison to NGRC, which had a 65-year range for age at diagnosis, the PAGE data sets had a narrower range of less than 30 years. Because PAGE participants were of similar age at study entry, the method of classifying the participants into prevalent PD (diagnosis before entry) vs incident PD (diagnosis after entry) inevitably assigned earlier ages at diagnosis to the prevalent group (PAGE$_{\text{P}}$) and later diagnoses to the incident group (PAGE$_{\text{I}}$). Mean age at diagnosis in PAGE$_{\text{P}}$ was 59.9 ± 6.6 years, which was similar to the NGRC (60.4 ± 11.4). PAGE$_{\text{I}}$ participants were on average 10 years older at diagnosis (69.4 ± 5.4, range 53–81 years). Given the disparity in the range and mean ages at diagnosis, we analyzed PAGE$_{\text{P}}$ and PAGE$_{\text{I}}$ separately.

Association of rs73656147 (block 1) with age at diagnosis replicated robustly (table 3). There was no evidence of heterogeneity between PAGE$_{\text{I}}$ and PAGE$_{\text{P}}$ in the association of rs73656147 with age at diagnosis, although the signal was stronger in PAGE$_{\text{P}}$ than in PAGE$_{\text{I}}$, which is not surprising, given that the former is enriched in cases with earlier age at diagnosis. Nor was there evidence of heterogeneity between PAGE and NGRC for the association of rs73656147 with age at diagnosis. Meta-analysis yielded HR = 2.14, *P* = 1E-3 for replication and HR = 2.00, *P* = 9E-9 for replication and discovery. Mean difference in age at diagnosis per copy of rs73656147 minor allele was

Table 3. Replication. Two SNPs with signals for association with age at diagnosis of PD in the NGRC data set (discovery) were genotyped and tested for association with age at diagnosis of PD in the PAGE dataset (replication). PAGE participants were designated as prevalent (PAGE$_P$) if they were diagnosed before study entry, or incident (PAGE$_I$) if they were diagnosed during the study. Cox regression was used to test association of SNP (additive model) with age at diagnosis (quantitative trait) and to calculate hazard ratios (HR) and corresponding significance (*P*). NGRC was adjusted for principal components (PC1-3) in GWAS and meta-analyses. Only a subset of PAGE had ancestry informative markers for which PC could be calculated; thus, results are shown for the full PAGE dataset without PC adjustment, and for PAGE subsample with PC adjustment. *P* values are two-sided for NGRC, and one-sided for PAGE due to the directionality of the hypothesis being replicated. Meta-analysis A: NGRC (PC1-3 adjusted), and PAGE (all data without PC adjustment). Meta-analysis B: NGRC (PC1-3 adjusted), and PAGE (subset of data adjusted for PC1-3). rs73656147 replicated robustly with no evidence of heterogeneity across datasets. rs17763929 replicated in PAGE$_P$ and showed significant heterogeneity between PAGE$_I$ and PAGE$_P$ or NGRC. Meta-analysis was conducted using the fixed effects model if there was no evidence for heterogeneity ($P \geq 0.1$), and the random effects model if there was heterogeneity ($P < 0.1$). Abbreviations: HR = hazard ratio; NGRC = NeuroGenetics Research Consortium; ns = not statistically significant; PAGE = Parkinson's, Genes and Environment; PC = principal component; PD = Parkinson disease.

		Age-at -diagnosis	Block 1 rs73656147		Block 2 rs17763929	
Datasets	N PD cases	Mean ± SD	HR	*P*	HR	*P*
NGRC (discovery)	1,950	60.4 ± 11.4	1.95	3E-6	1.88	5E-8
PAGE$_P$ (replication)	209	59.9 ± 6.6	2.88	7E-4	1.87	0.01
PAGE$_P$ with PC1-3	113	59.9 ± 6.8	2.17	0.05	3.03	4E-3
PAGE$_I$ (replication)	517	69.4 ± 5.4	1.62	0.07	1.04	0.41
PAGE$_I$ with PC1-3	283	69.2 ± 5.3	1.48	0.16	1.03	0.45
Meta-analysis A	Heterogeneity rs73656147	Heterogeneity rs17763929				
PAGE$_P$ & PAGE$_I$	ns	0.08	2.14	1E-3	1.34	0.31
NGRC & PAGE$_P$	ns	ns	2.08	2E-8	1.88	4E-9
NGRC & PAGE$_I$	ns	0.01	1.90	9E-7	1.42	0.23
NGRC & PAGE$_P$ & PAGE$_I$	ns	0.02	2.00	9E-9	1.53	0.04
Meta-analysis B						
PAGE$_P$ & PAGE$_I$	ns	0.02	1.73	0.07	1.67	0.34
NGRC & PAGE$_P$	ns	ns	1.97	6E-7	1.95	3E-9
NGRC & PAGE$_I$	ns	0.02	1.89	2E-6	1.43	0.23
NGRC & PAGE$_P$ & PAGE$_I$	ns	0.02	1.91	5E-7	1.68	0.05

−6.0 (95% confidence interval: −9.18 to −2.83) years in the NGRC, −5.53 (−9.72 to −1.34) in PAGE$_P$, −0.84 (−4.22 to 2.55) in PAGE$_I$, and −4.08 (−7.45 to −0.70) in the meta-analysis of the 3 data sets.

Association of rs17763929 (block 2) with age at diagnosis showed significant heterogeneity between PAGE$_I$ and PAGE$_P$ (table 3), as it had within the NGRC (table e-2, links.lww.com/NXG/A70). The association with rs17763929 replicated in PAGE$_P$ but not in PAGE$_I$. There was significant heterogeneity between PAGE$_I$ and NGRC, but not between PAGE$_P$ and NGRC. Meta-analysis of PAGE$_P$ and NGRC yielded HR = 1.88, *P* = 4E-9 for full PAGE data and HR = 1.95, *P* = 3E-9 for the PAGE subsample adjusted for PC1-3. Including PAGE$_I$ with PAGE$_P$ and NGRC in a random-effects meta-analysis diluted the effect size to HR = 1.53, *P* = 0.04. Mean difference in age at diagnosis per copy of rs17763929 minor allele was −5.65 (−8.20 to −3.11) years in the NGRC, −3.62 (−7.23 to −0.02) in PAGE$_P$, and 0.62 (−1.34 to 2.58) in PAGE$_I$.

Functional annotation

Hi-C analysis showed significant (FDR < 1E-6) chromatin interaction between the PD-associated *LPPR1* SNPs and promoters of *LPPR1* and several genes on chromosome 9 (figure 2, A). Some of the SNPs that were significant in Hi-C mapped to enhancers in the brain (table 4 and figure 1, G). Eleven of the genes identified through Hi-C are expressed in the brain: *LPPR1*, *SEC61B*, *MSANTD3-TMEFF1*, *TMEFF1*, *GALNT12*, *MURC*, *GRIN3A*, *NR4A3*, *ALG2*, *MRPL50*, and *ZNF189* (figure 2, B and C). The expression of *LPPR1* in the brain is the strongest in early prenatal stage and decreases with developmental stage and increasing age (figure 2, C).

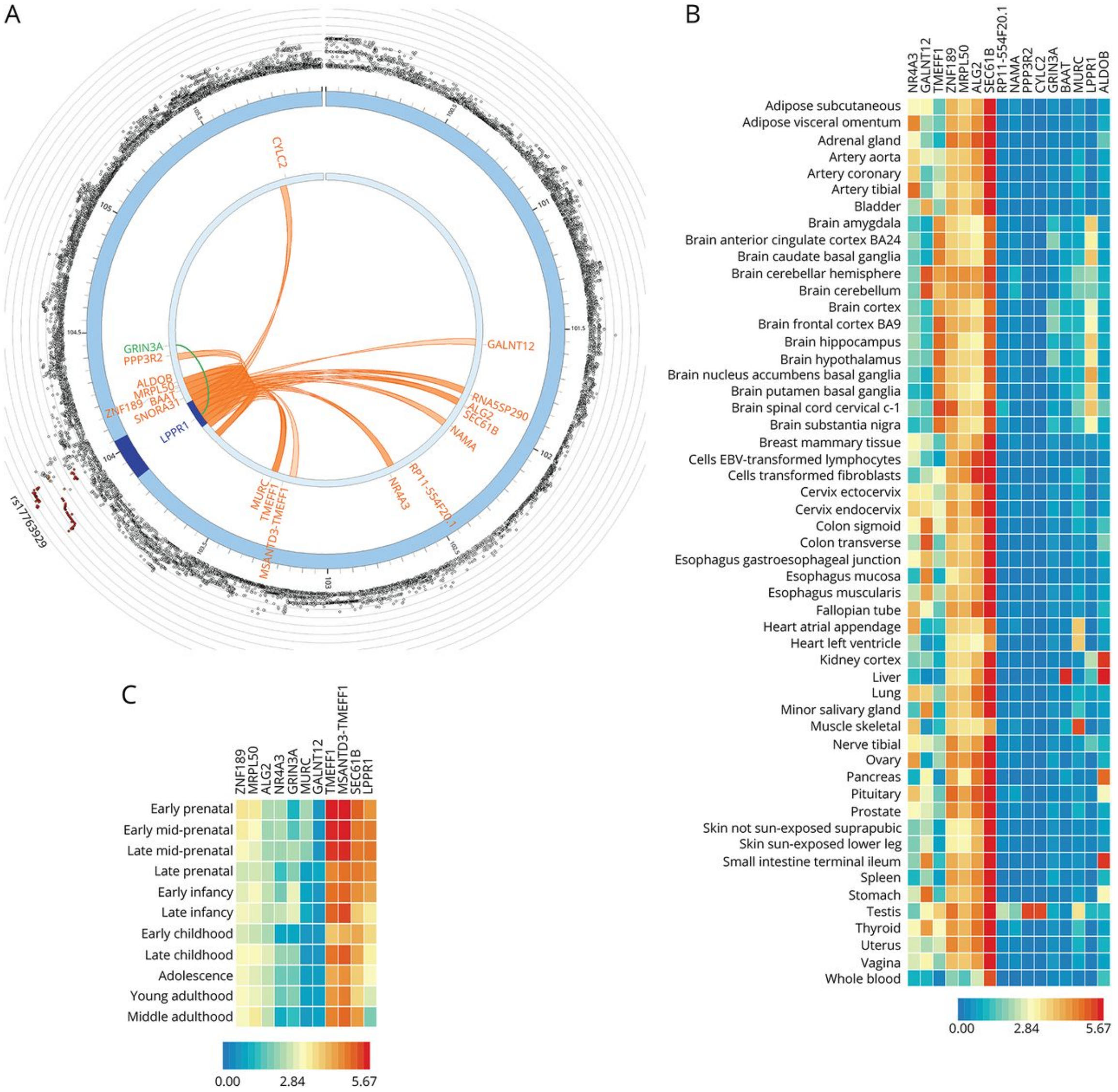


Figure 2: Functionally significant genes.

(A) 3D chromatin interaction (Hi-C) and eQTL analysis. Hi-C revealed significant interaction between GWAS variants in *LPPR1* and 17 other genes on chromosome 9 (FDR < 1E-6, shown in orange). An SNP in *LPPR1* was associated with the expression of *GRIN3A* (FDR = 4E-4, shown in green). (B) Tissue-specific expression of *LPPR1*, *GRIN3A*, and genes in Hi-C with *LPPR1*. Colors reflect average expression (log2 transformed) from highest (red) to lowest/absent (blue). (C) Age-specific expression of the genes in the brain. *LPPR1* expression decreases with age.

Table 4. Functionally significant variants. Functional annotation was conducted on SNPs with GWAS $P<1E\text{-}6$ and SNPs that were in high LD with them ($r^2>0.6$). Variants are shown if they are the lead SNP (most significant) for the block, or an eQTL (FDR=4E-4), or had a CADD score >10, or had both significance evidence for 3D chromatin interaction (Hi-C, FDR<1E-6) and overlapped with an enhancer in the brain. Block 1 is single block of SNPs in high LD. Block 2 has a complex LD structure with at least three sub-haplotypes (figure e1-C). Variants are shown with their rs accession number, chromosome position and the two alleles (major:minor), GWAS *P* value for association with age at diagnosis of PD, and their correlation (r^2) with the lead SNP of the block. eQTL**:** a SNP that is associated with gene expression, in this case, rs117451395 was associated with gene expression levels at *GRIN3A* (FDR=4E-4). CADD**:** a predictive score for the deleteriousness of a variant. A CADD score of 10 usually means the variant is among the top 10% of deleterious mutation in the genome. A CADD score of 20 puts the variant among the top 1% of deleterious mutations. Hi-C: SNPs with significant (FDR<1E-6) evidence for interacting with the promoter region of *LPPR1* or of another gene (see figure 2 for the genes). Hi-C/EnhBrain: the subset of Hi-C SNPs that map to enhancer regions of *LPPR1* in brain according to the Roadmap 111 epigenomes.

[a] One SNP shown to represent several variants in high LD ($r^2>0.9$) with similar MAF, GWAS *P* and Hi-C/EnhBrain evidence.

[b] This mutation yielded ΔΔG= -1.2 which predicts a destabilizing effect on the protein structure of LPPR1.

Block	GWAS SNP	position:alleles	GWAS *P*	r^2	eQTL	CADD	Hi-C/EnhBrain
1	rs77351585	9:103874925:C:T	2E-06	1	-	18	Hi-C/EnhBrain
1	rs73495940	9:103875807:G:C	9E-07	Lead	-	-	Hi-C
1	rs150164200	9:103875896:A:C	2E-06	1	-	10.4	-
1	rs117583993[a]	9:103876647:G:A	3E-06	1	-	-	Hi-C/EnhBrain
1	rs148874623	9:103939117:A:C	9E-06	1	-	12.1	-
1	rs117451395	9:103941039:C:T	1E-05	1	*GRIN3A*	-	Hi-C
1	rs41296085	9:103947810:T:G	2E-05	1	-	18 (missense)[b]	Hi-C
1	rs117900237	9:103959240:G:A	2E-05	1	-	10.5	Hi-C
2	rs17763929	9:103984900:A:G	5E-08	Lead	-	-	Hi-C
2	rs61188842	9:103988006:C:T	8E-05	0.6	-	-	Hi-C/EnhBrain
2	rs117058418	9:104011717:T:C	2E-07	1	-	10.4	Hi-C
2	rs117314512	9:104014244:G:A	2E-07	1	-	12.4	Hi-C
2	rs149155028	9:104032402:TTC:T	1E-05	0.7	-	18.6	Hi-C

CADD analysis, a scoring system for deleteriousness of genetic variants, identified 5 SNPs in block 1 and 3 in block 2, with CADD = 10–19 (table 4), which places them among the top 10% (CADD > 10) to 1% (CADD > 20) of most deleterious mutations in the genome.[27] rs41296085 (CADD = 18, in block 1) is a missense (p.Ser12Ala) in exon 2, predicted to structurally destabilize the LPPR1 protein ($\Delta\Delta G = -1.2$). The remainder of the variants with high CADD scores are in introns. eQTL analysis revealed an association between rs117451395 (block 1) with expression levels of *GRIN3A* (FDR = 4E-4).

DISCUSSION

There has been intense research on PD risk factors, which so far has resulted in identification of numerous causative genes, 40 susceptibility loci, several environmental factors, and a few genes that interact with the environmental factors to increase or reduce the risk of developing PD. In contrast, we know little about factors that affect the rate of disease progression. In this study, we attempted to identify genetic modifiers of age at diagnosis, a reflection of rate of progression, using an unbiased genome-wide approach, followed by independent replication, and functional annotation.

We uncovered evidence for association of genetic variants in neuronal plasticity-related gene 3 (*LPPR1*) with age at diagnosis of PD. Two signals of association were detected, each representing a haplotype block of SNPs. The variants that were associated with earlier age at diagnosis had low allele frequencies (MAF = 0.01–0.02), as were the variants that were previously found for age at onset of familial PD.[11] The low allele frequencies may be one reason why modifier genes have been more difficult to detect than common variants that associate with risk.

The association with block 1 replicated robustly in both PAGE_P and PAGE_I. Block 2 signal replicated in PAGE_P, but not in PAGE_I. Block 2 has a complex LD structure, with evidence of population substructure, which limits generalizability of results. Failure to capture a signal for block 2 in PAGE_I may be because we had genotype on only 1 SNP in block 2 for PAGE, which did not fully capture the complexity of block 2. PAGE_I participants being significantly older than NGRC and PAGE_P participants may also be a factor. *LPPR1* promotes neuroregeneration,[32–34] but its expression diminishes with age to nearly undetectable level by age 40 years (figure 2C). One can speculate that some detrimental variants may not have an effect after a certain age when the gene is no longer expressed.

Functional annotation of the PD-associated variants in *LPPR1* revealed the presence of several variants with predicted deleterious effects, including a missense that destabilizes the structure of LPPR1, a regulatory element that associates with expression levels of *GRIN3A*, and enhancers that interact with promoters of *LPPR1* and several other genes in the brain. Some of the candidate genes that were identified via interaction with *LPPR1* play key roles in pathways that are implicated in PD, including *GRIN3A* (which encodes a subunit of NMDA receptor involved in the glutamate-regulated ion channels in the brain), *SEC61B* (protein transport apparatus of the endoplasmic reticulum membrane), *MURC* (Rho kinase signaling), and *MRPL50* (mitochondrial ribosomal protein).

LPPR1 is one of the 5 members of a brain-specific gene family that modulates neuronal plasticity during development, aging, and after brain injury.[32–34] *LPPR1* is the strongest driver of axonal outgrowth in the gene family. Studies in mice have shown that after neuronal injury, overexpression of *LPPR1* enhances axonal growth, improves motor behavior, and promotes functional recovery.[33,34] Extrapolating to our findings, we propose that *LPPR1* is involved, not necessarily in the cause of PD, rather in response to damage, and influences the efficacy of regeneration and the subsequent rate of deterioration in preclinical PD. The actual cause of injury and neuronal death is not stipulated in this hypothesis; it could be head trauma, environmental toxins or genetic, but once the initial damage is incurred, it is the efficacy of intrinsic mechanisms of repair that determine the rate of disease progression. Present findings provide a strong foundation for mechanistic studies to investigate the role of *LPPR1* in PD and determine its potential as a therapeutic target to impede disease progression.

AUTHOR CONTRIBUTIONS

Z.D. Wallen: statistical analysis, review, and critique of the manuscript. H. Chen: creation of the PAGE data set, acquisition of data, review, and critique of the manuscript. E.M. Hill-Burns: assembly and QC of phenotype and genotype data, imputation, review, and critique of the manuscript. S.A. Factor and C.P. Zabetian: creation of the NGRC data set, review, and critique of the manuscript. H. Payami: study concept, design and execution, creation of the NGRC data set, wrote the manuscript, and obtained funding.

STUDY FUNDING

This work was supported by the National Institute of Neurological Disorders and Stroke grant R01NS036960. Additional support was provided by a Merit Review Award from the Department of Veterans Affairs grant 1I01BX000531; Office of Research &

Development, Clinical Sciences Research & Development Service, Department of Veteran Affairs; the Close to the Cure Foundation; and the Sartain Lanier Family Foundation. Genome-wide array genotyping was conducted by the Center for Inherited Disease Research, which is funded by the NIH grant HHSN268200782096C. Collection of samples and data for replication (Parkinson, Genes, and Environment (PAGE) study) was supported by the intramural research program of NIH National Institute of Environmental Health Sciences grant Z01 ES101986. Funding agencies did not have a role in the design or execution of the study.

DISCLOSURE

Z.D. Wallen, H. Chen, C.P. Zabetian, and E.M. Hill-Burns report no disclosures. S.A. Factor has received honoraria from Neurocrine, Lundbeck, Teva, Avanir, Sunovion Pharmaceuticals, Adamas, and UCB; has received research support from Ipsen, Medtronic, Teva, US WorldMeds, Sunovion Pharmaceuticals, Solstice, Vaccinex, Voyager, the CHDI Foundation, the Michael J. Fox Foundation, and the NIH; and receives publishing royalties from Demos, Blackwell Futura, and UpToDate. H. Payami has received research support from the NIH and the University of Alabama at Birmingham. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/NG.

REFERENCES

1. Obeso JA, Stamelou M, Goetz CG, et al. Past, present, and future of Parkinson's disease: a special essay on the 200th Anniversary of the Shaking Palsy. Mov Disord 2017;32:1264–1310.

2. Zareparsi S, Taylor TD, Harris EL, Payami H. Segregation analysis of Parkinson disease. Am J Med Genet 1998;80:410–417.

3. Maher NE, Currie LJ, Lazzarini AM, et al. Segregation analysis of Parkinson disease revealing evidence for a major causative gene. Am J Med Genet 2002;109:191–197.

4. McDonnell SK, Schaid DJ, Elbaz A, et al. Complex segregation analysis of Parkinson's disease: the Mayo clinic family study. Ann Neurol 2006;59:788–795.

5. Hamza TH, Payami H. The heritability of risk and age at onset of Parkinson's disease after accounting for known genetic risk factors J Hum Genet 2010;55:241–243.

6. Nalls MA, et alEscott-Price V, Consortium IPsDG, Nalls MA, et al. Polygenic risk of Parkinson disease is correlated with disease age at onset. Ann Neurol 2015;77:582–591.

7. Chang D, Nalls MA, Hallgrimsdottir IB, et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. Nat Genet 2017;49:1511–1516.

8. Nalls MA, Escott-Price V, Williams NM, et al. Genetic risk and age in Parkinson's disease: continuum not stratum. Mov Disord 2015;30:850–854.

9. Lill CM, Hansen J, Olsen JH, Binder H, Ritz B, Bertram L. Impact of Parkinson's disease risk loci on age at onset. Mov Disord 2015;30:847–850.

10. Pihlstrom L, Toft M. Cumulative genetic risk and age at onset in Parkinson's disease. Mov Disord 2015;30:1712–1713.

11. Hill-Burns EM, Ross OA, Wissemann WT, et al. Identification of genetic modifiers of age-at-onset for familial Parkinson's disease. Hum Mol Genet 2016;25:3849–3862.

12. Trinh J, Gustavsson EK, Vilarino-Guell C, et al. DNM3 and genetic modifiers of age of onset in LRRK2 Gly2019Ser parkinsonism: a genome-wide linkage and association study. Lancet Neurol 2016;15:1248–1256.

13. Hamza TH, Zabetian CP, Tenesa A, et al. Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. Nat Genet 2010;42:781–785.

14. Chen H, Huang X, Guo X, et al. Smoking duration, intensity, and risk of Parkinson disease. Neurology 2010;74:878–884.

15. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry 1992;55:181–184.Abstract/FREE Full Text

16. Schatzkin A, Subar AF, Thompson FE, et al. Design and serendipity in establishing a large cohort with wide dietary intake distributions: the National Institutes of Health-American Association of retired persons diet and Health study. Am J Epidemiol 2001;154:1119–1125.

17. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5:e1000529.

18. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909.

19. Aulchenko YS, Struchalin MV, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics 2010;11:134.

20. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017;8:1826.

21. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics 2007;23:1294–1296.

22. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010;26:2336–2337.

23. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–265.

24. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015;31:3555–3557.

25. Payami H, Kay DM, Zabetian CP, Schellenberg GD, Factor SA, McCulloch CC. Visualizing disease associations: graphic analysis of frequency distributions as a function of age using moving average plots (MAP) with application to Alzheimer's and Parkinson's disease. Genet Epidemiol 2010;34:92–99.

26. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 2015;11:e1004219.

27. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310–315.

28. Consortium G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648–660.Abstract/FREE Full Text

29. van Berkum NL, Lieberman-Aiden E, Williams L, et al. Hi-C: a method to study the three-dimensional architecture of genomes. J Vis Exp 2010:e-1869.

30. Consortium RE, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–330.

31. Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 2016;32:2936–2946.

32. Savaskan NE, Brauer AU, Nitsch R. Molecular cloning and expression regulation of PRG-3, a new member of the plasticity-related gene family. Eur J Neurosci 2004;19:212–220.

33. Broggini T, Schnell L, Ghoochani A, et al. Plasticity related gene 3 (PRG3) overcomes myelin-associated growth inhibition and promotes functional recovery after spinal cord injury. Aging (Albany NY) 2016;8:2463–2487.

34. Fink KL, Lopez-Giraldez F, Kim IJ, Strittmatter SM, Cafferty WBJ. Identification of intrinsic axon growth modulators for intact CNS neurons after injury. Cell Rep 2017;18:2687–2701.

APPLICATION OF SIXTEEN DIFFERENTIAL ABUNDANCE METHODS TO TWO LARGE PARKINSON DISEASE GUT MICROBIOME DATASETS

by

ZACHARY D. WALLEN AND HAYDEH PAYAMI

In preparation for *Microbiome*

Format adapted for dissertation

ABSTRACT

When studying the relationship between the microbiome and a particular disease, a common question asked is what individual microbes are differentially abundant between a disease and healthy state. Numerous differential abundance testing methods exist and range from standard statistical tests (e.g. Kruskal-Wallis rank sum test) to methods specifically designed for microbiome data (e.g. metagenomeSeq). Choice of differential abundance method can greatly influence what, and how many, significant differential abundance signatures are detected. To compare results of different differential abundance testing methods in our own data, we performed differential abundance analysis in two large Parkinson disease (PD)-gut microbiome datasets using 16 methods.

We found 16 differential abundance testing methods from the literature, tested association of PD with genera in two datasets (N=333 and 507) using all 16 methods, then compared results of each method within and across datasets. Concordances were calculated between each pair of method results and visualized using heatmaps. Hierarchical clustering was performed to determine if any groups of PD-genus associations were being agreed (or disagreed) upon by all or a subset of methods.

Pairwise concordances between method results ranged from 0.46-0.99 with the mean concordance being 0.76 per dataset. Mean concordance significantly dropped to 0.63 when incorporating information on which PD-genus associations replicated in both datasets. Variable effect of data filtering was observed for one method (ANCOM) when removal of rarer and unclassified genera before analysis drastically reduced the number of signals detected. Hierarchical clustering revealed three groups of PD-genus associations that were (1) more likely to be replicated by the majority of methods, (2)

replicated by little to no methods, and (3) more likely to be replicated by a subset of potentially more sensitive methods and included rarer genera enriched in PD.

Variation between method results was evident, especially when comparing results across two datasets. We observed that filtering taxonomic data before analysis has an opposite effect on ANCOM as compared to other methods, drastically decreasing the number of detected associations instead of increasing them. We found methods with previously reported low false positive rate (FPR) and false discovery rate (FDR) tended to be more similar to one another and replicated PD-genus associations more likely to be backed up by other methods. A subset of methods that included some previously shown to have high FPR/FDR, but also high sensitivity, detected and replicated a group of rarer genera, mostly enriched in PD, that might be of interest for future investigations.

INTRODUCTION

Microbiome research has gained immense traction in recent years driven primarily by technological advances in sequencing and exponential increase in computational resources and tools. The availability of these new tools and technologies have solidified a place for microbiome research in many fields of research including the biomedical research community where a large portion of the research effort is targeted at the gut microbiome [Schmidt, Raes & Bork 2018]. A number of diseases have been associated with alterations of individual taxa in the gut microbiome [Schmidt, Raes & Bork 2018], and these associations are usually made through a statistical analysis commonly referred to differential abundance testing [McMurdie & Holmes 2014]. Differential abundance testing involves the use of serial univariate statistical tests to determine if certain taxa are significantly different between groups [Thorsen et al. 2016]. Numerous differential abundance testing methods exist and include classical statistical tests (e.g. Kruskal-Wallis rank sum test), methods developed to detect differential expression of gene transcripts in RNA-Seq data (e.g. DESeq2, edgeR), methods specifically designed for detecting differentially abundant taxa in microbiome data (e.g. metagenomeSeq), and methods designed to detect differentially abundant features in compositional data (e.g. ALDEx2). Choice of differential abundance method can greatly influence what, and how many, differentially abundant taxa are detected in a disease state, and most, if not all, methods will respond differently to microbiome data due to differences in their underlying characteristics. Multiple studies have previously assessed and compared the performance of popularly used differential abundance testing methods, measuring their false positive rates (FPR), false discovery rates (FDR), sensitivities, and/or specificities from simulated data [McMurdie & Holmes 2014; Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019], with only one of these studies testing different methods on real data [Weiss et al. 2017].

There is little literature on how different differential abundance methods behave when performed on real, complex disease oriented gut microbiome datasets, therefore, we performed differential abundance testing using a variety of methods on two, large Parkinson disease (PD) – gut microbiome datasets in order to compare their results. We found 16 differential abundance testing methodologies from the literature, and used them to test for differentially abundant genera between PD and neurologically healthy controls in both datasets. We compared their results within and across datasets and found that concordances between methods varied as has been previously shown in method comparison studies. Methods that were previously shown to have lower FPR/FDR had the highest concordances on average, especially with one another. These methods also detected and replicated PD-genus associations more likely to be replicated by the majority of methods on average. Methods previously shown to have high FPR/FDR, but also higher sensitivity, produced the least concordant results on average with other methods, but managed to detect and replicate a subset of rarer genera enriched in PD that the lower FPR/FDR methods did not replicate.

METHODS

Subjects, metadata, gut microbiome

Study was approved by institutional review boards at all participating institutions. Subjects, metadata, and gut microbiome data of datasets 1 and 2 have been previously described [Hill-Burns et al. 2017; Wallen et al. 2020]. We enrolled subjects and collected metadata and fecal samples from 212 PD and 136 neurologically healthy control subjects for dataset 1, and 323 PD and 184 neurologically healthy controls for dataset 2. Dataset 1 subjects were enrolled in Seattle, WA, Albany, NY, and Atlanta, GA, while all dataset 2 subjects were enrolled in Birmingham, AL. Methods for enrollment and collection of metadata and fecal samples were uniform across enrollment sites. PD was diagnosed according to the UK Brain Bank criteria by movement disorder specialists. Controls were self-reported free of neurological disease. Metadata were collected using questionnaires, or extracted from medical records. Stool samples were collected at home using DNA/RNA-free sterile swabs and mailed through U.S. postal service. All subjects provided written informed consent for their participation in the study.

DNA was extracted from stool samples using the automated MoBio PowerMag Soil DNA Isolation Kit (dataset 1) or manual MoBio PowerSoil DNA Isolation Kit (dataset 2). Hypervariable region 4 (V4) of the 16S rRNA gene was amplified with primers 515F-806R. Paired-end 150 bp (dataset 1) or paired-end 250 bp (dataset 2) sequencing was performed on V4 amplicons using Illumina MiSeq. Fifteen samples in dataset 1 resulted in low sequence count and were excluded.

Bioinformatic processing of sequences was performed separately for each dataset. Primers were trimmed from sequences using cutadapt v 1.16 [Martin 2011]. DADA2 v 1.8 was used for quality trimming and filtering sequences, de-replicating sequences, inferring amplicon sequence variants (ASVs), merging of forward and reverse sequences, and detection and removal of chimeras [Callahan et al. 2016]. Final ASV tables for dataset 1 and dataset 2 contained 6,844 unique ASVs for 201 PD and 132 controls samples and 12,198 unique ASVs for 323 PD and 184 control samples respectively. Taxonomy was assigned to ASVs using DADA2's native implementation of the Ribosomal Database Project naïve Bayesian classifier with SILVA v 132 as reference and a bootstrap confidence of 80% [Wang et al. 2007]. Phylogenetic trees were constructed by first performing a multiple sequence alignment with DECIPHER v 2.8.1 [Wright et al. 2015], then building a phylogenetic tree with phangorn v 2.8.1 [Schliep et al. 2011]. Phyloseq v 1.24.2 was used to create a phyloseq object for each dataset containing their respective ASV table, taxonomy classifications, phylogenetic trees, and subject metadata [McMurdie & Holmes 2013]. To agglomerate ASV level phyloseq objects to genus level, the tax_glom function in phyloseq was used without removal of unclassified genera.

Total number of genera detected in dataset 1 was 445. Total number of genera detected in dataset 2 was 561.

Differential abundance testing

We tested for association between genera and PD in dataset 1 and 2 separately using 16 differential abundance methods. Method characteristics and parameters chosen for each method that differed from default can be found in Tables 1 and 2 respectively. Analyses using each method was performed as follows:

Kruskal-Wallis rank sum test [Kruskal & Wallis 1952]: Genera counts were transformed to relative abundance ($\frac{\text{genus count}}{\text{total sample count of all genera}}$), then unclassified genera, and genera present in < 10% of samples were excluded. The kruskal.test function from the stats R package was used to test for significant differences in genera relative abundances between PD and controls. *P* values were corrected for multiple testing using Benjamini-Hochberg (BH) false discovery rate (FDR) method implemented in the p.adjust function from stats package.

Welch's t-test with log transformation (log t-test) [Welch 1947]: Genera counts with a pseudo-count of 1 added were log transformed, then transformed to relative abundance. Unclassified genera, and genera present in < 10% of samples were excluded. The t.test function from the stats R package was used to test for significant differences in genera relative abundances between PD and controls. *P* values were corrected for multiple testing using BH FDR method from the p.adjust function.

Negative binomial generalized linear model with and without zero-inflation (GLM NBZI): Total sequence count was calculated for each sample. Unclassified genera, and genera present in < 10% of samples were then excluded. Using raw counts, a negative-binomial generalized linear model with and without a zero-inflation component was fitted for each genus with the glmmTMB R package v 0.2.2.0 using log(total sequence count) as an offset variable, and PD vs control as the independent variable. Results were extracted from the model with the lowest Akaike information criterion. *P* values were calculated using the base summary function in R and corrected for multiple testing using BH FDR method implemented in the p.adjust function from stats package.

Generalized linear model with centered log ratio transformed data (GLM CLR): Genera counts with a pseudo-count of 1 added were centered log ratio (clr) transformed, then unclassified genera, and genera present in < 10% of samples were excluded. A standard linear regression model using Gaussian distribution was fitted for each genus with the glm function from the R stats package with PD vs control as the independent variable. *P* values were calculated using the base summary function in R and corrected for multiple testing using BH FDR method implemented in the p.adjust function from stats package.

Analysis of Composition of Microbes (ANCOM) [Mandal et al. 2015]: ANCOM was ran twice, once excluding unclassified genera, and genera present in < 10% of samples (ANCOM filtered), and again using all genera (ANCOM unfiltered), due to the drastic decrease in significant signals observed for ANCOM when filtering genera before analysis. Raw counts of genera were used as input to the ANCOM.main function from the ANCOM v 2 R code (downloaded from https://sites.google.com/site/siddharthamandal1985/research). PD vs control was specified as the main variable. The taxa-wise FDR option (multcorr=2) was chosen for the multiple testing correction method. An FDR significance threshold of 0.05 was chosen for calculation of *W* statistics. *W* statistics greater than or equal to 80% of the total number of genera tested were considered significant.

metagenomeSeq fitZIG [Paulson et al. 2013]: Cumulative sum scaling (CSS) was applied to genera counts using the cumNorm function in metagenomeSeq R package v 1.22.0. Unclassified genera, and genera present in < 10% of samples were then excluded. A zero-inflated Gaussian model was fitted for each genus using function fitZig in metagenomeSeq. *P* values were corrected for multiple testing using BH FDR method implemented in the MRfulltable function in metagenomeSeq.

metagenomeSeq fitFeatureModel [Paulson et al. 2013]: CSS was applied to genera counts using cumNorm function in metagenomeSeq. Unclassified genera, and genera present in < 10% of samples were then excluded. A zero-inflated log-normal model was fitted for each genus using function fitFeatureModel in metagenomeSeq. *P* values were corrected for multiple testing using BH FDR method implemented in the MRfulltable function in metagenomeSeq.

edgeR exactTest-TMM (edgeR TMM) [Robinson & Smyth 2008]: Using raw genera counts, normalization factors were calculated with the trimmed mean of M-values (TMM) method using the calcNormFactors function in edgeR R package v 3.22.5. Common and tagwise dispersions were then estimated using estimateCommonDisp and estimateTagwiseDisp functions in edgeR. Unclassified genera, and genera present in < 10% of samples were then excluded. Testing for differential relative abundance between PD and controls was performed using exactTest function in edgeR. *P* values were corrected for multiple testing using BH FDR method implemented in the topTags function in edgeR.

edgeR exactTest-RLE (edgeR RLE) [Robinson & Smyth 2008]: Using genera counts with a pseudo-count of 1 added, normalization factors were calculated with the relative log expression (RLE) method using the calcNormFactors function in edgeR. The remaining steps were the same as exactTest-TMM.

DESeq2 nbinomWaldTest [Love et al. 2014]: Using raw genera counts, normalization factors were calculated using the function estimateSizeFactors in DESeq2 R package v 1.20.0 specifying type=”poscounts”. Unclassified genera, and genera present in < 10% of samples were then excluded. Testing for differential relative abundance between PD and controls was performed using the DESeq function in DESeq2 with default parameters. *P* values were corrected for multiple testing using BH FDR method implemented in the results function in DESeq2.

limma-voom [Ritchie et al. 2015]: Using raw genera counts, TMM values were calculated using the calcNormFactors function in edgeR. Log2 counts per million transformation and mean-variance trend estimation was performed using the voom function in limma R package v 3.36.5. Unclassified genera, and genera present in < 10% of samples were then excluded. Testing was performed by first fitting a linear model for each genus using function lmFit in limma, then testing for differential relative abundance between PD and controls using the eBayes function in limma. *P* values were corrected for multiple testing using BH FDR method implemented in the topTable function in limma.

baySeq [Hardcastle & Kelly 2010]: Total sequence count was calculated for each sample. Unclassified genera, and genera present in < 10% of samples were then excluded. PD and control designations were used as the replicate structure. A list of two group structures was created where one group structure specified all subjects belonged to the same group, and the other specified PD and control groups. The replicate structure, list of group structures, and raw genera counts were combined into a countData object. Total sequence counts were supplied to the countData object. Priors were estimated from a negative binomial distribution using the function getPriors.NB in baySeq R package v 2.14.0, then likelihoods were estimated using function getLikelihoods in baySeq. FDR values were calculated using the topCounts function in baySeq.

ALDEx2 [Fernandes et al. 2014]: Unclassified genera, and genera present in < 10% of samples were excluded. Raw genera counts were then used as input for the aldex function in ALDEx2 R package v 1.12.0 specifying 1000 Monte Carlo samples. Both Wilcoxon (ALDEx2 Wilcoxon) and t-test (ALDEx2 t-test) were used for testing differences in genera relative abundances between PD and controls. *P* values were corrected for multiple testing using BH FDR method implemented in the aldex function.

SAMseq [Li & Tibshirani 2013]: The SAM method for normalization of sequence counts was applied to genus counts using the samr.norm.data function in the samr R package v

3.0. Normalized values were rounded to the nearest integer. Unclassified genera, and genera present in < 10% of samples were excluded. Normalized genera counts were then used as input for the SAMseq function in the samr R package specifying “Two class unpaired” as the response type and the fdr.output as “1” in order to get full result list. FDR q-values were extracted from “siggenes.table” in the SAMseq output.

Linear discriminant analysis Effect Size (LEfSe) [Segata et al. 2011]: Genera counts were transformed to relative abundance, then unclassified genera, and genera present in < 10% of samples were excluded. Sample IDs, case/control class designations, and genera relative abundances were exported from R and used as input for LEfSe v 1.0.8.post1 (downloaded using LEfSe bioconda recipe https://bioconda.github.io/recipes/lefse/README.html). Only genus level taxonomy designations were included in the LEfSe input. The LEfSe input was formatted using the lefse-format_input.py script specifying the normalization value to be 1E6. LEfSe analysis was then ran on the formatted data using the run_lefse.py script with default parameters. Since LEfSe only outputs uncorrected *P* values for features that it finds significant, LEfSe analysis was ran again, but this time specifying parameters that would output all *P* values. The full range of LEfSe *P* values were multiple testing corrected using BH FDR method implemented in the p.adjust function from stats package. These corrected *P* values were substituted for the uncorrected *P* values outputted by the default LEfSe run.

After exclusion of unclassified genera and genera found in <10% of subjects, 109 and 163 genera remained for differential abundance testing in dataset 1 and 2 respectively, with 106 genera in common between both datasets. Unless otherwise

Table 1. Method characteristics.

[a] Thorsen et al. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome. 2016 Nov 25;4(1):62.

[b] Hawinkel et al. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief Bioinform. 2019 Jan 18;20(1):210-221.

[c] Weiss et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017 Mar 3;5(1):27.

[d] McMurdie PJ, Holmes S (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 10(4): e1003531.

*Results are for Wilcoxon rank sum test. Kruskal-Wallis rank sum test was not implemented in the comparison studies that measured FPR,[a] and FDR.[b,c]

**Results are for negative binomial GLM without zero-inflation component. Zero-inflation was not implemented in comparison study that measured FPR.[a]

GLM: generalized linear model; RLE: relative log expression; TMM: trimmed mean of M-values; CLR: centered log ratio; CPM: counts per million; TSS: total sum scaling, also referred to as relative abundance; CSS: cumulative sum scaling; Covar: can method handle covariates; Perm: does method use permutations or Monte Carlo simulations; FPR: False positive rate. Data for FPRs were derived from Figure 2 of ref [a]; FDR: False discovery rate. Data for FDRs were derived from Figure S16 of ref [b] and Additional file 7: Figure S6 of ref [c], and is the approximate range of the FDRs or average FDRs given in the figures. For both figures, focused on FDRs from the same normalization method as used in the present study when possible; AUC: area under the curve

Differential abundance method	Parametric or non-parametric	Pseudo-count added	Data transform	Normalization for total sequence depth	Accounts for compositio-nality	Predictor variable types	Covar	Perm	Error rates	AUCs
ALDEx2 t-test	parametric	Yes	CLR	CLR	Yes	Two-class	No	Yes	FPR=0 [a] FDR~0 [b]	0.55<AUC<0.65 [a] 0.52<AUC<0.8 [b]
ALDEx2 Wilcoxon	non-parametric	Yes	CLR & rank	CLR	Yes	Two-class	No	Yes	FPR=1E-3 [a]	0.55<AUC<0.63 [a]
ANCOM	non-parametric	Yes	Log ratio & rank	Log ratio	Yes	Two-class Multi-class Quantitative	Yes	No	FDR<0.05 [c]	-
Kruskal-Wallis rank-sum test	non-parametric	No	rank	None (used TSS in current study)	No	Two-class Multi-class	No	No	FPR=0.03 [a*] FDR≤0.15 [b*] FDR≤0.05 [c*]	0.55<AUC<0.65 [a*] 0.5<AUC<0.65 [b*]
metagenome-Seq fitFeature-Model	parametric	No	None, assumes zero-inflated log normal distribution	CSS	No	Two-class	No	No	FPR=0.02 [a] FDR≤0.05 [c]	0.57<AUC<0.7 [a]
Welch's t-test with log transform	parametric	Yes	Log	None (used TSS in current study)	No	Two-class	No	No	FPR=0.03 [a]	0.45<AUC<0.65 [a]
DESeq2 nbinomWald Test	parametric	No	None, assumes negative	Calculation of scaling factors	No	Two-class Multi-class Quantitative	Yes	No	FPR=0.04 [a] FDR≤0.1 [b] FDR<0.1 [c]	0.6<AUC<0.7 [a] 0.5<AUC<0.69 [b] 0.56<AUC<0.9 [d]

			binomial distribution	using median of ratios method						
limma-voom	parametric	Yes	log2 CPM	log2 CPM	No	Two-class Multi-class Quantitative	Yes	No	FDR<0.1 [b] FDR≤0.05 [c]	0.5<AUC<0.72 [b]
Negative binomial GLM with and without zero-inflation	parametric	No	None, assumes negative binomial distribution with or without zero-inflation	None (used log[total sequence count] as offset variable in current study)	No	Two-class Multi-class Quantitative	Yes	No	FPR=0.13 [a**]	0.58<AUC<0.74 [a]
baySeq	parametric	Yes	None, assumes negative binomial distribution	Total sequence count used as scaling factor	No	Two-class Multi-class	No	No	FPR=0.5 [a]	0.45<AUC<0.7 [a]
edgeR exactTest with RLE	parametric	Yes	None, assumes negative binomial distribution	Calculation of scaling factors using RLE method	No	Two-class	No	No	FDR≤0.5 [c]	0.56<AUC<0.99 [d]
edgeR exactTest with TMM	parametric	No	None, assumes negative binomial distribution	Calculation of scaling factors using TMM method	No	Two-class	No	No	FPR=0.31 [a] 0.6<FDR<0.9 [b]	0.6<AUC<0.78 [a] 0.55<AUC<0.8 [b]
metagenome-Seq zero-inflated Gaussian	parametric	No	None, assumes zero-inflated Gaussian distribution	CSS	No	Two-class Multi-class Quantitative	Yes	No	FPR=0.42 [a] 0.6<FDR<0.8 [b] FDR≤0.6 [c]	0.58<AUC<0.75 [a] 0.51<AUC<0.75 [b] 0.5<AUC<0.99 [d]
SAMseq	non-parametric	No	rank	Anscombe transformation then divide by square root of sequencing depth	No	Two-class Multi-class Quantitative	No	Yes	FDR≤0.9 [b]	-
GLM with CLR transform	parametric	Yes	CLR	CLR	Yes	Two-class Multi-class Quantitative	Yes	No	-	-
LEfSe	non-parametric	No	rank	None (used TSS in current study)	No	Two-class Multi-class	No	No	-	-

Table 2. List of chosen parameters for each function within each method that either differed from default, or did not have a set default. AIC: Akaike information criterion; FDR: False discovery rate; TMM: Trimmed mean of M-values; RLE: Relative log expression; LDA: Linear discriminant analysis

Method	Function	Role of function	Parameter	Parameter definition	Default or recommended option for parameter	Parameter choice for current study	Reasoning for parameter choice
KW	kruskal.test	Performs Kruskal-Wallis rank sum test.	no change of parameters from default				
log_t	t.test	Performs Welch's t-test.	no change of parameters from default				
GLM_NBZI	glmmTMB	Performs negative-binomial generalized linear model fitting with and without zero-inflation.	family	The underlying distribution of the data.	"gaussian"	"nbinom2"	Need to change distribution from gaussian to negative binomial distribution to run a negative binomial generalized linear model.
			ziformula	Formula for the zero-inflation model.	~0 (no zero-inflation)	~0 (no zero-inflation) & ~1 (zero-inflation for all observations)	Both zero and non-zero inflation was used and model fits compared using AIC, then model fit with lowest AIC was used.
GLM_CLR	glm	Performs generalized linear model fitting.	no change of parameters from default				
ANCOM	ANCOM.main	Performs full ANCOM workflow including data filtering, data transformation and normalization, then statistical testing.	multcorr	FDR multiple testing correction. Options are 1 (all tests: N taxa*[N taxa-1]), 2 (Taxa-wise: N taxa), or 3 (none).	2	2	Recommended option per ANCOMv2 documentation.

			sig	The level of significance. Used when calculating *W* statistics. Numeric value from 0-1.	None	0.05	Commonly used significance level cutoff and used in example of ANCOM in ANCOMv2 documentation
			prev.cut	Taxa with proportion of zeroes greater than prev.cut excluded. Numeric value from 0-1.	None	1	Causes all genera in the sample to be used for ANCOM analysis, ensuring the full sampling of the ecosystem is used for log ratio transformations and *W* statistic calculations. If this is lowered to 0.9 as shown in an example in ANCOMv2 documentation, results are drastically more conservative.
fitZIG	cumNorm	Performs cumulative sum scaling normalization.	no change of parameters from default				
	fitZig	Performs zero-inflated gaussian model fitting.	no change of parameters from default				
fitFeatMod	cumNorm	Performs cumulative sum scaling normalization.	no change of parameters from default				
	fitFeatureModel	Performs zero-inflated log-normal model fitting.	no change of parameters from default				
edgeR_TMM	calcNormFactors	Calculates normalization factors to scale the raw library sizes by.	no change of parameters from default				

	estimateCommonDisp	Estimates the common negative binomial dispersion value across all taxa, used in testing.	no change of parameters from default				
	estimateTagwiseDisp	Estimates the taxa specific negative binomial dispersion values, used in testing.	no change of parameters from default				
	exactTest	Peforms edgeR's exact test for differential abundance.	no change of parameters from default				
edgeR_RLE	calcNormFactors	Calculates normalization factors to scale the raw library sizes by.	method	The normalization method to be used.	"TMM"	"RLE"	In order to run edgeR exactTest with ratio log expression normalization, needed to change "TMM" to "RLE" to perform ratio log expression normalization.
	estimateCommonDisp	Estimates the common negative binomial dispersion value across all taxa, used in testing.	no change of parameters from default				
	estimateTagwiseDisp	Estimates the taxa specific negative binomial dispersion values, used in testing.	no change of parameters from default				
	exactTest	Peforms edgeR's exact test for differential abundance.	no change of parameters from default				

DESeq2	estimateSizeFactors	Calculates normalization factors to scale the raw library sizes by.	type	The method to be used for calculating normalization factors. Default is "ratio" (median ratio method introduced in original DESeq paper).	"ratio"	"poscounts"	Default method "ratio" cannot deal with zeros, which is rapant in metagenomic data, but "poscounts" has been designed to deal with zeros, evolving from cases where DESeq2 was being used to analyze metagenomic samples.
	DESeq	Performs full DESeq2 workflow including estimation of size factors (ignored here because it is done prior to running function), estimation of negative binomial dispersions, then performance of the negative binomial Wald test for differential abundance.	no change of parameters from default				
voom	calcNormFactors (edgeR package)	Calculates normalization factors to scale the raw library sizes by. Not part of limma package, but recommended to be done by limma-voom authors.	no change of parameters from default				

	voom (limma package)	Transform data to log2-counts per million and estimates mean-variance relationship to compute taxa weights, making data ready for linear models.	no change of parameters from default				
	lmFit (limma package)	Performs linear model fitting.	no change of parameters from default				
	eBayes (limma package)	Calculates statistics for assessing significance of fitted model.	no change of parameters from default				
baySeq	getPriors.NB	Estimates parameters for the underlying negative binomial distributions of taxa to be used in analysis.	cl	Object describing how to parallelize the analysis.	None	NULL	No parallelizing needed, function runs fast enough without it.
	getLikelihoods	Calculates likelihood of taxa belonging to a model of differential abundance.	cl	Object describing how to parallelize the analysis.	None	NULL	No parallelizing needed, function runs fast enough without it.
ALDEx2	aldex	Performs full ALDEx2 workflow including centered log-ratio transformation of data followed by statistical testing.	mc.samples	Number of Monte Carlo samples to use to estimate the underlying distributions.	128	1000	Per ALDEx2 documentation, 128 or more mc.samples for the t-test is recommended, 1000 mc.samples for a rigorous effect size calculation.

SAMseq	samr.norm.data	Normalizes data to look roughly like a Gaussian distribution with each sample having equal sequencing depth.	no change of parameters from default				
	SAMseq	Performs the statistical test for differential abundance.	resp.type	What type of data is the response variable. No default. Options are "Quantitative", "Two class unpaired", "Survival", "Multiclass", or "Two class paired"	None	"Two class unpaired"	The response variable (case vs control) is two classes and is not paired.
			fdr.output	False Discovery Rate cutoff for output in significant genes table.	0.2	1	Changed to 1 to ensure all results for all genera tested were outputted from analysis. Our cutoff for significance was FDR < 0.05.
LEfSe	run_lefse.py	Peforms statistical testing for differential abundance and calculates LDA effect sizes.	-a	Alpha value for the Anova test.	0.05	(a) 1 (b) 0.05	Performed two runs: Run (a) to capture results for all genera, and (b) to identify method designated significant genera.
			-w	Alpha value for the Wilcoxon test.	0.05	(a) 1 (b) 0.05	
			-l	Threshold for the absolute value of the logarithmic LDA score.	2	(a) 0 (b) 2	

mentioned above, significance was set at FDR < 0.05. PD – genus associations were considered replicated if they reached multiple testing corrected significance in both dataset 1 and 2.

Comparison of significant results for differential abundance methods

To measure similarity between method results, pairwise concordances were calculated between results for each method. For each dataset, a binary genus by method matrix was created with values denoting which methods did (1), or did not (0), detect a significant association between a genus and PD. Summing the two individual dataset matrices together resulted in a combined genus by method matrix with values denoting which methods detected a significant association between a genus and PD in no datasets (0), one dataset (1), or both datasets (2), effectively capturing information on replicated associations. Associations that had the same effect direction across datasets were then given a value of 3 to differentiate them from associations that were significant across datasets, but resulted in opposite effect directions (although significant in both datasets, not true replications since they have opposite direction of effects). Only tested genera that were in common between both datasets were included in the combined matrix (106 genera). For each genus by method matrix, a matrix of pairwise concordances between methods was calculated in R by summing the number of PD-genus associations that were called the same between two methods (i.e. both calling an association significant or not significant for individual datasets, or significant in either no datasets, one dataset, both datasets with opposite effect directions, or both datasets with same effect direction for combined results) and dividing by the total number of genera tested. Concordances were visualized as a heatmap using the heatmap.3 function (R code downloaded from https://raw.githubusercontent.com/obigriffith/biostar-tutorials/master/Heatmaps/heatmap.3.R on 10/21/2019).

To determine if any groups of PD-genus associations were being agreed (or disagreed) upon by all or a subset of methods, hierarchical clustering was performed to group genera based on similarities in association results between methods. A genus by method matrix of combined results was first created as stated above, then hierarchical clustering was performed and visualized in a heatmap using the heatmap.3 function with the default distance function, but specifying the hierarchical clustering function (hclustfun) to be diana from the cluster v 2.1.0 R package. DIANA performs a divisive hierarchical clustering algorithm [Kaufman & Rousseeuw 1990], which, in this situation, attempts to group genera based on the similarities in associations being detected in either no datasets, one dataset, both datasets with opposite effect directions, or both datasets with same effect direction between methods. The PD to control mean relative abundance ratios (MRAR) and control mean relative abundances (MRA) for each genus were also plotted next to the heatmap. MRARs were given a color gradient from red (lowest MRAR) to white (MRAR ~ 1) to blue (highest MRAR). Control MRAs were given a color gradient from white (lowest MRA) to dark green (highest MRA).

RESULTS

Method characteristics

Methods included in the present study span the fields of traditional statistics, RNA-Seq analysis, and microbiome analysis, and have varying underlying characteristics. A summary of method characteristics for the 16 differential abundance methods can be found in Table 1. The majority of methods included here utilized parametric statistical tests (assumes the data has some form of underlying distribution). Of these, the most commonly assumed data distribution was the negative binomial distribution (DESeq2, baySeq, edgeR RLE, edgeR TMM, GLM NBZI). No data transformations were performed for negative binomial methods, or metagenomeSeq methods, to try and bring the data to normality as non-normality of data is taken into account in their statistical models. The remaining parametric methods (ALDEx2 t-test, log t-test, limma-voom, GLM CLR) all used statistical tests that assumed a Gaussian distribution of the data, therefore, transformations were needed before analysis that included a log transform of some kind. Five methods (ALDEx2 Wilcoxon, ANCOM, Kruskal-Wallis, SAMseq, LEfSe) were considered non-parametric (assumes no underlying distribution of data) as they used statistical tests that transformed data to ranks. Methods also differed in what techniques were used to account for varying sequence depth between samples. Four of the five negative binomial methods (DESeq2, baySeq, edgeR RLE, edgeR TMM) calculated scaling factors for each sample to account for uneven sequence count. Cumulative sum scaling was used for both metagenomeSeq methods. Relative abundance transformations (also referred to as total sum scaling) were performed for three methods that did not have a built in technique to account for varying sequence depth (Kruskal-Wallis, log t-test, LEfSe) as it is a widely used normalization technique and recommended as the normalization technique of choice by LEfSe authors [Segata et al. 2011]. Log ratio transformations were used for ALDEx2 methods, GLM CLR, and ANCOM, which, in addition to normalizing to total sequence count, takes the compositionality of the data into account. The remaining 3 methods (limma voom, GLM NBZI, SAMseq) did not share normalization techniques with any other methods.

Pairwise concordances of significant PD-genus association calls between methods

The majority of the 16 methods being compared here have had their performances previously measured and compared on simulated microbiome data [McMurdie & Holmes 2014; Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019], therefore, it was of interest to see if similarities/dissimilarities in significant PD-genus association calls between methods obtained from analyses on our data reflected similarities/dissimilarities in previously reported performance metrics. On average, association calls between methods were moderately concordant, with pairwise concordances between method calls ranging from 0.46-0.99 with the mean concordance being 0.76 per dataset (Figure 1, Table 3). For both datasets, baySeq, GLM NBZI, fitZIG, limma voom, edgeR, and SAMseq made association calls that had the lowest concordances on average with other methods (mean concordance = 0.61-0.76). Kruskal-Wallis, log t-test, GLM CLR, fitFeatureModel, ALDEx2, DESeq2 and LEfSe made calls that had the highest concordances on average with other methods (mean concordance = 0.78-0.82). Interestingly, the lower concordant group of methods contained methods previously shown to have higher FPR and FDR, while the higher concordant group contained methods previously shown to have lower FPR and FDR (Table 1) [Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019]. When combining method calls for both datasets, effectively incorporating information on what significant association calls replicated in both datasets (see METHODS), the average overall concordance between

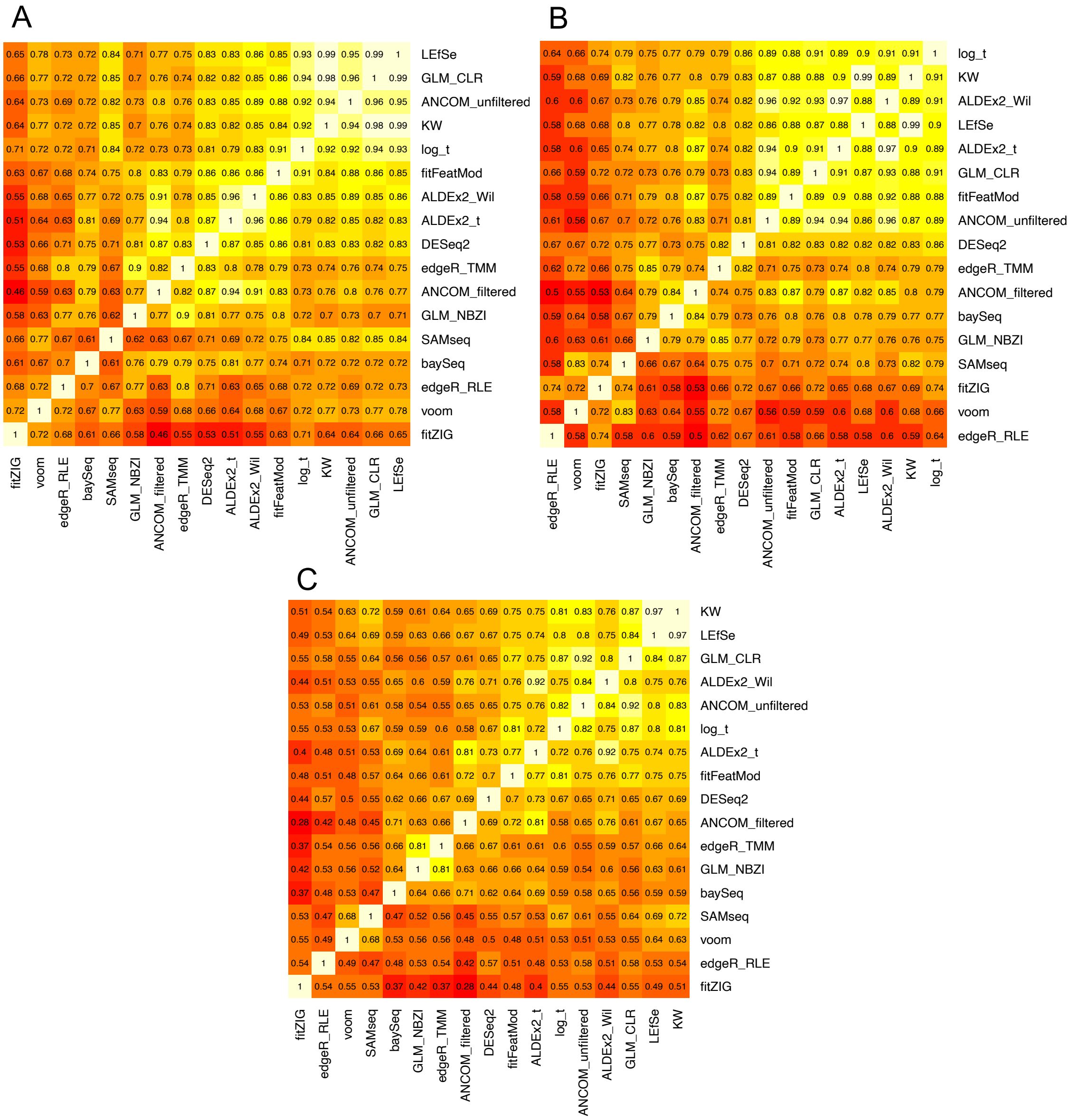


Figure 1: Pairwise concordances between method results.

Pairwise concordances were calculated between method results for (A) 109 genera in dataset 1, (B) 163 genera in dataset 2, and (C) combined results for 106 genera in common between both datasets. Values in heatmap cells correspond to the concordance between two methods. Concordances were calculated by summing the number of PD-genus associations that were called the same between two methods (i.e both calling an association significant or not significant for (A) and (B), or significant in either no datasets, one dataset, both datasets with opposite effect directions, or both datasets with same effect direction for (C)) and dividing by the total number of genera tested. Concordances for ANCOM_unfiltered were calculated after removing results for genera not tested by the other methods (those found in <10% of samples, and unclassified at genus level). Cells are colored by a red (lower concordance) to yellow (higher concordance) color gradient. Methods are ordered from lowest (bottom, left) to highest (top, right) mean concordance. KW: Kruskal-Wallis; GLM_CLR: generalized linear model with centered log ratio transformation; ALDEx2_Wil: ALDEx2 with Wilcoxon rank-sum test; log_t: Welch's t-test with log transformation; ALDEx2_t: ALDEx2 with t-test; fitFeatMod: fitFeatureModel from metagenomeSeq; edgeR_TMM: edgeR exact test with trimmed mean of M-values; GLM_NBZI: generalized linear model assuming negative binomial distribution with, or without, zero-inflation; edgeR_RLE: edgeR exact test with relative log expression

Table 3. Summary statistics for pairwise concordances between method results. Mean, minimum, and maximum concordance was calculated for each method in dataset 1, dataset 2, and when results across dataset 1 and 2 were combined. Methods are ordered by mean concordance for combined results (same order as rows in Figure 1, C). SD: Standard deviation

	Dataset 1				Dataset 2				Combined results			
Method	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
KW	0.82	0.10	0.64	0.99	0.82	0.10	0.59	0.99	0.71	0.12	0.51	0.97
LEfSe	0.82	0.10	0.65	0.99	0.81	0.10	0.58	0.99	0.70	0.12	0.49	0.97
GLM CLR	0.82	0.11	0.66	0.99	0.80	0.11	0.59	0.94	0.69	0.13	0.55	0.92
ANCOM unfiltered	0.82	0.10	0.64	0.96	0.80	0.12	0.56	0.96	0.68	0.13	0.51	0.92
ALDEx2 Wilcoxon	0.80	0.11	0.55	0.96	0.81	0.12	0.60	0.97	0.68	0.13	0.44	0.92
Log t-test	0.81	0.09	0.71	0.94	0.82	0.09	0.64	0.91	0.68	0.12	0.53	0.87
ALDEx2 t-test	0.79	0.12	0.51	0.96	0.81	0.12	0.58	0.97	0.68	0.14	0.40	0.92
fitFeature-Model	0.80	0.08	0.63	0.91	0.80	0.11	0.58	0.92	0.67	0.11	0.48	0.81
DESeq2	0.79	0.09	0.53	0.87	0.78	0.06	0.67	0.86	0.64	0.08	0.44	0.73
ANCOM filtered	0.75	0.12	0.46	0.94	0.75	0.12	0.50	0.87	0.61	0.14	0.28	0.81
edgeR TMM	0.76	0.08	0.55	0.90	0.75	0.06	0.62	0.85	0.60	0.09	0.37	0.81
GLM NBZI	0.73	0.08	0.58	0.90	0.73	0.07	0.60	0.85	0.60	0.08	0.42	0.81
baySeq	0.72	0.06	0.61	0.81	0.74	0.08	0.58	0.84	0.59	0.09	0.37	0.71
SAMseq	0.73	0.09	0.61	0.85	0.73	0.07	0.58	0.83	0.58	0.08	0.45	0.72
voom	0.70	0.06	0.59	0.78	0.64	0.07	0.55	0.83	0.55	0.06	0.48	0.68
edgeR RLE	0.70	0.05	0.63	0.80	0.61	0.05	0.50	0.74	0.52	0.04	0.42	0.58
fitZIG	0.61	0.07	0.46	0.72	0.67	0.06	0.53	0.74	0.47	0.08	0.28	0.55

method calls decreased (mean concordance = 0.63±0.1). As seen in individual datasets, baySeq, GLM NBZI, fitZIG, limma voom, edgeR, and SAMseq had the least similar calls on average with other methods (mean concordance = 0.47-0.60), and Kruskal-Wallis, log t-test, GLM CLR, fitFeatureModel, ALDEx2, DESeq2 and LEfSe had the most similar calls on average with other methods (mean concordance = 0.64-0.71). No significant difference between datasets was found for the average overall concordance between method calls (t-test *P* value=0.78), but the decrease in concordance between method calls when merging calls from both datasets was significant (t-test *P* value < 3E-6).

Variable effect of taxa filtering for ANCOM

When performing analyses with ANCOM on filtered genera data (referred to as ANCOM filtered here), we noticed that very few significant PD-genus associations were being detected, even in the larger, arguably more powered dataset 2. Association calls for ANCOM filtered also resulted in some of the lowest concordances observed among the methods, landing it in the lower concordant group of methods mentioned above for both datasets and when method calls of both datasets were combined (Figure 1). We found this odd as ANCOM was previously reported to have low FDR [Weiss et al. 2017], and shares similar characteristics with other low FPR/FDR methods included in this study (Table 1), therefore, we would expect it to be included with the other low FPR/FDR methods in the higher concordant group. We posited that this might be a result of taxa filtering before analysis. To investigate the effect of taxa filtering on ANCOM results, we performed ANCOM again using all genera in the analysis (referred to as ANCOM unfiltered here). In both datasets, comparing significant results between ANCOM filtered and ANCOM unfiltered showed a drastic reduction in significant associations when filtering genera before analysis (82% reduction in dataset 1 and 79% reduction in dataset 2, Table 4). As this was the opposite effect we expected to happen when filtering taxa

Table 4. Variable effect of taxa filtering for ANCOM in datasets 1 and 2. ANCOM was performed twice, once for genera that passed filtering criteria (present in at least 10% of samples and classified at genus level), and again for all genera to compare the effect of filtering vs not filtering on detection of significant signals. We observed the opposite effect from what we expected (should gain more significant signals when filtering, but actually produced less), therefore, we investigated effect of filtering on results of two other standard statistical methods (GLM CLR and Kruskal-Wallis) by performing each on filtered and unfiltered data. Column "Filter" contains significant results when implementing the 10% and unclassified genera filtering, while column "Unfilt" contains significant results when there was no filtering of genera. 201 PD patients and 132 controls were included in all analyses.
Hyphen (-); indicates that a genus did not result in a significant association with PD under that filtering condition.
NT; not tested due to removal of that genus during filtering.

ANCOM W			GLM CLR			Kruskal-Wallis		
	Significant *W* statistics			Significant FDR q-values			Significant FDR q-values	
Genera	Filter	Unfilt	Genera	Filter	Unfilt	Genera	Filter	Unfilt
Dataset 1								
Agathobacter	106	442	Lachnospiraceae_ ND3007_group	1E-04	4E-04	Lachnospiraceae_ ND3007_group	2E-04	1E-03
Roseburia	96	431	Lactobacillus	1E-04	4E-04	Lactobacillus	2E-04	1E-03
Lachnospira	92	429	Agathobacter	1E-04	5E-04	Agathobacter	2E-04	1E-03
Bifidobacterium	89	422	Blautia	1E-04	6E-04	Bifidobacterium	1E-03	5E-03
Lachnospiraceae_ ND3007_group	88	423	Lachnospira	2E-03	7E-03	Cloacibacillus	1E-03	6E-03
Blautia	-	425	Cloacibacillus	2E-03	8E-03	Faecalibacterium	1E-03	6E-03
Faecalibacterium	-	419	Bifidobacterium	2E-03	8E-03	Hungatella	1E-03	6E-03
Akkermansia	-	410	Hungatella	2E-03	8E-03	Lachnospira	1E-03	6E-03
Fusicatenibacter	-	408	Porphyromonas	2E-03	8E-03	Megasphaera	1E-03	6E-03
Lactobacillus	-	408	Roseburia	3E-03	0.01	Porphyromonas	1E-03	6E-03
Anaerostipes	-	404	Megasphaera	6E-03	0.03	Blautia	2E-03	9E-03
Butyricicoccus	-	401	Coprobacillus	7E-03	0.03	Coprobacillus	4E-03	0.01
Porphyromonas	-	399	Lachnospiraceae_ NK4B4_group	NT	0.03	Roseburia	4E-03	0.01
Prevotella	-	393	Prevotella	7E-03	0.03	Prevotella	6E-03	0.02
UBA1819	-	391	Faecalibacterium	7E-03	0.03	Akkermansia	7E-03	0.03
Lachnospiraceae_ UCG-004	-	390	Fusicatenibacter	7E-03	0.03	Butyricicoccus	7E-03	0.03
Hungatella	-	387	Butyricicoccus	9E-03	0.03	UBA1819	8E-03	0.03
Oscillospira	-	382	Anaeroplasma	NT	0.04	Ruminococcaceae _unclass	NT	0.04
Coprococcus_3	-	376	Lachnospiraceae_ UCG-004	0.01	0.04	Mobiluncus	NT	0.04

Cloacibacillus	-	371	Tannerellaceae_unclass	NT	0.04	Varibaculum	0.01	0.04
Ruminococcaceae_unclass	-	369	Haemophilus	0.01	0.05	Corynebacterium_1	0.01	0.04
Ezakiella	-	366	UBA1819	0.01	0.05	Ruminococcaceae_UCG-004	0.01	-
Desulfovibrio	-	362	Akkermansia	0.01	-	Fusicatenibacter	0.02	-
Corynebacterium_1	-	361	Anaerostipes	0.02	-	Lachnospiraceae_UCG-004	0.02	-
Haemophilus	-	360	Anaerotruncus	0.04	-	Oscillospira	0.02	-
Megasphaera	-	360	Bilophila	0.04	-	Anaerostipes	0.02	-
Ruminococcaceae_UCG-004	-	360	Coprococcus_3	0.03	-	Desulfovibrio	0.02	-
Coprobacillus	-	358	Corynebacterium_1	0.02	-	Anaerotruncus	0.03	-
			Desulfovibrio	0.04	-	Methanobrevibacter	0.03	-
			Ezakiella	0.02	-	Ezakiella	0.03	-
			Methanobrevibacter	0.02	-	Haemophilus	0.03	-
			Oscillospira	0.01	-	Bacteroides	0.04	-
			Ruminococcaceae_UCG-004	0.02	-	Bilophila	0.05	-
			Varibaculum	0.04	-			
Dataset 2								
Bifidobacterium	153	551	Lachnospiraceae_UCG-004	2E-06	8E-06	Bifidobacterium	6E-07	2E-06
Agathobacter	150	548	Blautia	2E-06	8E-06	Lachnospiraceae_UCG-004	1E-05	5E-05
Roseburia	145	543	Ruminococcaceae_UCG-013	2E-06	8E-06	Agathobacter	6E-05	2E-04
Faecalibacterium	143	541	Anaerostipes	4E-06	1E-05	Roseburia	3E-04	9E-04
Lachnospiraceae_UCG-004	143	541	Agathobacter	6E-06	2E-05	Eubacterium	3E-04	9E-04
Lachnospiraceae_ND3007_group	139	537	Roseburia	2E-05	6E-05	Lachnospiraceae_ND3007_group	6E-04	2E-03
Anaerostipes	137	536	Lachnospiraceae_ND3007_group	7E-05	2E-04	Ruminococcaceae_UCG-013	7E-04	2E-03
Lachnospira	136	535	Bifidobacterium	9E-05	3E-04	Anaerostipes	1E-03	5E-03
Ruminococcaceae_UCG-013	-	539	Eubacterium	2E-04	7E-04	Lawsonella	3E-03	0.01
Fusicatenibacter	-	530	Butyricicoccus	5E-04	2E-03	Faecalibacterium	3E-03	0.01
Ruminococcus_2	-	523	Lawsonella	2E-03	5E-03	Erysipelotrichaceae_unclass	NT	0.01
Blautia	-	522	Faecalibacterium	2E-03	5E-03	Turicibacter	6E-03	0.02
Eubacterium	-	521	Oscillospira	2E-03	5E-03	Pseudomonas	8E-03	0.02
Pseudomonas	-	520	Lachnospiraceae_unclass	NT	6E-03	UBA1819	8E-03	0.02
Oscillospira	-	513	Lachnospiraceae_UCG-001	3E-03	0.01	Corynebacterium_1	8E-03	0.02
Desulfovibrio	-	510	Lachnospira	4E-03	0.01	Erysipelotrichaceae_UCG-003	9E-03	0.03
Ruminococcus_1	-	505	Erysipelotrichaceae_UCG-003	4E-03	0.01	Anaerococcus	0.01	0.03

Butyricicoccus	-	504	Fusicatenibacter	5E-03	0.02	Lachnospiraceae_UCG-001	0.01	0.03
Methanobrevibacter	-	503	DTU089	5E-03	0.02	Desulfovibrio	0.01	0.04
Lachnospiraceae_UCG-001	-	503	Ruminococcus_2	7E-03	0.02	Chloroplast_unclass	NT	0.04
Turicibacter	-	501	Lachnoclostridium	8E-03	0.03	Lactobacillus	0.01	0.04
Corynebacterium_1	-	500	Turicibacter	8E-03	0.03	Lachnospira	0.01	0.04
Ruminiclostridium_6	-	495	Erysipelotrichaceae_unclass	NT	0.03	Oscillospira	0.01	0.04
DTU089	-	493	Chloroplast_unclass	NT	0.03	Varibaculum	0.01	0.04
Lawsonella	-	493	Acinetobacter	NT	0.04	Peptoniphilus	0.01	0.04
Anaerococcus	-	480	Candidatus_Soleaferrea	0.01	0.04	Hungatella	0.02	0.05
Lactobacillus	-	479	Ruminococcus_1	0.01	0.04	Acinetobacter	NT	0.05
Prevotella	-	477	Prevotella	0.01	0.04	Methanobrevibacter	0.02	-
Erysipelotrichaceae_unclass	NT	474	Parvimonas	0.01	0.04	Delftia	0.02	-
Erysipelotrichaceae_UCG-003	-	473	Desulfovibrio	0.02	0.05	Streptococcus	0.02	-
Porphyromonas	-	472	Corynebacterium_1	0.02	0.05	Porphyromonas	0.02	-
Lachnospiraceae_unclass	NT	469	Delftia	0.02	-	Prevotella	0.02	-
Ruminococcaceae_UCG-014	-	466	Porphyromonas	0.02	-	Fusicatenibacter	0.03	-
Veillonella	-	463	Anaerococcus	0.02	-	Anaerotruncus	0.03	-
Candidatus_Soleaferrea	-	460	Ruminiclostridium_6	0.02	-	Ruminococcus_2	0.03	-
Bacteroides	-	459	Methanobrevibacter	0.02	-	Parvimonas	0.03	-
Varibaculum	-	452	Pseudomonas	0.02	-	Mobiluncus	0.04	-
Phascolarctobacterium	-	449	Bacteroides	0.02	-	Actinomyces	0.04	-
			Veillonella	0.03	-	Finegoldia	0.04	-
			S5-A14a	0.03	-	S5-A14a	0.04	-
			Varibaculum	0.03	-	Blautia	0.04	-
			Victivallis	0.03	-	Murdochiella	0.04	-
			Peptoniphilus	0.04	-	Ezakiella	0.04	-
			Lactobacillus	0.04	-	DTU089	0.04	-
			Phocea	0.04	-			
			Ezakiella	0.04	-			
			Ruminococcaceae_UCG-014	0.04	-			
			Family_XIII_UCG-001	0.04	-			

data before analysis (expected to gain more significant signals when filtering), we investigated effect of filtering on results of two other standard statistical methods (GLM CLR and Kruskal-Wallis) by performing each on filtered and unfiltered data. In both datasets, results for these methods behaved as expected, producing more significant associations when data was filtered before analysis compared to when all genera were included in analysis (Table 4). The change in significant associations between analyses on filtered and unfiltered data were much less severe than ANCOM's with GLM CLR and Kruskal-Wallis gaining ~29% and ~33% more significant associations with filtered data in both datasets respectively. Concordances between association calls made by ANCOM unfiltered and other methods were also higher on average than those seen with ANCOM filtered, placing ANCOM unfiltered among the group of lower FPR/FDR methods with higher mean concordances (Figure 1, Table 3). Because ANCOM unfiltered behaves more like what we observed from other lower FPR/FDR methods, and has higher concordance to other methods on average, we used results from ANCOM unfiltered in place of ANCOM filtered for further comparisons between method results.

Hierarchical clustering of genera based on similarity in significant association calls between methods

To observe what groups of genera either all, or subsets, of methods were agreeing upon, we performed hierarchical clustering of genera based on similarities in significant PD-genus association calls between methods and visualized associations (both unreplicated and replicated) for all methods via heatmap (Figure 2).

The degree at which methods agreed upon PD-genus associations depended on the level, and tactic, used to observe the convergence of method results. At the most conservative level, only a minority of PD-genus associations were agreed upon by all 16 methods (3-4% of genera). The genera that were associated with PD by all 16 methods included *Agathobacter, Lachnospiraceae_ND3007_group,* and *Lactobacillus* in dataset 1 and *Agathobacter, Anaerococcus, Bifidobacterium, Lachnospiraceae_UCG-004, Porphyromonas, Prevotella,* and *Roseburia* in dataset 2. Only *Agathobacter*'s association was detected by all 16 methods in both datasets. At a more liberal threshold, 26 and 36 genera were associated with PD by >50% of methods in dataset 1 and dataset 2 respectively (22-24% of total genera). Of these genera, 17 were associated with PD by >50% of methods in both datasets and included *Agathobacter, Lachnospiraceae_ND3007_group*, *Lactobacillus*, *Bifidobacterium*, *Lachnospiraceae_UCG-004*, *Porphyromonas*, *Prevotella*, *Roseburia*, *Fusictenibacter, Lachnospira, Faecalibacterium, Butyricicoccus, Anaerostipes, Methanobrevibacter, Blautia, Ezakiella,* and *Hungatella.* At the most liberal threshold, 83 and 131 genera were associated with PD by at least one method in dataset 1 and dataset 2 respectively (76-80% of total genera). Of these genera, 49 were associated with PD by at least one method in both datasets, which is approximately half of the genera that were in common between datasets and eligible for replication.

Using hierarchical clustering, genera were clustered into three clear groups: (1) 24 genera (23% of genera in common between datasets) that were more likely to be associated with PD across methods in both datasets (10±3 methods on average; Figure 2, group 1), (2) 67 genera (63% of genera in common between datasets) that were associated with PD in both datasets by little to no methods (<1 methods on average; Figure 2, group 2), and (3) 15 genera (14% of genera in common between datasets),

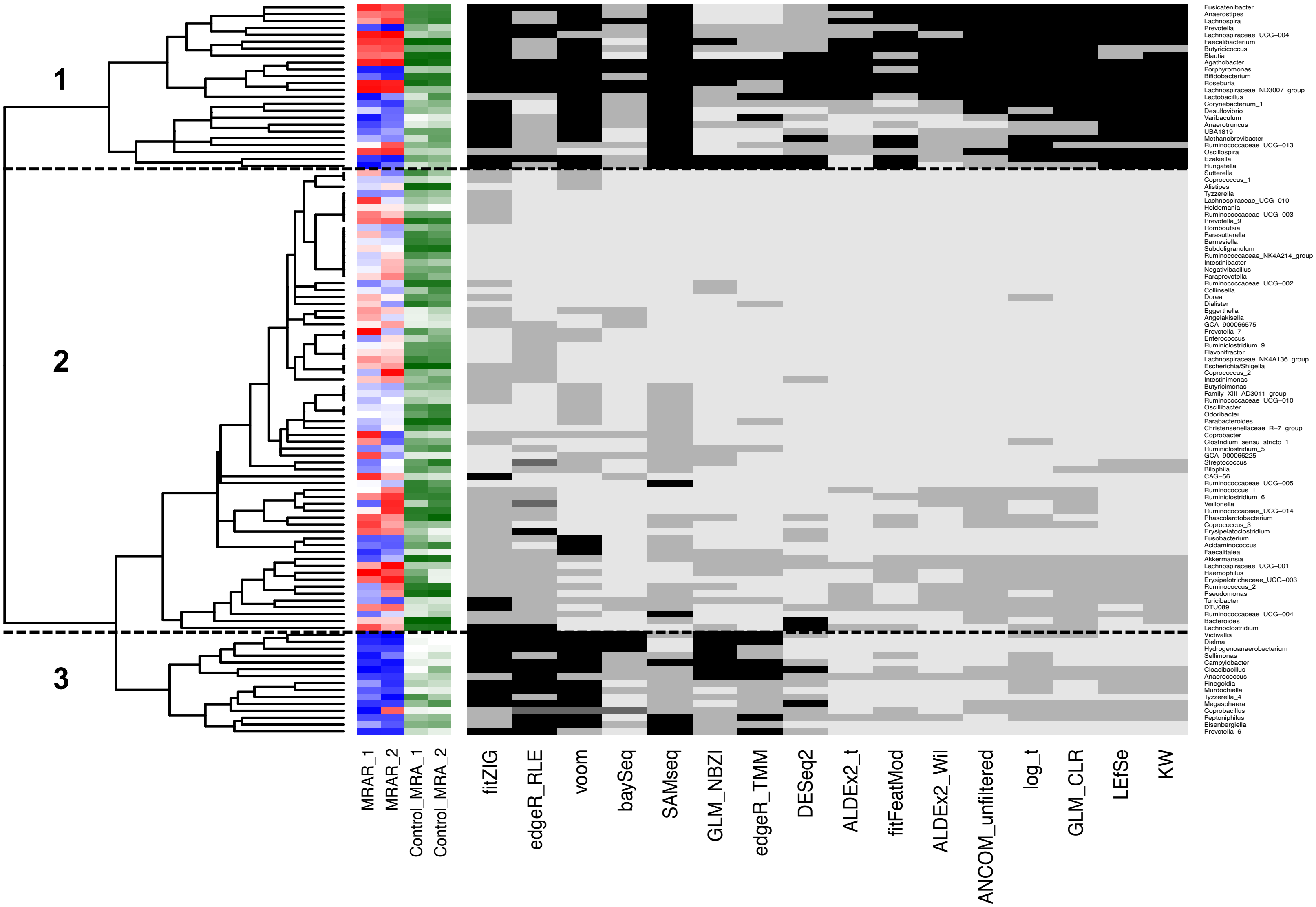

1
2
3
Fusicatenibacter
Anaerostipes
Lachnospira
Prevotella
Lachnospiraceae_UCG−004
Faecalibacterium
Butyricicoccus
Blautia
Agathobacter
Porphyromonas
Bifidobacterium
Roseburia
Lachnospiraceae_ND3007_group
Lactobacillus
Corynebacterium_1
Desulfovibrio
Varibaculum
Anaerotruncus
UBA1819
Methanobrevibacter
Ruminococcaceae_UCG−013
Oscillospira
Ezakiella
Hungatella
Sutterella
Coprococcus_1
Alistipes
Tyzzerella
Lachnospiraceae_UCG−010
Holdemania
Ruminococcaceae_UCG−003
Prevotella_9
Romboutsia
Parasutterella
Barnesiella
Subdoligranulum
Ruminococcaceae_NK4A214_group
Intestinibacter
Negativibacillus
Paraprevotella
Ruminococcaceae_UCG−002
Collinsella
Dorea
Dialister
Eggerthella
Angelakisella
GCA−900066575
Prevotella_7
Enterococcus
Ruminiclostridium_9
Flavonifractor
Lachnospiraceae_NK4A136_group
Escherichia/Shigella
Coprococcus_2
Intestinimonas
Butyricimonas
Family_XIII_AD3011_group
Ruminococcaceae_UCG−010
Oscillibacter
Odoribacter
Parabacteroides
Christensenellaceae_R−7_group
Coprobacter
Clostridium_sensu_stricto_1
Ruminiclostridium_5
GCA−900066225
Streptococcus
Bilophila
CAG−56
Ruminococcaceae_UCG−005
Ruminococcus_1
Ruminiclostridium_6
Veillonella
Ruminococcaceae_UCG−014
Phascolarctobacterium
Coprococcus_3
Erysipelatoclostridium
Fusobacterium
Acidaminococcus
Faecalitalea
Akkermansia
Lachnospiraceae_UCG−001
Haemophilus
Erysipelotrichaceae_UCG−003
Ruminococcus_2
Pseudomonas
Turicibacter
DTU089
Ruminococcaceae_UCG−004
Bacteroides
Lachnoclostridium
Victivallis
Dielma
Hydrogenoanaerobacterium
Sellimonas
Campylobacter
Cloacibacillus
Anaerococcus
Finegoldia
Murdochiella
Tyzzerella_4
Megasphaera
Coprobacillus
Peptoniphilus
Eisenbergiella
Prevotella_6
MRAR_1
MRAR_2
Control_MRA_1
Control_MRA_2
fitZIG
edgeR_RLE
voom
baySeq
SAMseq
GLM_NBZI
edgeR_TMM
DESeq2
ALDEx2_t
fitFeatMod
ALDEx2_Wil
ANCOM_unfiltered
log_t
GLM_CLR
LEfSe
KW

Figure 2: Hierarchical clustering of genera based on similarity in association results between methods.

Hierarchical clustering was performed to group genera (rows) based on similarities in association results between methods (columns) and was visualized via heatmap. Three obvious groups of genera were revealed by hierarchical clustering: (1) genera more likely to be replicated by majority of methods, (2) genera who were replicated by little to no methods, and (3) rarer genera enriched in PD who were replicated by a subset of methods, some of which were previously shown to have high sensitivity. Only results for genera tested and in common between datasets (106 genera) were included in hierarchical clustering and heatmap. Cells correspond to an association that was detected in no datasets (value=0, color=light gray), one dataset (value=1, color=gray), both datasets with opposite effect directions (value=2, color=dark grey), or both datasets with same effect directions (value=3, color=black). Mean relative abundance ratios for genera in dataset 1 (MRAR_1) and dataset 2 (MRAR_2) were plotted next to the heatmap, and given a color gradient from red (lowest MRAR) to white (MRAR ~ 1) to blue (highest MRAR). Control mean relative abundances for dataset 1 (Control_MRA_1) and dataset 2 (Control_MRA_2) were also plotted next to the heatmap, and given a color gradient from white (lowest MRA) to dark green (highest MRA). KW: Kruskal-Wallis; GLM_CLR: generalized linear model with centered log ratio transformation; ALDEx2_Wil: ALDEx2 with Wilcoxon rank-sum test; log_t: Welch’s t-test with log transformation; ALDEx2_t: ALDEx2 with t-test; fitFeatMod: fitFeatureModel from metagenomeSeq; edgeR_TMM: edgeR exact test with trimmed mean of M-values; GLM_NBZI: generalized linear model assuming negative binomial distribution with, or without, zero-inflation; edgeR_RLE: edgeR exact test with relative log expression mostly enriched in PD, who were more likely to be associated with PD by a specific subset of methods (4±2 methods on average; Figure 2, group 3). Group 1 included genera both enriched and depleted in PD that had a wide range of control MRAs and effect sizes ranging from highly prevalent genera with moderate effect sizes (e.g. *Agathobacter,* control MRA = 0.02-0.04, absolute fold change = 1.8 – 1.9) to rarer genera with larger effect sizes (e.g. *Prevotella*, control MRA = 6E-4 – 2E-3, absolute fold change = 2.6 – 4.4) with a mean control MRA of 6E-3 – 7E-3 and absolute fold change of 1.9 – 2.6. On average, group 2 included more prevalent genera (mean control MRA = 0.01) with smaller effect sizes (mean absolute fold change = 1.3 – 1.5). Group 3 was interesting as it contained genera mostly enriched in PD (mean absolute fold change for group = 3.2 – 4.5) and were made up of genera with very low control MRAs (mean control MRA = 6E-4 – 7E-4). The majority of PD-genus associations in group 3 were replicated by a subset of methods that have been previously shown to have high FPR/FDR, but also higher sensitivity [McMurdie & Holmes et al 2014; Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2017].

Comparison of method results across datasets

Datasets differed in size and had significant heterogeneity in microbiome composition [Wallen et al. 2020], therefore, we tested differences between method calls across datasets to see if there were any significant dataset differences in the number or proportion of genera being associated with PD by the 16 methods. For both datasets, approximately 80% of genera within each dataset were significantly associated with PD by at least one method (mean method per association = 6±5 dataset 1, 6±5 dataset 2).

Approximately 22-24% of genera were associated with PD by >50% of the methods (mean method per association = 12±2 dataset 1, 13±2 dataset 2), and 3-4% were associated with PD by all 16 methods. No significant difference was found between datasets for the average number of methods that detected a PD-genus association (t-test *P* value=0.67). For dataset 1, the maximum number of PD-genus associations detected was 64 (fitZIG, encompassing 59% of genera), the minimum detected was 11 (ALDEx2 t-test, encompassing 11% of genera), and the mean per method was 31±14 (encompassing on average 28% of genera). For dataset 2, the maximum number of PD-genus associations detected was 90 (edgeR RLE, encompassing 55% of genera), the minimum detected was 30 (ALDEx2 t-test, encompassing 18% of genera), and the mean per method was 50±21 (encompassing on average 31% of genera). Methods on average detected a significantly higher number of PD-genus associations in the larger dataset 2 (t-test *P* value=0.007), but no significant difference was found between datasets when the association count for each method was normalized by the number of genera tested in analysis (t-test *P* value=0.72). Overall, despite differences in the size of datasets and heterogeneity in microbiome composition, we observed no significant differences between datasets in the proportion of genera being associated with PD on average. The number of genera being associated with PD was significantly increased in dataset 2, which is to be expected since it is the larger, and potentially more powered dataset.

DISCUSSION

In summary, we found 16 differential abundance testing methods in the literature and used them to detect differentially abundant genera in PD patients versus healthy controls in two large PD-gut microbiome datasets. Methods spanned multiple fields and had both common and unique characteristics when compared to other methods. Significant PD-genus associations were detected by all methods and the number of associations detected by each method ranged from a small subset of genera to over half of the genera tested. Concordances between significant method calls varied overall. Methods previously shown to have lower FPR/FDR consistently resulted in higher overall concordances with other methods, while methods previously shown to have higher FPR/FDR consistently resulted in lower overall concordances with other methods. For one method (ANCOM), we detected an unorthodox effect of taxa filtering, where filtering of genera before analysis drastically reduced the number of significant PD-genus associations compared to when all genera were included in the analysis. When grouping genera based on similarities in significant PD-genus association calls between methods, we found that three clear groups of genera were formed: (1) those more likely to be replicated in both datasets by the majority of methods, (2) those associated with PD in both datasets by little to no methods, and (3) those more likely to be associated with PD by a specific subset of methods with potentially higher sensitivity. Although datasets were heterogeneous in microbiome composition, we observed no significant differences between datasets in the proportion of genera being associated with PD on average.

The variation between method results reported here aligns with the variation between differential abundance testing method performances previously reported in method comparison studies [Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019]. Although performance of methods could not be assessed here, as analyses were conducted on real datasets where the true answers are unknown, we observed that methods with similar previously reported performance metrics seemed to group together based on their concordances with one another and the PD-genus associations they detected. Methods with previously reported low FPR/FDR (i.e. group of methods containing Kruskal-Wallis, log t-test, ALDEx2, fitFeatureModel, ANCOM, DESeq2) not only had the highest average concordances across methods, but also had the highest concordances among each other (mean concordance ~ 0.9 per dataset). This observation makes logical sense for two reasons: (1) methods with lower FPR/FDR have also been previously reported to be conservative in their performance [Thorsen et al. 2016], detecting less taxa as differentially abundant compared to higher FPR/FDR methods (which we also observed in our data), and (2) they seemed to detect and replicate more robust PD-genus associations (e.g. genera shown in Figure 2, group 1 that were more likely to be agreed upon by the majority of methods). Taken together, we can extrapolate that methods with lower FPR/FDR would be more likely to converge on the same taxa because they are detecting less taxa as differentially abundant overall, and the signatures they do detect are those that tend to be more robust to methodological variation.

A surprising finding from this study was the variable effect of taxa filtering prior to testing with ANCOM. Usually, filtering of taxa before analysis would increase the number of significant associations detected by a method, mostly due to the decreased burden of multiple testing correction at the FDR calculation step. However with ANCOM, filtering of taxa before analysis greatly decreased the number of significant associations. This might be due to the different statistics used by ANCOM compared to the standard FDR q-value. To determine significance, ANCOM calculates a *W* statistic, which is the number of times the log ratio of a taxon with every other taxon being tested was detected to be significantly different across groups (in this case PD vs control) [Mandal et al. 2015]. Because *W* statistics are based on pairwise comparisons between all taxa being tested, they will automatically decrease overall if less taxa are included in the analysis, and the threshold range for significant *W* statistics will also decrease. On top of that, if low prevalent taxa are being removed, this will not only decrease the *W* statistics overall, but now *W* statistic calculation might become more conservative since higher prevalent, potentially more stable taxa have been selected for, the ratios of which might not differ enough to be detected as significant at a particular *W* statistic threshold.

A finding from this study that not only helped illustrate the behavior of the methods on our data, but is relevant to PD itself was the detection of two groups of genera that were converged upon by either the majority or subset of methods used here. Hierarchical clustering of genera based on similarity in method results showed one group of genera that were more likely to be replicated by the majority of methods on average (Figure 2, group 1). Theoretically, this group might be looked at as the "high confidence" group, as methods from across the spectrum tended to not only detect, but replicate the associations in both datasets. This group included genera previously associated with PD such as *Bifidobacterium*, *Lactobacillus*, and short-chain fatty-acid producing bacteria *Faecalibacterium*, *Roseburia*, *Blautia*, and other members of the *Lachnospiraceae* family. Hierarchical clustering also revealed a second group of genera that were only detected and replicated by a subset of methods (fitZIG, edgeR, limma-voom, baySeq, SAMseq, GLM NBZI, DESeq2; Figure 2, group 3). This group is intriguing because it mostly contains genera enriched in PD that have low control MRAs, higher effect sizes on average, and was more likely to be replicated by methods previously reported to have higher sensitivity [McMurdie & Holmes 2014; Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019]. Without the use of more sensitive methods, this group of genera would have been missed, arguing that, although these methods were previously reported to have higher FPR/FDR, they might be useful in detecting rarer taxa. As done here, the use of a second replication dataset could help curve the number of false positives these methods might detect.

One of the biggest limitations of this study is the lack of ability to test the actual performance of these methods, as no simulations were performed and only real data was used, so the true answers are unknown. As a proxy, we attempted to match patterns seen in our data to performance metrics previously reported in the literature [McMurdie & Holmes 2014; Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019]. Unfortunately, not all methods implemented in this study had previously reported performance metrics (i.e. LEfSe, GLM CLR), therefore, we can only superimpose previously reported performance metrics from methods they were most concordant with. A methodological limitation of this study is the choice of parameters used for each method. Each method contains multiple functions with multiple parameters, and it was beyond the scope of this study to try different combinations of parameters to fully optimize each method. We attempted to make parameter choices based on what was default for the method and/or what was recommended for the method especially in the context of microbiome data analysis, but this process is inherently biased as we did not attempt to try every combination of parameter choice possible for each method. Also, the genera detected in this study provide interesting leads for future studies in PD, but should be interpreted with caution as the goal of this study was mainly to compare different differential abundance methods in our data, and did not take into account any confounding variables that might drive false signals.

In conclusion, we performed 16 differential abundance testing methods in two large PD-gut microbiome datasets and compared their results. We found results varied between methods, but methods previously reported to have lower FPR/FDR tended to have higher overall concordance. Filtering of taxa before analysis with ANCOM drastically reduces the number of significant associations detected, most likely due to the way *W* statistics are calculated and used for significance. This suggests it might be more advantageous to supply ANCOM with unfiltered taxa abundances for higher taxonomic levels such as genus, and only filter out very rare taxa (that is most likely just noise) for lower taxonomic levels such as OTUs/ASVs. The majority of methods converged on a group of PD-genus associations that seemed more robust to inter-methodological differences, while a subset of higher sensitivity methods converged on a smaller, second group of rarer genera enriched in PD. This study fills a void in the literature on how different differential abundance methods behave when performed on real, complex disease oriented gut microbiome datasets, and we hope that it helps to inform future studies looking to perform these types of analyses.

DATA AVAILABILITY

Individual-level raw sequences and basic metadata are publicly available at NCBI Sequence Read Archive (SRA) BioProject ID PRJNA601994.

CODE AVAILABILITY

No custom codes were used. All software and packages, their versions, relevant specification and parameters are stated in the "METHODS" section.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of Neurological Disorders and Stroke grant R01 NS036960 (to H.P.), The US Army Medical Research Materiel Command endorsed by the US Army through the Parkinson's Research Program Investigator-Initiated Research Award under Award number W81XWH1810508 (to H.P.), and NIH Training Grant T32NS095775 (to Z.D.W.). Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the US Army or the NIH.

COMPETING INTERESTS

The authors declare no competing interests.

AUTHORS CONTRIBUTIONS

Z.D.W and H.P. both contributed to the conception and design of the work, and analysis and interpretation of the data. Z.D.W drafted the work.

REFERENCES

Schmidt, T. S. B., Raes, J. & Bork, P. The human gut microbiome: from association to modulation. Cell 172, 1198–1215 (2018).

McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* **10**, e1003531, doi:10.1371/journal.pcbi.1003531 (2014).

Thorsen et al. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome. 2016 Nov 25;4(1):62.

Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27, doi:10.1186/s40168-017-0237-y (2017).

Hawinkel et al. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief Bioinform. 2019 Jan 18;20(1):210-221.

Hill-Burns, E. M. *et al.* Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Mov Disord* **32**, 739-749, doi:10.1002/mds.26942 (2017).

Wallen, Z.D., Appah, M., Dean, M.N. et al. Characterizing dysbiosis of gut microbiome in PD: evidence for overabundance of opportunistic pathogens. npj Parkinsons Dis. 6, 11 (2020). https://doi.org/10.1038/s41531-020-0112-6.

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **17**, 10-12 (2011).

Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581-583, doi:10.1038/nmeth.3869 (2016).

Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**, 5261-5267 (2007).

Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (2011).

McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217, doi:10.1371/journal.pone.0061217 (2013).

Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association. Vol 47, No 260 (Dec., 1952), pp. 583-621.

Welch, BL. THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARIANCES ARE INVOLVED. Biometrika, Volume 34, Issue 1-2, January 1947, Pages 28–35.

Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* **26**, 27663, doi:10.3402/mehd.v26.27663 (2015).

Paulson JN, et al. Robust methods for differential abundance analysis in marker gene surveys. Nat Methods. 2013 Dec; 10(12): 1200–1202.

Paulson JN, et al. (2013). metagenomeSeq: Statistical analysis for sparse high-throughput sequncing.. Bioconductor package, http://www.cbcb.umd.edu/software/metagenomeSeq.

Robinson MD and Smyth GK (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics, 9, 321-332.

Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12): 550.

Ritchie, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 2015, Vol. 43, No. 7 e47.

Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:422.

Fernandes AD, et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014 May 5;2:15.

Li J and Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013 Oct; 22(5): 519-536.

Segata N et al. Metagenomic biomarker discovery and explanation. Genom Biol. 2011. 12(6): R60.

Kaufman L and Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. 8 March 1990. John Wiley & Sons, Inc.

CHARACTERIZING DYSBIOSIS OF GUT MICROBIOME IN PD: EVIDENCE FOR OVERABUNDANCE OF OPPORTUNISTIC PATHOGENS

by

ZACHARY D. WALLEN, MARY APPAH, MARISSA N. DEAN, CHERYL L. SESLER, STEWART A. FACTOR, ERIC MOLHO, CYRUS P. ZABETIAN, DAVID G. STANDAERT AND HAYDEH PAYAMI

npj Parkinsons Dis. 6, 11 (2020). doi: 10.1038/s41531-020-0112-6
PMID: 32566740 PMCID: PMC7293233

Format adapted for dissertation

ABSTRACT

In Parkinson's disease (PD), gastrointestinal features are common and often precede the motor signs. Braak and colleagues proposed that PD may start in the gut, triggered by a pathogen, and spread to the brain. Numerous studies have examined the gut microbiome in PD; all found it to be altered, but found inconsistent results on associated microorganisms. Studies to date have been small (N = 20 to 306) and are difficult to compare or combine due to varied methodology. We conducted a microbiome-wide association study (MWAS) with two large datasets for internal replication (N = 333 and 507). We used uniform methodology when possible, interrogated confounders, and applied two statistical tests for concordance, followed by correlation network analysis to infer interactions. Fifteen genera were associated with PD at a microbiome-wide significance level, in both datasets, with both methods, with or without covariate adjustment. The associations were not independent, rather they represented three clusters of co-occurring microorganisms. Cluster 1 was composed of opportunistic pathogens and all were elevated in PD. Cluster 2 was short-chain fatty acid (SCFA)-producing bacteria and all were reduced in PD. Cluster 3 was carbohydrate-metabolizing probiotics and were elevated in PD. Depletion of anti-inflammatory SCFA-producing bacteria and elevated levels of probiotics are confirmatory. Overabundance of opportunistic pathogens is an original finding and their identity provides a lead to experimentally test their role in PD.

INTRODUCTION

Parkinson's disease (PD) is a common, progressive, and debilitating disease, which currently cannot be prevented or cured. With the exception of rare genetic forms, the cause of PD is unknown. Many susceptibility loci[1] and environmental risk factors[2] have been identified, but each has a modest effect on risk and none is sufficient to cause disease. Gene–environment interaction studies have not been able to identify a causative combination[3,4,5,6]. The triggers that cause PD are unknown.

The emerging information about the importance of the gut microbiome in human health and disease[7], together with the well-established connection between PD and the gut including common and early occurrence of constipation[8], inflammation[9], and increased gut membrane permeability[10], have raised the possibility that microorganisms in the gut may play a role in PD pathogenesis and prompted a fast growing literature on studies conducted in humans and animal models[11-30]. Every study that has compared the global composition of the gut microbiome in PD vs. controls found it to be significantly altered; in contrast, attempts to identify PD-associated microorganisms have produced inconsistent results[31,32]. Low reproducibility has been attributed to small sample sizes (missing true associations due to low power), relaxed statistical thresholds (inflating false-positive results), and publishing without a replication dataset (required for genomic studies). Differences in methods of sample collection, transportation and storage, DNA extraction, sequencing, bioinformatics, and statistics can all contribute to inter-study variations. The choice of taxonomic resolution for analysis (PD has been tested at all levels from phylum to species) and the inconsistent taxonomic assignments and nomenclature used in various reference databases add to the confusion when comparing results. Last but not least is confounding by heterogeneity in the populations that were studied: PD is heterogenous and so is the microbiome. PD subtypes cannot be readily identified; thus, patient populations are inevitably varied. A myriad of factors can affect the microbiome ranging from diet, health, and medication to cultural habits, lifestyles, race, and geography[33,34].

Identifying microorganisms involved in the dysbiosis of the microbiome is essential for understanding their role in disease. We conducted a hypothesis-free microbiome-wide association study (MWAS) modeled after and using the standards of rigors that are used in genome-wide association studies (GWAS), but with analytic methods that are appropriate for the high-dimensionality and compositionality of the microbiome data. We used two datasets to allow internal replication. The sample sizes in prior PD-microbiome studies have ranged from 10 to 197 PD cases and 10 to 130 controls[32]. The largest published study (197 cases and 130 controls) is the dataset 1 in the present study, re-analyzed here with a more advanced bioinformatics pipeline than we previously published[16]. In addition, we present an unpublished independent dataset with 323 cases of PD and 184 controls, analyzed in parallel to dataset 1. Two large datasets allowed for internal replication and power to detect both rare and common signals. We standardized data collection and processing as much as possible across the two datasets, and for variations that could not be handled in study design, we used statistical techniques to make appropriate adjustments. We used two different statistical tests for MWAS and focused only on results that were reproducibly significant across methods and across datasets. We employed correlation network analysis to infer interactions among PD-associated microorganisms. We were able to confirm some of the previously reported associations with common taxa, and, in addition, identified associations with rare microorganisms that are commensal, but can become opportunistic pathogens in immune-compromised hosts.

RESULTS

Dramatic difference between datasets

We discovered a remarkable difference between the two datasets, despite efforts to standardize data collection and analysis (Fig. 1). All subjects lived in the United States. Diagnosis, subject selection, and data collection were performed by the NeuroGenetics Research Consortium (NGRC) investigators at the four NGRC-affiliated movement disorder clinics, using standardized methods. Dataset 1 (212 PD and 136 controls) was collected in Seattle, WA, Albany, NY, and Atlanta, GA, in 2014. Dataset 2 (323 PD and 184 controls) was collected in Birmingham, AL, during 2015–2017. We used uniform protocols for sample collection, transportation, and storage for the two datasets. Stool was collected using the same kit, DNA was extracted using the same chemistry, and the 16S rRNA gene V4 region was sequenced using the same primers, but in different laboratories, resulting in 10× greater sequence depth in dataset 2 than dataset 1. The same pipeline was used on the two datasets to process the sequences and assign taxonomic classification. Yet, principal component analysis (PCA)[35] revealed the composition of the microbiome of the samples to be strikingly different in the two datasets (Fig. 1) and the difference was statistically significant ($P < 1\text{E} - 5$, tested using permutational multivariate analysis of variance (PERMANOVA)). The separation of datasets was evident in cases and in controls, in the same pattern. Greater sequence depth in dataset 2 was a significant contributor to this disparity, but not the sole explanation, because the difference between datasets was still significant once sequence depth was adjusted for (PERMANOVA $P < 1E-5$). For all statistical tests (global composition, MWAS, correlations, and network analysis), the two datasets were analyzed separately for two reasons: (i) for independent validation and (ii) to avoid confounding by mixing two clearly different datasets.

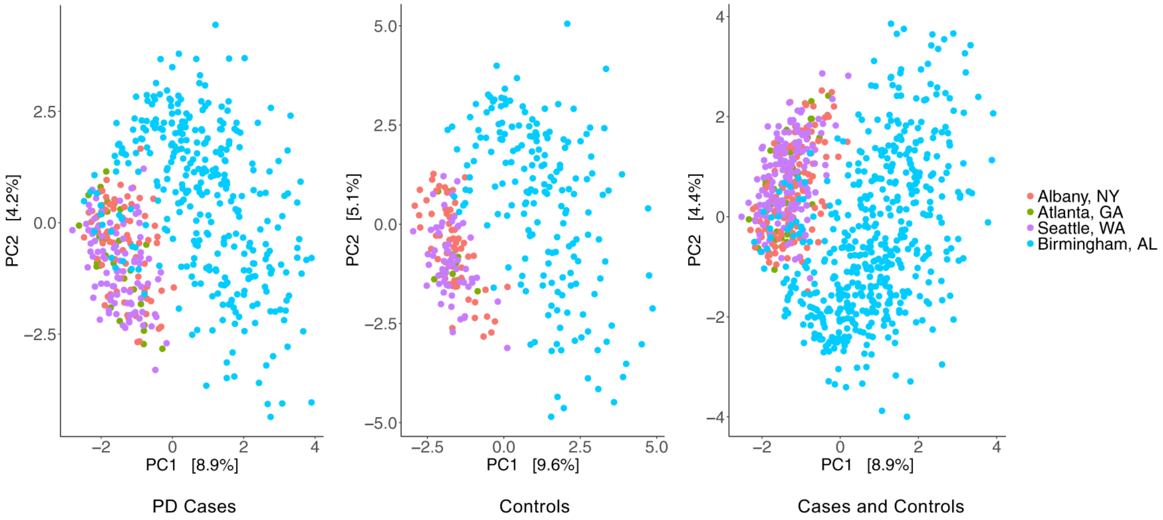


Figure 1: The gut microbiome compositions of the two dataset differed significantly.

Principal component (PC) analysis was used to generate the graphs for PD cases (left, N = 522), controls (middle, N = 316), and cases and controls combined (right, N = 838), where each point represents the composition of the gut microbiome of one individual and distances indicate degree of similarity to other individuals. Percentages on the x-axis and y-axis correspond to the percent variation in gut microbiome compositions explained by PC1 and PC2. The difference between dataset 1 and dataset 2 was formally tested using PERMANOVA and was significant ($P < 1E\text{-}5$). Dataset 1: red (Albany, NY), purple (Seattle, WA), and green (Atlanta, GA). Dataset 2: blue (Birmingham, AL).

Metadata and confounders

Metadata were collected using two self-administered questionnaires and medical records (Supplementary Table 1). An Environmental and Family History Questionnaire

(EFQ)[4,36] was used to collect data relevant to PD. A Gut Microbiome Questionnaire (GMQ)[16] was completed immediately after stool collection and gathered data relevant to the microbiome including diet, gastrointestinal problems, medical conditions, and use of medications. PD medications that subjects were taking at the time of stool collection were extracted from medical records by clinical investigators. The aim of this study was to identify reproducible signals of association between PD and microbiota, and to that end, metadata were used as potential confounders, not as research questions. For example, we did not set out to test the effects of constipation, levodopa, or any of the 47 variables listed in Supplementary Table 1 on the microbiome, because, although of interest, that was not the primary aim of the study and doing so would have reduced the power for the primary aim.

To identify which of the variables might confound the study, we tested the distribution of each variable in cases vs. controls and those that differed at a conservative uncorrected $P < 0.05$ in at least one dataset were tagged as potential confounders (Supplementary Table 1). These included, most notably, constipation in the past 3 months (more common in PD, $P = 6E - 16$ dataset 1, $P = 6E - 10$ dataset 2) and gastrointestinal discomfort on the day of stool collection (more common in PD, $P = 2E - 9$ dataset 1, $P = 4E - 6$ dataset 2), as well as sex and age, body mass index (BMI), weight loss, fruits or vegetable intake, alcohol use, and stool sample travel time. These variables and geographic site were tested along with case–control status in PERMANOVA (global composition test) and those that were significant were used as covariates in analysis of composition of microbiomes (ANCOM) (differential abundance test for MWAS). Thus, the results on both the global composition test and PD-associated taxa in MWAS have been adjusted for known potential confounders, except PD medications, which had to be handled differently because of collinearity with PD (see section on "Cause of disease or consequence of medication").

Global composition of microbiome

First, we tested the difference between PD and controls in the global composition of the gut microbiome (β-diversity, Table 1). Case vs. control status was tested once by itself, once with all potential confounders in the model in a marginal test where each variable was tested while being adjusted for all others in the model, and once stratified by PD medication (Table 1). To gauge the effect of distance metric on the results, all tests were repeated with Aitchison[35], generalized UniFrac (GUniFrac)[37], and Canberra[38] distances. Tests were conducted using PERMANOVA[39] with 99,999 permutations limiting maximum achievable significance to $P = 1\text{E} - 5$.

PD microbiomes differed significantly from control microbiomes, in both datasets, with every distance metric measured ($P < 1\text{E} - 5$, Table 1). The PD effect was significant and independent of all analyzed confounders, including geography, constipation, gastrointestinal discomfort, sex, age, BMI, fruit or vegetable intake, alcohol use, and stool sample travel time.

Results were in agreement with population studies in detecting significant effects of sex, age, BMI, gastrointestinal issues, and diet on the microbiome33,34, and with other PD studies in detecting evidence for dysbiosis in PD[11-30].

Table 1. Effect of PD and other key variables on the global composition of gut microbiome. Model A tested PD vs. control without any other variable in the model. Sample size for Model A was 201 cases and 132 controls in dataset 1 and 323 cases and 184 controls in dataset 2. Model B included 11 variables (including case/control) and each variable was tested while adjusting for the other 10, without priority. Model B included subset of samples that had complete data on all 11 variables: N = 160 cases and 111 controls in dataset 1 and 283 cases, and 167 controls in dataset 2. For Model C, patients were stratified by each PD medication they were taking at the time of stool collection; those not on medication (varying N for different medications, see Supplementary Table 1) were tested against controls (N = 132 in dataset 1 and 184 in dataset 2). Power was low for patients not on L-dopa (N patients <50) and patients not on any PD medication (<20) due to small sample sizes, but not for other medications (N patients not on medication = 88–179 in dataset 1 and 153–312 in dataset 2). All analyses were repeated with three different distance measures: Aitchison, Canberra, and GUniFrac (generalized UniFrac). % var was the inter-individual variation explained by each variable. *P* value was calculated using 99,999 permutations, setting the highest achievable significance at *P* = 1E − 05.

	Dataset 1						Dataset 2					
	Aitchison		GUniFrac		Canberra		Aitchison		GUniFrac		Canberra	
	%var	*P*	%var	*P*	%var	*P*	%var	*P*	%var	*P*	%var	*P*
Model A. All PD vs Control	0.71	<1E-05	1.38	<1E-05	0.57	<1E-05	0.56	<1E-05	0.89	<1E-05	0.38	<1E-05
Model B. PD and confounders												
Geography (Seattle, Atlanta, Albany)	0.99	2E-03	1.10	0.02	0.84	2E-03	-	-	-	-	-	-
PD (case vs. control)	0.58	1E-03	1.12	7E-05	0.53	4E-05	0.48	<1E-05	0.62	9E-05	0.32	2E-05
Sex (male vs female)	0.51	9E-03	0.52	0.08	0.49	2E-04	0.48	2E-05	0.49	2E-03	0.34	2E-05
Age (continuous)	0.45	0.04	0.76	5E-03	0.43	0.01	0.45	<1E-05	0.62	1E-04	0.34	3E-05
GI discomfort on day of stool collection (yes vs no)	0.45	0.04	0.40	0.26	0.43	9E-03	0.24	0.2	0.22	0.39	0.23	0.2
Fruits or vegetables daily (yes vs no)	0.38	0.3	0.55	0.05	0.42	0.02	-	-	-	-	-	-
Constipation in the past three months (yes vs no)	0.34	0.77	0.38	0.35	0.37	0.39	0.26	0.06	0.38	0.02	0.24	0.05
BMI (continuous)	0.40	0.21	0.48	0.12	0.39	0.13	0.33	3E-03	0.34	0.04	0.27	6E-03
Drinks alcohol (yes vs no)	0.35	0.66	0.31	0.64	0.37	0.35	0.26	0.07	0.28	0.15	0.24	0.1
Lost >10 pounds in past year (yes vs no)	0.34	0.71	0.36	0.42	0.36	0.64	0.20	0.87	0.15	0.91	0.21	0.71
Stool sample travel time (continuous)	0.35	0.66	0.70	0.01	0.36	0.58	0.23	0.26	0.3	0.09	0.24	0.11
Model C. Removing PD medications												
PD not on levodopa vs control	0.93	0.01	1.12	0.04	0.78	0.02	0.48	0.17	0.54	0.16	0.47	0.11
PD not on COMT inhibitors vs control	0.66	9E-05	1.27	<1E-05	0.56	<1E-05	0.55	<1E-05	0.88	<1E-05	0.38	<1E-05
PD not on anticholinergics vs control	0.73	<1E-05	1.31	<1E-05	0.58	<1E-05	0.57	<1E-05	0.92	2E-05	0.39	<1E-05

PD not on MAO-B inhibitors vs control	0.81	<1E-05	1.50	3E-05	0.66	<1E-05	0.71	<1E-05	1.07	<1E-05	0.45	<1E-05
PD not on dopamine agonists vs control	0.81	2E-04	1.51	3E-05	0.70	<1E-05	0.57	1E-04	0.80	3E-04	0.44	4E-05
PD not on amantadine vs control	0.73	3E-05	1.37	<1E-05	0.60	<1E-05	0.48	3E-05	0.74	3E-05	0.37	<1E-05
PD not on any PD drug vs control	1.00	0.07	0.89	0.22	0.82	0.06	0.48	0.58	0.52	0.37	0.48	0.79

Identification of PD-associated microorganisms

To identify PD-associated microorganisms, we conducted MWAS, testing differences between cases and controls in the relative abundances of genera. We conducted MWAS on each dataset separately to test whether results replicate and also to avoid confounding by the heterogeneity between datasets. Each dataset was tested with two methods to test analytic concordance: once using ANCOM[40] and again using Kruskal–Wallis (KW) rank sum test[41]. We chose ANCOM, because among the numerous methods that have been proposed, ANCOM singularly met three key criteria: incorporates compositionality of the eco-system, allows covariate adjustment, and keeps false-positive rate low while maintaining power[40,42]. Differential abundance was tested hypothesis-free microbiome-wide: ANCOM included all 445 genera detected in dataset 1 and 561 genera in dataset 2; KW included 109 genera in dataset 1 and 163 in dataset 2 (excluding unassigned genera and genera present in <10% of samples). In ANCOM, dataset-specific covariates were included and adjusted for (see MWAS section in Methods). Resulting significance metrics were corrected for multiple testing, using false discovery rate (FDR)-corrected P-values to calculate *W* in ANCOM and Benjamini–Hochberg FDR in KW.

We detected association signals for 15 genera that were microbiome-wide significant by both methods and reproduced robustly in the two datasets, with or without covariate adjustment (Table 2 and Fig. 2). Five genera had higher abundances in PD than in controls: *Porphyromonas*, *Prevotella*, *Corynebacterium_1*, *Bifidobacterium*, and *Lactobacillus*. Ten genera had lower abundances in PD than controls: *Faecalibacterium*, *Agathobacter*, *Blautia*, *Roseburia*, *Fusicatenibacter*, *Lachnospira*, *Butyricicoccus*,

Table 2. PD-associated genera identified via MWAS. MWAS was conducted in two datasets independently, testing differential abundance of genera in PD vs. controls, using two statistical methods (ANCOM and KW). The 15 genera shown are those that achieved microbiome-wide significance for association with PD in both datasets and by both methods, with (ANCOM) and without (KW) covariate adjustment (see "METHODS" for covariates). Sample size: ANCOM included subset of subjects for whom complete data were available on all covariates tested: N = 171 cases and 117 controls in dataset 1 and 306 cases and 177 controls in dataset 2. KW included all subjects: N = 201 cases and 132 controls in dataset 1, and 323 cases and 184 controls in dataset 2. Clusters were identified hypothesis-free using correlation network analysis (Fig. 3). PubMed search was conducted after analyses were completed using genus and species name as search term (Supplementary Table 6). Function (opportunistic pathogen, SCFA, probiotic) was taken strictly from PubMed and is likely oversimplified. Microbiota have been studied under a narrow lens of what is already known about them. Opportunistic pathogens are often looked for in clinical specimen with infection, SCFA bacteria are studied intensively for their anti-inflammatory and other protective effects, and probiotics are understudied but highly advertised. The full function of the microbiota are not yet fully understood. In comparing results across published studies, note that a "genus" classified by one study may not be the same as the genus by the same name in another study. Taxonomic classifications and nomenclature are not standardized across reference databases, e.g., "*Prevotella*", as annotated in some databases including NCBI, is further divided by SILVA (used here) into several non-monophyletic groups that SILVA calls, *Prevotella_2*, *Prevotella_6*, *Prevotella_7*, *Prevotella_9*, and *Prevotella* (see Discussion).

ANCOM analysis of composition of microbiomes. *FC* fold change in patients (MRA in patients/MRA in controls). *FDR* Benjamini–Hochberg false discovery rate (multiple testing corrected P-value). *KW* Kruskal–Wallis. *MWAS* microbiome-wide association study. *MRA* mean relative abundance in controls. *NC* not uncultured (uncharacterized). *Opp path* opportunistic pathogen (often commensal microorganism that can become pathogenic in immune-compromised individuals). *Probiotic* carbohydrate-metabolizing bacteria commonly known as probiotics. *SCFA* short-chain fatty acid-producing bacteria. *W* ANCOM score indicating the number of times a genus achieved FDR 0.05 as compared with other genera (maximum *W* possible: 444 in dataset 1, 560 in dataset 2, threshold 0.8 was used for significance, all shown genera were above significance threshold).

PD-associated genera					MWAS significant in Dataset 1				MWAS significant in Dataset 2				Cluster	PubMed
Phylum	Class	Order	Family	Genus	MRA	FC	ANCOM (*W*)	KW (FDR)	MRA	FC	ANCOM (*W*)	KW (FDR)		
Bacteroidetes	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	*Porphyromonas*	0.001	4.20	406	1E-03	0.001	2.94	468	2E-02	1	Opp path
Bacteroidetes	*Bacteroidia*	*Bacteroidales*	*Prevotellaceae*	*Prevotella*	0.002	2.56	400	6E-03	0.001	4.39	463	2E-02	1	Opp path
Actinobacteria	*Actinobacteria*	*Corynebacteriales*	*Corynebacteriaceae*	*Corynebacterium_1*	0.001	1.96	360	1E-02	0.002	2.53	465	8E-03	1	Opp path
Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Faecalibacterium*	0.06	0.63	411	1E-03	0.04	0.66	535	3E-03	2	SCFA
Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Agathobacter*	0.04	0.53	441	2E-04	0.02	0.56	545	6E-05	2	SCFA
Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Blautia*	0.02	0.68	410	2E-03	0.02	0.79	533	4E-02	2	SCFA
Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Roseburia*	0.02	0.48	391	4E-03	0.01	0.60	541	3E-04	2	SCFA
Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Fusicatenibacter*	0.004	0.56	388	2E-02	0.005	0.69	521	3E-02	2	SCFA
Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Lachnospira*	0.004	0.80	426	1E-03	0.005	0.68	521	1E-02	2	SCFA
Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Butyricicoccus*	0.002	0.66	382	7E-03	0.002	0.68	505	6E-02	2	SCFA

Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Lachnospiraceae _ND3007*	0.001	0.37	418	2E-04	0.001	0.59	538	6E-04	2	NC
Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Lachnospiraceae _UCG-004*	0.001	0.48	384	2E-02	0.001	0.38	544	1E-05	2	NC
Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Oscillospira*	6E-4	0.65	367	2E-02	5E-4	0.64	525	1E-02	2	NC
Actinobacteria	*Actinobacteria*	*Bifidobacteriales*	*Bifidobacteriaceae*	*Bifidobacterium*	0.01	1.83	410	1E-03	0.01	2.72	553	6E-07	3	Probiotic
Firmicutes	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	4E-4	6.61	407	2E-04	0.004	1.57	458	1E-02	3	Probiotic

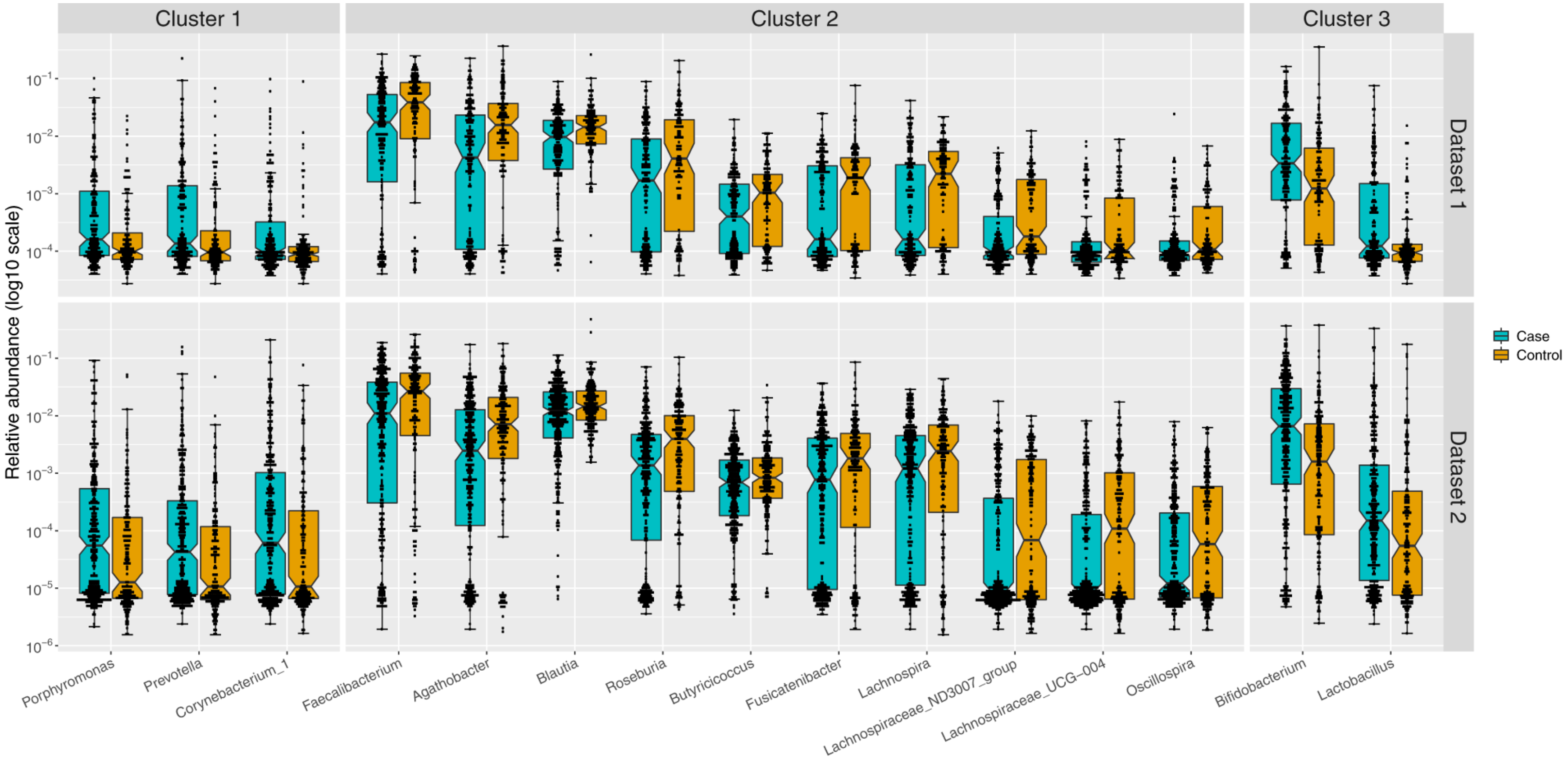


Fig. 2: Differential abundances of 15 PD-associated genera replicated in two datasets.

Relative abundances in PD cases (blue) and controls (orange) were plotted as log10 scale on the *y*-axis. Sample size was 201 cases and 132 controls in dataset 1, and 323 cases and 184 controls in dataset 2. Each dot represents a sample, plotted according to the relative abundance of the genus in the sample. The notch in each box indicates the confidence interval of the median. The bottom, middle, and top boundaries of each box represent the first, second (median), and third quartiles of the relative abundances. The whiskers (lines extending from the top and bottom of the box and ending in horizontal cap) extend to points within 1.5 times the interquartile range. The points extending above the whiskers are outliers.

Lachnospiraceae_ND3007_group, *Lachnospiraceae_UCG-004*, and *Oscillospira*. Complete MWAS results are in Supplementary Tables 2–5.

Correlation network analysis

We questioned whether the 15 association signals were independent. We used hypothesis-free correlation network analysis[43] to infer ecological networks of interacting organisms microbiome-wide (Fig. 3 and Supplementary Fig. 1). The PD-associated genera mapped to three polymicrobial clusters. *Porphyromonas*, *Prevotella*, and *Corynebacterium_1*, which were elevated in PD, mapped to a community of highly correlated organisms, which we denoted as cluster 1. Cluster 1 was the most distinct cluster in the microbiome with correlations reaching $r = 0.82$ ($P < 3\text{E} - 4$), the highest in the microbiome in our data. The ten genera that were depleted in PD formed cluster 2, where eight of them clustered at $r \geq 0.4$ ($P < 3\text{E} - 4$), and the remaining two (*Oscillospira* and *Lachnospiraceae_UCG-004*), clustered with the others at $r = 0.25$ ($P < 3\text{E} - 4$) and $r = 0.35$ ($P < 3\text{E} - 4$). *Lactobacillus* and *Bifidobacterium*, both elevated in PD, were correlated with each other at $r = 0.33$ ($P < 3\text{E} - 4$), which we denoted as cluster 3. Correlations within each cluster were all in the positive direction, i.e., members of clusters 1 tended to increase in abundance together, cluster 2 decreased together, and cluster 3 increased together.

Functional characteristics

Analyses so far were all hypothesis-free, data-driven, and blind to the functional relevance of the microorganisms. Having identified the associations and their

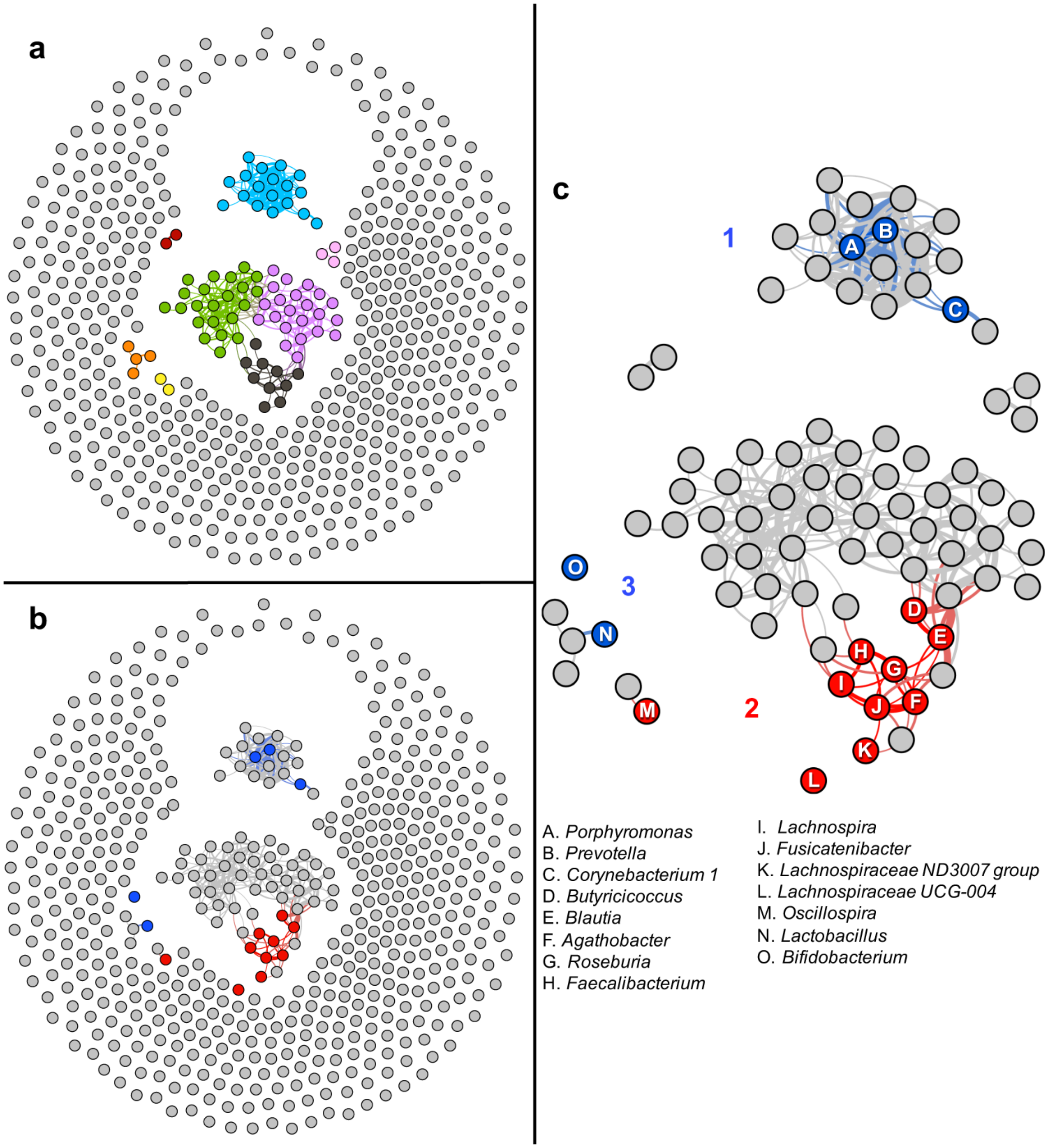


Figure 3: Correlation network analysis mapped PD-associated genera to three polymicrobial clusters.

Pairwise correlations in relative abundances were calculated for all genera microbiome-wide and were used to detect clusters of co-occurring microorganisms. To display, we used an arbitrary correlation coefficient threshold at $r \geq |0.4|$ to connect the genera that were correlated. All correlations noted were significant at $P < 3E-4$ (the limit for 3000 permutations). Here we show the result for PD cases in dataset 2, because it had larger sample size ($N = 323$ cases) and greater sequencing depth than dataset 1 (see Supplementary Fig. 1 for cases and controls in dataset 1 and dataset 2). (a) Algorithm-detected clusters shown in different colors. (b) The algorithm-detected clusters, as in a but shown in gray, and PD-associated genera highlighted in blue (if increased in PD) or red (if decreased in PD). (c) Zoomed in version of (b). The 15 PD-associated genera fell in three clusters. Cluster 1 was a tightly correlated cluster of microorganisms (r approaching 0.8), which included *Porphyromonas*, *Prevotella*, and *Corynebacterium_1* (all elevated in PD). Cluster 2 included the ten genera that were reduced in PD, eight of which are shown connected at $r \geq 0.4$, and two are unconnected but correlated significantly ($P = 3E - 4$) with the others in the cluster at $r = 0.25$ and $r = 0.35$. *Lactobacillus* and *Bifidobacterium* (correlated at $r = 0.33$ ($P < 3E - 4$)) were denoted cluster 3. For unconnected genera ($r < 0.4$), the proximity between nodules does not imply relatedness, e.g., *Oscillospira* (M) falls closer to *Lactobacillus* (N) than to *Roseburia* (G) but it is correlated significantly with *Roseburia* ($r = 0.25$, $P < 3E - 4$) and not with *Lactobacillus* ($r = 0.04$, $P = 0.44$).

corresponding clusters, we broke the blind by searching PubMed. PubMed results on functional characteristics converged on clusters defined by agnostic network analysis.

PubMed results suggest genera in cluster 1 are opportunistic pathogens. *Porphyromonas* and *Prevotella* are anaerobic, Gram-negative bacteria with lipopolysaccharides (endotoxins) in their outer membrane. They are commensal to the human gastrointestinal and urogenital tracts. *Corynebacterium* are aerobic, Gram-positive, and have a higher abundance in the skin microbiota than the gut. Although commensal and often harmless, *Porphyromonas*, *Prevotella*, and *Corynebacterium* are opportunistic pathogens capable of causing infections in immune-compromised individuals or if they gain access to sterile sites via compromised membranes, post surgery, bites, or wounds[44,45,46].

Many, but not all species of *Porphyromonas*, *Prevotella*, and *Corynebacterium* are pathogens. *Corynebacterium diphtheriae* is the leading cause of diphtheria. *Porphyromonas gingivalis* causes periodontal disease. We did not detect *C. diphtheriae* and *P. gingivalis* was extremely rare in our samples. We were interested in knowing the species that made up these three genera in our PD samples. The bioinformatic pipeline used in our study (DADA2 with SILVA as reference database) assigned the detected sequences (amplicon sequence variants (ASVs)) to species if the sequences were 100% identical; otherwise, the ASV was unassigned to species. To confirm and expand on DADA2-SILVA assignments, we blasted all the ASVs that made up each of the three genera against the NCBI 16S rRNA database, focusing only on matches that were >99–100% identical to a species with high statistical confidence. In PD patients, we found that 80% of *Corynebacterium_1* was composed of one unique ASV with 100% identity to *Corynebacterium amycolatum* and *Corynebacterium lactis*; 96% of *Porphyromonas* was composed of ASVs that matched *Porphyromonas asaccharolytica*, *Porphyromonas bennonis*, *Porphyromonas somerae*, or *Porphyromonas uenonis* with >99–100% identity, and 98% of *Prevotella* was composed of ASVs that matched *Prevotella bivia*, *Prevotella buccalis*, *Prevotella disiens*, or *Prevotella timonensis* with >99–100% identity (83% of *Prevotella* matched *P. bivia*, *P. buccalis*, *P. disiens*, or *P. timonensis* at 100% identity). We conducted a PubMed search for each of these ten species, using genus and species name as the key word (ex. *Corynebacterium amycolatum*), with search filters as follows: Humans, English, and Title/Abstract. Excluding method papers, PubMed returned 104 articles that addressed function, characteristics, or relevance to human health, and every article was about the microorganism (search term) as a pathogen in clinical specimens from various infections (Supplementary Table 6).

Clinical specimen from chronic wounds, infections, and inflammations are often polymicrobial[44,45,46]. *Porphyromonas*, *Prevotella*, *Corynebacterium*, and other members of cluster 1 are often observed together in these polymicrobial infections[44,45,46]. With the newly acquired knowledge on the potential biological significance of cluster 1, we questioned whether this polymicrobial group as a whole may be relevant to PD. The co-occurring organisms in cluster 1 (defined by correlation $r \geq 0.4$) were *Anaerococcus*, *Campylobacter*, *Ezakiella*, *Finegoldia*, *Murdochiella*, *Peptoniphilus*, *Porphyromonas*, *Prevotella*, and *Varibaculum* in dataset 1, and *Anaerococcus*, *Campylobacter*, *Corynebacterium_1*, *Ezakiella*, *Fastidiosipila*, *Finegoldia*, *Lawsonella*, *Mobiluncus*, *Mogibacterium*, *Murdochiella*, *Negativicoccus*, *Peptoniphilus*, *Porphyromonas*, *Prevotella*, *Prevotella_6*, *S5-A14a*, *Varibaculum*, and unclassified *Corynebacteriaceae* in dataset 2. Most of these organisms are rare and may have been missed in MWAS. We conducted another MWAS where we collapsed the nonsignificant members of cluster 1 into one group (partial cluster 1), leaving *Porphyromonas*, *Prevotella*, and *Corynebacterium_1* as individual genera along with the rest of the genera in MWAS. As expected, we recaptured all 15 PD-associated genera, as well as an additional signal for the partial cluster 1 that was ANCOM and KW significant in both datasets (dataset 1: 2.9-fold increased abundance in PD, ANCOM $W = 392$, KW FDR = 0.03; dataset 2: 2.5-fold increased abundance in PD, ANCOM $W = 480$, KW FDR = 0.002).

Most (possibly all) genera in cluster 2 produce short-chain fatty acids (SCFAs). Of the ten PD-associated genera in cluster 2, three (*Oscillospira*, *Lachnospiraceae_UCG-004*, and *Lachnospiraceae_ND3007_group*) have been detected only by sequencing and not yet been cultured. The rest (*Agathobacter*, *Blautia*, *Butyricicoccus*, *Faecalibacterium*, *Fusicatenibacter*, *Lachnospira*, and *Roseburia*) are all anaerobic, Gram-positive bacteria in the *Ruminococcaceae* and *Lachnospiraceae* families. They are best known for producing SCFAs, mainly butyrate, which help maintain integrity of the gut membrane and have anti-inflammatory properties[47,48].

The literature on genera in cluster 3 suggest they are probiotic, but with the potential of becoming opportunistic pathogens and immunogenic. *Lactobacillus*[49] and *Bifidobacteria*[50] are anaerobic Gram-positive bacteria. They are among the ubiquitous inhabitants of the human gastrointestinal microbiome. They metabolize carbohydrates in plants and dairy, and are considered probiotic for their health benefits[51,52], although they have also been implicated as cause of infection and excessive immune stimulation in susceptible individuals[52,53].

Cause of disease or consequence of medication

Human association studies are powerful tools for identifying disease-relevant leads and to generate hypotheses that can then be tested experimentally. Even if we find a strong candidate that blurs the line between association and causality, we cannot prove that it preceded PD, because there are decades of preclinical and prodromal disease, and we do not know when it all begins. Although cause cannot be proven in these studies, we can sometimes tease out consequence.

Medications have profound effects on the microbiome[33]. Levodopa is the most commonly used PD medication (>85% of PD patients were on varying doses of levodopa). To gauge if the association of PD with any of the 15 genera was a consequence of levodopa treatment, we tested whether the change in the differential abundance of the 15 genera correlated with increasing levodopa dose.

We found no significant evidence to suggest that the increasing abundance of *Porphyromonas*, *Prevotella*, or *Corynebacterium_1* (cluster 1) correlated with levodopa therapy. We did find significant evidence (two-sided P value < 0.05) in dataset 2 to suggest that increasing doses of levodopa were correlated with decreasing levels of SCFA-producing organisms (*Faecalibacterium* $P = 0.01$, *Agathobacter* $P = 0.02$, *Blautia* $P = 5\text{E} - 4$, *Roseburia* $P = 0.02$, *Fusicatenibacter* $P = 0.01$, *Lachnospira* $P = 5\text{E} - 3$, *Lachnospiraceae_ND3007_group* $P = 5\text{E} - 3$, *Lachnospiraceae_UCG-004* $P = 0.03$). A similar pattern was present in dataset 1, albeit most did not reach statistical significance possibly due to the smaller sample size of dataset 1. We also detected significant correlation between increasing levodopa dose and increasing levels of *Bifidobacterium* (dataset 1 $P = 5\text{E} - 3$, dataset 2 $P = 2\text{E} - 6$) and *Lactobacillus* (dataset 2 $P = 4\text{E} - 3$). These data suggest that the increase in abundance of cluster 1 (opportunistic pathogens) is independent of levodopa, but that the reduction in cluster 2 (SCFA) and increase in cluster 3 (probiotics), if not solely a consequence of medication, worsen with increasing doses of levodopa.

DISCUSSION

To summarize, we first confirmed that the gut microbiome is altered in PD and showed that the PD effect on the global composition of the gut microbiome is independent of the effects of sex, age, BMI, constipation, gastrointestinal discomfort, geography, and diet. Next, using hypothesis-free microbiome-wide association studies we identified 15 PD-associated genera that achieved microbiome-wide significance in both datasets, with two methods, and with or without covariate adjustment. The 15 association signals were robust to the dramatic population-specific differences in the composition of microbiomes of the two datasets. We used hypothesis-free correlation network analysis to infer interactions and to identify communities of co-occurring microorganisms. Using this agnostic approach, we learned that the 15 PD-associated genera represent three polymicrobial clusters. Review of the literature revealed that the clusters, as defined by agnostic network analysis, also share functional characteristics. Our results suggest the gut microbiomes of persons with PD can present with (1) an overabundance of a polymicrobial cluster of opportunistic pathogens, (2) reduced levels of SCFA-producing bacteria, and/or (3) elevated levels of carbohydrate metabolizers commonly known as probiotics.

Our data align with and expand on PD-microbiome literature. Reduced levels of SCFA-producing bacteria[12,14,16,18,19,21,26,27] and elevated levels of probiotic bacteria in PD[14,16,18,21,25,26,27] have been reported before, and thus are confirmatory. Overabundance of opportunistic pathogens, however, had not been reported before. We suspect the reason we were able to detect these microorganisms is because they are rare (Fig. 2) and we had a much larger sample size and power than prior studies. The microorganisms identified in prior PD studies were among the more abundant microorganisms in the gut. There have been two systematic reviews of PD-microbiome studies, which clearly show the vast disparity in the findings, but also reveal few findings that have emerged in more than one study[31,32]. The most recent review highlighted six associations that were significant in more than one study: *Faecalibacterium*, *Roseburia*, *Bifidobacterium*, *Lactobacillus*, *Akkemansia*, and *Prevotella*[32]. We confirmed the reduction in *Faecalibacterium* and *Roseburia* (cluster 2), and the increase in *Bifidobacterium* and *Lactobacillus* (cluster 3). We also confirmed increased *Akkermansia* in both datasets but it was only significant in dataset 1. *Prevotella* results are interesting, with Scheperjans et al.[11] and Petrov et al.[18] reporting it decreased in PD, whereas we find it elevated in both datasets. The apparent inconsistency may be simply because what is being referred to as "*Prevotella*" is not the same in these studies. We all used different taxonomic classification: Scheperjans et al.[11] reported at the family level (*Prevotellaceae*), we at genus level (*Prevotella*), and Petrov et al.[18] at species level (*Prevotella copri*). The SILVA database we used here, classified family *Prevotellaceae* into 11 genera. The more common genera in the *Prevotellaceae* family (*Paraprevotella*, *Prevotella_9*, and *Prevotella_7*) did in fact have lower frequencies in PD than in controls, as Scheperjans et al.[11] observed, but the difference was not significant in our datasets (FDR > 0.6 in both datasets). Species *P. copri*, which Petrov et al.[18] found reduced in PD, was the main species of the *Prevotella_9* genus, which was reduced in our PD samples as well but not significantly (FDR > 0.8 in both datasets). We found instead elevated levels of the less common genus *Prevotella* (FDR = 0.006 in dataset 1 and FDR = 0.02 in dataset 2). These findings suggest family *Prevotellaceae* may be heterogenous in its association with PD. When comparing studies, another important consideration is the reference database: there are many and they have varied phylogenetic resolution and nomenclature. For example, genus *Corynebacterium* in NCBI is divided into two non-monophyletic genera in SILVA: *Corynebacterium_1* and *Corynebacterium*. Similarly, what is called genus *Prevotella* in NCBI, is divided into multiple non-monophyletic genera in SILVA (we detected *Prevotella*, *Prevotella_2*, *Prevotella_6*, *Prevotella_7*, and *Prevotella_9*). The varying resolution at which the tests are conducted and the reference databases used cause confusion in the literature.

The evidence for overabundance of opportunistic pathogens in PD gut microbiome was potentially the most exciting finding of this study. Braak et al.[54,55]

originally hypothesized that non-inherited forms of PD are caused by a pathogen that can pass through the mucosal barrier of the gastrointestinal tract and spread to the brain through the enteric nervous system. Although many aspects of Braak's hypothesis have gained support in recent years, there is no direct evidence that a pathogen is involved. Presence of α-synuclein in the gastrointestinal tract has been documented in persons with established Lewy body disease[56], as well as those with rapid eye movement sleep behavior disorder, which is considered prodromal PD[57]. Epidemiological studies suggest that truncal vagotomy if conducted decades before onset of PD reduces risk of developing PD[58,59]. In a mouse model, α-synuclein fibrils injected into the gut induced α-synuclein pathology which spread to the brain resulting in Parkinsonian neurodegeneration and behavioral phenotype; whereas truncal vagotomy and α-synuclein deficiency prevented the gut-to-brain spread and the associated neurodegeneration[60]. Human studies unrelated to PD have shown that infection in the gut or the olfactory system induce α-synuclein expression, and the increased abundance of α-synuclein mobilizes the immune system to fight the pathogen[61,62]. It was also shown in a genetic model of PD (*Pink1* knockout mice) that intestinal infection by pathogens elicits activation of cytotoxic T cells in the periphery and the brain, and leads to deterioration of dopaminergic cells and motor impairment, suggesting that intestinal infection acts as a triggering event in PD[63]. Despite the increasing evidence linking the gut, α-synuclein, and inflammation to PD, there was no direct evidence that a pathogen is responsible for the pathology. Here, we present evidence from human samples indicating an overabundance of opportunistic pathogens in the gut microbiome of persons with PD. The three genera that rose to significance (*Porphyromonas*, *Prevotella*, or *Corynebacterium_1*) represented a larger polymicrobial cluster of opportunistic pathogens that co-occur in controls as well as in patients (although at much lower abundances in healthy gut). Per literature, these opportunist pathogens are often harmless, but can grow and cause infections if the immune system is compromised or if they penetrate sterile sites through, e.g., compromised membranes[44,45,46]. The exciting question is whether these are Braak's pathogens capable of triggering PD, or they are irrelevant to PD but are able to penetrate the gut and grow, because the gut lining is compromised in PD. We re-emphasize that no claims can be made on function based solely on association. The knowledge on the function of microorganisms in the gut is currently limited. Although there may be a large body of literature, each organism has been studied with a narrow lens. Organisms that are known to be opportunistic pathogens are being looked for in clinical specimen, whether they have other critical functions is not known. The identity of these microorganisms will enable experimental studies to determine if and how they play a role in PD.

Our second main finding was a polymicrobial cluster of ten genera whose relative abundances were reduced in PD. All ten genera belong to the *Lachnospiraceae* and *Ruminococcaceae* families, well-known for producing SCFA. Several studies had found reduced levels of different SCFA-producing bacteria in PD patients[12,14,16,18,19,21,26,27]. Our finding is therefore confirmatory and expands on the list of PD-associated genera in these two taxonomic families. We and others noted that the decreasing levels of *Lachnospiraceae* correlate with increasing daily dose of levodopa, disease duration[12], disease severity and motor impairment[26], which suggest SCFA-producing microorganisms diminish as a consequence of medication and/or advancing disease. SCFA promote gastrointestinal motility, maintain integrity of the gut lining, and control inflammation in the gut and the brain[47,48,64,65,66], all of which are compromised in PD. It is important to note, however, that reduced levels of SCFA in the gut has been documented in many inflammatory disorders[67,68,69,70,71], and is not specific to PD.

We also found elevated levels of *Bifidobacterium* and *Lactobacillus* in PD, which are generally considered as probiotics. Increased *Bifidobacterium* and *Lactobacillus* have been noted in some of the prior PD studies, albeit not consistently[14,16,18,21,25,26,27]. Both are ubiquitous inhabitants of human gut and metabolize carbohydrates derived from plants and dairy[49,50]. We found a significant correlation between increasing levodopa dose and increasing *Bifidobacterium* and *Lactobacillus* levels. *Lactobacillus* produce a bacterial enzyme that metabolizes levodopa into dopamine before it can reach the brain, reducing efficacy of the drug and requiring higher doses, which in feedback causes further growth of the bacteria[72,73]. Ironically, *Bifidobacterium* and *Lactobacillus* are sold in stores as probiotics, and a clinical trial has reported fermented milk, which contained *Bifidobacterium*, *Lactobacillus*, and fiber, among other active ingredients, improved constipation in PD[74]. Although generally believed to be safe, and possibly beneficial for the healthy population, they can act as opportunistic pathogens and cause infection and excessive immune stimulation in immune-compromised individuals[52,53]. It is important to understand why *Bifidobacterium* and *Lactobacillus* are elevated in PD and if they are beneficial (a compensatory mechanism to overcome the dysbiosis) or detrimental (feedback of levodopa).

There were limitations in this study that should be considered in designing follow-up studies. The sample size, although the largest PD-microbiome study to date, was not sufficiently powered to detect rare microorganisms. If PD is indeed associated with polymicrobial clusters of rare opportunistic pathogens, larger sample sizes are needed to tease out the microorganisms individually. In addition to larger sample size, identifying the microorganisms will require shotgun metagenomic sequencing. The 16S amplicon sequencing used here was sufficient for exploratory MWAS, but did not provide the resolution to species, strain and gene level. We also lacked ability to detect viruses and fungi. Since this study was launched in 2014, the field has advanced rapidly. To maintain uniformity in data collection, we did not change the method of stool collection mid-study from sterile swabs to preservative solutions, but employed the latest advances if they could be applied to both datasets uniformly, notably in bioinformatics and statistics, and took analytic measures to identify potential confounders. We made certain decisions for data analyses, such as using stringent criteria to declare significance, and the choice of parameters used to define networks and clusters. We have made both the raw data and summary statistics publicly available so they can be analyzed with any methods and specifications.

In conclusion, we uncovered robust and reproducible signals, which reaffirm (SCFA and probiotics) and generate leads (opportunistic pathogens) for experimentation into cause and effect, disease progression, and therapeutic targets. This study was limited by its singular and precise focus and intentionally conservative analytic execution. There is more to be learned with larger sample sizes with greater power, longitudinal studies to track change from prodromal to advanced disease, and by next-generation metagenome sequencing to broaden the scope from bacteria and archaea to include viruses and fungi, and improve the resolution to strain and gene level.

METHODS

Subjects and data collection

The study was approved by institutional review boards for ethical conduct of human subject research at all participating institutions, namely New York State Department of Health, University of Alabama at Birmingham, VA Puget Sound Health Care System, Emory University, and Albany Medical Center. All subjects provided written informed consent for their participation.

Subjects characteristics are provided in Supplementary Table 1. Subjects were enrolled by NGRC investigators, using standardized methods, at four NGRC-affiliated movement disorder clinics in the United States. Dataset 1 was collected in Seattle, WA, Albany, NY, and Atlanta, GA, in 2014 and included 212 persons with PD and 136 controls[16]. Dataset 2 was collected in Birmingham, AL, during 2015–2017 and included 323 PD and 184 controls (unpublished). PD was diagnosed by a movement disorder specialist using UK Brain Bank criteria[75], and controls were self-reported free of neurological disease. Each individual represents a distinct and unique data point (no repeated measurements were used).

Metadata are provided in Supplementary Table 1. Data were collected using two self-administered questionnaires: an EFQ and GMQ[4,16,36]. EFQ covered sex, age, ancestry, and lifetime exposure data on PD-related risk factors. GMQ covered information pertinent to microbiome analysis and was filled out immediately after stool sample collection. PD medications that subjects were taking at the time of sample collection were extracted from medical records by clinical investigators.

Stool samples were collected by the subjects at home using DNA/RNA-free sterile cotton swabs (BD BBL CultureSwab Sterile/Media-free Swabs, Fisher Scientific, Pittsburgh, PA). The sample was collected from excreted stool (the kit was not a rectal swab), thus minimizing contamination by skin microbiota, water, and urine. The stool samples were shipped immediately via standard US postal service at ambient temperature and stored at −20 °C upon arrival. The collection kit chosen was the most reasonable option at the time (2014). We did not use stabilizing solution, because collection kits with stabilizing solutions (e.g., OMNIgene GUT by DNA Genotek) were first introduced in 2015–2016. Immediate freezing was not feasible because we could not collect stool from over 800 participants, most of whom suffer constipation, while in clinic, nor was it acceptable to the participants to place their stool in their home freezer before shipping. We tested the effect of stool sample travel time on the results as follows. Subjects recorded the collection date and we recorded when it was placed in −20 °C freezer, the difference was calculated as the stool sample travel time. We tested the stool sample travel time in cases vs. controls (Supplementary Table 1). We adjusted the PERMANOVA and MWAS for stool sample travel time.

DNA extraction and sequencing

DNA extraction and sequencing of datasets were done in different laboratories (the Knight Lab at University of California San Diego for dataset 1[16] and HudsonAlpha Institute for Biotechnology for dataset 2), keeping methods uniform as possible. Negative controls were included in both datasets. DNA was extracted using MoBio PowerMag Soil DNA Isolation Kit for dataset 1 and MoBio PowerSoil DNA Isolation Kit for dataset 2, both kits using equivalent chemistries (MoBio Industries, Carlsbad, CA). Case and control samples were randomized on plates for sequencing to avoid batch effect. Hypervariable region 4 (V4) of the bacterial/archaeal 16S rRNA gene was PCR amplified using primers 515F (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACHVGGGTWTCTAAT-3′) and sequenced using Illumina MiSeq. For dataset 1, paired-end 150 bp was used and all samples were sequenced in one run. For dataset 2, paired-end 250 bp was used and samples were sequenced in six runs. Sequence files were de-multiplexed using QIIME2 (core distribution 2018.6)[76] for dataset 1 and Illumina's BCL2FASTQ software on BaseSpace for dataset 2. Fifteen samples in dataset 1 had low sequencing counts and were excluded for present analysis.

Bioinformatics

Forward and reverse primers were trimmed from the 5′-end of sequences using cutadapt v 1.16[77]. After primer trimming, only sequences with lengths of 147–151 bp in dataset 1 and 230–233 bp in dataset 2 were retained. DADA2 R package v 1.8[78] was used for the remaining bioinformatics with default parameters unless when specified. Sequences were quality trimmed and filtered using the filterAndTrim function: trimming 3′-ends to 147 bp (forward) and 147 bp (reverse) in dataset 1, and 228 bp (forward) and 203 bp (reverse) in dataset 2, and removing sequences if they exceeded a maximum of two expected errors.

ASVs were inferred and ASV tables were constructed as follows. For each sequencing run (a) a model for sequencing error was constructed using the learnErrors function specifying that all bases in all sequences be used for constructing the model, (b) sequences were de-replicated to find unique sequences using the derepFastq function, (c) ASVs were inferred from de-replicated sequences using the dada function, (d) forward and reverse sequences were merged using the mergePairs function, and (e) sequences with <250 bp or >256 bp were removed. This resulted in one ASV table for dataset 1 and six ASV tables for dataset 2. The six ASV tables of dataset 2 were merged using the mergeSequenceTables function. Chimeras were detected and removed using the removeBimeraDenovo function.

The following data transformation procedures were used to account for variable sequence depth. Sequence counts were normalized to relative abundances (calculated by dividing the number of sequences that were assigned to a unique ASV or to a genus by the total sequence count in the sample) for PERMANOVA when using Canberra or GUniFrac distance, for MWAS when using KW, and for testing correlation with levodopa drug dose. Centered-log ratio (clr) transformation (using the transform function of the microbiome v 1.2.1 R package (http://microbiome.github.com/microbiome)) was used for PCA and for PERMANOVA when using Aitchison distance. Log ratios (implemented internally in ANCOM and SparCC) were used when using ANCOM for MWAS and for correlation network analysis. Earlier microbiome studies (including our first study conducted with dataset 1)[16] often used rarefaction to normalize the sequence count. Although not as efficient as the other methods due to data loss[79], for added assurance, we rarefied the data, repeated the MWAS with ANCOM and were able to recover all 15 significant PD-associated genera.

Taxonomic assignments were made using SILVA (v 132) in DADA2. MWAS and correlation network analysis were conducted at genus level. To define genera, first each unique ASV was assigned to a genus using the assignTaxonomy function, which performs DADA2's native implementation of the Ribosomal Database Project naive Bayesian classifier[80], using SILVA v 132 as reference and a bootstrap confidence of 80%. Then, each genus (including the unclassified genera) was formed by agglomerating all ASVs that were assigned to that genus using the tax_glom function in phyloseq.

Post MWAS, we explored PD-associated genera at the species level. DADA2 pipeline assigns ASVs to species only if the sequences match 100%. We used the addSpecies function in DADA2 with SILVA as reference and addMultiple=TRUE, first finding 100% matches, then filtering out those matches that did not correspond to the genus given by the assignTaxonomy function. To confirm and expand on DADA2-SILVA species assignments, we BLASTed ASVs against the NCBI 16S rRNA gene sequence database (downloaded on 12/3/2019), and extracted taxonomic designations with the most significant *E* value. Nucleotide BLAST search was performed using the BLAST + executables v 2.9.0 with default parameters[81] (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/).

A phylogenetic tree of ASVs was constructed for each dataset, as described by Callahan et al.[82]. Briefly, multiple sequence alignment of ASVs was performed using the AlignSeqs function from the DECIPHER R package v 2.8.1[83]. Aligned ASVs were then used to build a phylogenetic tree using the phangorn R package v 2.5.3[84].

A phyloseq object was created for each dataset for use in conducting statistical analyses. For each dataset, the ASV table, taxonomic assignments, phylogenetic tree and metadata were merged into a single file, using phyloseq function in phyloseq R package v 1.24.2[85].

Data analysis and statistics

PCA was performed on the clr transformed ASV data[35] using the ordinate function in phyloseq. PC1 and PC2 were plotted using the plot_ordination function in phyloseq (Fig. 1).

We interrogated 47 variables as potential confounders (Supplementary Table 1). In each dataset, we first tested the distribution of each variable in cases vs controls, using Fisher's exact test (fisher.test function in R) for categorical variables, and Mann–Whitney *U* (wilcox.test function in R) for quantitative variables. Variables that differed between cases and control at uncorrected two-sided $P < 0.05$ were tagged as potential confounders, and were then included in PERMANOVA, along with case–control status, and tested for their effects on microbiome composition (Table 1). As PERMANOVA was conducted using marginal effects model without rank (see below), simultaneous inclusion of case–control and other variables allowed testing the association of each variable with microbiome composition while adjusting for all other variables in the model. Thus, PD effect on microbiome composition (β-diversity) was adjusted for variables that differed between cases and controls. Next, variables that were associated with microbiome composition at PERMANOVA $P < 0.05$ were included as covariates in MWAS. Thus, variables that could have led to spurious taxa-disease association because they differed between cases and controls and were also associated with microbiome, were adjusted for in MWAS.

PD medications (also potential confounders) were present only in PD cases and could not be included as covariates in PERMANOVA or MWAS. To gauge the effect of PD on β-diversity independent of each medication, we performed PERMANOVA using cases not on PD medication vs. controls (Table 1). The potential confounding effect of medication on differential abundance of genera was tested post MWAS. For each genus whose relative abundance was associated with PD, we tested the correlation between relative abundance of the genus with daily dose of levodopa (mg/day) using Spearman correlation (two-sided *P* value) implemented in the cor.test function in R.

To investigate changes in the global composition of microbiome (β-diversity) PERMANOVA was used to identify variables that had a significant effect on β-diversity (Table 1). Tests were conducted using adonis2 function in vegan v 2.5.3 (https://CRAN.R-project.org/package=vegan). *P* values were generated by 99,999 permutations which caps at $P < 1E - 5$ as highest significance. Three models were tested as follows:

(Model A) PD vs. control: [Distance ~ case/control]

(Model B) PD vs. control and all variables tagged as potential confounders:

Dataset 1: [Distance ~ case/control + sex + age + geography + BMI + loss of 10 lbs in past year + gastrointestinal discomfort on day of stool collection + constipation in past 3 months + alcohol use + fruits or vegetables daily + stool sample travel time]

Dataset 2: [Distance ~ case/control + sex + age + BMI + loss of 10 lbs in past year + gastrointestinal discomfort on day of stool collection + constipation in past 3 months + alcohol use + stool sample travel time]

(Model C) Subset of PD cases not on a given PD medication vs controls: [Distance ~ case/control] where distance (a measure of (dis)similarity between pairs of samples), age (in years), BMI (kg/m^2), and stool sample travel time (in days) were continuous variables and the remaining variables were categorical. We tested marginal effects, so that each variable was tested while being adjusted for all others in the model, without priority.

To gauge the effect of the distance measure on the results, all three models were tested using Aitchison[35], GUniFrac[37], and Canberra[38] distances. Aitchison distances were calculated by first transforming the ASV data using clr, and then calculating the Euclidean distances using the vegdist function. To calculate GUniFrac distances, unrooted ASV phylogenetic trees were rooted using the root function in the ape v 5.3 R package[86] specifying the unique ASV with the highest raw count as the root, then data were transformed to relative abundances and distances were calculated using the GUniFrac function in the R package GUniFrac v 1.1[37], specifying α to be 0.5. To calculate Canberra distances, data were transformed to relative abundances and distances were calculated using the vegdist function in vegan.

We conducted MWAS to identify the genera whose abundances differed in cases vs. controls. We chose genus classification, because it is the highest resolution attainable with high confidence from 16S sequencing. For statistical analysis of MWAS, we used ANCOM (Table 2 and Supplementary Tables 2–3). We chose ANCOM, because it incorporates compositionality of the microbiome data, has low false-positive rate, and allows covariate adjustment[40,42]. ANCOM was run using ANCOM.main function from the ANCOMv2 R code (https://sites.google.com/site/siddharthamandal1985/research). All genera that were detected in each dataset were included in ANCOM MWAS. Sequence counts were transformed to log ratios, as implemented in ANCOM.

Case/control status was specified as the main variable. For each dataset, the variables that were significant at $P < 0.05$ in PERMANOVA were included as covariates to be adjusted, as follows:

Dataset 1: [Genus ~ case/control + sex + age + geography + gastrointestinal discomfort on day of stool collection + fruits or vegetables daily + stool sample travel time]

Dataset 2: [Genus ~ case/control + sex + age + BMI + constipation in past 3 months]

where genus (ASV counts assigned to a genus, transformed to log ratios by ANCOM), age (in years), BMI (kg/m^2), and stool sample travel time (in days) were continuous variables and the remaining variables were categorical. We used the taxa-wise FDR option (multcorr = 2) and set significance level to FDR < 0.05 to generate *W* statistics, and threshold of 0.8 for declaring an association as significant.

For comparison, we repeated the MWAS using KW as statistical test (Table 2 and Supplementary Tables 4–5). For KW, genera counts were transformed to genera relative abundances. Unclassified genera, and genera present in <10% of samples were excluded from KW MWAS. KW does not allow covariate adjustment. The kruskal.test function from the stats R package was used to test for significance. *P* values were two-sided and corrected for multiple testing using Benjamini–Hochberg FDR method implemented in the p.adjust function from stats package.

To visualize the distribution of genera that were significant in MWAS (Fig. 2), boxplots were created using ggplot2 v 3.1.0 (https://ggplot2.tidyverse.org) with a pseudo-count of 1 added to counts before transforming to relative abundances to avoid taking the log of zero during plotting.

Correlation network analysis was performed for each dataset, and for cases and controls separately (Fig. 3 and Supplementary Fig. 1). Pairwise correlations were calculated between all genera, microbiome-wide, using log-ratio transformed relative abundances as implemented in the SparCC[43] (https://bitbucket.org/yonatanf/sparcc). Significance of each correlation was determined by pseudo P values based on 3000 permutations. Correlation networks were visualized by plotting all genera, microbiome-wide, and connecting correlated genera with an edge, using the program Gephi v 0.9.2[87]. We chose a minimum correlation (r) of 0.4 to connect two genera with an edge to create the graphic. All correlations $r \geq 0.4$ were significant at $P < 3\text{E} - 4$, which is the maximum significance attainable with 3000 permutations. To better visualize networks of connected genera, we first used the force-directed algorithm, Force Atlas 2[88], then a community detection algorithm[89] as implemented in Gephi's modularity function.

DATA AVAILABILITY

Individual-level raw sequences and basic metadata are publicly available at NCBI Sequence Read Archive (SRA) BioProject ID PRJNA601994. Summary statistics are provided in Supplementary Tables 2–5.

CODE AVAILABILITY

No custom codes were used. All software and packages, their versions, relevant specification and parameters are stated in the "METHODS" section.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of Neurological Disorders and Stroke grant R01 NS036960 (to H.P.), The US Army Medical Research Materiel Command endorsed by the US Army through the Parkinson’s Research Program Investigator-Initiated Research Award under Award number W81XWH1810508 (to H.P.); NIH Udall grant P50 NS062684 (to C.P.Z.) and P50 NS108675 (to D.G.S.), and NIH Training Grant T32NS095775 (to Z.D.W.). Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the US Army or the NIH.

COMPETING INTERESTS

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

All authors met all four criteria (1) substantial contributions to the conception (H.P.) or design (H.P., Z.D.W., S.A.F., E.M., C.P.Z., and D.G.S.) of the work or the acquisition (H.P., S.A.F., E.M., C.P.Z., D.G.S., and M.N.D.), analysis (H.P., Z.D.W., M.A., and C.L.S.) or interpretation of the data (H.P., Z.D.W., S.A.F., E.M., C.P.Z., D.G.S., and C.L.S.); (2) drafting the work (H.P. and Z.D.W.) or revising it critically for important intellectual content (all authors); (3) final approval of the completed version (all authors); (4) accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved (all authors).

REFERENCES

1. Chang, D. et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson’s disease risk loci. Nat. Genet. 49, 1511–1516 (2017).

2. Tanner, C. M. Advances in environmental epidemiology. Mov. Disord. 25(Suppl 1), S58–S62 (2010) .

3. Cannon, J. R. & Greenamyre, J. T. Gene-environment interactions in Parkinson’s disease: specific evidence in humans and mammalian models. Neurobiol. Dis. 57, 38–46 (2013).

4. Hamza, T. H. et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson’s disease modifier gene via interaction with coffee. PLoS Genet. 7, e1002237 (2011).

5. Hill-Burns, E. M. et al. A genetic basis for the variable effect of smoking/nicotine on Parkinson’s disease. Pharmacogenomics J. 13, 530–537 (2013).

6. Biernacka, J. M. et al. Genome-wide gene-environment interaction analysis of pesticide exposure and risk of Parkinson’s disease. Parkinsonism Relat. Disord. 32, 25–30 (2016).

7. Schmidt, T. S. B., Raes, J. & Bork, P. The human gut microbiome: from association to modulation. Cell 172, 1198–1215 (2018).

8. Chen, H. et al. Meta-analyses on prevalence of selected Parkinson’s nonmotor symptoms before and after diagnosis. Transl. Neurodegener. 4, 1 (2015).

9. Houser, M. C. et al. Stool immune profiles evince gastrointestinal inflammation in Parkinson’s disease. Mov. Disord. 33, 793–804 (2018).

10. Forsyth, C. B. et al. Increased intestinal permeability correlates with sigmoid mucosa alpha-synuclein staining and endotoxin exposure markers in early Par- kinson’s disease. PLoS ONE 6, e28032 (2011).

11. Scheperjans, F. et al. Gut microbiota are related to Parkinson’s disease and clinical phenotype. Mov. Disord. 30, 350–358 (2015).

12. Keshavarzian, A. et al. Colonic bacterial composition in Parkinson’s disease. Mov. Disord. 30, 1351–1360 (2015).

13. Hasegawa, S. et al. Intestinal dysbiosis and lowered serum lipopolysaccharide-binding potein in Parkinson’s disease. PLoS ONE 10, e0142164 (2015).

14. Unger, M. M. et al. Short chain fatty acids and gut microbiota differ between patients with Parkinson's disease and age-matched controls. Parkinsonism Relat. Disord. https://doi.org/10.1016/j.parkreldis.2016.08.019 (2016).

15. Sampson, T. R. et al. Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease. Cell 167, 1469–1480. e1412 (2016).

16. Hill-Burns, E. M. et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. Mov. Disord. 32, 739–749 (2017).

17. Bedarf, J. R. et al. Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naive Parkinson's disease patients. Genome Med. 9, 39 (2017).

18. Petrov, V. A. et al. Analysis of gut microbiota in patients with Parkinson's disease. Bull. Exp. Biol. Med. 162, 734–737 (2017).

19. Li, W. et al. Structural changes of gut microbiota in Parkinson's disease and its correlation with clinical features. Sci. China Life Sci. 60, 1223–1233 (2017).

20. Hopfner, F. et al. Gut microbiota in Parkinson disease in a northern German cohort. Brain Res. 1667, 41–45 (2017).

21. Lin, A. et al. Gut microbiota in patients with Parkinson's disease in southern China. Parkinsonism Relat. Disord. 53, 82–88 (2018).

22. Qian, Y. et al. Alteration of the fecal microbiota in Chinese patients with Parkinson's disease. Brain Behav. Immun. 70, 194–202 (2018).

23. Heintz-Buschart, A. et al. The nasal and gut microbiome in Parkinson's disease and idiopathic rapid eye movement sleep behavior disorder. Mov. Disord. 33, 88–98 (2018).

24. Weis, S. et al. Effect of Parkinson's disease and related medications on the composition of the fecal bacterial microbiota. NPJ Parkinsons Dis. 5, 28 (2019).

25. Barichella, M. et al. Unraveling gut microbiota in Parkinson's disease and atypical parkinsonism. Mov. Disord. 34, 396–405 (2019).

26. Pietrucci, D. et al. Dysbiosis of gut microbiota in a selected population of Parkinson's patients. Parkinsonism Relat. Disord. https://doi.org/10.1016/j.parkreldis.2019.06.003 (2019).

27. Aho, V. T. E. et al. Gut microbiota in Parkinson's disease: temporal stability and relations to disease progression. EBioMedicine 44, 691–707 (2019).

28. Lin, C. H. et al. Altered gut microbiota and inflammatory cytokine responses in patients with Parkinson's disease. J. Neuroinflammation 16, 129 (2019).

29. Li, F. et al. Alteration of the fecal microbiota in North-Eastern Han Chinese population with sporadic Parkinson's disease. Neurosci. Lett. 707, 134297 (2019).

30. Li, C. et al. Gut microbiota differs between Parkinson's disease patients and healthy controls in Northeast China. Front Mol. Neurosci. 12, 171 (2019).

31. Gerhardt, S. & Mohajeri, M. H. Changes of colonic bacterial composition in Parkinson's disease and other neurodegenerative diseases. Nutrients 10, https://doi.org/10.3390/nu10060708 (2018).

32. Boertien, J. M., Pereira, P. A. B., Aho, V. T. E. & Scheperjans, F. Increasing comparability and utility of gut microbiome studies in Parkinson's disease: a systematic review. J. Parkinsons Dis. 9, S297–S312 (2019).

33. Falony, G. et al. Population-level analysis of gut microbiome variation. Science 352, 560–564 (2016).

34. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352, 565–569 (2016).

35. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. Front. Microbiol. 8, 2224 (2017).

36. Powers, K. et al. Combined effects of smoking, coffee and NSAIDs on Parkinson's disease risk. Mov. Disord. 23, 88–95 (2008).

37. Chen, J. et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28, 2106–2113 (2012).

38. Lance, G. N. & Williams, W. T. Computer programs for hierarchical polythetic classification ("similarity analyses"). Computer J. 9, 60–64 (1966).

39. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26, 32–46 (2001).

40. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. Micro. Ecol. Health Dis. 26, 27663 (2015).

41. Hollander, M. & Wolfe, D. A. Nonparametric Statistical Methods 115–120 (John Wiley & Sons, 1973).

42. Weiss, S. et al. Normalization and microbial differential abundance strategies

depend upon data characteristics. Microbiome 5, 27 (2017).

43. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. PLoS Comput. Biol. 8, e1002687 (2012).

44. Citron, D. M., Goldstein, E. J., Merriam, C. V., Lipsky, B. A. & Abramson, M. A. Bacteriology of moderate-to-severe diabetic foot infections and in vitro activity of antimicrobial agents. J. Clin. Microbiol. 45, 2819–2828 (2007).

45. Wagner Mackenzie, B. et al. Bacterial community collapse: a meta-analysis of the sinonasal microbiota in chronic rhinosinusitis. Environ. Microbiol. 19, 381–392 (2017).

46. Choi, Y. et al. Co-occurrence of anaerobes in human chronic wounds. Micro. Ecol. 77, 808–820 (2019).

47. Hamer, H. M. et al. Review article: the role of butyrate on colonic function. Aliment Pharm. Ther. 27, 104–119 (2008).

48. Canani, R. B. et al. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J. Gastroenterol. 17, 1519–1528 (2011).

49. O'Callaghan, J. & O'Toole, P. W. Lactobacillus: host-microbe relationships. Curr. Top. Microbiol. Immunol. 358, 119–154 (2013).

50. O'Callaghan, A. & van Sinderen, D. Bifidobacteria and their role as members of the human gut microbiota. Front. Microbiol. 7, 925 (2016).

51. Reid, G. The scientific basis for probiotic strains of Lactobacillus. Appl Environ. Microbiol. 65, 3763–3766 (1999).

52. Suez, J., Zmora, N., Segal, E. & Elinav, E. The pros, cons, and many unknowns of probiotics. Nat. Med. 25, 716–729 (2019).

53. Doron, S. & Snydman, D. R. Risk and safety of probiotics. Clin. Infect. Dis. 60(Suppl 2), S129–S134 (2015).

54. Braak, H. et al. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol. Aging 24, 197–211 (2003).

55. Braak, H., Rub, U., Gai, W. P. & Del Tredici, K. Idiopathic Parkinson's disease: possible routes by which vulnerable neuronal types may be subject to neu- roinvasion by an unknown pathogen. J. Neural Transm. (Vienna) 110, 517–536 (2003).

56. Breen, D. P., Halliday, G. M. & Lang, A. E. Gut-brain axis and the spread of alpha-synuclein pathology: Vagal highway or dead end? Mov. Disord. 34, 307–316 (2019).

57. Knudsen, K. et al. In-vivo staging of pathology in REM sleep behaviour disorder: a multimodality imaging case-control study. Lancet Neurol. 17, 618–628 (2018).

58. Svensson, E. et al. Vagotomy and subsequent risk of Parkinson's disease. Ann. Neurol. 78, 522–529 (2015).

59. Liu, B. et al. Vagotomy and Parkinson disease: a Swedish register-based matched-cohort study. Neurology 88, 1996–2002 (2017).

60. Kim, S. et al. Transneuronal propagation of pathologic alpha-synuclein from the gut to the brain models Parkinson's disease. Neuron 103, 627–641 e627 (2019).

61. Stolzenberg, E. et al. A role for neuronal alpha-synuclein in gastrointestinal immunity. J. Innate Immun. https://doi.org/10.1159/000477990 (2017).

62. Tomlinson, J. J. et al. Holocranohistochemistry enables the visualization of alpha-synuclein expression in the murine olfactory system and discovery of its systemic anti-microbial effects. J. Neural Transm. (Vienna) 124, 721–738 (2017).

63. Matheoud, D. et al. Intestinal infection triggers Parkinson's disease-like symptoms in Pink1(-/-) mice. Nature 571, 565–569 (2019).

64. Park, J., Wang, Q., Wu, Q., Mao-Draayer, Y. & Kim, C. H. Bidirectional regulatory potentials of short-chain fatty acids and their G-protein-coupled receptors in autoimmune neuroinflammation. Sci. Rep. 9, 8837 (2019).

65. Haase, S., Haghikia, A., Wilck, N., Muller, D. N. & Linker, R. A. Impacts of microbiome metabolites on immune regulation and autoimmunity. Immunology 154, 230–238 (2018).

66. Furusawa, Y. et al. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. Nature 504, 446–450 (2013).

67. Kang, C. et al. Gut microbiota mediates the protective effects of dietary capsaicin against chronic low-grade inflammation and associated obesity induced by high- fat diet. MBio 8, https://doi.org/10.1128/mBio.00470-17 (2017).

68. Sun, Q., Jia, Q., Song, L. & Duan, L. Alterations in fecal short-chain fatty acids in patients with irritable bowel syndrome: a systematic review and meta-analysis. Med. (Baltimore) 98, e14513 (2019).

69. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55–60 (2012).

70. Guo, Z. et al. Intestinal microbiota distinguish gout patients from healthy humans. Sci. Rep. 6, 20602 (2016).

71. Yamada, T. et al. Rapid and sustained long-term decrease of fecal short-chain fatty acids in critically ill patients with systemic inflammatory response syndrome. JPEN J. Parenter. Enter. Nutr. 39, 569–577 (2015).

72. Zhang, K. & Ni, Y. Tyrosine decarboxylase from Lactobacillus brevis: soluble expression and characterization. Protein Expr. Purif. 94, 33–39 (2014).

73. Maini Rekdal, V., Bess, E. N., Bisanz, J. E., Turnbaugh, P. J. & Balskus, E. P. Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. Science 364, https://doi.org/10.1126/science.aau6323 (2019).

74. Barichella, M. et al. Probiotics and prebiotic fiber for constipation associated with Parkinson disease: An RCT. Neurology 87, 1274–1280 (2016).

75. Gibb, W. R. G. & Lee, A. J. A comparison of clinical and pathological features of young- and old-onset Parkinson disease. Neurology 38 (1988).

76. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857 (2019).

77. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet 17, 10–12 (2011).

78. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583 (2016).

79. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 10, e1003531 (2014).

80. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267 (2007).

81. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. J. Comput Biol. 7, 203–214 (2000).

82. Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J. & Holmes, S. P. Bioconductor workflow for microbiome data analysis: from raw reads to com- munity analyses. F1000Res 5, 1492 (2016).

83. Wright, E. S. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinformatics 16, 322 (2015).

84. Schliep, K. P. phangorn: phylogenetic analysis in R. Bioinformatics 27, 592–593 (2011).

85. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8, e61217 (2013).

86. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289–290 (2004).

87. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. North America. Available at: https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154 (2009).

88. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS ONE 9, e98679 (2014).

89. Blondel, V., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. 2008, P10008 (2008).

EVIDENCE FOR INTERACTION BETWEEN GENETIC VARIATION IN THE *SNCA* LOCUS AND ABUNDANCE OF OPPORTUNISTIC PATHOGENS IN THE PARKINSON DISEASE GUT MICROBIOME

by

ZACHARY D. WALLEN, STEWART A. FACTOR*, ERIC MOLHO*, CYRUS P. ZABETIAN*, DAVID G. STANDAERT* AND HAYDEH PAYAMI

*Authors listed are prospective co-authors for the manuscript in preparation. They have not seen the data nor approved of the text written here.

Manuscript in preparation for *npj Parkinsons Disease*

Format adapted for dissertation

ABSTRACT

Parkinson disease (PD), a progressive, neurodegenerative disease with no current treatments, has been associated with a dysbiotic gut microbiome by a number of studies in human. We recently performed the largest PD-gut microbiome study to date where we detected significant enrichment of three bacterial genera in PD (*Corynebacterium_1, Porphyromonas, Prevotella*) who were part of a highly correlated poly-microbial group of genera also enriched in PD. These genera were defined as potential opportunistic pathogens through literature search. As both presence of pathogens in the gut and genetic variants in and around the *SNCA* gene have been previously shown to increase *SNCA* expression, and increasing dosages of *SNCA* is important in PD pathogenesis as seen in *SNCA* duplication and triplication cases, we hypothesized that the combination of both opportunistic pathogens in the gut and genetic variation in the *SNCA* region might increase the risk of PD. To establish a connection between previously reported opportunistic pathogens and genetic variation in the *SNCA* region, we aimed to detect genetic variation in the *SNCA* region that moderates the associations between PD and previously reported opportunistic pathogens, and then, test if presence of opportunistic pathogens enhanced the detected genetic variants' associations with PD.

Using two PD-gut microbiome datasets with both gut microbiome data and human genotype data (N = 319 and 486), we performed a genetic scan of the *SNCA* region in search of candidate SNPs that moderate the association between PD and the three previously detected opportunistic pathogens, and their poly-microbial group, followed by meta-analysis of results. We then tested candidate SNPs for association with

PD in individuals who were positive for these genera, and in those negative for genera, to determine if SNP associations with PD were enhanced in the presence of targeted genera.

Top hits for candidate SNPs were identified in the 3′ end of *SNCA* for all microbial groups tested. Visualization and quantification of the interactions between SNPs, PD, and tested microbial groups showed obvious interactions for three out of the four microbial groups tested. Testing association of SNPs with PD in the presence or absence of these microbial groups showed an increase in association strength of SNPs with PD in subjects positive for microbes.

INTRODUCTION

Parkinson disease (PD) is a progressive, neurodegenerative disease with currently no disease modifying treatments. Mendelian forms of PD exist, caused by rare mutations in a number of genes, but the vast majority of PD cases are idiopathic. Both genetic [Chang et al. 2017; Nalls et al. 2019] and environmental [Tanner 2010] factors have been identified that associate with increased risk of PD, but none have large enough effect sizes individually, or in combination, to fully explain the cause of PD. Interaction between genetic and environmental factors have also been explored in PD, but these also have failed to fully encapsulate the cause of PD [Hamza et al. 2011; Cannon & Greenamyre 2013; Hill-Burns et al. 2013; Biernacka et al. 2016]. The search continues for the cause of idiopathic PD.

One area of research that the PD field has turned to recently is the gut microbiome. Multiple studies in human have associated a dysbiotic gut microbiome with PD, all finding individual microorganisms significantly enriched or depleted in PD, albeit with varying results [Gerhardt & Mohajeri 2018; Boertien 2019]. Recently, we performed a gut microbiome-wide association study of PD in two of the largest PD-gut microbiome datasets to date where we detected and replicated 15 PD-microorganism associations across the two datasets [Wallen et al. 2020]. Three of the associations were with the bacterial genera *Corynebacterium_1*, *Porphyromonas*, and *Prevotella* (as defined by SILVA v 132 reference database) who were enriched in PD, and, through literature search, were defined as opportunistic pathogens. These genera belonged to a larger poly-microbial group of correlated genera (termed "cluster 1"), which on its own was found to be significantly enriched in PD. Detection of overabundance of opportunistic pathogens in the PD gut was an interesting finding as it harks back to Braak's hypothesis that a yet to be identified pathogen is responsible for causing non-familial forms of PD by invading the brain through the gastrointestinal tract and enteric neurons [Braak et al. 2003; Braak et al. 2003]. This still begs the question of how pathogens in the gut might cause, or at least increase the risk of, PD.

We hypothesized that the combination of both opportunistic pathogens in the gut and genetic variation in and around the *SNCA* gene might increase the risk of PD. Genetic variants in and around *SNCA*, which codes for α-synuclein, the pathological hallmark of PD, are the most highly associated variants with increased PD risk [Chang et al. 2017; Nalls et al. 2019], and have been shown to increase the expression of *SNCA* [GTEx Consortium 2015; Soldner et al. 2016; Emelyanov et al. 2016]. This provides a plausible mechanism of interaction between gut pathogens and genetic risk in the *SNCA* region, as overexpression of α-synuclein has also been seen with infections unrelated to PD [Stolzenberg et al. 2017; Tomlinson et al. 2017]. *SNCA* dosage is important in PD pathogenesis as seen in PD cases with *SNCA* duplications and triplications [Devine, Gwinn, Singleton & Hardy 2011], therefore, having two hits of overexpression of α-synuclein (one from gut, the other from host genetic variation) might increase the risk of disease.

To investigate connections between genetic variation in the *SNCA* region, our previously detected opportunistic pathogens, and PD, we first identified candidate single nucleotide polymorphisms (SNPs) by investigating if any SNPs in the *SNCA* region moderated the association between PD and *Corynebacterium_1*, *Porphyromonas*, *Prevotella,* or cluster 1, then tested those SNPs for association with PD in the presence or absence of *Corynebacterium_1*, *Porphyromonas*, *Prevotella,* or cluster 1. Following this schema, we detected SNPs at the 3′ end of *SNCA* whose associations with PD were strengthened when subjects were positive for target genera compared to those who were negative, or the sample population as a whole, providing the first evidence for a potential gene-gut microbiome interaction in PD.

METHODS

Study approval and participant consents

The study of both human genetic and microbiome data was approved by the institutional review boards at all participating institutions. Written informed consent was obtained from all participants in this study.

Subjects and data collection

This study included two cohorts of PD patients and neurologically healthy controls (referred to as dataset 1 and dataset 2) enrolled as part of the NeuroGenetics Research Consortium (NGRC). Subjects were enrolled from four NGRC-affiliated movement disorder clinics using standardized protocols. For dataset 1, 212 PD patients and 136 controls were enrolled from movement disorder clinics and surrounding areas in Atlanta, GA, Albany, NY, and Seattle, WA in 2014 [Hill-Burns et al. 2017; Wallen et al. 2020]. For dataset 2, 323 PD patients and 184 controls were enrolled from the movement disorder clinic at the University of Alabama at Birmingham and surrounding area [Wallen et al. 2020]. PD was diagnosed using UK Brain Bank criteria by a movement disorder specialist [Gibb & Lee 1988]. Controls were self-reported free of neurological disease.

A summary of metadata collected for dataset 1 and 2 subjects has been previously reported [Wallen et al. 2020; Supplementary Table 1]. Metadata were collected using two questionnaires filled out by each subject: a Gut Microbiome questionnaire (GMQ) and an Environmental and Family History questionnaire (EFQ) [Hill-Burns et al. 2017; Hamza et al. 2011; Powers et al. 2008]. The GMQ, completed shortly after stool sample collection, collected information on variables that might influence results in microbiome analysis including dietary information and gastrointestinal health, while the EFQ collected information on exposure to environmental factors related to PD-risk, family history of PD, and ancestry. Both questionnaires collected basic demographics such as sex and age. Information on PD medication use at the time of stool sample collection was extracted from patients' medical records by clinical investigators.

Data used in this study were derived from two sample types collected from subjects: stool samples to derive data on gut microbial abundances, and blood/saliva samples to derive human genotype data. Stool samples were collected by subjects at home, swabbing excreted stool with a DNA/RNA-free, sterile cotton swab (BD BBL CultureSwab Sterile/Media-free Swabs, Fisher Scientific, Pittsburgh, PA), then immediately shipping the sample via United States Postal Service in ambient temperature. Stool samples were immediately placed in −20 °C upon arrival, and stored there until DNA extraction.

Extraction and sequencing of microbial DNA

Extraction and sequencing of microbial DNA for datasets 1 and 2 were completed in different laboratories, but methods used were matched to the best of our ability to minimize any technical variation. For dataset 1, the MoBio PowerMag Soil DNA Isolation Kit (optimized for KingFisher) was used for extraction of DNA from stool, while the MoBio PowerSoil DNA Isolation Kit was used for dataset 2 extractions (MoBio Industries, Carlsbad, CA). Both kits have equivalent chemistries. To avoid any batch effects from sequencing, samples from PD patients and control subjects were randomized when prepping for sequencing. Sequencing was performed on the MiSeq platform (Illumina, San Diego, CA) targeting PCR amplicons of the bacterial/archaeal 16S rRNA gene hypervariable region 4 (V4; using primers 515F and 806R). Paired end 150bp and 250bp sequencing was used for dataset 1 and 2 respectively. Samples for dataset 1 were all sequenced in one run, while samples for dataset 2 were sequenced in 6 runs resulting in ~10x greater sequencing depth per sample in dataset 2 when compared to dataset 1 samples. QIIME2 (core distribution 2018.6) [Boleyn et al. 2019] and BCL2FASTQ (Illumina, San Deigo, CA) were used to demultiplex pooled sequence files for dataset 1 and 2 respectively. Fifteen samples in dataset 1 resulted in too low of sequences to analyze, therefore, they were excluded before beginning bioinformatics of sequences.

Bioinformatics of microbial amplicon sequences

The bioinformatic pipeline for processing dataset 1 and 2 16S rRNA V4 amplicons has been previously described in detail [Wallen et al. 2020]. Bioinformatics were performed separately for each dataset. The major bioinformatic steps for processing amplicon sequences to unique amplicon sequence variants (ASVs) and corresponding abundances included the following: (1) removal of remaining PCR primers from sequences using cutadapt v 1.16 [Martin 2011], (2) quality trimming and filtering of sequences using the filterAndTrim function from DADA2 v 1.8 [Callahan et al. 2016], (3) inference of unique ASVs and their abundances using the learnErrors, derepFastq, and dada functions from DADA2, (4) merging of forward and reverse sequence pairs using mergePairs function from DADA2, (5) filtering merged sequences for those between 250-256 bp in length, and (6) removal of chimeric sequences using the removeBimeraDenovo function from DADA2. Unique ASVs were then given taxonomic assignments via the assignTaxonomy function from DADA2 using SILVA v 132 as a reference. For each dataset, unique ASV abundances, taxonomy assignments, subject metadata, and a phylogenetic tree of unique ASVs (created using DECIPHER v 2.8.1 [Wright 2015] and phangorn v 2.5.3 [Schliep 2011]) were merged into a single phyloseq object file using the phyloseq function from phyloseq v 1.24.2 [McMurdie & Holmes 2013].

All analyses reported in this study were performed at the genus level for three specific genera (*Corynebacterium_1*, *Porphyromonas*, and *Prevotella*) and one poly-microbial group of correlated genera (previously defined in Wallen et al. 2020) that will be referred to as "cluster 1". Members of cluster 1 for dataset 1 included *Porphyromonas, Prevotella, Anaerococcus, Ezakiella, Varibaculum, Campylobacter, Peptoniphilus, Murdochiella,* and *Finegoldia.* Members of cluster 1 for dataset 2 included genera listed for dataset 1 and additionally *Corynebacterium_1, Fastidiosipila, Lawsonella, Mogibacterium, Negativicoccus, Mobiluncus, S5-A14a, Prevotella_6,* and unclassified *Corynebacteriaceae.* To define genera, unique ASVs and their corresponding abundances were agglomerated into their assigned genus using the tax_glom function from phyloseq without removal of unclassified genera. Then, genera abundances with a pseudo-count of 1 added were either transformed using the centered-log ratio (clr) transformation [Aitchison 1986] when performing interaction analyses with genotype data, or transformed to relative abundances when creating plots. The clr transformation was chosen for interaction analyses because it accounts for multiple characteristics of microbiome data that make it difficult to analyze with standard statistical methods (non-normality, inter-sample variation in sequencing depth, compositionality) [Gloor et al. 2017]. Downstream statistical analyses were then performed for *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, and cluster 1 separately (see sub-section "*Statistical analysis"*).

Human genotype data and quality control

Genome-wide genotypes were generated for dataset 1 and 2 subjects from human DNA extracted from blood or saliva samples using three different genotyping arrays: HumanOmni1-Quad_v1-0_B BeadChip, Infinium Multi-Ethnic EUR/EAS/SAS-8 Kit, and Infinium Global Diversity Array-8 v1.0 Kit (Illumina, San Diego, CA). Genotyping and quality control (QC) of SNP genotypes are described below for each of the array groups separately. Unless otherwise specified, QC was performed using PLINK 1.9 (v1.90b6.16) [Chang et al. 2015].

HumanOmni1-Quad_v1-0_B BeadChip: Approximately 70% of dataset 1 subjects (N=244; referred to as dataset 1.1) were previously genome-wide genotyped using the HumanOmni1-Quad_v1-0_B BeadChip for a GWAS of PD in 2010 resulting in genotypes for 1,012,895 SNPs after removal of failed SNPs [Hamza et al. 2010]. In addition to this, subjects were also genotyped using the Illumina Immunochip resulting in genotypes for 202,798 SNPs. Genotyping for both arrays was performed at the Johns Hopkins Center for Inherited Disease Research (CIDR). Quality control of genotype data had been previously performed using PLINK v1.07 [Hamza et al. 2010], therefore, this process was redone using an updated version of PLINK (v1.9). The mean non-Y chromosome call rate for samples in both arrays was 99.9%. Calculation of identity-by-descent in PLINK using HumanOmni genotypes revealed no cryptic relatedness between samples (PI_HAT > 0.15). A subset of SNP mappings were in NCBI36/hg18 build, and were converted to GRCh37/hg19 using the liftOver executable and hg18ToHg19.over.chain.gz chain file from UCSC genome browser (downloaded from https://hgdownload.soe.ucsc.edu/downloads.html). SNP filtering for both HumanOmni and Immunochip genotypes included removal of SNPs with call rate < 99%, Hardy-Weinberg equilibrium (HWE) *P* value < 1E-6, minor allele frequency (MAF) < 0.01, and MAF difference between sexes > 0.15. HumanOmni and Immunochip data were then merged, and SNPs with significant differences in PD patient and control missing rates (*P* < 1E-5) and duplicate SNPs were removed. To remove duplicate SNPs, we first checked the genotype concordance between duplicated SNPs. If duplicate SNPs were concordant, we took the SNP with the lowest missing rate, or the first listed SNP if missing rates were the same. If duplicate SNPs were discordant, we removed both SNPs as we do not know which SNP is correct. After QC, the remaining number of SNPs for dataset 1.1 was 910,083 with a mean call rate of 99.8%.

Infinium Multi-Ethnic EUR/EAS/SAS-8 Kit: (Illumina, San Diego, CA) Approximately 30% of dataset 1 subjects (N=90; referred to as dataset 1.2) were not included in the 2010 PD GWAS. These samples were genome-wide genotyped at a later time using the Infinium Multi-Ethnic EUR/EAS/SAS-8 array at HudsonAlpha Institute for Biotechnology. Raw genotyping intensity files were received from the genotyping laboratory and uploaded to GenomeStudio v 2.0.4 (Illumina, San Diego, CA) where genotype cluster definitions and calls were determined for each SNP using intensity data from all samples. The GenCall (genotype quality score) threshold for calling SNP genotypes was set at 0.15, and SNPs that resulted in a genotype cluster separation < 0.2 were zeroed out for their genotype. Genotypes for 1,649,668 SNPs were then exported from GenomeStudio using the PLINK plugin v 2.1.4, and converted to PLINK binary files for further QC. The mean non-Y chromosome call rate for samples was 99.8%. Calculation of identity-by-descent revealed no cryptic relatedness among samples (PI_HAT < 0.15). A subset of SNP mappings were in GRCh38/hg38 build, and were converted to GRCh37/hg19 using the liftOver executable and hg38ToHg19.over.chain.gz chain file. The same SNP filtering criteria was implemented here as it was for dataset 1.1 genotypes: call rate < 99%, HWE *P* value < 1E-6, MAF < 0.01, MAF difference between sexes > 0.15, significant differences in PD patient and control missing rates (*P* < 1E-5), and duplicate SNPs. The remaining number of SNPs for dataset 1.2 was 749,362 with a mean call rate of 100%. To avoid batch effects due to differences in the genotyping array used for dataset 1.2, we did not attempt to merge genotype data for dataset 1 subjects and instead analyzed them separately for analyses involving genotype data as dataset 1.1 and dataset 1.2. Results from these two groups, along with dataset 2, were later combined together through meta-analyses.

Infinium Global Diversity Array-8 v1.0 Kit: (Illumina, San Diego, CA) A subset of dataset 2 subjects (N=486) were genotyped at CIDR using the Infinium Global Diversity Array, the newest genome-wide genotyping array from Illumina and the commercial version of the microarray chosen by the All of Us Research Program (https://allofus.nih.gov/). Genotype clusters were defined using GenomeStudio v 2011.1 and 99% of the genotyped samples. Genotypes were not called for SNPs with GenCall score <0.15, and failure criteria for autosomal and X chromosome SNPs included the following: call rate < 85%, MAF ≤ 1% and call rate < 95%, heterozygote rate ≥ 80%, cluster separation < 0.2, any positive control replicate errors, absolute difference in call rate between genders > 10% (autosomal only), absolute difference in heterozygote rate between genders > 30% (autosomal only), and male heterozygote rate greater than 1% (X only). All Y chromosome, XY pseudo-autosomal region (PAR), and mitochondrial SNPs were manually reviewed. Genotypes for 1,827,062 SNPs were released in the form of PLINK binary files. The mean non-Y chromosome call rate for samples was 99.2%. Calculation of identity-by-descent showed two subjects were genetically related as a parent and offspring (PI_HAT = 0.5), which we had already noted previously. We decided not to exclude these subjects from the study as the target genera of this study, and the majority of the gut microbiome as a whole, were not found to have high heritability in a recent, large meta-analysis of the gut microbiome and human genetics [Kurilshikov et al. 2020], therefore, analyses would most likely not be affected by the genetic relationship between these two subjects. The same SNP filtering criteria was implemented here as it was for dataset 1 genotypes: call rate < 99%, HWE *P* value < 1E-6, MAF < 0.01, MAF difference between sexes > 0.15, significant differences in PD patient and control missing rates (*P* < 1E-5), and duplicate SNPs. The remaining number of SNPs for dataset 2 was 783,263 with a mean call rate of 99.9%.

After QC, three genotype datasets were available for analysis and will be referred to as dataset 1.1 (dataset 1 subjects genotyped using HumanOmni1-Quad_v1-0_B BeadChip and Immunochip), dataset 1.2 (dataset 1 subjects genotyped using the Infinium Multi-Ethnic EUR/EAS/SAS-8 array), and dataset 2 from this point on.

Principal component analysis (PCA) with 1000 Genomes Phase 3 reference genotypes was performed separately for datasets 1.1, 1.2, and 2 genotypes in order to determine genetic ancestry of subjects, and to view any potential outlying samples. Study genotypes were first merged with 1000 Genomes Phase 3 genotypes (previously filtered for non-triallelic SNPs and SNPs with MAF > 5%) using GenotypeHarmonizer v 1.4.23 [Deelen et al. 2014] and PLINK. Merged genotypes were then linkage disequilibrium (LD) pruned as previously described [Hamza et al. 2010], resulting in a mean LD pruned subset of SNP of ~148,000. Principal components were then calculated using pruned SNPs and the top two PCs were plotted for each genotype dataset using ggplot2 (Figure 1). PCA reflected what was previously recorded for subject's self-reported race [reported in Wallen et al. 2020, Supplementary Table 1]. All samples fell within a defined 1000 Genomes superpopulation with no evident outlying samples. No samples were removed based on PCAs as the majority of samples fell within the European superpopulation and the remaining samples (1% of the total samples across datasets), although differing in ancestral origin, were unlikely to have a significant influence on results, again, due to the low heritability of the gut microbiome [Kurilshikov et al. 2020].

Imputation of genotypes

To increase the breadth of SNPs available for analysis, genotypes for each genotyping dataset were submitted for imputation of additional SNP genotypes using the Trans-Omics for Precision Medicine (TOPMed) Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov), which uses the newest and largest reference panel to date derived from the TOPMed program [Taliun et al. 2019], and is based off of the widely used Michigan Imputation Server that implements Minimac4 for imputation [Das et al. 2016]. The TOPMed reference panel only included autosomal and

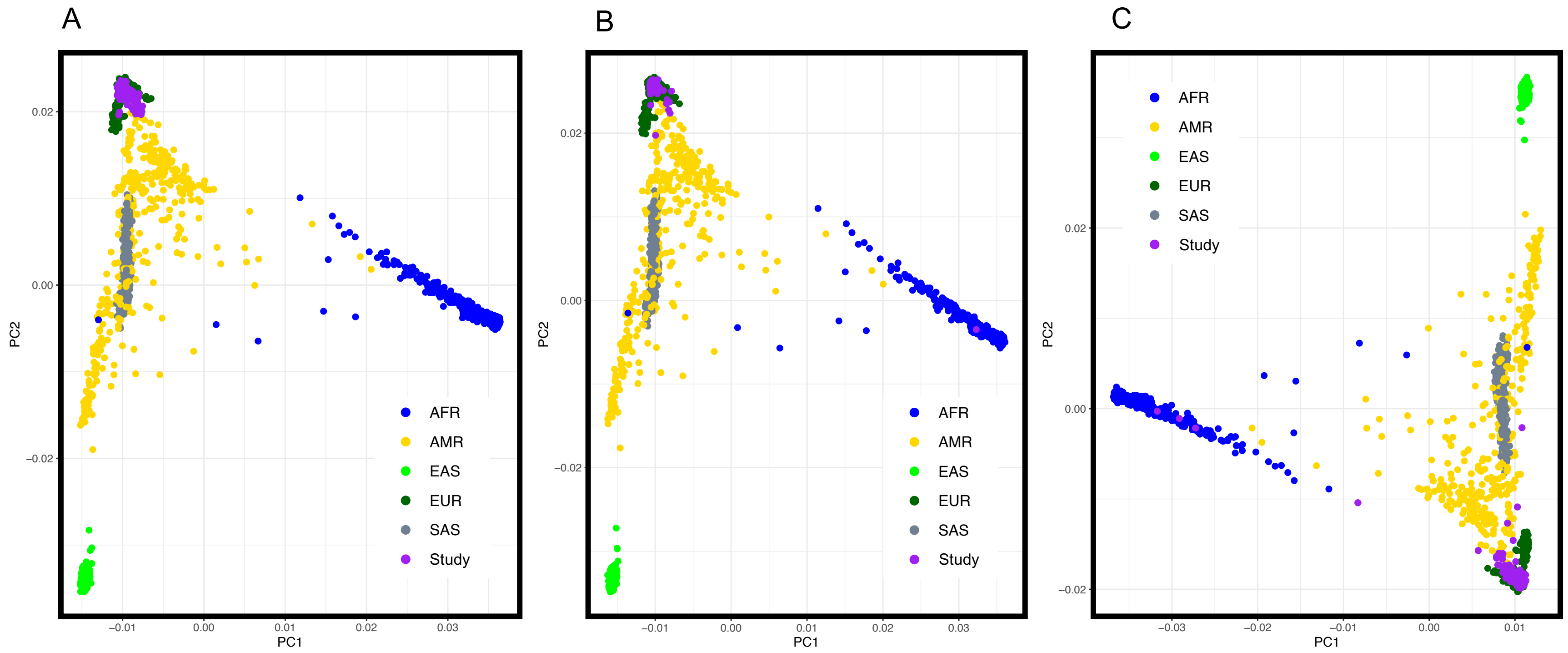


Figure 1. Principal component analysis of study genotypes with 1000 Genome Phase 3 superpopulations.

Principal component analysis was performed on merged, LD pruned study and 1000 Genome Phase 3 genotypes for dataset 1.1 genotypes (A), dataset 1.2 genotypes (B), and dataset 2 genotypes (C). All study subjects fall within a defined 1000 Genome Phase 3 superpopulation with no obvious outlying samples. AFR: African superpopulation from 1000 Genomes; AMR: Admixed American superpopulation from 1000 Genomes; EAS: East Asian superpopulation from 1000 Genomes; EUR: European superpopulation from 1000 Genomes; SAS: South Asian from 1000 Genomes; Study: study genotypes from either dataset 1.1, 1.2, or 2.

X chromosomes, therefore, for each genotype dataset, SNPs in the PAR of chromosome X were merged with the rest of the X chromosome SNPs. Coordinates for SNPs in all datasets were converted to GRCh38/hg38 using the liftOver executable and hg19ToHg38.over.chain.gz chain file as the TOPMed reference panel is only available in GRCh38/hg38 coordinates. SNP mappings were then checked and corrected for use with TOPMed reference panels using the utility scripts HRC-1000G-check-bim.pl (v4.3.0) and CreateTOPMed.pl (downloaded from https://www.well.ox.ac.uk/~wrayner/tools/), and a TOPMed reference file ALL.TOPMed_freeze5_hg38_dbSNP.vcf.gz (downloaded from https://bravo.sph.umich.edu/freeze5/hg38/download). Running of these utility scripts resulted in a series of PLINK commands to correct genotypes files for concordance with TOPMed by excluding SNPs that did not have a match in TOPMed, mitochondrial SNPs, palindromic SNPs with frequency > 0.4, SNPs with non-matching alleles to TOPMed, indels, and duplicates. Genotype files were then converted to variant call format (VCF) for submission to the TOPMed Imputation Server.

Genotype VCF files were submitted to the TOPMed Imputation Server using the following parameters: reference panel TOPMed version r2 2020, array build GRCh38/hg38, r^2 filter threshold 0.3, Eagle v2.4 for phasing, skip QC frequency check, and run in QC & imputation mode. With these parameters, the imputation server first performed quality control of SNPs excluding those that contained invalid alleles, duplicates, indels, monomorphic sites, those with allele mismatches between TOPMed panel and submitted data, and those with < 90% call rate. Phasing was then performed using Eagle v2.4 [Loh et al. 2016], which estimated haplotype phases using the Haplotype Reference Consortium reference panel [McCarthy et al. 2016], followed by imputation using Minimac4 [Howie et al. 2012] with TOPMed reference panel. Imputed SNPs were then filtered based on imputation quality score (*r*2), removing those with $r^2 <$ 0.3, which is a commonly used exclusion threshold and the chosen threshold to use in the original manuscript introducing this quality metric [Li et al. 2010]. As stated in the original manuscript, at an r^2 threshold of 0.3 we can expect to remove the majority of low quality imputed SNPs (approximately 70% of these SNPs) while minimizing the removal of higher quality imputed SNPs (approximately 0.5% of these SNPs) [Li et al. 2010]. VCF files with genotypes and imputed dosage data were then outputted by the imputation server and used in statistical analyses. Imputation was successful for all chromosomes in all genotype datasets, and after applying the quality score filter, resulted in 12.3 – 20.7 million imputed SNPs for datasets. Imputed SNPs for all datasets were of high quality, the majority reaching $r^2 > 0.9$ for every chromosome (Table 1).

Statistical analysis

We performed per dataset interaction analyses followed by meta-analysis and visualization of detected interactions to determine if the association of PD with *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, or cluster 1 was moderated by any *SNCA* SNP genotypes. Subjects that had both microbiome and genotype data available were included in the analysis (dataset 1.1 N=231, dataset 1.2 N=88, dataset 2 N=486). The genomic region defined for analysis was 89.6 Mb – 89.9 Mb on chromosome 4 (GRCh38/hg38), which covers the entire *SNCA* gene with a base pair window of approximately ±100 kb. Only SNPs that had an MAF > 5% in their respective datasets were included in the analysis.

Table 1. Total and per chromosome SNP counts for genotype datasets. Chr: Chromosome; r^2: imputation quality metric given by Minimac4 ranging from 0-1.

	Total	Chr 1	Chr 2	Chr 3	Chr 4	Chr 5	Chr 6	Chr 7	Chr 8	Chr 9	Chr 10	Chr 11	Chr 12	Chr 13	Chr 14	Chr 15	Chr 16	Chr 17	Chr 18	Chr 19	Chr 20	Chr 21	Chr 22	Chr X
Dataset 1.1																								
Genotyped	873695	73275	72203	57398	49564	52849	66283	45095	45911	39249	45753	42929	42503	30265	27986	24134	26713	23824	23597	17923	23134	11916	12694	18497
Imputed	17211150	1310484	1414385	1195271	1207220	1081065	1053226	977740	916364	723813	829887	824277	806591	608089	542240	476909	525051	456912	477747	381351	373846	229128	235107	564447
r^2>0.9	14554014 (81%)	1115623 (81%)	1213095 (82%)	1029277 (82%)	1017706 (81%)	934039 (82%)	928528 (83%)	823327 (80%)	783219 (81%)	623382 (82%)	707514 (81%)	703123 (81%)	694564 (82%)	525879 (82%)	465342 (82%)	398931 (80%)	428496 (78%)	368867 (77%)	402920 (80%)	311048 (78%)	313053 (79%)	190176 (79%)	187950 (76%)	387955 (67%)
Dataset 1.2																								
Genotyped	692382	53665	58103	49621	45196	41287	48519	37460	36480	29781	34048	33404	31264	24490	22084	20678	22592	19467	20236	13735	17036	9557	9577	14102
Imputed	12339948	943298	999942	857059	867918	778730	782916	698809	660634	516650	621891	594311	571269	438352	395185	341479	358551	321148	345824	277987	269257	167402	167744	363592
r^2>0.9	8679157 (67%)	672217 (67%)	706478 (67%)	621337 (69%)	621403 (68%)	549126 (67%)	589973 (71%)	486654 (66%)	467481 (67%)	363325 (66%)	455885 (70%)	425640 (68%)	407154 (68%)	313262 (68%)	283530 (68%)	240111 (66%)	236104 (62%)	222619 (65%)	246122 (67%)	178859 (61%)	185955 (65%)	116154 (66%)	112689 (64%)	177079 (47%)
Dataset 2																								
Genotyped	719329	53738	59816	51004	46822	42748	50850	38326	36809	30280	34167	34074	32193	25355	22398	20868	22365	19403	20831	13658	17142	9766	9565	27151
Imputed	20669655	1592836	1702791	1409899	1424937	1298273	1261069	1165353	1098577	866215	1002198	959592	966973	743888	651082	590920	619413	567004	573486	461103	450835	277712	289980	695519
r^2>0.9	12149004 (57%)	926485 (56%)	1016912 (58%)	839488 (57%)	850086 (58%)	760159 (57%)	783657 (60%)	684569 (57%)	645088 (57%)	495860 (55%)	611749 (59%)	564539 (57%)	571512 (57%)	449237 (58%)	390528 (58%)	352109 (58%)	334629 (52%)	320353 (55%)	339817 (57%)	252753 (53%)	257520 (55%)	162150 (56%)	171819 (57%)	367985 (51%)

For each genotype dataset separately, linear regression was performed to test the interaction effect between case/control status and SNP genotype dosages on *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, or cluster 1 adjusting for sex, age, and main effects of case/control status and SNP genotype dosages. PLINK 2's --glm function was used to perform the analysis specifying the model to be the following:

[Taxon ~ SNP + sex + age + case/control + SNP x case/control]

where taxon (clr transformed abundances of either *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, or cluster 1), SNP (additive model; dosages of the minor allele ranging from 0 – 2), and age (in years) were continuous variables and the remaining variables were categorical. PLINK 2 (v2.3 alpha) was used instead of PLINK 1.9 due to PLINK 1.9 not being able to handle genotype dosages with its --linear function. Results for the SNP x case/control interaction variable were extracted from PLINK result outputs, including both betas (β) and corresponding standard errors, in order to be used as input for meta-analysis. Per dataset interaction analyses were then repeated using the same parameters and model specification with the exception of treating SNP genotypes as a dominant genetic model (coded as 0 for homozygous major allele and 1 for minor allele carrier).

Once per dataset interaction tests were completed, meta-analyses for *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, and cluster 1 were performed on SNP x case/control interaction betas and standard errors of the three genotype datasets using METASOFT v2.0.1. Both fixed- and random-effects models based on inverse-variance-weighted effect size were performed along with an additional random-effects model optimized for detecting associations when there is heterogeneity between studies [Han & Eskin 2012]. Heterogeneity estimates (Cochran's Q and *P* value and I^2) were also calculated and outputted with the meta-analysis results. To visualize results and top interactions of the meta-analyses, results were uploaded to LocusZoom [Pruim et al. 2010] using fixed-effects model results if no evidence for heterogeneity between studies was present (Cochran's Q $P \geq 0.1$), else random-effects model results were uploaded. Linkage disequilibrium between visualized SNPs in LocusZoom was based on the "EUR" LD population. Pairwise LD estimates between top meta-analysis SNPs was performed using the LDpair tool with 1000 Genome phase 3 European data from LDlink v4.1 [Machiela & Chanock 2015].

To visualize the interaction between case/control status, SNP genotype, and *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, or cluster 1, boxplots of PD and control relative abundances stratified by the number of minor allele copies (for additive model) or presence/absence of the minor allele (dominant model) were created for top meta-analysis interactions using ggplot2 v3.1.0. Boxplots were created for datasets 1 and 2, and for datasets pooled together. Due to the low relative abundance of genera included in this study [Wallen et al. 2020], and reduced concern of batch effects for analyses with single SNPs, all dataset 1 subjects were included in the same plot to aid in visualization of the interactions. A pseudo-count of 1 was added to genera and cluster 1 counts before transforming to relative abundances to avoid taking the log of zero during plotting. Hard call genotypes were extracted from VCF files for each SNP using PLINK in order to have discrete genotype groups for stratification of boxplots. For each SNP x case/control group, the geometric mean and its standard error were calculated and superimposed over the boxplots. The geometric mean was chosen instead of the standard arithmetic mean as it more accurately describes the central tendency of heavily skewed data, and in practice has shown to be more robust to outlying datapoints [Clark-Carter 2010]. To quantitate any visual differences observed between PD and control groups in the boxplots, the mean relative abundance ratio (MRAR) between PD and control groups was calculated for each genotype group. Wilcoxon rank sum tests were used to test differences between PD and control relative abundances within each genotype group.

To test if genotypes of SNPs in top meta-analysis hits associated with PD irrespective of *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, or cluster 1, Firth's penalized logistic regression was performed using the following model:

[case/control ~ SNP + sex + age]

where SNP (additive model; hard call genotypes ranging from 0 – 2 copies of the minor allele) and age (in years) were continuous variables and the remaining variables were categorical. To test if the association between PD and top meta-analysis SNP genotypes were moderated by presence or absence of *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, or cluster 1, subjects were stratified into two groups: those whose samples were positive for *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, or cluster 1 and those whose samples were negative for the same. Firth's penalized logistic regression was then performed for each group using the same model as stated above. This process was repeated treating SNP genotypes as a dominant genetic model. To maximize power, as testing a dichotomous outcome tends to have lower power than testing a quantitative outcome [Altman & Royston 2006], all logistic regressions were performed using pooled datasets.

RESULTS

Genetic interaction analyses

For all genera and cluster 1, the top meta-analysis hits detected in the 89.6 Mb – 89.9 Mb region of chromosome 4 were located in the 3′ region of *SNCA*, which is the region of *SNCA* most highly associated with risk of idiopathic PD [Nalls et al. 2019]. The most significant top hit using an additive genetic model belonged to *Corynebacterium_1* and rs356229 ($\beta_{\text{dataset 1.1}} = 0.71$, $P_{\text{dataset 1.1}} = 0.02$; $\beta_{\text{dataset 1.2}} = 0.48$, $P_{\text{dataset 1.2}} = 0.23$; $\beta_{\text{dataset 2}} = 0.52$, $P_{\text{dataset 2}} = 0.15$; $\beta_{\text{meta}} = 0.59$, $P_{\text{meta}} = 3\text{E-}3$), followed by *Porphyromonas* and rs10029694 ($\beta_{\text{dataset 1.1}} = 1.06$, $P_{\text{dataset 1.1}} = 0.08$; $\beta_{\text{dataset 1.2}} = 1.22$, $P_{\text{dataset 1.2}} = 0.09$; $\beta_{\text{dataset 2}} = 0.79$, $P_{\text{dataset 2}} = 0.14$; $\beta_{\text{meta}} = 0.98$, $P_{\text{meta}} = 5\text{E-}3$), cluster 1 and rs10029694 ($\beta_{\text{dataset 1.1}} = 1.58$, $P_{\text{dataset 1.1}} = 0.06$; $\beta_{\text{dataset 1.2}} = 0.77$, $P_{\text{dataset 1.2}} = 0.44$; $\beta_{\text{dataset 2}} = 0.88$, $P_{\text{dataset 2}} = 0.13$; $\beta_{\text{meta}} = 1.05$, $P_{\text{meta}} = 0.01$), and *Prevotella* and rs356183 ($\beta_{\text{dataset 1.1}} = -0.61$, $P_{\text{dataset 1.1}} = 0.08$; $\beta_{\text{dataset 1.2}} = -0.37$, $P_{\text{dataset 1.2}} = 0.49$; $\beta_{\text{dataset 2}} = -0.42$, $P_{\text{dataset 2}} = 0.2$; $\beta_{\text{meta}} = -0.49$, $P_{\text{meta}} = 0.03$) (Figure 2). All meta-analysis results were also confirmed by the additional Han & Eskin random-effects meta-analysis performed by METASOFT. Interaction analyses using a dominant genetic model resulted in the same top meta-analysis hits as additive model with similar effect sizes and *P* values. An exception to this was *Prevotella*, whose meta-analysis of dominant model results resulted in a different SNP being tagged as a top hit (rs356228; meta-analysis $\beta = 0.82$, $P = 0.02$) (Figure 3). This SNP had an opposite effect direction from the SNP tagged when using the additive model, potentially showing some heterogeneity in the interaction between PD and SNPs in this region on *Prevotella* relative abundance. SNPs that were tagged in top meta-analysis hits for

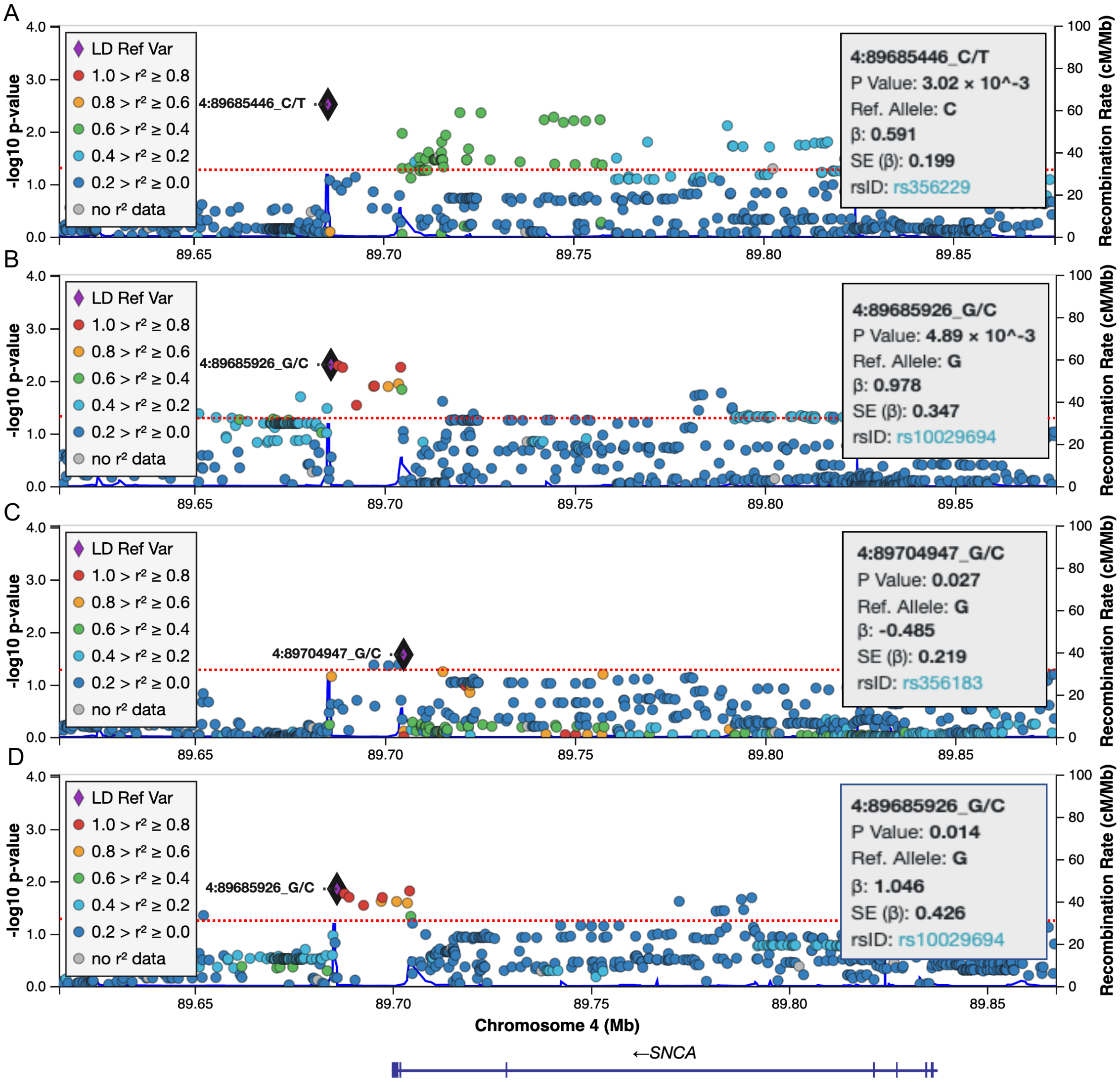


Figure 2: Top hits for interaction meta-analyses in 3′ *SNCA* region under an additive genetic model.

Under an additive model for SNP genotypes, per dataset linear regression was performed to test the interaction effect between case/control status and SNP genotype dosages on *Corynebacterium_1* (A), *Porphyromonas* (B), *Prevotella* (C), or cluster 1 (D) followed by meta-analysis. Sex, age, and main effects of case/control status and SNP genotype were adjusted for in the analyses. Total sample size for the meta-analyses was 513 cases and 292 controls. Results for meta-analyses were visualized using LocusZoom, where each dot represents a SNP plotted according to its -log10(*P* value) and base pair position. Colors of dots correspond to the level of LD (r^2) shared with the top SNP (marked with a diamond), and is detailed in the left legend of each plot. The red dotted line marks the point on the x-axis where $P = 0.05$. LD: linkage disequilibrium; Mb: Megabase; P value: *P* value from meta-analysis; β: beta coefficient from meta-analysis; SE (β): standard error of the meta-analysis beta coefficient; rsID: reference SNP ID for the marked SNPs

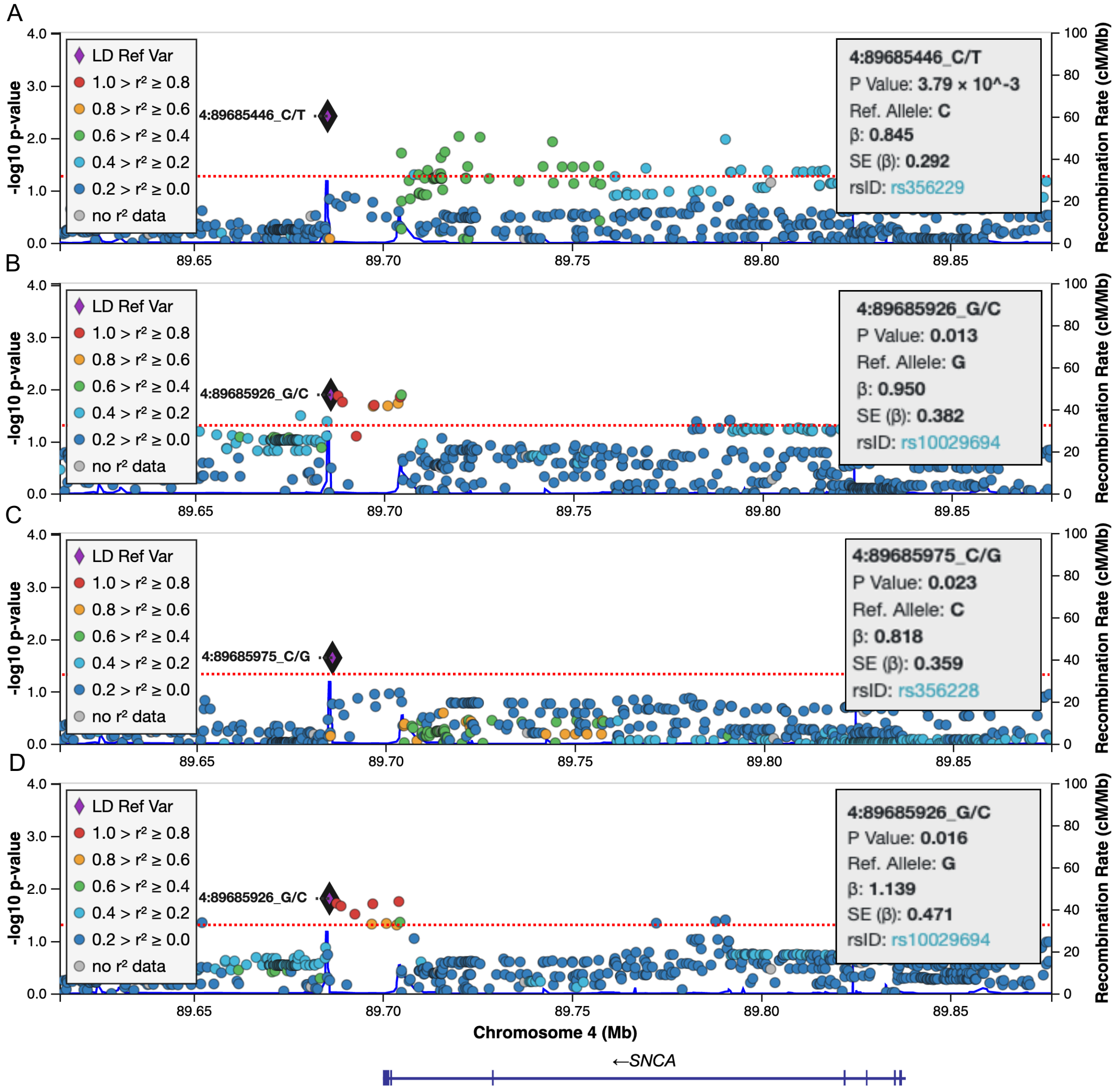


Figure 3: Top hits for interaction meta-analyses in 3′ *SNCA* region under a dominant genetic model.

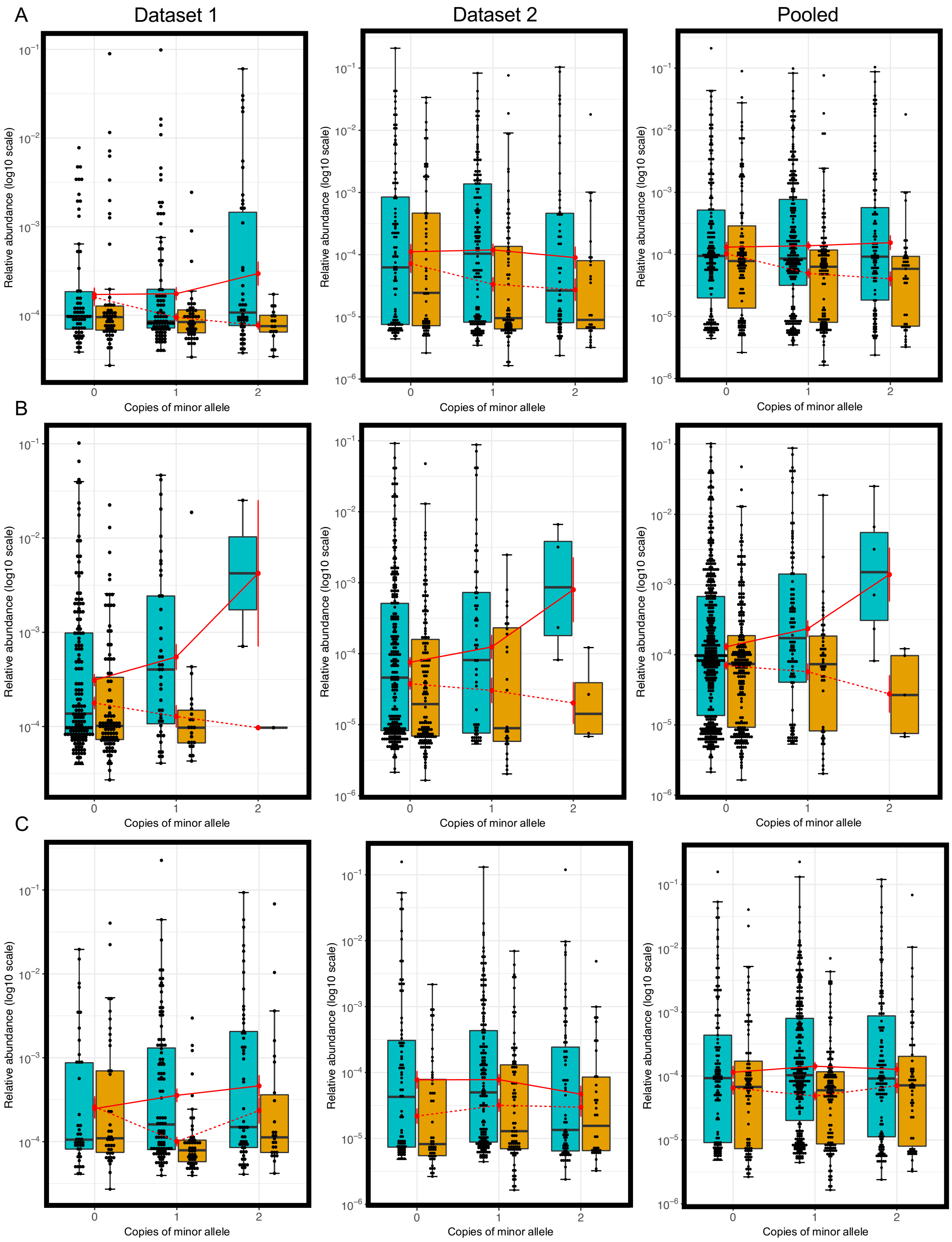
Under a dominant model for SNP genotypes, per dataset linear regression was performed to test the interaction effect between case/control status and SNP genotype on *Corynebacterium_1* (A), *Porphyromonas* (B), *Prevotella* (C), or cluster 1 (D) followed by meta-analysis. Sex, age, and main effects of case/control status and SNP genotype were adjusted for in the analyses. Total sample size for the meta-analyses was 513 cases and 292 controls. Results for meta-analyses were visualized using LocusZoom, where each dot represents a SNP plotted according to its -log10(*P* value) and base pair position. Colors of dots correspond to the level of LD (r^2) shared with the top SNP (marked with a diamond), and is detailed in the left legend of each plot. The red dotted line marks the point on the x-axis where $P = 0.05$. LD: linkage disequilibrium; Mb: Megabase; P value: *P* value from meta-analysis; β: beta coefficient from meta-analysis; SE (β): standard error of the meta-analysis beta coefficient; rsID: reference SNP ID for the marked SNPs

Corynebacterium_1 and *Prevotella* were common (MAF = 0.4 – 0.49) and in LD with one another (D′ = 0.74 – 0.97, r^2 = 0.42 – 0.69). The same SNP was tagged in the top meta-analysis hits for *Porphyromonas* and cluster 1. This SNP was more uncommon (MAF ~ 0.1), and was in LD with rs356228 tagged in dominant model for *Prevotella* (D′ = 1, r^2 = 0.16), but not in LD with the other top SNPs (D′ = 4E-3 – 0.57, r^2 = 0 – 0.05) indicating that this might be a second, but slightly correlated, 3′ *SNCA* signal not only involving the top SNP (rs10029694), but five additional SNPs with similar effect sizes and *P* values (rs3857048, rs3910106, rs59923547, rs3906628, rs3857050) the majority of which were in perfect LD with the top SNP (D′ = 1, r^2 = 1). One round of LD pruning of SNPs in the targeted genomic region resulted in 28 independent SNPs in common between genotype datasets, therefore, a multiple testing corrected study-wide significance threshold for each genus and cluster 1 would be approximately P < 1.8E-3 (0.05 / 28 independent tests). No top meta-analysis hit surpassed this threshold, therefore, the interaction results described here should be considered suggestive as they require further replication.

Visualization of interactions and stratified analysis

Boxplots of interactions between case/control status, SNP genotypes, and genera or cluster 1 relative abundances showed obvious interactions between case/control status and SNP genotype (under both additive and dominant models) on the relative abundances of *Corynebacterium_1, Porphyromonas,* and cluster 1 (Figure 4, A,B,D; Figure 5, A,B,D). In all boxplots, differences between cases and controls in the relative abundances of *Corynebacterium_1, Porphyromonas,* and cluster 1 seemed to be exacerbated with increasing copies of the minor allele, with cases usually having an increase in relative abundances of these taxa with increasing copies of the minor allele, and controls having a decrease. Indeed, quantification of observed differences for each genotype group showed an increase in the MRAR of these taxa between cases and controls with increasing minor allele copies (Table 2) or just for being a carrier of the minor allele (Table 3). The only obvious exception was for dataset 1 relative abundances of cluster 1, where the MRAR of homozygous minor allele cases and controls was actually similar to that of homozygous major allele cases and controls, but this might have been due to extremely low numbers in these groups (N = 2 cases and 1 control) as the dominant model showed an almost 4x increase in MRAR of minor allele carrying cases and controls (Table 3). Unlike the MRARs, the significance of case and control differences, tested using Wilcoxon rank sum tests, did not always increase with increasing copies of the minor allele, which might have been due to lower numbers and higher variances in the homozygous minor allele groups. When carriers of minor alleles were collapsed into one group under a dominant model, differences between cases and controls always resulted in a more significant *P* value than the homozygous major allele groups (Table 3).

For *Prevotella*, both additive and dominant model tagged SNPs did not show an obvious interaction when stratified under an additive model (Figure 4, C), which was reflected in the quantifications of MRARs (Table 2). A slight interaction was observed when stratified under a dominant model for dataset 1 and pooled datasets (Figure 5, C; Table 3), but dataset 2 actually goes in the opposite direction. This again emphasized the potential for heterogeneity in interactions with 3′ *SNCA* genetic variants for *Prevotella*.


A
Dataset 1
Dataset 2
Pooled
B
C
Relative abundance (log10 scale)
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
0
1
2
Copies of minor allele

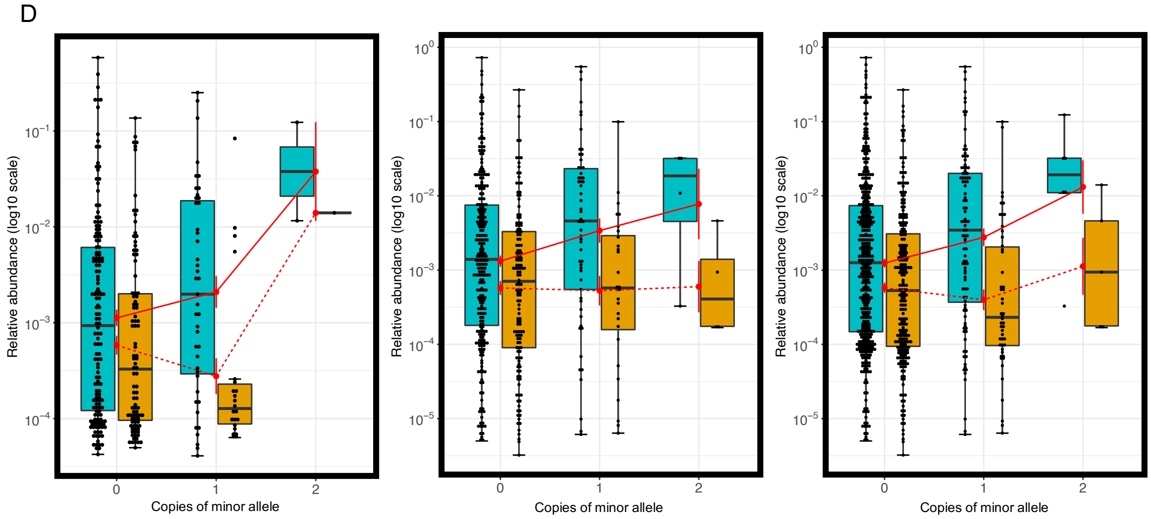
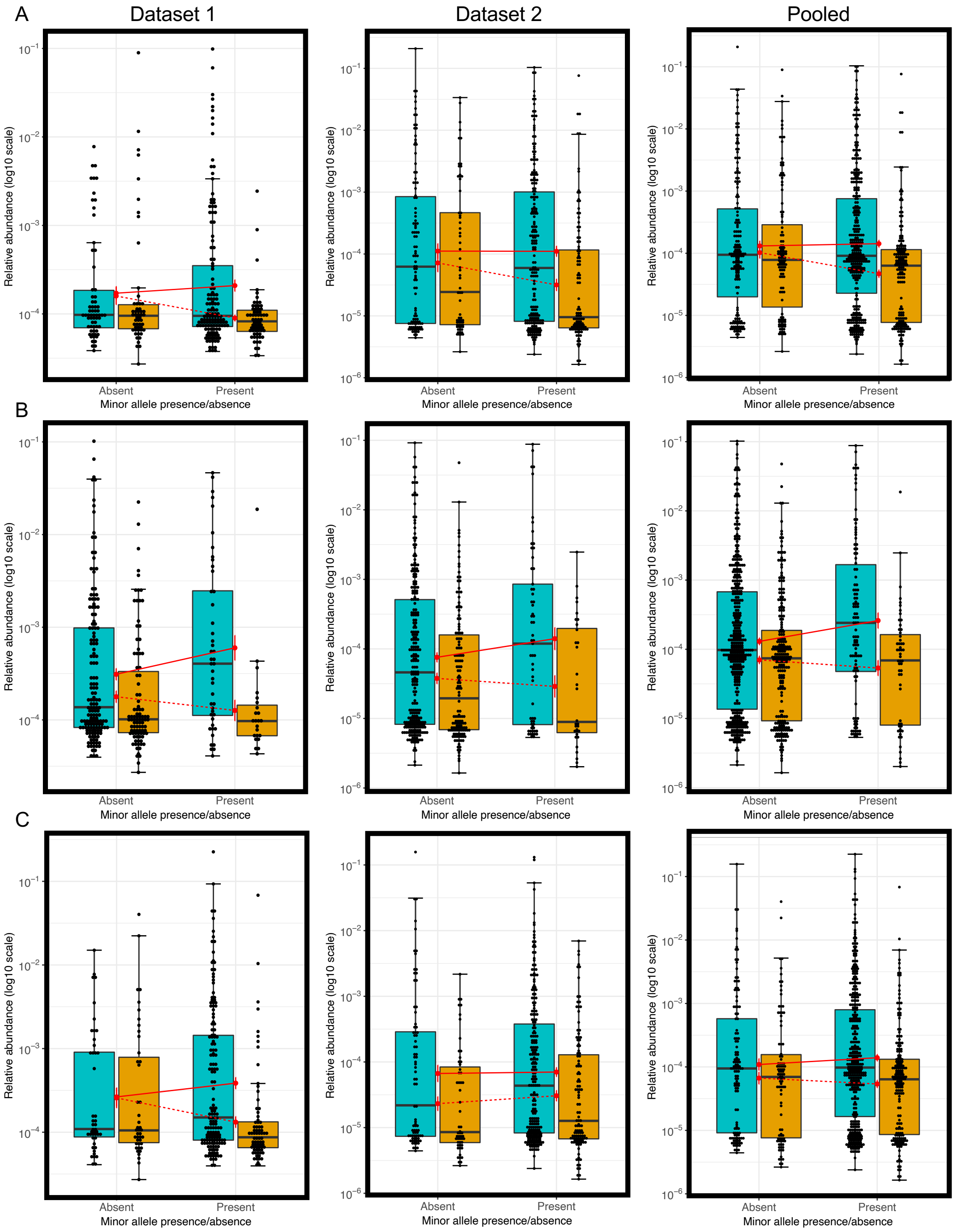


Figure 4: Case and control relative abundances of *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, and cluster 1 stratified by copies of the minor allele of SNPs in top meta-analysis hits.

Case (blue) and control (orange) relative abundances for *Corynebacterium_1* (A), *Porphyromonas* (B), *Prevotella* (C), and cluster 1 (D), stratified by minor allele copy number of top meta-analysis SNPs for each respective genus or cluster 1, were plotted to visualize interactions. Relative abundances were stratified for SNPs rs356229 (A), rs10029694 (B and D), and rs356183 (C). Plots for *Prevotella*'s top SNPs (rs356183 and rs356228) were very similar, therefore, only plots stratified by rs356183 minor allele copies are shown as it was the top hit for additive model meta-analysis. Relative abundances (y-axis) were plotted on log10 scale. Sample size was 201 cases and 118 controls for dataset 1, 312 cases and 174 controls for dataset 2, and 513 cases and 292 controls for pooled datasets. Each dot corresponds to a single subject, plotted according to the relative abundance of either *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, and cluster 1 in their stool sample. The top, middle, and bottom horizontal lines in the boxes represent the first, second (median), and third quartiles of the groups relative abundances. The lines extending from the top and bottom of the boxes (ending in a horizontal cap) reach to the furthest point that is within 1.5x the interquartile range. Any points outside of the lines are considered outliers. Red dots correspond to the geometric means of groups, and are connected by either a solid line (cases) or dashed line (controls). The vertical lines extending from the red dots represent the standard error of the geometric means.


A
Dataset 1
Dataset 2
Pooled
B
C
Relative abundance (log10 scale)
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
Absent
Present
Minor allele presence/absence

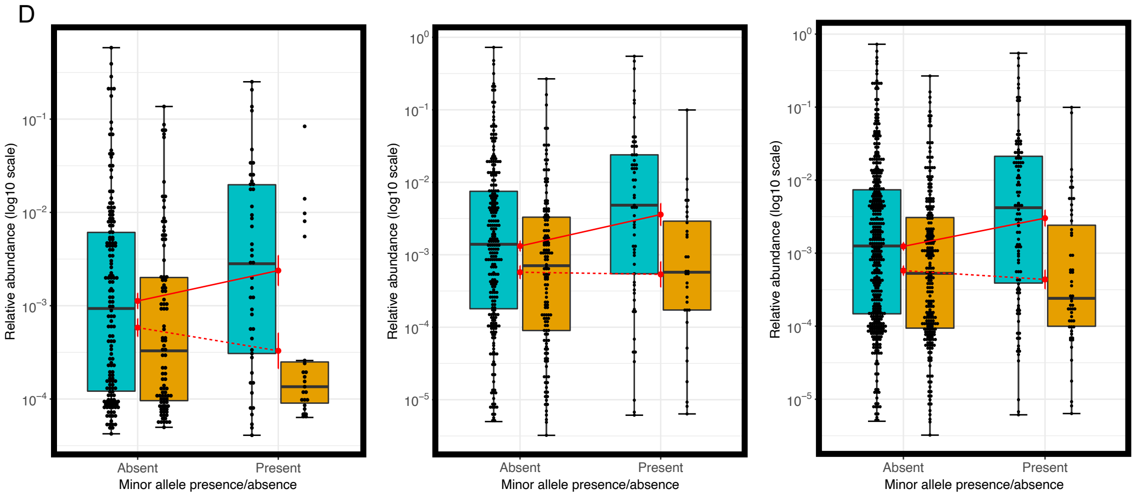


Figure 5: Case and control relative abundances of *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, and cluster 1 stratified by presence or absence of the minor allele for SNPs in top meta-analysis hits.

Case (blue) and control (orange) relative abundances for *Corynebacterium_1* (A), *Porphyromonas* (B), *Prevotella* (C), and cluster 1 (D), stratified by presence/absence of the minor allele for top meta-analysis SNPs for each respective genus or cluster 1, were plotted to visualize interactions. Relative abundances were stratified for SNPs rs356229 (A), rs10029694 (B and D), and rs356228 (C). Plots for *Prevotella*'s top SNPs (rs356183 and rs356228) were very similar, therefore, only plots stratified by rs356228 minor allele presence/absence are shown as it was the top hit for dominant model meta-analysis. Relative abundances (y-axis) were plotted on log10 scale. Sample size was 201 cases and 118 controls for dataset 1, 312 cases and 174 controls for dataset 2, and 513 cases and 292 controls for pooled datasets. Each dot corresponds to a single subject, plotted according to the relative abundance of either *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, and cluster 1 in their stool sample. The top, middle, and bottom horizontal lines in the boxes represent the first, second (median), and third quartiles of the groups relative abundances. The lines extending from the top and bottom of the boxes (ending in a horizontal cap) reach to the furthest point that is within 1.5x the interquartile range. Any points outside of the lines are considered outliers. Red dots correspond to the geometric means of groups, and are connected by either a solid line (cases) or dashed line (controls). The vertical lines extending from the red dots represent the standard error of the geometric means.

Table 2. Quantification of differences in relative abundance between PD and controls stratified by copies of the minor allele of top meta-analysis SNPs. For each genus or cluster 1, mean relative abundance ratios (MRAR) between PD and controls were calculated for each genotype group using the geometric means of the groups. Differences in relative abundance between PD and controls were tested using Wilcoxon rank sum test. Both top SNPs for *Prevotella* had similar results, therefore, only SNP rs356183 is shown for *Prevotella* as it was the top hit for additive model meta-analysis. PD: Parkinson disease; N: number of subjects; MRAR: mean relative abundance ratio; *P*: *P* value derived from Wilcoxon rank sum test.

	No copies of the minor allele				One copy of the minor allele				Two copies of the minor allele			
	PD N	Control N	MRAR	*P*	PD N	Control N	MRAR	*P*	PD N	Control N	MRAR	*P*
Corynebacterium_1 and rs356229												
Dataset 1	65	53	1.08	0.57	90	48	1.85	0.09	46	17	3.84	0.02
Dataset 2	107	66	1.55	0.34	150	80	3.58	8E-04	55	28	3.28	0.05
Pooled	172	119	1.28	0.39	240	128	2.79	1E-04	101	45	3.82	3E-03
Porphyromonas and rs10029694												
Dataset 1	156	95	1.75	0.02	43	22	4.27	2E-03	2	1	43.47	0.67
Dataset 2	251	142	2.01	6E-03	57	28	4.09	0.02	4	4	39.38	0.06
Pooled	407	237	1.85	2E-03	100	50	4.11	1E-03	6	5	50.18	2E-02
Prevotella and rs356183												
Dataset 1	42	41	0.95	0.99	104	51	3.56	5E-06	55	26	1.97	0.24
Dataset 2	83	52	3.55	2E-03	158	86	2.45	4E-03	71	36	1.58	0.55
Pooled	125	93	1.76	0.07	262	137	2.92	1E-05	126	62	1.80	0.17
Cluster 1 and rs10029694												
Dataset 1	156	95	1.93	0.03	43	22	7.56	1E-03	2	1	2.70	1.00
Dataset 2	251	142	2.30	3E-03	57	28	6.44	2E-03	4	4	12.96	0.11
Pooled	407	237	2.15	2E-04	100	50	6.94	2E-05	6	5	11.69	0.08

Table 3. Quantification of differences in relative abundance between PD and controls stratified by presence or absence of the minor allele for top meta-analysis SNPs. For each genus or cluster 1, mean relative abundance ratios (MRAR) between PD and controls were calculated for those with and without the minor allele of the associated SNP using the geometric means of the groups. Differences in relative abundance between PD and controls were tested using Wilcoxon rank sum test. Both top SNPs for *Prevotella* had similar results, therefore, only SNP rs356228 is shown for *Prevotella* as it was the top hit for additive model meta-analysis. PD: Parkinson disease; N: number of subjects; MRAR: mean relative abundance ratio; *P*: *P* value derived from Wilcoxon rank sum test.

	Minor allele absent				Minor allele present			
	PD N	Control N	MRAR	*P*	PD N	Control N	MRAR	*P*
Corynebacterium_1 and rs356229								
Dataset1	65	53	1.08	0.57	136	65	2.32	6E-03
Dataset2	107	66	1.55	0.34	205	108	3.49	1E-04
Pooled	172	119	1.28	0.39	341	173	3.04	2E-06
Porphyromonas and rs10029694								
Dataset1	156	95	1.75	0.02	45	23	4.73	8E-04
Dataset2	251	142	2.01	6E-03	61	32	4.86	7E-03
Pooled	407	237	1.85	2E-03	106	55	4.85	2E-04
Prevotella and rs356228								
Dataset1	45	43	1.03	0.70	156	75	2.92	3E-05
Dataset2	81	55	2.88	7E-03	231	119	2.30	2E-03
Pooled	126	98	1.64	0.08	387	194	2.59	9E-06
Cluster 1 and rs10029694								
Dataset1	156	95	1.93	0.03	45	23	7.25	2E-03
Dataset2	251	142	2.30	3E-03	61	32	6.69	8E-04
Pooled	407	237	2.15	2E-04	106	55	6.90	8E-06

Association of SNCA genetic variants with PD modified by gut microbiome

When testing SNP association with PD in pooled datasets, all but one SNP (rs10029694) was associated with increased risk of PD (OR > 1) when tested irrespective of *Corynebacterium_1*, *Porphyromonas*, *Prevotella*, and cluster 1 presence or absence (Table 4). These SNPs (rs356229, rs356183, and rs356228) have been previously associated with increased risk of PD in a large GWAS meta-analysis of PD risk (odds ratio (OR) ~ 1.3, *P* = 3E-50 – 3E-32; pdgene.org) whose results matched the ORs we detected in the current dataset. SNP rs10029694 resulted in ORs > 1 suggesting a trend towards increased risk of PD, but the effect sizes were small ($OR_{additive}$ = 1.1, $OR_{dominant}$ = 1.2) and associations were not significant (*P* > 0.39) (Table 4). For all genera and cluster 1, again with the exception of *Prevotella*, testing SNP association with PD in the presence of each genus or cluster 1 resulted in an increased OR for both additive and dominant models corresponding to a 20 – 50% increase in PD risk per copy of the minor allele, or a 30 – 70% increase in risk for being a minor allele carrier, compared to results obtained irrespective of *Corynebacterium_1*, *Porphyromonas*, and cluster1 (Table 4). *P* values for associations stayed relatively the same for rs356229 in the presence of *Corynebacterium_1*, while *P* values for association of rs10029694 with PD became significant in the presence of *Porphyromonas*. Although an increase in ORs was observed for both additive and dominant models for cluster 1, the associations did not reach significance (*P* < 0.05) raising the hypothesis that not all members of cluster 1 might be needed to see an association with PD, and including all members might actually be washing out a significant association signal. Indeed, when removing members

Table 4. Association of top meta-analysis SNPs with PD stratified by presence or absence of *Corynebacterium_1*, *Porphyromonas,* or cluster 1. Firth's penalized logistic regression was performed for SNP association with PD in pooled datasets irrespective of genera or cluster 1, and stratified by genera or cluster 1 presence/absence. Analyses were adjusted for age and sex. Analyses were performed once using an additive model for SNP genotype and once using a dominant model for SNP genotype. N: number of subjects; OR: odds ratio from Firth's logistic regression; CI: confidence interval; *P*: *P* value from Firth's logistic regression; PD: Parkinson disease

	Total N	Copies of the minor allele			Additive model		Dominant model	
		0	1	2	OR [95% CI]	*P*	OR [95% CI]	*P*
Corynebacterium_1 and rs356229								
Association of rs356229 with PD risk irrespective of *Corynebacterium_1*								
PD	513	172	240	101	1.3 [1.0-1.6]	0.02	1.4 [1.0-1.9]	0.03
Control	292	119	128	45				
Association of rs356229 with PD risk in the presence of *Corynebacterium*_1								
PD	235	79	111	45	1.5 [1.0-2.2]	0.03	1.7 [1.0-2.8]	0.04
Control	94	44	38	12				
Association of rs356229 with PD risk in the absence of *Corynebacterium_1*								
PD	278	93	129	56	1.2 [0.9-1.6]	0.17	1.3 [0.9-1.9]	0.19
Control	198	75	90	33				
Porphyromonas and rs10029694								
Association of rs10029694 with PD risk irrespective of *Porphyromonas*								
PD	513	407	100	6	1.1 [0.8-1.6]	0.43	1.2 [0.8-1.7]	0.39
Control	292	237	50	5				
Association of rs10029694 with PD risk in the presence of *Porphyromonas*								
PD	276	210	60	6	1.6 [1.0-2.7]	0.05	1.9 [1.1-3.4]	0.03
Control	119	100	16	3				
Association of rs10029694 with PD risk in the absence of *Porphyromonas*								
PD	237	197	40	0	0.7 [0.4-1.2]	0.19	0.7 [0.4-1.2]	0.25
Control	173	137	34	2				
Cluster 1 and rs10029694								
Association of rs10029694 with PD risk irrespective of cluster 1								
PD	513	407	100	6	1.1 [0.8-1.6]	0.43	1.2 [0.8-1.7]	0.39
Control	292	237	50	5				
Association of rs10029694 with PD risk in the presence of cluster 1								
PD	427	335	86	6	1.3 [0.9-2.0]	0.12	1.5 [1.0-2.3]	0.07
Control	222	185	32	5				
Association of rs10029694 with PD risk in the absence of cluster 1								
PD	86	72	14	0	0.5 [0.2-1.2]	0.14	0.5 [0.2-1.2]	0.14
Control	70	52	18	0				

of cluster 1 one by one and retesting association between rs10029694 and PD, reducing cluster 1 down to *Porphyromonas*, *Prevotella*, *Varibaculum*, *Anaerococcus*, *Peptoniphilus*, and *Lawsonella* (dataset 2 only) increased the ORs by ~ 13% ($OR_{additive}$ = 1.5 [1.0-2.2], $OR_{dominant}$ = 1.7 [1.1-2.7]) and lowered the *P* values for additive model to borderline significance (*P* = 0.06) and dominant model to significance (*P* = 0.03). When SNPs were tested for association with PD in the absence of *Corynebacterium_1*, *Porphyromonas*, and cluster 1, ORs either decreased or stayed relatively the same compared to those derived from testing irrespective of *Corynebacterium_1*, *Porphyromonas*, and cluster 1, and associations were not significant (Table 4).

For *Prevotella*, stratifying by *Prevotella* presence or absence did not result in an obvious decrease or increase in ORs for additive or dominant model top SNPs, and testing in the *Prevotella* positive group usually resulted in non-significant associations compared to testing in *Prevotella* negative group and irrespective of *Prevotella*.

DISCUSSION

Numerous studies have been performed in human on the association of gut microbiome or genetic variants with PD, but this is the first, to our knowledge, that has attempted to study the interaction between the two. This study was a candidate gene, candidate taxa study where we used prior knowledge of gut opportunistic pathogens that associated with PD in a previous study [Wallen et al. 2020] and searched for genetic modifiers of these associations in the top genomic locus associated with PD risk through large genetic studies [Nalls et al. 2019]. Performing a statistical genetic scan of the *SNCA* gene and surrounding area, we uncovered evidence for potential genetic modifiers of the association between *Corynebacterium_1*, *Porphyromonas*, and cluster 1 with PD in the 3′ *SNCA* region. Differences in the relative abundances of *Corynebacterium_1*, *Porphyromonas*, and cluster 1 between PD and control subjects increased with increasing copies of the minor allele of these genetic variants. Increasing differences were usually due to an increase in relative abundances in PD and either a decrease or relatively stable level in controls with increasing copies of the minor allele. We tested association of SNPs with PD in the presence or absence of *Corynebacterium_1*, *Porphyromonas*, or cluster 1 and found that the strength of association between these SNPs and PD were increased in subjects where *Corynebacterium_1*, *Porphyromonas*, or cluster 1 was detected in their stool samples. Results might be a reflection of a higher risk of PD when both the minor alleles of these SNPs and *Corynebacterium_1, Porphyromonas,* and cluster 1 are present, but also might be driven by the disease itself, consequences of underlying genetic variation, or a combination of both. This study is only associative in nature, therefore, causal inference cannot be made here, which would require further functional and longitudinal studies.

This study was motivated by the hypothesis that the presence of both pathogens (opportunistic or not) in the gut and overexpression of α-synuclein, which has been associated with genetic variants in the 3′ *SNCA* region, might increase risk of PD. This stems from several previous studies showing connections between the gut and α-synuclein, and α-synuclein involvement in response to pathogens. Presence of α-synuclein has been shown in the gastrointestinal tract of persons with early PD [Shannon et al. 2012], Lewy body disease [Breen, Halliday & Lang 2019], and rapid eye movement disorder [Knudsen et al. 2018], which has a high conversion rate to PD. Large epidemiological studies have suggested a reduction in PD risk for those who have undergone truncal vagotomy years before PD onset [Svensson et al. 2015; Liu et al. 2017], and a study in mouse saw that truncal vagotomy and endogenous α-synuclein deficiency prevented gut to brain spread of injected preformed α-synuclein fibrils and development of PD-like neurodegeneration and behavioral deficits [Kim et al. 2019]. Studies in human have shown a role of α-synuclein in pathogen response where infection of the gut or olfactory system triggered α-synuclein expression, which in turn mobilized the immune system to respond to the infection [Stolzenberg et al. 2017; Tomlinson et al. 2017]. Experimentally, it has been shown in a *Pink1* knockout mouse model of PD that intestinal infection may act as a trigger for dopaminergic cell loss and motor impairment through activation of T cells in the periphery [Matheoud et al. 2019]. A hypothesis that has gained popularity in recent years, termed "Braak's hypothesis", states that non-inherited forms of PD may be caused by a yet to be identified pathogen that invades the gastrointestinal tract and, through the enteric nervous system, makes its way to the brain [Braak et al. 2003; Braak et al. 2003]. This hypothesis has been further modified to state that it may not be an actual pathogen making its way to the brain, but pathogenic species of α-synuclein initiated in the gut by a pathogen, or altered microbial state, and traveling to the brain. We previously detected a significant enrichment of three genera (*Corynebacterium_1*, *Porphyromonas*, *Prevotella*) and a poly-microbial group of genera in PD that, per the literature, are often harmless, commensal members of the gut and oral microbiome, but can become pathogenic in certain scenarios [Wallen et al. 2020; Citron et al. 2007; Wagner et al. 2017; Choi et al. 2019]. Here, we attempted to connect the overabundance of these opportunistic pathogens to the most highly associated portion of the genome with increased risk of PD, and in the end, found evidence that the presence of at least a subset of these genera in the gut might increase the risk of PD in conjunction with genetic variants tied to α-synuclein. In addition, the minor allele of at least one of these variants (rs356229) was found previously to be associated with higher expression of *SNCA* (*P* = 9E-5; https://www.gtexportal.org/home/snp/rs356229), which might provide a lead on the mechanism of how PD risk may be escalated in the presence of these genera and 3′ *SNCA* genetic variants. However, as this study is only associative in nature, the opposite also might be true, where changes in relative abundances of genera studied here are actually due to variables relating to the disease itself (e.g. compromised gut lining and/or immune system allowing these genera to increase in number), underlying consequences of genetic variation (e.g. dysfunctional α-synuclein that is unable to perform normal duties related to immunity), or a combination of both. Further experimental functional studies in PD animal models and potentially longitudinal studies in human would be required to confirm any of the above mentioned scenarios.

The obvious limitation of this study was the sample size. Sample sizes for genetic studies, and meta-analyses of such, usually range in the thousands whereas here we analyzed samples from approximately 800 subjects. This is most likely the reason no interactions for meta-analysis reached the study-wide multiple corrected threshold for significance, and why interactions detailed here should be considered suggestive. Additional analyses need to be performed on larger samples sizes to replicate and confirm these findings in human. Additionally, experimental studies should be conducted to tease out if there is a true biological interaction between genera targeted in this study, α-synuclein, and/or overexpression of *SNCA*. Another limitation of this study is the nature of the interaction and association analyses performed. In this study, we have merely provided suggestive evidence for an interaction between targeted genera, PD, and *SNCA* genetic variants. Although we can specify statistical models a certain way (with genera as outcome, or PD as outcome) this does not imply true causation, therefore, we can only speculate about what may be occurring biologically. The changes in relative abundances of genera studied here might not be involved in the pathogenesis of disease, but a response to disease, and/or the biological effect produced by the genetic variants. Regardless, this study provides interesting leads for follow-up investigations in both human and in experimental models.

In conclusion, we detected potential genetic modifiers for the associations between *Corynebacterium_1*, *Porphyromonas*, and a group of poly-microbial opportunistic pathogens and PD in a highly relevant genomic region for risk of developing PD. Differences in relative abundances of these genera between PD and controls were increased with increasing copies of the minor allele of detected SNPs. We then provided evidence for a potential increase in genetic risk of PD when both the minor allele and genera were present compared to the sample population as a whole, or when genera were absent. These results provide interesting leads for future human and animal studies.

DATA AVAILABILITY

For microbiome data, individual-level raw sequences and basic metadata are publicly available at NCBI Sequence Read Archive (SRA) BioProject ID PRJNA601994. Genotypes will be deposited in dbGaP once data has been accepted for publication.

CODE AVAILABILITY

No custom codes were used. All software and packages, their versions, relevant specification and parameters are stated in the "METHODS" section.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of Neurological Disorders and Stroke grant R01 NS036960 (to H.P.), The US Army Medical Research Materiel Command endorsed by the US Army through the Parkinson's Research Program Investigator-Initiated Research Award under Award number W81XWH1810508 (to H.P.); NIH Udall grant P50 NS062684 (to C.P.Z.) and P50 NS108675 (to D.G.S.), and NIH Training Grant T32NS095775 (to Z.D.W.). Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the US Army or the NIH.

REFERENCES

Chang, D. et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson’s disease risk loci. Nat. Genet. 49, 1511–1516 (2017).

Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 18(12):1091-1102 (2019). doi:10.1016/S1474-4422(19)30320-5

Tanner, C. M. Advances in environmental epidemiology. Mov. Disord. 25(Suppl 1), S58–S62 (2010) .

Hamza, T. H. et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson’s disease modifier gene via interaction with coffee. PLoS Genet. 7, e1002237 (2011).

Cannon, J. R. & Greenamyre, J. T. Gene-environment interactions in Parkinson’s disease: specific evidence in humans and mammalian models. Neurobiol. Dis. 57, 38–46 (2013).

Hill-Burns, E. M. et al. A genetic basis for the variable effect of smoking/nicotine on Parkinson’s disease. Pharmacogenomics J. 13, 530–537 (2013).

Biernacka, J. M. et al. Genome-wide gene-environment interaction analysis of pesticide exposure and risk of Parkinson’s disease. Parkinsonism Relat. Disord. 32, 25–30 (2016).

Gerhardt, S. & Mohajeri, M. H. Changes of colonic bacterial composition in Parkinson’s disease and other neurodegenerative diseases. Nutrients 10, https://doi.org/10.3390/nu10060708 (2018).

Boertien, J. M., Pereira, P. A. B., Aho, V. T. E. & Scheperjans, F. Increasing comparability and utility of gut microbiome studies in Parkinson’s disease: a systematic review. J. Parkinsons Dis. 9, S297–S312 (2019).

Wallen, Z.D., Appah, M., Dean, M.N. et al. Characterizing dysbiosis of gut microbiome in PD: evidence for overabundance of opportunistic pathogens. npj Parkinsons Dis. 6, 11 (2020). https://doi.org/10.1038/s41531-020-0112-6.

Braak, H. et al. Staging of brain pathology related to sporadic Parkinson’s disease. Neurobiol. Aging 24, 197–211 (2003).

Braak, H., Rub, U., Gai, W. P. & Del Tredici, K. Idiopathic Parkinson’s disease: possible routes by which vulnerable neuronal types may be subject to neuroinvasion by an unknown pathogen. J. Neural Transm. (Vienna) 110, 517–536 (2003).

GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348(6235):648-660 (2015).

Soldner, F., Stelzer, Y., Shivalila, C. et al. Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. Nature 533, 95–99 (2016).

Emelyanov, A.K., Andoskin, P.A., Miliukhina, I.V. et al. SNCA alleles rs356219 and rs356165 are associated with Parkinson's disease and increased α-synuclein gene expression in CD45+ blood cells. Cell Tiss. Biol. 10, 277–283 (2016).

Devine MJ, Gwinn K, Singleton A, Hardy J. Parkinson's disease and α-synuclein expression. Mov Disord. 26(12):2160-2168 (2011).

Stolzenberg, E. et al. A role for neuronal alpha-synuclein in gastrointestinal immunity. J. Innate Immun. https://doi.org/10.1159/000477990 (2017).

Tomlinson, J. J. et al. Holocranohistochemistry enables the visualization of alpha-synuclein expression in the murine olfactory system and discovery of its systemic anti-microbial effects. J. Neural Transm. (Vienna) 124, 721–738 (2017).

Hill-Burns, E. M. *et al.* Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Mov Disord* **32**, 739-749, doi:10.1002/mds.26942 (2017).

Gibb, W. R. G. & Lee, A. J. A comparison of clinical and pathological features of young- and old-onset Parkinson disease. Neurology 38 (1988).

Hamza, T. H. et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. PLoS Genet. 7, e1002237 (2011).

Powers, K. et al. Combined effects of smoking, coffee and NSAIDs on Parkinson's disease risk. Mov. Disord. 23, 88–95 (2008).

Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857 (2019).

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet 17, 10–12 (2011).

Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583 (2016).

Wright, E. S. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinformatics 16, 322 (2015).

Schliep, K. P. phangorn: phylogenetic analysis in R. Bioinformatics 27, 592–593 (2011).

McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8, e61217 (2013).

Aitchison, J. The Statistical Analysis of Compositional Data. Royal Statistical Society 44, 2, 139-160 (1986).

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. Front. Microbiol. 8, 2224 (2017).

Chang C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets, GigaScience 4, 1 (2015). https://doi.org/10.1186/s13742-015-0047-8

Kurilshikov, A. et al. Genetics of human gut microbiome composition. bioRxiv 2020.06.26.173724; doi: https://doi.org/10.1101/2020.06.26.173724

Deelen P, Bonder MJ, van der Velde KJ, et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Res Notes. 7, 901 (2014). doi:10.1186/1756-0500-7-901

Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv 563866; doi: https://doi.org/10.1101/563866

Das S, et al. Next-generation genotype imputation service and methods. Nature Genetics 48, 1284–1287 (2016).

Loh, P., Danecek, P., Palamara, P. et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 48, 1443–1448 (2016). https://doi.org/10.1038/ng.3679

McCarthy, S., Das, S., Kretzschmar, W. et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48, 1279–1283 (2016). https://doi.org/10.1038/ng.3643

Howie B, et al. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).

Li, Y., et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol., 34: 816-834 (2010). doi:10.1002/gepi.20533

Han B, Eskin E. Interpreting Meta-Analyses of Genome-Wide Association Studies. PLoS Genet 8(3): e1002555 (2012). https://doi.org/10.1371/journal.pgen.1002555

Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 26(18):2336-2337 (2010). doi:10.1093/bioinformatics/btq419

Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31:3555–3557 (2015).

Clark-Carter, D. Measures of Central Tendency. International Encyclopedia of Education (Third Edition), Elsevier, Pages 264-266 (2010). https://doi.org/10.1016/B978-0-08-044894-7.01343-9.

Altman, D.G. and Royston, P. The cost of dichotomising continuous variables. BMJ (Clinical research ed.) vol. 332,7549 (2006). doi:10.1136/bmj.332.7549.1080

Shannon KM, Keshavarzian A, Mutlu E, et al. Alpha-synuclein in colonic submucosa in early untreated Parkinson's disease. Mov Disord. 27(6):709-715 (2012). doi:10.1002/mds.23838

Breen, D. P., Halliday, G. M. & Lang, A. E. Gut-brain axis and the spread of alpha-synuclein pathology: Vagal highway or dead end? Mov. Disord. 34, 307–316 (2019).

Knudsen, K. et al. In-vivo staging of pathology in REM sleep behaviour disorder: a multimodality imaging case-control study. Lancet Neurol. 17, 618–628 (2018).

Svensson, E. et al. Vagotomy and subsequent risk of Parkinson’s disease. Ann. Neurol. 78, 522–529 (2015).

Liu, B. et al. Vagotomy and Parkinson disease: a Swedish register-based matched-cohort study. Neurology 88, 1996–2002 (2017).

Kim, S. et al. Transneuronal propagation of pathologic alpha-synuclein from the gut to the brain models Parkinson’s disease. Neuron 103, 627–641 e627 (2019).

Matheoud, D. et al. Intestinal infection triggers Parkinson’s disease-like symptoms in Pink1(-/-) mice. Nature 571, 565–569 (2019).

Citron, D. M., Goldstein, E. J., Merriam, C. V., Lipsky, B. A. & Abramson, M. A. Bacteriology of moderate-to-severe diabetic foot infections and in vitro activity of antimicrobial agents. J. Clin. Microbiol. 45, 2819–2828 (2007).

Wagner Mackenzie, B. et al. Bacterial community collapse: a meta-analysis of the sinonasal microbiota in chronic rhinosinusitis. Environ. Microbiol. 19, 381–392 (2017).

Choi, Y. et al. Co-occurrence of anaerobes in human chronic wounds. Micro. Ecol. 77, 808–820 (2019).

SUMMARY AND DISCUSSION

This dissertation addressed two areas of ongoing PD research: the identification of age at diagnosis modifiers in PD and continued characterization of gut microbiome changes in PD and their potential role in PD risk. We first were able to identify two new genetic modifiers of age at diagnosis of PD that accounted for an earlier age at diagnosis of about 6 years. Within these genetic signals, we uncovered functionally relevant variants that might have a deleterious effect on a putative PD modifier gene *LPPR1* and influence the expression of another gene *GRIN3A*. We next tested and compared 16 differential abundance testing methods on two large PD-gut microbiome datasets, providing, to our knowledge, the first example of a differential abundance comparison study performed on real gut microbiome data from a complex disease. We then performed a study characterizing the PD gut microbiome in two of the largest PD-gut microbiome datasets to date, using larger sample sizes, robust statistical methods, stringent statistical criteria, and a replication paradigm to detect robust associations between PD and 15 bacterial genera who belonged to 3 poly-microbial groups of correlated genera. Lastly, we investigated interaction between PD, genetic variants in the *SNCA* locus, and three genera and one poly-microbial group of genera, found enriched in PD and to be opportunistic pathogens. We uncovered that genetic variants at the 3′ end of *SNCA* moderated the associations between PD and these genera, and associations between detected genetic variants and PD were enhanced when tested using subjects positive for genera, potentially reflecting an increase in PD risk when both are present. Results presented in this dissertation provide novel insights into age at diagnosis modification in PD and potential interaction between the gut microbiome, PD, and genetic risk of PD. Results will hopefully be useful in guiding future research in both humans and animals to replicate, confirm, and tease apart the biology behind these findings.

LPPR1, a potential target for modifiying PD disease progression?

In the first chapter, we uncovered evidence for association of genetic variants in neuronal plasticity-related gene 3 (*LPPR1*) with age at diagnosis of PD. Two signals were detected, each tagging a block of SNPs in high LD. These variants had low allele frequencies (MAF = 0.01–0.02), similarly to previously found variants for age at onset of familial PD [Hill-Burns et al. 2016]. The first LD block of SNPs replicated robustly in both prevalent PD cases (PAGE_{P}) and incident PD cases (PAGE_{I}) in PAGE. The second LD block replicated in PAGE_{P}, but not in PAGE_{I} showing some potential heterogeneity in PAGE subjects. Functional annotation of the PD-associated variants in *LPPR1* revealed several variants had predicted deleterious effects, including a missense that destabilizes the structure of LPPR1, a regulatory element that associates with expression levels of *GRIN3A*, and enhancers that interact with promoters of *LPPR1* and several other genes in the brain. *LPPR1* is one of 5 members of a brain-specific gene family that modulates neuronal plasticity during development, aging, and after brain injury [Savaskan, Brauer & Nitsch 2004; Broggini et al. 2016; Fink et al. 2017]. *LPPR1* is the strongest driver of axonal outgrowth in the gene family. Studies in mice have shown that after neuronal injury, overexpression of *LPPR1* enhances axonal growth, improves motor behavior, and promotes functional recovery [Broggini et al. 2016; Fink et al. 2017]. Extrapolating to our findings, we posit that *LPPR1* is not necessarily involved in the cause of PD, but might be involved in the response to neuronal damage, and influences how well neuronal cells respond to injury and the rate at which neurons deteriorate in preclinical PD. The actual cause of neuronal death in PD may be initiated by environmental exposures such as toxins, genetics, or a combination of both, but once the initial insult has occurred the rate of neuronal deterioration and disease progression is dependent upon how well intra-neuronal mechanisms of repair are able to mitigate the damage done by the initial insult(s).

How LPPR1 elicits its protective effects is still poorly understood, but it might be through promotion of cell adhesion and inhibition of GTPase RhoA activity to resist neurite outgrowth inhibition [Broggini et al. 2016; Iweka et al. 2019]. LPPR1 was previously shown to reduce activity of RhoA through interaction with PIP_2 (Phosphatidylinositol 4,5-bisphosphate) [Broggini et al. 2016], and a Rho family-specific guanine nucleotide dissociation inhibitor (RhoGDI) [Iweka et al. 2019] when RhoA activating molecules were introduced to neural cells. Activation of RhoA, which primarily acts upon the cytoskeleton of a cell, has been previously implicated in the formation of stress fibers and focal adhesions and, through growth cone collapse, can inhibit neurite outgrowth [Kalpachidou, Spiecker, Kress & Quarta 2019]. While this dissertation describes the first study to directly connect *LPPR1* to PD, evidence exists for potential RhoA involvement in the pathogenesis of PD. RhoA activity was previously shown to be increased by a number of mechanisms relevant to PD including exposure to toxins such as rotenone and 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) [Villar-Cheda et al. 2012; Mattii et al. 2019], recombinant human α-synuclein through binding of integrin CD11b (which can also bind to other damaging molecules such as bacterial lipopolysaccharide and amyloid beta) [Hou et al. 2018], deficiency in Parkin expression (mutations in which cause a recessively inherited form of PD) [Bogetofte et al. 2019], and can even be activated by certain bacterial toxins [Musilli et al. 2016]. In these same studies, inhibition of RhoA activity through various mechanisms was able to rescue detrimental phenotypes including MPTP induced dopaminergic cell death and microglial activation [Villar-Cheda et al. 2012], α-synuclein based activation of reactive oxygen species forming enzyme NOX2 [Hou et al. 2018], enhanced neuronal migration and neurite outgrowth deficiency caused by mutated Parkin [Bogetofte et al. 2019], and rotenone induced damage and morphological changes to dopaminergic neurons [Mattii et al. 2019]. These studies show a beneficial effect of RhoA inhibition in a variety of PD related pathogenesis routes, induced by toxins to genetics, therefore, if upregulation of *LPPR1* truly has an inhibitory effect on RhoA activity, it could potentially be a useful therapy in a wide variety of PD cases.

A lot of ground needs to be covered before deciding if *LPPR1* is a good candidate to pursue as a potential PD therapeutic. The function of LPPR1 and its role in neuroprotection still remains largely uncharacterized, and it’s supposed neuroprotective effects have not been experimentally tested in a PD relevant model system. Even if LPPR1 is eventually shown to have neuroprotective effects in respect to PD, there still is the issue that, so far, no report in the literature exists of a compound that increases expression of *LPPR1*. This means time would need to be spent screening for potential compounds that influence *LPPR1* expression in neuronal cells, or designing a safe and effective viral construct for delivery of a more active *LPPR1* gene, which has shown to be possible for other potential targets for PD therapeutics [Axelson & Woldbye 2018]. More characterization of interacting partners of LPPR1 is also needed to get a fuller picture of the potential systemic effects that modulation of *LPPR1* expression would cause, and to reveal other potential therapeutic targets such as we have with *GRIN3A*. However, even with the obvious hurdles, based on previous literature and results from our study, we believe *LPPR1* is a promising lead for future investigations into its therapeutic benefits for PD.

Differential abundance methods vary in results, but can still come to agreements

The next chapter detailed a comparison study of 16 differential abundance testing methods on a real gut microbiome dataset derived from a complex, heterogeneous disease. We detected variations between method results, which aligned with the variation between differential abundance testing method performances previously reported in method comparison studies [Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019]. Although we could not assess method performance due to the true answers being unknown, we found methods with similar previously reported performance metrics grouped together based on their concordances with one another, and the PD-genus associations they detected. Methods with previously reported low FPR/FDR had the highest average concordances across methods, and the highest concordances among each other. We found a variable effect of taxa filtering for ANCOM, which might be due to the different statistics used by ANCOM compared to the standard FDR q-value. We also detected two groups of genera through hierarchical clustering based on similarity in method results. One group of genera were more likely to be replicated by the majority of methods on average and included genera previously associated with PD such as *Bifidobacterium*, *Lactobacillus*, and short-chain fatty-acid producing bacteria *Faecalibacterium*, *Roseburia*, *Blautia*, and other members of the *Lachnospiraceae* family. The second group of genera were only detected and replicated by a subset of methods (fitZIG, edgeR, limma-voom, baySeq, SAMseq, GLM NBZI, DESeq2). This group mostly contained genera enriched in PD that had low control MRAs, higher effect sizes on average, and were more likely to be replicated by methods previously reported to have higher sensitivity [McMurdie & Holmes 2014; Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019]. Although the two PD-gut microbiome datasets used in this study were found to be heterogeneous in microbiome composition, we observed no significant differences between datasets in the proportion of genera being associated with PD on average.

In previous method comparison studies, variation in results between different differential abundance testing methods was evident [Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019], which aligns with our results, but we also observed some similarities between methods through calculation of method result concordances. Method performances could not be assessed in this study as analyses were performed on real data where true answers are unknown, but by using performance metrics calculated in previous studies as a proxy [Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019], we observed that methods with similar performance metrics showed similar patterns in their concordances with one another. Methods that were previously reported to have lower FPR and FDR, which included the group of methods containing Kruskal-Wallis, t-test with log transform, ALDEx2, fitFeatureModel, ANCOM, and DESeq2, had the highest average concordances across all methods, and had even higher average concordances among each other (a mean concordance of ~0.9 per dataset compared to the overall concordance average of 0.76 per dataset). Methods that were previously reported to have a higher FPR and FDR, which included the group of methods containing baySeq, GLM NBZI, fitZIG, and edgeR, had the lowest average concordances across all methods. These observations make sense, as methods with lower FPR/FDR were more likely to agree on the same PD-genus associations. They detected less overall significant associations compared to higher FPR/FDR methods, and, for the most part, only detected and replicated PD-genus associations that seemed to be robust to inter-method variation. This would naturally increase the odds of lower FPR/FDR methods converging on the same differential abundance signatures compared to higher FPR/FDR methods, who detected more overall significant associations, a larger number of which were not agreed upon by the majority of methods.

A finding of this study that was not expected was the variable effect of taxa filtering prior to analysis with ANCOM. Decreasing the number of taxa that is included in a differential abundance analysis, to a certain extent, usually results in increased number of significant associations detected by a method. This is usually due to the decreased burden of multiple testing correction when calculating FDR q-values. With ANCOM, however, filtering of taxa before analysis greatly decreased the number of significant associations. The reason behind this unorthodox effect of taxa filtering might be due to the statistics used by ANCOM to determine significance, which differs from the standard FDR correction used by the other methods included in this study. ANCOM calculates a *W* statistic, which is the number of times the log ratio of a taxon with every other taxon being tested was detected as significantly different across groups [Mandal et al. 2015]. Because ANCOM's *W* statistics are based off of pairwise combinations between all taxa tested, they will automatically decrease overall if less taxa are being analyzed, which may also decrease the range of significant *W* statistics at a particular threshold. If lower prevalent taxa are being removed, this might make the *W* statistic calculation more conservative since more prevalent, and potentially stable, taxa have been selected for whose ratios with one another might not differ enough to be detected as significant at a particular significance threshold. These findings suggest that it might be beneficial to perform ANCOM using all detected taxa, or at least the majority of detected taxa, only removing very rare taxa (those whose detection was most likely due to technical errors) to reduce noise.

Although methods overall varied in the PD-genus associations they detected, two groups of genera seemed to be converged upon by the majority or a subset of methods used in this study. Through hierarchical clustering of genera based on similarities in method results, a group genera was revealed whose associations with PD seemed to be more robust to inter-method variation as they were more likely to be replicated by the majority of methods on average. This group contained genera previously associated with PD including *Bifidobacterium, Lactobacillus, Faecalibacterium, Roseburia, Blautia,* and other members of the *Lachnospiraceae* family. This might suggest that at least some of the PD-genus associations placed into this group were more robust to inter-method variation not just because they were strong associations or ideally met the assumptions of multiple methods in our data, but because they are biologically associated with PD. Regardless of why these PD-genus associations were able to be detected across the spectrum of methods implemented in this study, it shows that even with obvious differences in results between methods, different differential abundance methods, even with their different underlying characteristics, can converge on the same answers especially when utilizing a second dataset for replication of associations. Hierarchical clustering also revealed a second group of genera whose associations with PD were only replicated by a subset of methods that were previously shown to have higher sensitivities (fitZIG, edgeR, limma-voom, baySeq, SAMseq, GLM NBZI, DESeq2) [McMurdie & Holmes 2014; Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019]. This group was interesting because it contains only genera increased in PD (with the exception of one) who had lower control mean relative abundances and resulted in higher fold changes on average when compared to other genera. This group of genera was not replicated by any of the more conservative lower FPR/FDR methods, therefore, without the use of more sensitive methods, this group of PD-genus associations would have gone unnoticed. This potentially argues that, although some of these methods were previously reported to have high FPR/FDR, they could prove useful for detecting rarer taxa. The use of a second dataset, as done with this study, might help to mitigate some of the potential false positives detected by these methods.

Characterization of gut microbiota dysbiosis in PD resulted in detection of familiar signatures, and generation of new leads

The final two chapters of the dissertation dealt with characterizing the gut microbiome of PD patients in two large datasets of PD and control subjects, followed by an interaction analysis between three genera and a poly-microbial group of genera, PD, and genetic variants in the *SNCA* locus. In the characterization study, 15 genera were detected as significantly associated with PD, but the majority of these associations were confirmatory and had been previously detected enriched (*Bifidobacterium, Lactobacillus*) or depleted (SCFA producing bacteria) in PD. The novel finding was three genera enriched in PD who, through literature search, were found to be opportunistic pathogens in the right conditions (*Porphyromonas, Corynebacterium_1, Prevotella*). These three genera then mapped to a poly-microbial cluster of correlated genera whose members were also found to be majorly opportunistic pathogens. This was of interest to us as it harks back to Braak's hypothesis that states a yet to be identified pathogen enters the gastrointestinal tract and invades the brain to cause PD [Braak et al. 2003; Braak et al. 2003]. We decided these genera might be the most relevant to increasing PD risk and carried them forward to interaction analysis. To decide on the genomic region for interaction analysis, we referred back to the literature detailing the involvement of pathological α-synuclein in the gut and how α-synuclein might play a role in the immune response to infection [Stolzenberg et al. 2017; Tomlinson et al. 2017]. We posited that combined presence of opportunistic pathogens in the gut with genetic risk of PD in the *SNCA* locus might interact together to increase risk of PD. Through interaction analysis of *SNCA* SNPs, PD, and genera, we detected SNPs at the 3′ end of *SNCA* that moderated the associations between PD and these genera. We then found associations between detected SNPs and PD were enhanced when tested using subjects positive for genera. As stated in the last chapter, no interaction analyses reached a multiple testing corrected significance, so results are considered suggestive and need to be replicated in larger cohorts.

Multiple PD-gut microbiome studies in human have previously associated decreased SCFA producing bacteria with PD [Keshavarzian et al. 2015; Unger et al. 2016; Hill-Burns et al. 2017; Petrov et al. 2017; Li et al. 2017; Lin, A et al. 2018; Pietrucci et al. 2019; Aho et al. 2019], which our findings confirmed. The ten genera we detected reduced in PD belonged to the bacterial families *Lachnospiraceae* and *Ruminococcaceae*, whose members have been among those detected in previous PD-gut microbiome studies, and four of which were even confirmed in a recent meta-analysis of five PD-gut microbiome studies [Nishiwaki et al. 2020]. In chapter 3 of this dissertation (and its corresponding publication Wallen et al. 2020) and in previous work by our group, we reported correlation between decreased genera of the *Lachnospiraceae* family and increasing daily dose of levodopa [Wallen et al. 2020] and disease duration [Hill-Burns et al. 2017]. Others have also found correlations between decreased members of this bacterial family and disease duration [Keshavarzian et al. 2015], along with disease severity and motor impairment [Pietrucci et al. 2019]. This might suggest SCFA producing bacteria are depleted in PD as a consequence of disease and/or use of disease related medication. However, this does not mean that reduction of SCFA producing bacteria has no potential role in disease pathogenesis. SCFAs have been previously shown to help with reducing inflammation in the gut and brain, enhancing gut epithelial health, and promoting increased gastrointestinal motility [Hamer et al. 2008; Canani et al. 2011; Park et al. 2019; Haase et al. 2018; Furusawa et al. 2013], all of which are issues observed in PD. The absence of SCFA producing bacteria might not be playing a large role in the initiation of PD, but could play a role in enhancing disease progession, and/or symptoms, which might be improved by replenishing SCFA producing bacterial numbers, or supplementation with SCFAs. Pre-clinical work has already been ongoing to see if there is any benefit to supplementation with SCFAs, mainly butyrate. More research into the role of SCFAs in PD is needed, however, as some studies report beneficial effects of SCFAs in PD model systems [St. Laurent et al. 2013; Paiva et al. 2017; Zhou et al. 2011; Kidd et al. 2010; Salama et al. 2015], while others have associated them with worse PD relevant outcomes [Sampson et al. 2016] and actually found some SCFAs to be increased in PD patients [Shin et al. 2020]. It is also important to note that deficiency in SCFA producing bacteria is not specific to PD, rather, it has been found in multiple disorders that have an inflammatory component [Kang et al. 2017; Sun et al. 2019; Qin et al. 2012; Guo et al. 2016; Yamada et al. 2015] including other neurological diseases such as multiple systems atrophy [Engen et al. 2017], multiple sclerosis [Cantarel et al. 2015; Tremlett et al. 2016; Jangi et al. 2017], and amyotrophic lateral sclerosis [Fang et al. 2016], therefore, decrease of SCFA bacteria might be a general signature for diseases involving inflammation.

Other PD-gut microbiome signatures that have been replicated in a number of previous studies, including our own, have been enrichment of *Bifidobacterium* [Unger et al. 2016; Hill-Burns et al. 2017; Petrov et al. 2017; Barichella et al. 2018; Lin, A et al. 2018; Aho et al. 2019; Wallen et al. 2020] and *Lactobacillus* (or its family classification

Lactobacillaceae) [Hasegawa et al. 2015; Hill-Burns et al. 2017; Petrov et al. 2017; Aho et al. 2019; Lin CH, et al. 2019; Wallen et al. 2020; Scheperjans et al. 2015; Hopfner et al. 2017; Barichella et al. 2018; Pietrucci et al. 2019; Nishiwaki et al. 2020], although some heterogeneity exists between different populations for the association of *Lactobacillus* [Qian et al. 2018; Li, C et al. 2019]. Bacteria belonging to both of these taxa are commonly referred to as probiotics, and are normal inhabitants of the gut who assist with digestion of carbohydrates from plants and dairy [O'Callaghan & O'Toole 2013; O'Callaghan & van Sinderen 2016]. Similarly to SCFA producing bacteria, we found both genera significantly correlated with daily levodopa dose, increasing in abundance with increasing doses of levodopa [Wallen et al. 2020], which again suggests that alterations of these genera might be a consequence of disease progression and/or PD medication use. Interestingly, *Lactobacillus* species have been shown previously to produce an enzyme, bacterial tyrosine decarboxylase, that can metabolize levodopa into dopamine before it reaches the brain [Zhang & Ni 2014; van Kessel et al. 2019]. Metabolic activity of this enzyme on levodopa reduces its bioavailability, which hinders its effectiveness and requires higher doses of levodopa to reach intended therapeutic effects [van Kessel et al. 2019; Maini Rekdal et al. 2019]. In addition, the human DOPA decarboxylase inhibitor carbidopa was ineffective at inhibiting the activity of bacterial tyrosine decarboxylase [van Kessel et al. 2019; Maini Rekdal et al. 2019], therefore, current regiments of carbidopa/levodopa do not influence bacterial metabolic activity on levodopa. Based on the above evidence for *Lactobacillus* involvement in levodopa metabolism, increased levels of *Lactobacillus* could prove to be a key explanatory factor for PD patients needing increased carbidopa/levodopa dosages. Ironically, *Lactobacillus*, along with *Bifidobacterium*, are commonly included in food products as probiotics, and a clinical trial using fermented milk containing *Lactobacillus, Bifidobacterium*, fiber, and other active ingredients showed benefits for constipation in PD [Barichella et al. 2016]. In certain scenarios, *Lactobacillus* and *Bifidobacterium* can act as opportunistic pathogens, resulting in infection and excessive immune system stimulation in immune-compromised individuals [Suez et al. 2019; Doron & Snydman 2015]. Clearly more work needs to be completed to fully understand the role of increased levels of *Bifidobacterium* and *Lactobacillus* in PD and if they play a beneficial role (potentially as compensatory mechanisms to overcome an unhealthy gut), and/or detrimental role (creating a need for further increases in levodopa dose).

The most novel findings from the final two chapters of this dissertation involved the detection of overabundance of opportunistic pathogens in the PD gut [Wallen et al. 2020], and the establishment of a potential connection between detected opportunistic pathogens and genetic variation in the 3′ end of *SNCA*. These results are intriguing as they hark back to Braak's hypothesis, that non-inherited forms of PD are caused by a pathogen that can pass through the gastrointestinal tract lining and spread to the brain [Braak et al. 2003; Braak et al. 2003], and its derivative that gut infection, and/or imbalance in the gut microbiota, might attribute to formation of pathogenic α-synuclein that travels from the gut to the brain via the vagal nerve. These hypotheses have gained traction in recent years with continued evidence supporting a connection between pathogenic manifestations in the gut and brain in PD, but direct evidence for gut pathogen involvement in PD and/or interaction with α-synuclein has been lacking. Presence of α-synuclein has been shown in the gastrointestinal tract of persons with early

PD [Shannon et al. 2012], Lewy body disease [Breen, Halliday & Lang 2019], and rapid eye movement disorder [Knudsen et al. 2018], which has a high conversion rate to PD. Large epidemiological studies conducted in Scandinavia have suggested a reduction in PD risk for those who have undergone truncal vagotomy years before PD onset [Svensson et al. 2015; Liu et al. 2017]. The protective effect of truncal vagotomy was experimentally supported by a study in mouse that observed truncal vagotomy and endogenous α-synuclein deficiency prevented gut to brain spread of injected preformed α-synuclein fibrils and development of PD-like neurodegeneration and behavioral deficits [Kim et al. 2019]. Studies in human have shown a role of α-synuclein in pathogen response where infection of the gut or olfactory system triggered α-synuclein expression, which in turn mobilized the immune system to respond to the infection [Stolzenberg et al. 2017; Tomlinson et al. 2017]. Connections between pathogenic manifestations in the gut and brain in PD have also been supported experimentally through multiple studies in mice. It has been shown in a *Pink1* knockout mouse model of PD that intestinal infection may act as a trigger for dopaminergic cell loss and motor impairment through activation of T cells in the periphery [Matheoud et al. 2019]. Mice overexpressing a pathogenic mutation in α-synuclein (A53T) showed earlier age at onset of motor dysfunction, and exacerbated dopaminergic neuron death and α-synuclein pathology when mild chronic gut inflammation was induced compared to the same mice without gut inflammation [Kishimoto et al. 2019]. Inoculation of the mouse duodenal intestinal lining with α-synuclein preformed fibrils induced formation of phosphorylated α-synuclein inclusions and promoted inflammation in the gut, disrupted enteric nervous system connectivity, and promoted progression of α-synuclein pathology from the gut to the brain in aged mice

[Challis et al. 2020]. Despite the increasing evidence for a connection between the gut, α-synuclein, and PD, no direct evidence in human has been shown for pathogen involvement. In our studies, we provide direct evidence from human samples that a subset of PD gut microbiomes present with an overabundance of opportunistic pathogens and that the presence of these opportunistic pathogens might increase the risk of PD in combination with genetic variation in the 3′ *SNCA* region. The three genera that we detected as overabundant in PD included *Corynebacterium_1, Porphyromonas,* and *Prevotella*, which represented an even larger polymicrobial cluster. The co-occurrence of these taxa, who per literature search were deemed to be opportunistic pathogens, were also found to co-occur in control subjects at a much lower abundance, suggesting that they can be naturally present in healthier guts, but become overabundant in the right scenarios. Indeed, these taxa have been shown to grow and cause infections if the immune system is compromised or if they are able to penetrate into sterile sites through, for example, compromised membranes [Citron et al. 2007; Wagner Mackenzie et al. 2017; Choi et al. 2019]. As shown in the last chapter of this dissertation, genetic variation in the *SNCA* region might contribute another scenario that allows for overabundance of these opportunistic pathogens as differences between PD patients and healthy individuals in abundances of these taxa were exacerbated with additional copies of particular alleles. Although an initial connection has been made between opportunistic pathogens, α-synuclein, and PD, it is important to emphasize that no claims can be made to the true functional role of these opportunistic pathogens in PD as all analyses performed in these studies are associative in nature, and therefore, do not imply a causal direction. The knowledge on the function of microorganisms in the gut in respect to PD, and in general, is still currently limited, and these microorganisms are usually studied under narrow lenses (e.g. SCFA producing bacteria are deemed anti-inflammatory, therefore, are studied for their beneficial properties, while bacteria deemed opportunistic pathogens are looked for in clinical specimens for potential infections). Finding the identities of these opportunistic pathogens will enable future experimental studies to determine if they play a functional role in PD. Uncovering a connection between these opportunistic pathogens and genetic variation in the *SNCA* region provides the first specific lead for follow up studies in human to confirm these observations, and in experimental models of PD to determine if a true biological interaction between these taxa and genetic variants exist and what the direction of effect is. Hopefully with future experimental studies we can answer trailing questions: Do these opportunistic pathogens play a role in the pathogenesis of PD, and if so, do they influence the progression of PD and its pathological manifestations? Does enrichment of these opportunistic pathogens in conjunction with genetic variation in *SNCA* region make PD initiation more likely? Is increase in these opportunistic pathogens a consequence of disease or disease related factors? Is the increase due to the biological implications of genetic variation in the *SNCA* region, and/or the disease in which the genetic variants make an individual more prone to?

Conclusion

The main theme throughout the studies performed in this dissertation is the generation of new leads for further study, completed for both the variability in PD disease progression (i.e. newly identified genetic modifiers of age at diagnosis) and role of the gut microbiome in PD (i.e. detection of overabundance of opportunistic pathogens in PD gut and establishment of a connection between these taxa and genetic variation in the *SNCA* region in PD). While going into each study with our own hypotheses, we uncovered new leads and additional hypotheses by implementing mostly unbiased, hypothesis-free methodology that allowed signals to reveal themselves to us in absence of our influence. Results described here have importance for both the basic understanding of potential mechanisms for resistance against disease progression, and understanding of the alterations reproducibly occurring in the gut microbiome of PD and potential interactions those alterations have with host genetics. Results from these studies have implications for potential future therapeutics of PD (e.g. increasing neuronal resistance to damage and neuronal regeneration through upregulation of *LPPR1*, or mitigating changes in gut microbiota alterations through supplementation of SCFA bacteria and/or specific antibiotic targeting of opportunistic pathogens) and potential biomarkers for PD prediction (e.g. input of microbiome alterations in conjunction with subject data and host genetic information to machine learning algorithms for predicting likelihood of developing PD). As with most research, much has been learned, but even more questions have now been generated that needs answering. There is more knowledge available for unraveling using additional human studies with even larger sample sizes to increase power and allow analysis of genetic variants genome-wide with gut microbes microbiome-wide, longitudinal studies in both human and animal to track changes in prodromal to advanced disease, experimental studies to assess and tease apart biological functions, and more advanced methodologies (such as shotgun metagenomics, metatranscriptomics, and metabolomics) to broaden the scope of study to other microorganisms (viruses and eukaryotes) and functional outputs (active transcription of microbial genes and metabolite production) as well as improving resolution of microbial detection to strain and gene level. A lot of work is still left to be done, but the potential for new, important discoveries seems to become increasingly more tangible as the work pushes on.

LIST OF GENERAL REFERENCES

de Lau, L.M. & Breteler, M.M. Epidemiology of Parkinson's disease. Lancet Neurol 5, 525-35 (2006).

Kowal, S.L., Dall, T.M., Chakrabarti, R., Storm, M.V. & Jain, A. The current and projected economic burden of Parkinson's disease in the United States. *Mov Disord* **28**, 311-8 (2013).

Pang, S.Y., Ho, P.W., Liu, H. et al. The interplay of aging, genetics and environmental factors in the pathogenesis of Parkinson’s disease. Transl Neurodegener 8, 23 (2019).

Yang, W., Hamilton, J.L., Kopil, C. et al. Current and projected future economic burden of Parkinson’s disease in the U.S.. npj Parkinsons Dis. 6, 15 (2020).

Obeso, J A et al. Past, present, and future of Parkinson's disease: A special essay on the 200th Anniversary of the Shaking Palsy. Mov Disord. 32,9, 1264-1310 (2017). doi:10.1002/mds.27115

Postuma, R.B. et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord. 30(12):1591-601 (2015).

Michel, P.P., Hirsch, E.C. & Hunot, S. Understanding Dopaminergic Cell Death Pathways in Parkinson Disease. Neuron 90, 4, 675-691, (2016).

Dickson DW, Braak H, Duda JE, et al. Neuropathological assessment of Parkinson's disease: refining the diagnostic criteria. Lancet Neurol. 8(12) 1150-1157 (2009).

Kalia LV, Lang AE. Parkinson's disease. Lancet.386(9996):896-912 (2015).

Pont-Sunyer C, Hotter A, Gaig C, et al. The onset of nonmotor symptoms in Parkinson's disease (the ONSET PD study). Mov Disord. 30(2):229-237 (2015).

Tanner, C.M. Advances in Environmental Epidemiology. Mov Disord. 25(1):S58-S62 (2010).

Powers, K. et al. Combined effects of smoking, coffee and NSAIDs on Parkinson’s disease risk. Mov. Disord. 23, 88–95 (2008).

Chang, D. et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson’s disease risk loci. Nat. Genet. 49, 1511–1516 (2017).

Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 18(12):1091-1102 (2019). doi:10.1016/S1474-4422(19)30320-5

Cannon, J. R. & Greenamyre, J. T. Gene-environment interactions in Parkinson’s disease: specific evidence in humans and mammalian models. Neurobiol. Dis. 57, 38–46 (2013).

Hamza, T. H. et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson’s disease modifier gene via interaction with coffee. PLoS Genet. 7, e1002237 (2011).

Hill-Burns, E. M. et al. A genetic basis for the variable effect of smoking/nicotine on Parkinson’s disease. Pharmacogenomics J. 13, 530–537 (2013).

Biernacka, J. M. et al. Genome-wide gene-environment interaction analysis of pesticide exposure and risk of Parkinson’s disease. Parkinsonism Relat. Disord. 32, 25–30 (2016).

Blauwendraat C, Nalls MA, Singleton AB. The genetic architecture of Parkinson's disease. Lancet Neurol. 19(2):170-178 (2020).

Kumar KR, Djarmati-Westenberger A, Grunewald A. Genetics of Parkinson’s disease. Semin Neurol. 31(5):433–40 (2011).

Klein C, Schlossmacher MG. The genetics of Parkinson disease: Implications for neurological care. Nat Clin Pract Neurol. 2(3):136-146 (2006).

Hernandez DG, Reed X, Singleton AB. Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. J Neurochem. 139 Suppl 1(Suppl 1):59-74 (2016).

Zareparsi S, Taylor TD, Harris EL, Payami H. Segregation analysis of Parkinson disease. Am J Med Genet 80:410–417 (1998).

Maher NE, Currie LJ, Lazzarini AM, et al. Segregation analysis of Parkinson disease revealing evidence for a major causative gene. Am J Med Genet 109:191–197 (2002).

McDonnell SK, Schaid DJ, Elbaz A, et al. Complex segregation analysis of Parkinson's disease: the Mayo clinic family study. Ann Neurol 59:788–795 (2006).

Hamza TH, Payami H. The heritability of risk and age at onset of Parkinson's disease after accounting for known genetic risk factors. J Hum Genet. 55:241–243 (2010).

Nalls MA, et al. Polygenic risk of Parkinson disease is correlated with disease age at onset. Ann Neurol 77:582–591 (2015).

Blauwendraat C, Heilbron K, Vallerga CL, et al. Parkinson's disease age at onset genome-wide association study: Defining heritability, genetic loci, and α-synuclein mechanisms. Mov Disord. 34(6):866-875 (2019).

Hill-Burns EM, Ross OA, Wissemann WT, et al. Identification of genetic modifiers of age-at-onset for familial Parkinson's disease. Hum Mol Genet 25:3849–3862 (2016).

Wallen ZD, Chen H, Hill-Burns EM, Factor SA, Zabetian CP, Payami H. Plasticity-related gene 3 (LPPR1) and age at diagnosis of Parkinson disease. Neurol Genet. 4(5):e271 (2018).

Chen H, Huang X, Guo X, et al. Smoking duration, intensity, and risk of Parkinson disease. Neurology 74:878–884 (2010).

Lloyd-Price, J., Abu-Ali, G. & Huttenhower, C. The healthy human microbiome. Genome Med 8, 51 (2016).

Liang, D., Leung, R.K., Guan, W. et al. Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities. Gut Pathog 10, 3 (2018).

Schmidt, T. S. B., Raes, J. & Bork, P. The human gut microbiome: from association to modulation. Cell 172, 1198–1215 (2018).

Segata, N., Haake, S.K., Mannon, P. et al. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 13, R42 (2012).

Venter JC, Remington K, Heidelberg JF, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 304(5667):66-74 (2004).

Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci. 5:209 (2014).

Qin J, et al., MetaHIT Consortium. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464, pp. 59-65 (2010).

Li J, et al., MetaHIT Consortium An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol., 32, pp. 834-841 (2014).

Nelson, K.E. et al. A catalog of reference genomes from the human microbiome Science, 328, pp. 994-999 (2010).

Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature, 486, pp. 207-214 (2012).

Quast C, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596 (2013).

Yilmaz P, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucl. Acids Res. 42:D643-D648 (2014).

Balvočiūtė, M., Huson, D.H. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare?. BMC Genomics 18, 114 (2017).

Pace N. R., Stahl D. A., Lane D. J., Olsen G. J. The analysis of natural microbial populations by ribosomal RNA sequences. Adv. Microb. Ecol. 9 1–55 (1986).

Hugenholtz P, Pace NR. Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. Trends Biotechnol. 14(6):190-7 (1996).

Callahan, B., McMurdie, P. & Holmes, S. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11, 2639–2643 (2017).

Bolyen, E., Rideout, J.R., Dillon, M.R. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857 (2019).

Johnson, J.S., Spakowicz, D.J., Hong, B. et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10, 5029 (2019).

Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 71(12):8228–8235 (2005).

Chen, J. et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28, 2106–2113 (2012).

Robinson MD and Smyth GK (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics, 9, 321-332.

Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12): 550.

Thorsen et al. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome. Nov 25;4(1):62 (2016).

Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27, doi:10.1186/s40168-017-0237-y (2017).

Hawinkel et al. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief Bioinform. 2019 Jan 18;20(1):210-221.

Buttigieg PL, Ramette A. A Guide to Statistical Analysis in Microbial Ecology: a community-focused, living review of multivariate data analyses. FEMS Microbiol Ecol. 90: 543–550 (2014).

Anderson M. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26: 32–46 (2001).

Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. PLoS Comput Biol 8(9): e1002687 (2012).

Blondel, V., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. 2008, P10008 (2008).

Falony, G. et al. Population-level analysis of gut microbiome variation. Science 352, 560–564 (2016).

Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352, 565–569 (2016).

Choo, J., Leong, L. & Rogers, G. Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 5, 16350 (2015).

Wagner Mackenzie B, Waite DW, Taylor MW. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. Front Microbiol. 6:130 (2015).

Yu, Z., García-González, R., Schanbacher, F. L. & Morrison, M. Evaluations of different hypervariable regions of archaeal 16S rRNA genes in profiling of methanogens by Archaea-specific PCR and denaturing gradient gel electrophoresis. Applied and Environmental Microbiology 74, 889–893 (2008).

Klindworth, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research 41, e1–e1 (2013).

Yang, B., Wang, Y. & Qian, P. Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17, 135 (2016).

Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS ONE 15(1): e0227434 (2020).

Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. It's all relative: analyzing microbiome data as compositions. Ann Epidemiol. 26(5):322-329 (2016).

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 8:2224 (2017).

Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 26:27663 (2015).

Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2:15 (2014).

Cersosimo MG, Raina GB, Pecci C, et al. Gastrointestinal manifestations in Parkinson's disease: prevalence and occurrence before motor symptoms. J Neurol. 260(5):1332–1338 (2013).

Chen, H. et al. Meta-analyses on prevalence of selected Parkinson's nonmotor symptoms before and after diagnosis. Transl. Neurodegener. 4, 1 (2015).

Shannon KM, Keshavarzian A, Mutlu E, et al. Alpha-synuclein in colonic submucosa in early untreated Parkinson's disease. Mov Disord. 27(6):709-715 (2012).

Breen, D. P., Halliday, G. M. & Lang, A. E. Gut-brain axis and the spread of alpha-synuclein pathology: Vagal highway or dead end? Mov. Disord. 34, 307–316 (2019).

Knudsen, K. et al. In-vivo staging of pathology in REM sleep behaviour disorder: a multimodality imaging case-control study. Lancet Neurol. 17, 618–628 (2018).

Forsyth CB, Shannon KM, Kordower JH, et al. Increased intestinal permeability correlates with sigmoid mucosa alpha-synuclein staining and endotoxin exposure markers in early Parkinson's disease. PLoS One. 6(12):e28032 (2011).

Svensson, E. et al. Vagotomy and subsequent risk of Parkinson's disease. Ann. Neurol. 78, 522–529 (2015).

Liu, B. et al. Vagotomy and Parkinson disease: a Swedish register-based matched- cohort study. Neurology 88, 1996–2002 (2017).

Kim, S. et al. Transneuronal propagation of pathologic alpha-synuclein from the gut to the brain models Parkinson’s disease. Neuron 103, 627–641 e627 (2019).

Stolzenberg, E. et al. A role for neuronal alpha-synuclein in gastrointestinal immunity. J. Innate Immun. https://doi.org/10.1159/000477990 (2017).

Tomlinson, J. J. et al. Holocranohistochemistry enables the visualization of alpha-synuclein expression in the murine olfactory system and discovery of its systemic anti-microbial effects. J. Neural Transm. (Vienna) 124, 721–738 (2017).

Matheoud, D. et al. Intestinal infection triggers Parkinson’s disease-like symptoms in Pink1(-/-) mice. Nature 571, 565–569 (2019).

Braak, H. et al. Staging of brain pathology related to sporadic Parkinson’s disease. Neurobiol. Aging 24, 197–211 (2003).

Braak, H., Rub, U., Gai, W. P. & Del Tredici, K. Idiopathic Parkinson’s disease: possible routes by which vulnerable neuronal types may be subject to neuroinvasion by an unknown pathogen. J. Neural Transm. (Vienna) 110, 517–536 (2003).

Gerhardt, S. & Mohajeri, M. H. Changes of colonic bacterial composition in Parkinson’s disease and other neurodegenerative diseases. Nutrients 10, https://doi.org/10.3390/nu10060708 (2018).

Boertien, J. M., Pereira, P. A. B., Aho, V. T. E. & Scheperjans, F. Increasing comparability and utility of gut microbiome studies in Parkinson’s disease: a systematic review. J. Parkinsons Dis. 9, S297–S312 (2019).

Sampson TR, Debelius JW, Thron T, et al. Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. Cell. 2016;167(6):1469-1480.e12.

GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348(6235):648-660 (2015).

Soldner, F., Stelzer, Y., Shivalila, C. et al. Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. Nature 533, 95–99 (2016).

Emelyanov, A.K., Andoskin, P.A., Miliukhina, I.V. et al. SNCA alleles rs356219 and rs356165 are associated with Parkinson’s disease and increased α-synuclein gene expression in CD45+ blood cells. Cell Tiss. Biol. 10, 277–283 (2016).

Devine MJ, Gwinn K, Singleton A, Hardy J. Parkinson's disease and α-synuclein expression. Mov Disord. 26(12):2160-2168 (2011).

Savaskan NE, Brauer AU, Nitsch R. Molecular cloning and expression regulation of PRG-3, a new member of the plasticity-related gene family. Eur J Neurosci 2004;19:212–220.

Broggini T, Schnell L, Ghoochani A, et al. Plasticity related gene 3 (PRG3) overcomes myelin-associated growth inhibition and promotes functional recovery after spinal cord injury. Aging (Albany NY) 2016;8:2463–2487.

Fink KL, Lopez-Giraldez F, Kim IJ, Strittmatter SM, Cafferty WBJ. Identification of intrinsic axon growth modulators for intact CNS neurons after injury. Cell Rep 2017;18:2687–2701.

Iweka CA, Tilve S, Mencio C, et al. The lipid phosphatase-like protein PLPPR1 increases cell adhesion by modulating RhoA/Rac1 activity. bioRxiv 470914 (2019); doi: https://doi.org/10.1101/470914

Kalpachidou T, Spiecker L, Kress M, Quarta S. Rho GTPases in the Physiology and Pathophysiology of Peripheral Sensory Neurons. Cells. 8(6):591 (2019).

Villar-Cheda B, Dominguez-Meijide A, Joglar B, et al. Involvement of microglial RhoA/Rho-Kinase pathway activation in the dopaminergic neuron death. Role of angiotensin via angiotensin type 1 receptors. Neurobiol. Dis. 47(2): 268-279 (2012).

Mattii L, Pardini C, Ippolito C, et al. Rho-inhibition and neuroprotective effect on rotenone-treated dopaminergic neurons in vitro. Neurotoxicology. 2019;72:51-60.

Hou L, Bao X, Zang C, et al. Integrin CD11b mediates α-synuclein-induced activation of NADPH oxidase through a Rho-dependent pathway. Redox Biol. 2018;14:600-608.

Bogetofte H, Jensen P, Okarmus J, et al. Perturbations in RhoA signalling cause altered migration and impaired neuritogenesis in human iPSC-derived neural cells with PARK2 mutation. Neurobiol Dis. 2019;132:104581.

Musilli M, Ciotti MT, Pieri M, et al. Therapeutic effects of the Rho GTPase modulator CNF1 in a model of Parkinson's disease. Neuropharmacology. 2016;109:357-365.

Axelsen TM, Woldbye DPD. Gene Therapy for Parkinson's Disease, An Update. J Parkinsons Dis. 2018;8(2):195-215.

Keshavarzian, A. et al. Colonic bacterial composition in Parkinson's disease. Mov. Disord. 30, 1351–1360 (2015).

Unger, M. M. et al. Short chain fatty acids and gut microbiota differ between patients with Parkinson's disease and age-matched controls. Parkinsonism Relat. Disord. https://doi.org/10.1016/j.parkreldis.2016.08.019 (2016).

Hill-Burns, E. M. et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. Mov. Disord. 32, 739–749 (2017).

Petrov, V. A. et al. Analysis of gut microbiota in patients with Parkinson's disease. Bull. Exp. Biol. Med. 162, 734–737 (2017).

Li, W. et al. Structural changes of gut microbiota in Parkinson's disease and its correlation with clinical features. Sci. China Life Sci. 60, 1223–1233 (2017).

Lin, A. et al. Gut microbiota in patients with Parkinson's disease in southern China. Parkinsonism Relat. Disord. 53, 82–88 (2018).

Pietrucci, D. et al. Dysbiosis of gut microbiota in a selected population of Par- kinson's patients. Parkinsonism Relat. Disord. https://doi.org/10.1016/j. parkreldis.2019.06.003 (2019).

Aho, V. T. E. et al. Gut microbiota in Parkinson's disease: temporal stability and relations to disease progression. EBioMedicine 44, 691–707 (2019).

Nishiwaki H, Ito M, Ishida T, et al. Meta-Analysis of Gut Dysbiosis in Parkinson's Disease [published online ahead of print, 2020 Jun 18]. Mov Disord. 2020; doi:10.1002/mds.28119

Wallen, Z.D., Appah, M., Dean, M.N. et al. Characterizing dysbiosis of gut microbiome in PD: evidence for overabundance of opportunistic pathogens. npj Parkinsons Dis. 6, 11 (2020).

Barichella, M. et al. Unraveling gut microbiota in Parkinson's disease and atypical parkinsonism. Mov. Disord. 34, 396–405 (2019).

Hamer, H. M. et al. Review article: the role of butyrate on colonic function. Aliment Pharm. Ther. 27, 104–119 (2008).

Canani, R. B. et al. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J. Gastroenterol. 17, 1519–1528 (2011).

Park, J., Wang, Q., Wu, Q., Mao-Draayer, Y. & Kim, C. H. Bidirectional regulatory potentials of short-chain fatty acids and their G-protein-coupled receptors in autoimmune neuroinflammation. Sci. Rep. 9, 8837 (2019).

Haase, S., Haghikia, A., Wilck, N., Muller, D. N. & Linker, R. A. Impacts of microbiome metabolites on immune regulation and autoimmunity. Immunology 154, 230–238 (2018).

Furusawa, Y. et al. Commensal microbe-derived butyrate induces the differ- entiation of colonic regulatory T cells. Nature 504, 446–450 (2013).

St. Laurent R, O'Brien LM, Ahmad ST. Sodium butyrate improves locomotor impairment and early mortality in a rotenone-induced Drosophila model of Parkinson's disease. Neuroscience. (2013) 246:382–90.

Paiva I, Pinho R, Pavlou MA, et al. . Sodium butyrate rescues dopaminergic cells from alpha-synuclein-induced transcriptional deregulation and DNA damage. Hum Mol Genet. (2017) 26:2231–46.

Zhou W, Bercury K, Cummiskey J, et al. Phenylbutyrate up-regulates the DJ-1 protein and protects neurons in cell culture and in animal models of Parkinson disease. J Biol Chem. (2011) 286:14941–51.

Kidd SK, Schneider JS. Protection of dopaminergic cells from MPP+-mediated toxicity by histone deacetylase inhibition. Brain Res. (2010) 1354:172–8.

Salama AF, Ibrahim W, Tousson E, et al. Epigenetic study of Parkinson's disease in experimental animal model. Int J Clin Exp Neurol. (2015) 3:11–20.

Shin C, Lim Y, Lim H, Ahn TB. Plasma Short-Chain Fatty Acids in Patients With Parkinson's Disease. Movement Disorders : Official Journal of the Movement Disorder Society. 2020 Jun;35(6):1021-1027.

Kang, C. et al. Gut microbiota mediates the protective effects of dietary capsaicin against chronic low-grade inflammation and associated obesity induced by high- fat diet. MBio 8, https://doi.org/10.1128/mBio.00470-17 (2017).

Sun, Q., Jia, Q., Song, L. & Duan, L. Alterations in fecal short-chain fatty acids in patients with irritable bowel syndrome: a systematic review and meta-analysis. Med. (Baltimore) 98, e14513 (2019).

Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55–60 (2012).

Guo, Z. et al. Intestinal microbiota distinguish gout patients from healthy humans. Sci. Rep. 6, 20602 (2016).

Yamada, T. et al. Rapid and sustained long-term decrease of fecal short-chain fatty acids in critically ill patients with systemic inflammatory response syndrome. JPEN J. Parenter. Enter. Nutr. 39, 569–577 (2015).

Engen PA, Dodiya HB, Naqib A, et al. The potential role of gut-derived inflammation in multiple system atrophy. J. Parkinsons Dis. 2017, 7, 331–346.

Cantarel BL, Waubant E, Chehoud C, et al. Gut microbiota in multiple sclerosis: possible influence of immunomodulators. J Investig Med. 2015;63(5):729-734.

Tremlett H, Fadrosh DW, Faruqi AA, et al. Gut microbiota in early pediatric multiple sclerosis: a case-control study. Eur J Neurol. 2016;23(8):1308-1321.

Jangi S, Gandhi R, Cox LM, et al. Alterations of the human gut microbiome in multiple sclerosis. Nat Commun. 2016;7:12015.

Fang X, Wang X, Yang S, et al. Evaluation of the Microbial Diversity in Amyotrophic Lateral Sclerosis Using High-Throughput Sequencing. Front Microbiol. 2016;7:1479.

Hasegawa, S. et al. Intestinal dysbiosis and lowered serum lipopolysaccharide- binding protein in Parkinson’s disease. PLoS ONE 10, e0142164 (2015).

Lin, C. H. et al. Altered gut microbiota and inflammatory cytokine responses in patients with Parkinson’s disease. J. Neuroinflammation 16, 129 (2019).

Scheperjans, F. et al. Gut microbiota are related to Parkinson’s disease and clinical phenotype. Mov. Disord. 30, 350–358 (2015).

Hopfner, F. et al. Gut microbiota in Parkinson disease in a northern German cohort. Brain Res. 1667, 41–45 (2017).

Qian, Y. et al. Alteration of the fecal microbiota in Chinese patients with Par- kinson’s disease. Brain Behav. Immun. 70, 194–202 (2018).

Li, C. et al. Gut microbiota differs between Parkinson’s disease patients and healthy controls in Northeast China. Front Mol. Neurosci. 12, 171 (2019).

O’Callaghan, J. & O’Toole, P. W. Lactobacillus: host-microbe relationships. Curr. Top. Microbiol. Immunol. 358, 119–154 (2013).

O’Callaghan, A. & van Sinderen, D. Bifidobacteria and their role as members of the human gut microbiota. Front. Microbiol. 7, 925 (2016).

Zhang, K. & Ni, Y. Tyrosine decarboxylase from Lactobacillus brevis: soluble expression and characterization. Protein Expr. Purif. 94, 33–39 (2014).

van Kessel SP, Frye AK, El-Gendy AO, et al. Gut bacterial tyrosine decarboxylases restrict levels of levodopa in the treatment of Parkinson's disease. Nat Commun. 2019;10(1):310.

Maini Rekdal, V., Bess, E. N., Bisanz, J. E., Turnbaugh, P. J. & Balskus, E. P. Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. Science 364, https://doi.org/10.1126/science.aau6323 (2019).

Barichella, M. et al. Probiotics and prebiotic fiber for constipation associated with Parkinson disease: An RCT. Neurology 87, 1274–1280 (2016).

Kishimoto Y, Zhu W, Hosoda W, Sen JM, Mattson MP. Chronic Mild Gut Inflammation Accelerates Brain Neuropathology and Motor Dysfunction in α-Synuclein Mutant Mice. Neuromolecular Med. 2019;21(3):239-249.

Challis C, Hori A, Sampson TR, et al. Gut-seeded α-synuclein fibrils promote gut dysfunction and brain pathology specifically in aged mice. Nat Neurosci. 2020;23(3):327-336.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR “PLASTICITY-RELATED GENE 3 (*LPPR1*) AND AGE AT DIAGNOSIS OF PARKINSON DISEASE”

Table e-1. Assessing interdependence of the two association signals using conditional analysis

			Main effect		Conditional effect		
		Test SNP	HR	*P*	Covariate SNP	HR	*P*
NGRC	Block 1	rs73656147	1.95	3E-6	rs17763929	1.35	0.10
	Block 2	rs17763929	1.88	5E-8	rs73656147	1.62	1E-3
PAGE Prevalent	Block 1	rs73656147	2.88	7E-4	rs17763929	4.75	3E-3
		Adj PC1-3	2.17	0.05	Adj PC1-3	1.88	0.23
	Block 2	rs17763929	1.87	0.01	rs73656147	0.94	0.44
		Adj PC1-3	3.03	4E-3	Adj PC1-3	2.24	0.10
PAGE Incident	Block 1	rs73656147	1.62	0.07	rs17763929	1.66	0.07
		Adj PC1-3	1.48	0.16	Adj PC1-3	1.50	0.16
	Block 2	rs17763929	1.04	0.41	rs73656147	0.96	0.42
		Adj PC1-3	1.03	0.45	Adj PC1-3	0.97	0.45

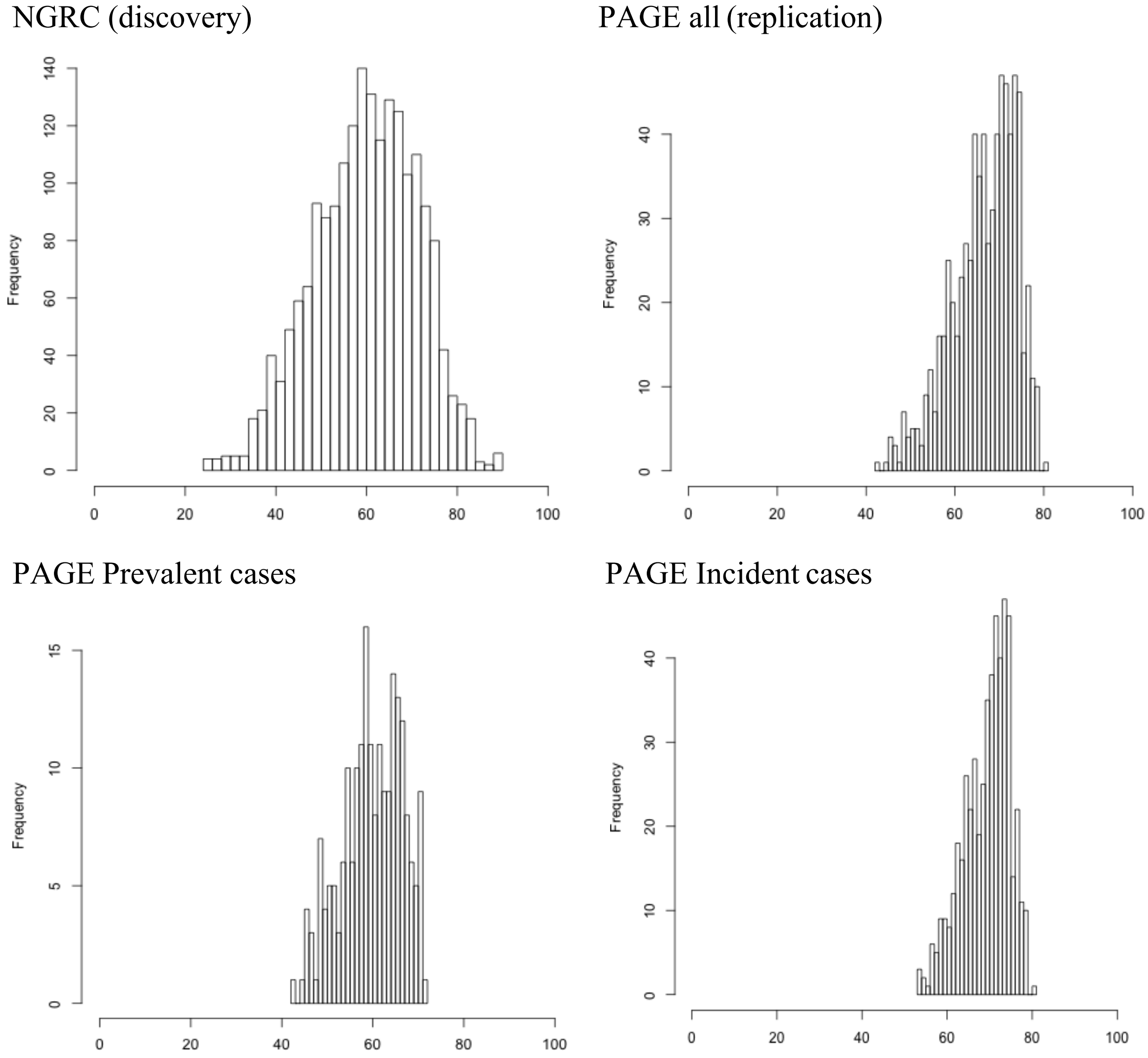
Main effect is the association of test SNP with age-at-diagnosis. Conditional effect is the association of test SNP with age-at-diagnosis adjusted for covariate SNP. HR (hazard ratio) is the age-for-age increase in the odds of event (PD diagnosis) per copy of the minor allele estimated using Cox regression, with its associated significance (P). NGRC was adjusted for PC1-3. PAGE had ancestry informative markers (AIMs) for about half of the participants. For PAGE, the first row for each SNP is using the total sample size without adjusting for PCs, and the second row for each SNP is using the subset of samples with AIMs and adjusting for PC1-3. P values are two- sided for NGRC, and one-sided for PAGE.

Table e-2. Investigating robustness and heterogeneity in the association signals.

		Block 1				Block 2			
		rs73656147				rs17763929			
	N	MAF	HR	P	P Het	MAF	HR	P	P Het
All	1950	0.013	1.95	3E-06		0.022	1.88	5E-08	
PD-associated risk factors									
Familial	424	0.013	1.84	0.05		0.023	2.09	3E-03	
Sporadic	1526	0.013	2.00	2E-05	0.81	0.022	1.84	4E-06	0.65
Male	1312	0.012	2.24	8E-06		0.020	1.81	9E-05	
Female	638	0.015	1.59	0.05	0.25	0.028	1.94	2E-04	0.77
Smokers	724	0.016	1.83	4E-03		0.027	1.89	2E-04	
Non-smokers	852	0.012	1.87	6E-03	0.94	0.019	1.74	5E-03	0.76
Coffee High	610	0.014	1.88	0.01		0.027	1.69	6E-03	
Coffee Low	828	0.013	1.86	4E-03	0.98	0.020	2.07	2E-04	0.47
OTC NSAIDs - ever	963	0.015	2.05	2E-04		0.026	1.70	7E-04	
OTC NSAIDs - never	582	0.011	1.63	0.08	0.50	0.016	1.79	0.02	0.86
Recruitment site									
New York	410	0.016	2.86	2E-04		0.024	2.83	1E-05	
Oregon	473	0.012	2.39	5E-03		0.025	1.88	4E-03	
Georgia	230	0.024	1.17	0.62		0.027	0.89	0.71	
Washington	837	0.010	1.93	9E-03	0.11	0.019	2.04	3E-04	3E-03
Ashkenazi Jewish									
Yes	88	0.017	2.88	0.09		0.069	1.24	0.51	
No	1862	0.013	1.95	6E-06	0.54	0.020	2.05	7E-09	0.15
Paternal or maternal ancestry									
Great Britain	527	0.009	2.67	3E-03		0.014	2.59	7E-04	
Germany / Austria	434	0.012	1.88	0.05		0.022	2.54	1E-04	
Ireland	245	0.019	1.22	0.57		0.038	1.37	0.22	
Scandinavia	223	0.005	1.83	0.38		0.013	2.28	0.09	
Eastern Europe	91	0.027	4.37	3E-03		0.033	2.39	0.06	
Italy	89	0.034	4.45	1E-03		0.041	4.14	6E-04	
France	80	0.016	4.79	0.03		0.028	5.53	2E-03	
Russia	62	0.008	4.28	0.16	0.35	0.049	0.69	0.44	1E-04

The NGRC dataset was stratified by variables relevant to PD, association of each SNP with age-at-diagnosis was tested within each stratum adjusted for PC 1-3, using Cox regression to generate hazard ratio (HR) and significance (P). The results across strata were then compared for evidence of heterogeneity (P Het). Association signal for block 1 (rs73656147) was relatively consistent across strata with so significant evidence for heterogeneity (this block was also robustly replicated in PAGE dataset with no evidence of heterogeneity). The signal for block 2 (rs17763929) varied significantly as a function of geographic origin of participants, both for their current residence within US and their European country of origin (this SNP was significantly associated with PC1 and PC3, and gave evidence for heterogeneity in PAGE dataset as well). MAF=minor allele frequency. HR=hazard ratio using Cox regression. P=statistical significance of HR. P Het=statistical significance of heterogeneity across strata. Early-onset PD: age ≤50 years at onset of motor symptom. Late-onset PD: age >50 years at onset of motor symptom. Smoker: ≥100 cigarettes in lifetime. Coffee: Number of cups of caffeinated coffee drank per day multiplied by the number of years of consumption; high and low divided at the median in NGRC participants with PD. OTC NSAIDs: Ever or never use of over the counter non-steroidal anti-inflammatory drugs. Jewish/Non-Jewish: Defined by self-report, and verified by principal component analysis (the core of the Jewish cluster was defined within 0.04≤PC1≤0.055 and 0.001≤PC2≤0.013). Recruitment site: US states where participants were recruited from. Paternal or maternal ancestry: Self reports of the countries from which ancestors immigrated to US.

Figure e-1. Distribution of age-at-diagnosis.



NGRC (discovery)

PAGE all (replication)

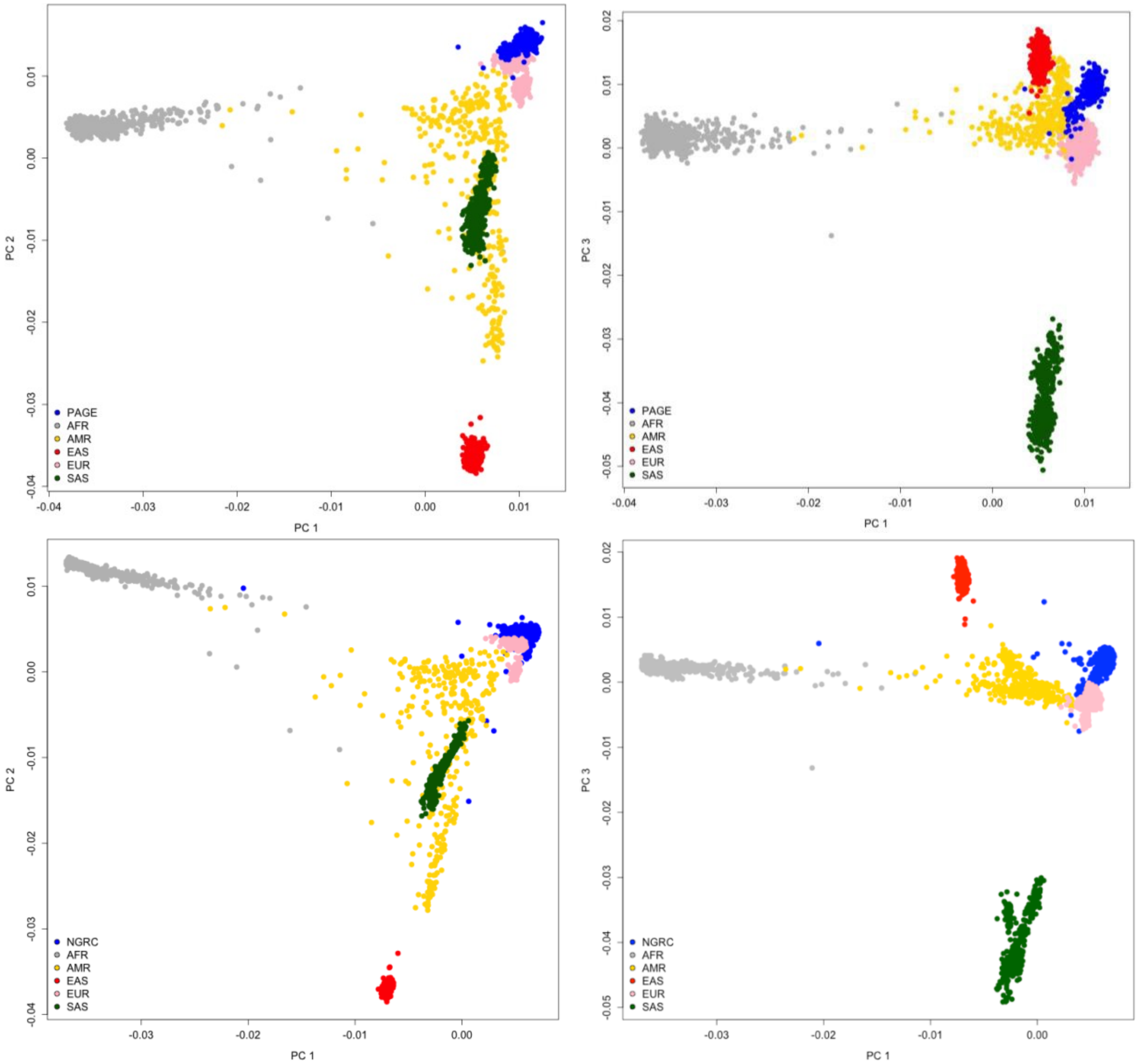
PAGE Prevalent cases

PAGE Incident cases

Figure e-2. Principal Component Analysis (PCA) plots.

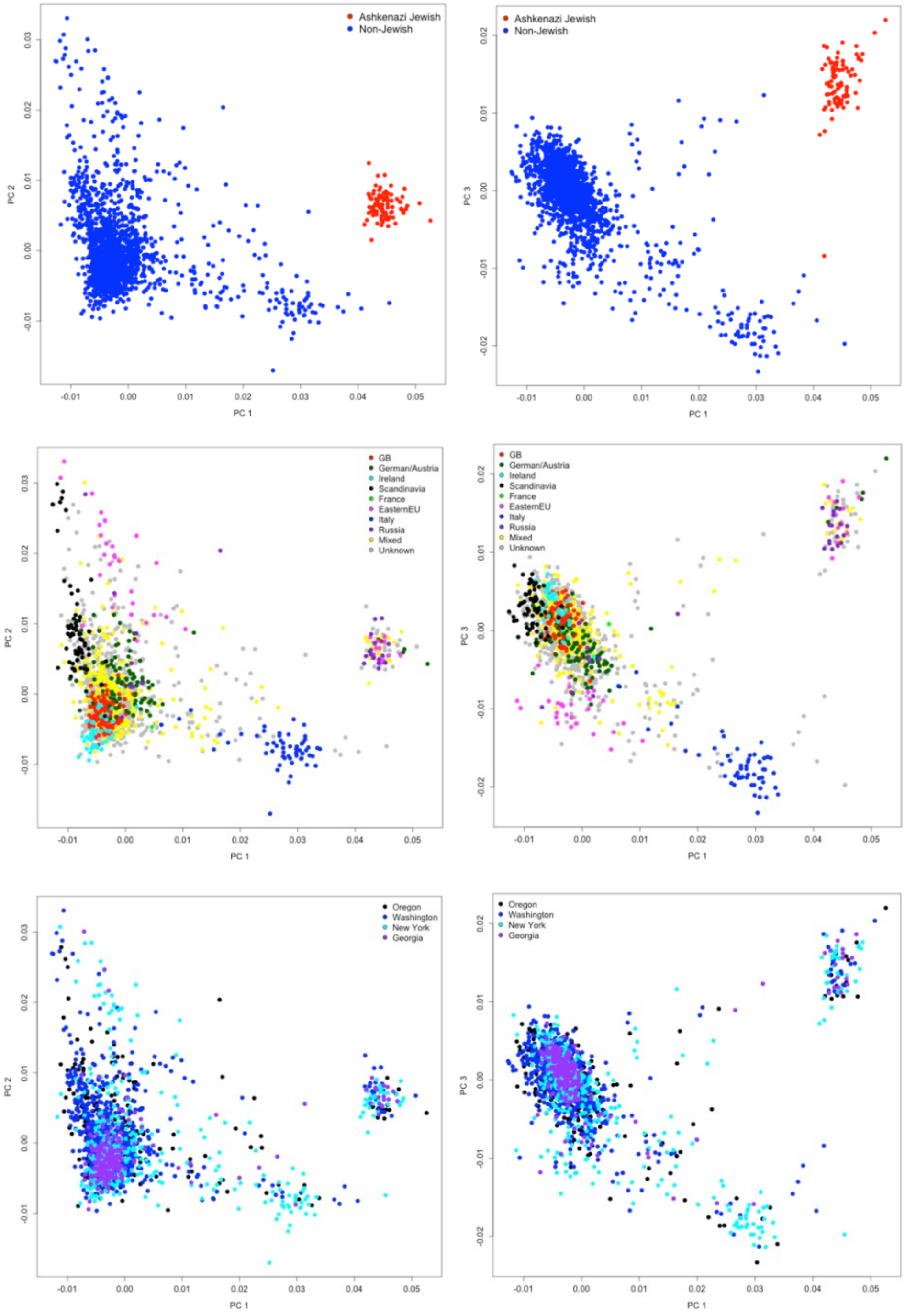
A. NGRC and PAGE cluster with Europeans in 1000G_Phase_3.

PC1 vs. PC2 (left panel) and PC1 vs. PC3 (right panel) were plotted for NGRC participants with PD (top row, N=1950, using 100K pruned SNPs), and for subset of PAGE participants for whom PCs could be calculated (bottom row, N=396, using 20K pruned SNPs). Blue: NGRC (first row) or PAGE (second row). Pink: European. Yellow: Americas. Green: South Asia. Red: East Asia. Grey: African.



B. PCA of NGRC by Jewish ancestry, European country of origin, and US enrollment site.

First row shows clustering of Jewish ancestry (red) in NGRC. The core of the Jewish cluster was defined within 0.04≤PC1≤0.055 and 0.001≤PC2≤0.013. Second row is the self-reported European country of ancestry, which follows the map of Europe. In the third row, NGRC participants are colored according to the state where they were enrolled, showing no particular clustering for enrollment site.



C. Sub-haplotypes & ancestral origins.

As there seems to be an ancestral effect, we check whether different sub- haplotypes are present in these blocks for the different ancestral groups in NGRC participants with PD. Block 1 (left triangle in the haploview) is a single block (LD>0.9) represented by rs73495940. Block 2 (right triangle in the haploview) is composed of several sub-haplotypes (LD≤0.7). To tag these haplotypes, while incorporating functional information, we used the 5 variants in block 2 that had significant evidence for functional relevance (either HiC FDR<1E-6 and enhancer in brain, or CADD>10, or eQTL with FDR=4E-4 see table 4). As shown in the LD table below, 3 of the 5 variants are in close LD (r2>0.95) which can be tagged by one SNP (rs17763929). We defined 3 sub-haplotypes in block 2 represented by rs17763929, rs6118842, and rs149155028. We identified the individuals who carried the minor allele of each variant, and mapped them to PC1 vs. PC2 (top row) and PC1 vs. PC3 (bottom row). There is no pattern that would suggest sub- haplotypes cluster with different ancestral group.

LD in Block 2	rs17763929	rs61188842	rs117058418	rs117314512	rs149155028
rs17763929	-	0.58	0.96	0.96	0.66
rs61188842		-	0.57	0.57	0.39
rs117058418			-	1	0.7
rs117314512				-	0.7
rs149155028					-

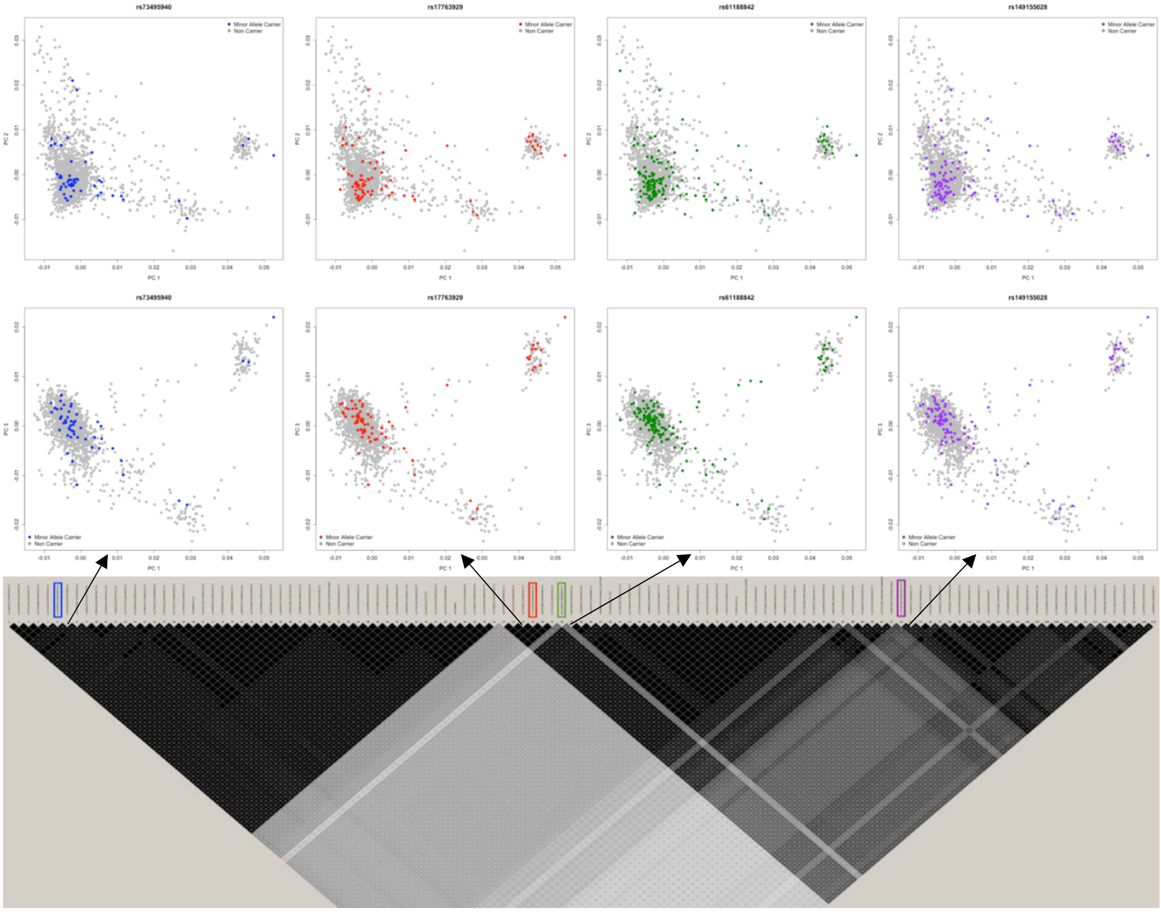
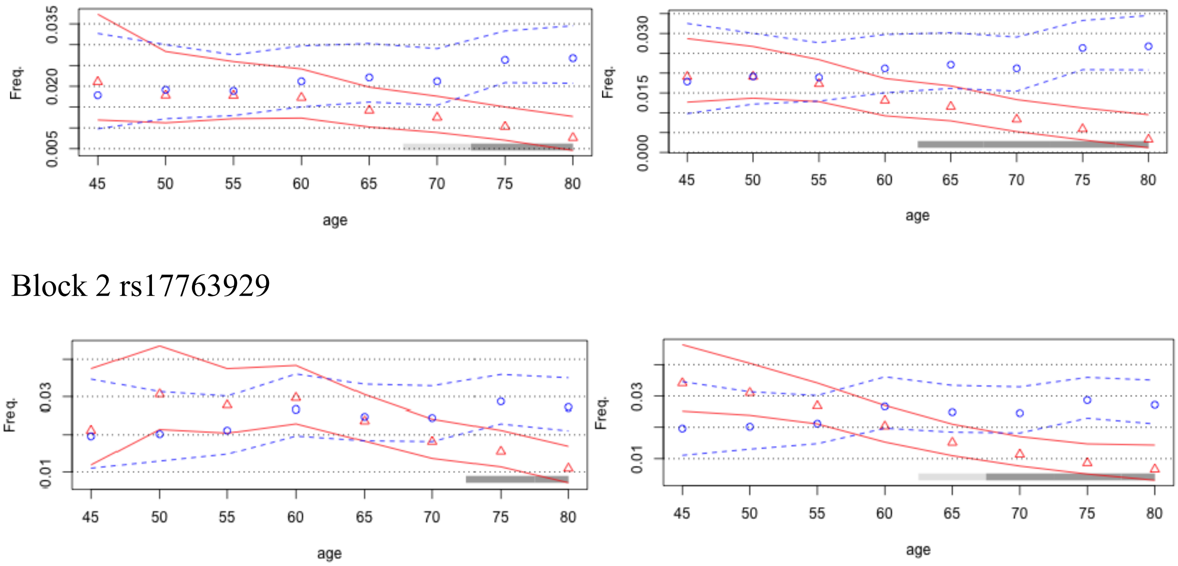


Figure e-3. Moving average allele frequency plots (MAP) of PD-associated variants in *LPPR1*



A. Age in cases (red) vs. age in controls (blue), N=2000 PD, 1986 controls

B. Age at diagnosis in cases (red) vs. age in controls (blue), N=1950 cases, 1986 controls

Block 1 rs73656147

Block 2 rs17763929

To visualize the dynamics of allele frequency changes as a function of age and age-at-diagnosis, average minor allele frequencies (MAF) were plotted in a moving window across the age spectrum, using MAP software (freqMAP_v_0.2 in R). NGRC dataset was used. Average MAF (and 95% central posterior interval) was plotted by age in controls (blue circles, N=1986), and age (panel A, N=2000) or age-at-diagnosis (panel B, N=1950) in patients (red triangles). Ages and ages-at-diagnosis ≤45 were collapsed to 45, and ≥80 were collapsed to 80. Significance of difference in MAF between cases and controls is shown by a light gray bar (≥95% posterior probability) or dark gray bar (≥99% posterior probability). The patterns show minor allele frequencies declined by increasing age and age-at-diagnosis in cases, but not in controls, which is consistent with pattern expected for a modifier. Specifically, the MAF for rs73656147 started at ~0.02 at age 45 in both cases and controls and declined steadily as a function of age and age-at-diagnosis in cases, but not in controls, ending at age 80, with MAF~0.009 in cases and MAF~0.028 in controls. The MAF for rs17763929 started at ~0.03 in cases and ~0.02 in controls, decreased by age and age-at-diagnosis in cases but not in controls, ending by age 80 at MAF~0.01 in cases and ~0.03 in controls. Based on statistics (grey bars here, and conditional analysis in table 2), the decline in MAF is significant in cases for both SNPs, is driven by age-at-diagnosis, and retains significance when adjusted for age. Conditional analysis (table 2) suggests the primary driver of allele frequency decline is the association with age- at-diagnosis, and that the age effect in cases is a by-product of the correlation between age and age-at-diagnosis.

APPENDIX B

SUPPLEMENTARY MATERIAL FOR "CHARACTERIZING DYSBIOSIS OF GUT MICROBIOME IN PD: EVIDENCE FOR OVERABUNDANCE OF OPPORTUNISTIC PATHOGENS"

Supplementary Table 1. Subject Data

			Dataset 1					Dataset 2				
			PD		Control			PD		Control		
			N with data	Summary statistics	N with data	Summary statistics	P	N with data	Summary statistics	N with data	Summary statistics	P
		Number of subjects enrolled with complete data	212	-	136	-	-	323	-	184	-	-
	Microbiome	Number of subjects whose 16S sequences passed QC	201	-	132	-	-	323	-	184	-	
		Number of unique ASVs detected	201	4,863	132	3,315	-	323	9,188	184	6,667	
		Number of genera detected	201	404	132	333	-	323	527	184	441	
	Metadata	Number of subjects who passed sequence and metadata QC	199	-	132	-	-	323	-	184	-	
1	Age & Sex	Age	199	68.3±9.2	132	70.2±8.6	0.04	323	67.7±9.0	184	66.4±8.3	0.05
2		Sex (N & % male)	199	133 (67%)	132	52 (39%)	1E-06	323	206 (64%)	184	55 (30%)	2E-13
3	Geography	Seattle, WA	199	93	132	58	-	323	0	184	0	-
		Albany, NY		75		62	-		0		0	-
		Atlanta, GA		31		12	-		0		0	-
		Birmingham, AL		0		0	-		323		184	-
4		Stool sample travel time in days	190	3.3±1.9	129	2.6±1.5	2E-03	314	5.2±3.3	183	5.0±2.6	0.73
5	Race	Race (N & %White)	199	196 (98%)	132	132 (100%)	0.28	321	317 (99%)	184	183 (>99%)	0.66
6	Weight	BMI	192	26.6±5.5	128	28.3±5.7	0.02	312	27.4±5.0	180	27.9±5.9	0.62
7		Lost >10 pounds in past year	195	45 (23%)	126	15 (12%)	0.01	316	79 (25%)	181	21 (12%)	3E-04
8		Gained >10 pounds in past year	196	26 (13%)	129	10 (8%)	0.15	309	45 (15%)	179	20 (11%)	0.33
9	Diet	Fruits or vegetables daily	194	151 (78%)	131	116 (89%)	0.02	-	-	-	-	-
10		Meat, fish, poultry daily	193	110 (57%)	131	82 (63%)	0.36	-	-	-	-	-
11		Nuts daily	194	43 (22%)	130	36 (28%)	0.29	-	-	-	-	-
12		Yogurt at least a few times a week	191	68 (36%)	128	57 (45%)	0.13	-	-	-	-	-
13		Grains daily	192	132 (69%)	129	86 (67%)	0.72	-	-	-	-	-
14		Alcohol	194	116 (60%)	131	93 (71%)	0.05	320	133 (42%)	181	101 (56%)	3E-03
15		Tobacco	196	14 (7%)	131	5 (4%)	0.24	321	13 (4%)	183	13 (7%)	0.15
16		Caffeine	193	137 (71%)	131	100 (76%)	0.31	319	273 (86%)	183	161 (88%)	0.50

17	GI Health	Constipation (no bowel movement) in ≥3 days prior to stool collection	196	29 (15%)	129	2 (2%)	3E-05	302	54 (18%)	176	8 (5%)	2E-05
18		Diarrhea on the day of stool collection	196	6 (3%)	129	2 (2%)	0.49	301	12 (4%)	178	5 (3%)	0.61
19		GI pain on the day of stool collection	164	14 (9%)	120	8 (7%)	0.66	303	27 (9%)	179	3 (2%)	1E-03
20		Excess gas on the day of stool collection	196	27 (14%)	130	2 (2%)	9E-05	303	47 (16%)	180	8 (4%)	2E-04
21		Bloating on the day of stool collection	197	20 (10%)	131	3 (2%)	7E-03	305	36 (12%)	179	9 (5%)	0.01
22		GI discomfort on the day of stool collection (yes to any item 17-21)	183	104 (57%)	119	26 (22%)	2E-09	305	103 (34%)	176	26 (15%)	4E-06
23		Constipation (<3 bowel movements per week) in the past 3 months	191	82 (43%)	130	6 (5%)	6E-16	312	138 (44%)	180	31 (17%)	6E-10
24		Diarrhea in the past 3 months	189	32 (17%)	127	28 (22%)	0.31	306	80 (26%)	181	54 (30%)	0.40
25		Colitis	192	9 (5%)	130	2 (2%)	0.21	316	54 (17%)	180	24 (13%)	0.31
26		IBS	191	14 (7%)	130	8 (6%)	0.82	312	17 (5%)	178	14 (8%)	0.34
27		Crohn's disease	193	4 (2%)	131	1 (1%)	0.65	314	3 (1%)	180	0 (0%)	0.56
28		IBD	193	5 (3%)	130	2 (2%)	0.71	307	9 (3%)	178	4 (2%)	0.78
29		Ulcers	192	18 (9%)	130	9 (7%)	0.54	314	6 (2%)	180	4 (2%)	1.00
30		SIBO	-	-	-	-	-	305	0 (0%)	177	0 (0%)	1.00
31		Celiac	-	-	-	-	-	314	0 (0%)	177	0 (0%)	1.00
32		GI cancer	-	-	-	-	-	315	1 (<1%)	179	1 (<1%)	1.00
33		Intestinal disease (yes to any item 25-32)	193	38 (20%)	131	19 (15%)	0.24	298	80 (27%)	173	41 (24%)	0.51
34	Medications	Currently taking digestive medication	192	60 (31%)	126	22 (17%)	6E-03	-	-	-	-	-
35		Currently taking antibiotics	193	8 (4%)	130	3 (2%)	0.54	315	13 (4%)	179	7 (4%)	1.00
36		Taken antibiotics in past 3 months	190	24 (13%)	130	22 (17%)	0.33	308	63 (20%)	170	33 (19%)	0.81
37		Currently taking anti-inflammatory drugs	190	77 (41%)	128	56 (44%)	0.64	-	-	-	-	-
38		Currently taking probiotics	184	42 (23%)	128	33 (26%)	0.59	-	-	-	-	-
39	Parkinson Disease	Disease duration in years	199	13.8±6.7	-	-	-	323	9.2±7.1	-	-	-
40		Patients on carbidopa/levodopa	187	170 (91%)	-	-	-	313	266 (85%)	-	-	-
41		Levodopa dose, mg/day	181	764±574	-	-	-	313	563±443	-	-	-

42	Duration & Medications	Patients on dopamine agonist	187	99 (53%)	-	-	-	303	153 (50%)	-	-	-
43		Patients on MAO-B inhibitor	187	71 (38%)	-	-	-	318	86 (27%)	-	-	-
44		Patients on amantadine	187	49 (26%)	-	-	-	315	60 (19%)	-	-	-
45		Patients on COMT inhibitor	187	37 (20%)	-	-	-	320	13 (4%)	-	-	-
46		Patients on anticholinergics	187	8 (4%)	-	-	-	322	10 (3%)	-	-	-
47		Patients not on PD medication	187	3 (2%)	-	-	-	316	17 (5%)	-	-	-

Column "N with data" shows the number of individuals for whom data on the specified variable was available; for all metadata, only subjects who passed both sequence and metadata quality control (QC) were considered. "Summary statistics" for metadata are shown as mean±SD for quantitative traits, and number and percentage of individuals with positive response (yes) for dichotomous traits. 15 samples (all in dataset 1) yielded no or too few 16S sequences to be analyzed and were removed. Two subjects had unreliable self-reported metadata; they were included in analyses that required only sequences and case-control status but were excluded from all analyses that required any metadata (these subjects are identified as 10122.FP0016201 and 10122.GMWA.1090 in the dataset on NCBI SRA). P-values are two-sided testing the difference in the distribution of each variable in PD vs. control. Variables that differed in PD vs. control at a conservatively uncorrected two-sided P<0.05 were carried forward and included with case-control status in PERMANOVA and tested for their effects on inter-individual differences in microbiome composition (β diversity). Constipation (no bowel movement) in ≥3 days prior to stool collection, GI pain on day of stool collection, Excess gas on day of stool collection, and Bloating on day of stool collection were captured by GI discomfort on day of stool collection, hence only GI discomfort on day of stool collection was carried forward to PERMANOVA. Currently taking digestive medication (mainly laxatives or antacid) was not carried to PERMANOVA because it was no longer significant when adjusted for GI discomfort on day of stool collection.

Supplementary Table 2. MWAS of dataset 1 conducted using ANCOM

Sample size for ANCOM included subset of samples that had complete data on all covariates tested: N= 171 cases and 117 controls in dataset 1.
W= ANCOM score indicating the number of times a genus achieved FDR<0.05 as compared to other genera (maximum W possible: 444 in dataset 1, 560 in dataset 2).
0.8= Threshold at which results were considered significant (TRUE).

W	0.8	Kingdom	Phylum	Class	Order	Family	Genus
441	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Agathobacter
426	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospira
418	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_ND3007_group
411	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
410	TRUE	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
410	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia
407	TRUE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
406	TRUE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas
400	TRUE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella
393	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Hungatella
391	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
388	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Fusicatenibacter
384	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-004
382	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Butyricicoccus
378	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Ezakiella
376	TRUE	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Cloacibacillus
374	TRUE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Megasphaera
372	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_3
368	TRUE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Coprobacillus
367	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillospira
365	TRUE	Bacteria	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Akkermansiaceae	Akkermansia
360	TRUE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium_1
356	TRUE	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
347	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerostipes
331	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	NA
327	FALSE	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	Methanobrevibacter
326	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	UBA1819
323	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-013

319	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Anaerococcus
318	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-004
306	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus
302	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Varibaculum
293	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	NA
275	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Mobiluncus
263	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	NA	NA
252	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_NK4B4_group
249	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Peptoniphilus
65	FALSE	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	NA
31	FALSE	Bacteria	Tenericutes	Mollicutes	Anaeroplasmatales	Anaeroplasmataceae	Anaeroplasma
20	FALSE	NA	NA	NA	NA	NA	NA
19	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Fournierella
17	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Alcaligenes
16	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotellaceae_UCG-001
15	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_6
15	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	NA	NA
15	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	NA
14	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	NA
14	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Bilophila
14	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptoclostridium
14	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	CHKCI002
14	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Paenalcaligenes
14	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	env.OPS_17	NA
14	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Pseudocitrobacter
14	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Leuconostoc
14	FALSE	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Nocardioides
14	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium
14	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Pelomonas
14	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter
14	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelothrix
14	FALSE	Bacteria	Dependentiae	Babeliae	Babeliales	Vermiphilaceae	NA
14	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Intrasporangiaceae	Ornithinimicrobium
14	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Melissococcus

14	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylobacillus
14	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Nocardiaceae	Rhodococcus
14	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Anaerofustis
14	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Tannerellaceae	NA
13	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
13	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Sellimonas
13	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Crocinitomicaceae	NA
13	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Yersinia
13	FALSE	Bacteria	Actinobacteria	Actinobacteria	Streptosporangiales	Nocardiopsaceae	Nocardiopsis
13	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Hafnia-Obesumbacterium
13	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Glutamicibacter
13	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Parvimonas
13	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Xylophilus
13	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Pseudoclavibacter
13	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotellaceae_Ga6A1_group
13	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	Bradyrhizobium
13	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Leucobacter
13	FALSE	Bacteria	Actinobacteria	Actinobacteria	Pseudonocardiales	Pseudonocardiaceae	Pseudonocardia
13	FALSE	Bacteria	Actinobacteria	Actinobacteria	NA	NA	NA
13	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Paenarthrobacter
13	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	NA
13	FALSE	Bacteria	Actinobacteria	Actinobacteria	Frankiales	Geodermatophilaceae	Blastococcus
13	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Family_X	Thermicanus
13	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	Empedobacter
13	FALSE	Eukaryota	NA	NA	NA	NA	NA
13	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	NA
13	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	NA	NA	NA
13	FALSE	Bacteria	Entotheonellaeota	Entotheonellia	Entotheonellales	Entotheonellaceae	NA
13	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	NA
13	FALSE	Bacteria	Cyanobacteria	Melainabacteria	Obscuribacterales	NA	NA
12	FALSE	Bacteria	Firmicutes	Clostridia	NA	NA	NA
12	FALSE	Bacteria	Epsilonbacteraeota	Campylobacteria	Campylobacterales	Campylobacteraceae	Campylobacter
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Comamonas
12	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	Paenibacillus

12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Paracaedibacterales	Paracaedibacteraceae	Candidatus_Odyssella
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Lawsonella
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	Sphingobacteriaceae	NA
12	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	NA
12	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Catenisphaera
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Rhodocyclaceae	NA
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Methylophilaceae	Methylophilus
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Herminiimonas
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Phocea
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Family_XIII_UCG-001
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	Sphingobacteriaceae	Pedobacter
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Moryella
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Altererythrobacter
12	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Mitsuokella
12	FALSE	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Cetobacterium
12	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Bdellovibrionales	Bdellovibrionaceae	Bdellovibrio
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_FCS020_group
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium
12	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Bdellovibrionales	Bacteriovoracaceae	Peredibacter
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-008
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotellaceae_UCG-003
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinotignum
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Ochrobactrum
12	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Facklamia
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Salinisphaerales	Solimonadaceae	Nevskia
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Simplicispira
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptostreptococcus
12	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Enterorhabdus
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Rikenella
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	NA
12	FALSE	Bacteria	Tenericutes	Mollicutes	NA	NA	NA
12	FALSE	Bacteria	Lentisphaerae	Oligosphaeria	Oligosphaerales	Oligosphaeraceae	Z20
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Herbinix

12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	Elizabethkingia
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Helcococcus
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	NA
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Rhodocyclaceae	Methyloversatilis
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Heliobacteriaceae	Hydrogenispora
12	FALSE	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	NA	NA
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Verticia
12	FALSE	Bacteria	Spirochaetes	Brachyspirae	Brachyspirales	Brachyspiraceae	Brachyspira
12	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Anaerovibrio
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Bordetella
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Robinsoniella
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	NA
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Arcanobacterium
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Brucella
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Nesterenkonia
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	GCA-900066755
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	JTB23	NA	NA
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Anaerovorax
12	FALSE	Bacteria	Kiritimatiellaeota	Kiritimatiellae	WCHB1-41	NA	NA
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Dyadobacter
12	FALSE	Bacteria	Bacteroidetes	Rhodothermia	Rhodothermales	Rhodothermaceae	NA
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Amnibacterium
12	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Granulicatella
12	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	NA	NA
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Trueperella
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Promicromonosporaceae	Cellulosimicrobium
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Harryflintia
12	FALSE	Bacteria	Patescibacteria	Saccharimonadia	Saccharimonadales	Saccharimonadaceae	NA
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Neisseriaceae	Neisseria
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Epulopiscium
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	NA
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Micrococcus
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Clostridioides
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Cuneatibacter

12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	NA
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Acetanaerobacterium
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	NA
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Paenochrobactrum
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Shinella
12	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	NA
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Candidatus_Soleaferrea
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Saprospiraceae	NA
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Parapusillimonas
12	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Succiniclasticum
12	FALSE	Bacteria	Verrucomicrobia	Verrucomicrobiae	Opitutales	Puniceicoccaceae	NA
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Ralstonia
12	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	Ammoniphilus
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	NA	NA
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Thermomonas
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	Desulfotomaculum
12	FALSE	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	Methanobacterium
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Dermabacteraceae	Dermabacter
12	FALSE	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	NA
12	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Jeotgalicoccus
12	FALSE	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	Rhodopirellula
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Neorhizobium
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	NA
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Methylobacterium
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Paeniclostridium
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Tolumonas
12	FALSE	Bacteria	Chloroflexi	Dehalococcoidia	SAR202_clade	NA	NA
12	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Oceanobacillus
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Dysgonomonadaceae	Proteiniphilum
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Pseudoramibacter
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	NA
12	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Aerococcus
12	FALSE	Bacteria	Firmicutes	Bacilli	NA	NA	NA
12	FALSE	Bacteria	Bacteroidetes	NA	NA	NA	NA

12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Rhodanobacteraceae	Rhodanobacter
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	Moheibacter
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	NA	NA
12	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Selenomonas_4
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	NA	NA	NA
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Defluviitaleaceae	NA
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NA
12	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	mle1-27	NA
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Syntrophomonadaceae	NA
12	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinobaculum
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Rhabdobacter
12	FALSE	Bacteria	Actinobacteria	NA	NA	NA	NA
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Pseudoflavonifractor
12	FALSE	Bacteria	Patescibacteria	Saccharimonadia	Saccharimonadales	NA	NA
12	FALSE	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	NA
12	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Asteroleplasma
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Mesorhizobium
12	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Rhodocyclaceae	Dechloromonas
12	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Flavihumibacter
12	FALSE	Bacteria	Verrucomicrobia	Verrucomicrobiae	NA	NA	NA
11	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Providencia
11	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas
11	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Achromobacter
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	S5-A14a
11	FALSE	Bacteria	Epsilonbacteraeota	Campylobacteria	Campylobacterales	Arcobacteraceae	Arcobacter
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	UC5-1-2E3
11	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Pseudoglutamicibacter
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Shuttleworthia
11	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingopyxis
11	FALSE	Bacteria	Actinobacteria	Actinobacteria	Streptomycetales	Streptomycetaceae	Streptomyces
11	FALSE	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Synergistes
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerosporobacter
11	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Massilia

11	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Pseudochrobactrum
11	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	Sphingobacteriaceae	Nubsella
11	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Weissella
11	FALSE	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Gardnerella
11	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Caulobacter
11	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Solobacterium
11	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Senegalimassilia
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-011
11	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Cupriavidus
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Defluviitaleaceae	Defluviitaleaceae_UCG-011
11	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Rothia
11	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium
11	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Aquabacterium
11	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	NA
11	FALSE	Bacteria	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	NA	NA
10	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Succinivibrionaceae	Succinivibrio
10	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Succinivibrionaceae	NA
10	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	NA
10	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella_3
10	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	Lysinibacillus
10	FALSE	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Leptotrichiaceae	Sneathia
10	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium
10	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Caproiciproducens
10	FALSE	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Alloscardovia
10	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Adlercreutzia
10	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Millionella
10	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Sarcina
10	FALSE	Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	NA	NA
10	FALSE	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Pyramidobacter
10	FALSE	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Jonquetella
10	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Olsenella
10	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Family_XI	Gemella
10	FALSE	Bacteria	Elusimicrobia	Elusimicrobia	Elusimicrobiales	Elusimicrobiaceae	Elusimicrobium
10	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Paracoccus

10	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	NA
10	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	NA	NA
10	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Cosenzaea
10	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	NA	NA
10	FALSE	Bacteria	Proteobacteria	NA	NA	NA	NA
10	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA
10	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	NA
9	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
9	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Alloprevotella
9	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-003
9	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	Coprobacter
9	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Murdochiella
9	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Ottowia
9	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Merdibacter
9	FALSE	Bacteria	Firmicutes	Clostridia	DTU014	NA	NA
9	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerofilum
9	FALSE	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	Methanosphaera
9	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	NA
9	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium
8	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-005
8	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Variovorax
8	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Oligella
8	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	NA
8	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Muribaculaceae	CAG-873
8	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	NA
8	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	Brevibacillus
8	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Oxalobacter
8	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	NA
8	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	Rummeliibacillus
8	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Faecalicoccus
8	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-009
8	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_13
8	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Cellulosilyticum
8	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Mailhella

8	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Gordonibacter
8	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Oribacterium
8	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Nosocomiicoccus
8	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	NA	NA
8	FALSE	Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	Victivallaceae	NA
8	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Allobaculum
8	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas
8	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	NA
7	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Dorea
7	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotellaceae_NK3B31_group
7	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Desulfovibrio
7	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae_UCG-004
7	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Candidatus_Stoquefichus
7	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Holdemania
7	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Acidovorax
7	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	GCA-900066575
7	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	W5053
7	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Negativicoccus
7	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Anaeroglobus
7	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Sanguibacteroides
7	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	NA
6	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Intestinimonas
6	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	NA
6	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Pediococcus
6	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Fastidiosipila
6	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Gallicola
6	FALSE	Archaea	Euryarchaeota	Thermoplasmata	Methanomassiliicoccales	Methanomassiliicoccaceae	Methanomassiliicoccus
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoclostridium
5	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium
5	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	NA
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiales_vadinBB60_group	NA
5	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Megamonas
5	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelatoclostridium
5	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	NA

5	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Dysgonomonadaceae	Dysgonomonas
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Terrisporobacter
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-010
5	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	NA
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	GCA-900066225
5	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Pseudoxanthomonas
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Marvinbryantia
5	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus
5	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	NA
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Angelakisella
5	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Allisonella
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	Peptococcus
5	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	NA
5	FALSE	Archaea	Euryarchaeota	Thermoplasmata	Methanomassiliicoccales	Methanomethylophilaceae	NA
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Howardella
5	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Brevibacteriaceae	Brevibacterium
5	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Eubacterium
5	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Atopobium
5	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Slackia
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillibacter
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_1
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Finegoldia
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Family_XIII_AD3011_group
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	CAG-56
4	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Dielma
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_1
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	NA
4	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	Sphingobacteriaceae	Sphingobacterium
4	FALSE	Bacteria	Bacteroidetes	Bacteroidia	NA	NA	NA
4	FALSE	Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	Victivallaceae	Victivallis
4	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriales_Incertae_Sedis	NA
4	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	Cloacibacterium
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Hydrogenoanaerobacterium

4	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Eggerthella
4	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	NA
4	FALSE	Bacteria	Firmicutes	NA	NA	NA	NA
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Mogibacterium
4	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus
4	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	NA
3	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Proteus
3	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Acidaminococcus
3	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Christensenellaceae_R-7_group
3	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-003
3	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Morganella
3	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_NK4A214_group
3	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Odoribacter
3	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Intestinibacter
3	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	CAG-352
3	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_7
3	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Rikenellaceae_RC9_gut_group
3	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Catenibacterium
3	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_2
3	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Faecalitalea
3	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Butyricimonas
3	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	Chryseobacterium
3	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_2
3	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Veillonella
3	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	NA
3	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	NA	NA
3	FALSE	Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	vadinBE97	NA
3	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	NA
2	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Phascolarctobacterium
2	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Romboutsia
2	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_5
2	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Turicibacter
2	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Stenotrophomonas

2	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-001
2	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	DTU089
2	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	NA	NA	NA
1	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides
1	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Dialister
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-002
1	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Parasutterella
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Butyrivibrio
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_NK4A136_group
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Eisenbergiella
1	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Holdemanella
1	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella
1	FALSE	Bacteria	Tenericutes	Mollicutes	Izimaplasmatales	NA	NA
1	FALSE	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium
1	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Brevundimonas
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_1
1	FALSE	Bacteria	Cyanobacteria	Melainabacteria	Gastranaerophilales	NA	NA
1	FALSE	Bacteria	NA	NA	NA	NA	NA
0	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
0	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	NA
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus_2
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
0	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_9
0	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas
0	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella_4
0	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_6
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus_1
0	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae_UCG-003
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Flavonifractor
0	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella
0	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Delftia

0	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Sutterella
0	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	Barnesiella
0	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	NA	NA
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_9
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-014
0	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Paraprevotella
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Negativibacillus
0	FALSE	Bacteria	Tenericutes	Mollicutes	Mollicutes_RF39	NA	NA
0	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
0	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Muribaculaceae	NA
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-010

Supplementary Table 3. MWAS of dataset 2 conducted using ANCOM

Sample size for ANCOM included subset of samples that had complete data on all covariates tested: N= 306 cases and 177 controls in dataset 2.
W= ANCOM score indicating the number of times a genus achieved FDR<0.05 as compared to other genera (maximum W possible: 444 in dataset 1, 560 in dataset 2).
0.8= Threshold at which results were considered significant (TRUE).

W	0.8	Kingdom	Phylum	Class	Order	Family	Genus
553	TRUE	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
545	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Agathobacter
544	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-004
541	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
541	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-013
538	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_ND3007_group
536	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerostipes
535	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
533	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia
530	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Eubacterium
525	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillospira
524	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus_2
521	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Fusicatenibacter
521	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospira
521	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_6
521	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus_1
505	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Butyricicoccus
505	TRUE	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas
503	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-001
496	TRUE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Lawsonella
493	TRUE	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Desulfovibrio
493	TRUE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Turicibacter
491	TRUE	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	Methanobrevibacter

479	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	DTU089
477	TRUE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae_UCG-003
468	TRUE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas
465	TRUE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium_1
463	TRUE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella
459	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoclostridium
458	TRUE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
458	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-014
454	TRUE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Veillonella
452	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	NA
449	TRUE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Candidatus_Soleaferrea
440	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter
439	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	NA
438	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Intestinimonas
433	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_9
428	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Anaerococcus
425	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelatoclostridium
422	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Family_XIII_UCG-001
411	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Phocea
408	FALSE	Bacteria	Cyanobacteria	Oxyphotobacteria	Chloroplast	NA	NA
408	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	S5-A14a
407	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
405	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Holdemania
399	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Granulicatella
395	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Cuneatibacter
388	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-003
386	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	GCA-900066575
364	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Delftia

356	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Peptoniphilus
350	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-008
348	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Parvimonas
344	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Family_XI	Gemella
307	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Varibaculum
299	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Stenotrophomonas
268	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus
218	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Selenomonas_3
209	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Asteroleplasma
62	FALSE	Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	Victivallaceae	Victivallis
58	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Olsenella
54	FALSE	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Tessaracoccus
50	FALSE	Bacteria	Epsilonbacteraeota	Campylobacteria	Campylobacterales	Campylobacteraceae	Campylobacter
49	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Ezakiella
47	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Bosea
47	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Methylobacterium
46	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	NA
44	FALSE	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Cutibacterium
44	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	NA
43	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_13
41	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Pediococcus
40	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Fastidiosipila
40	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Mobiluncus
40	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Pseudoclavibacter
39	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Gallicola
39	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	NA
38	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	NA
38	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	Sphingobacteriaceae	Sphingobacterium

37	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Murdochiella
37	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-009
37	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-011
37	FALSE Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Scardovia
36	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingopyxis
35	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Achromobacter
35	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Clostridioides
35	FALSE Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Dermabacteraceae	Dermabacter
35	FALSE Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Jonquetella
34	FALSE Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	Chryseobacterium
34	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-008
33	FALSE Bacteria	Tenericutes	Mollicutes	Anaeroplasmatales	Anaeroplasmataceae	Anaeroplasma
32	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus
32	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Finegoldia
32	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Hafnia-Obesumbacterium
31	FALSE Bacteria	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Carnobacterium
31	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Howardella
31	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Hungatella
30	FALSE Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriales_Incertae_Sedis	Raoultibacter
30	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	UBA1819
29	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerosporobacter
29	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Citrobacter
29	FALSE Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Cryptobacterium
29	FALSE Bacteria	Actinobacteria	Acidimicrobiia	Microtrichales	NA	NA
29	FALSE Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Nocardiaceae	Rhodococcus
29	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Acetobacterales	Acetobacteraceae	Roseomonas
28	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	28-4
28	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Acetatifactor

28	FALSE Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Acidipropionibacterium
28	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Acidovorax
28	FALSE Bacteria	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	Ammoniphilus
28	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Brevundimonas
28	FALSE Bacteria	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	Cohnella
28	FALSE Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	Domibacillus
28	FALSE Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Facklamia
28	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Kerstersia
28	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-006
28	FALSE Bacteria	Actinobacteria	Actinobacteria	Micromonosporales	Micromonosporaceae	Micromonospora
28	FALSE Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Muribaculaceae	Muribaculum
28	FALSE Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	NA	NA
28	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Acetobacterales	Acetobacteraceae	NA
28	FALSE Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	NA
28	FALSE Bacteria	Firmicutes	Bacilli	NA	NA	NA
28	FALSE Bacteria	Verrucomicrobia	Verrucomicrobiae	Opitutales	NA	NA
28	FALSE Bacteria	Cyanobacteria	Oxyphotobacteria	Phormidesmiales	Nodosilineaceae	NA
28	FALSE Bacteria	Firmicutes	Negativicutes	Selenomonadales	NA	NA
28	FALSE Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	NA	NA
28	FALSE Bacteria	Acidobacteria	FFCH5909	NA	NA	NA
28	FALSE Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Nocardioides
28	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Paraclostridium
28	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Psychroglaciecola
28	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	SN8
28	FALSE Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Terribacillus
28	FALSE Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Trueperella
27	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas
27	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Alcanivoracaceae	Alcanivorax

27	FALSE Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Allobaculum
27	FALSE Bacteria	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Alloiococcus
27	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Aminobacter
27	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Succinivibrionaceae	Anaerobiospirillum
27	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerocolumna
27	FALSE Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Arcanobacterium
27	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	ATCC-39006
27	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Aureimonas
27	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Blastomonas
27	FALSE Bacteria	Spirochaetes	Brachyspirae	Brachyspirales	Brachyspiraceae	Brachyspira
27	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Caldicoprobacteraceae	Caldicoprobacter
27	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Caedibacterales	Caedibacteraceae	Candidatus_Nucleicultrix
27	FALSE Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Centipeda
27	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Christensenella
27	FALSE Bacteria	Cyanobacteria	Oxyphotobacteria	Nostocales	Chroococcidiopsaceae	Chroococcidiopsis_SAG_2023
27	FALSE Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	Cloacibacterium
27	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Cosenzaea
27	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Cupriavidus
27	FALSE Bacteria	Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	Deinococcus
27	FALSE Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Denitrobacterium
27	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	Desulfitibacter
27	FALSE Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobulbaceae	Desulfobulbus
27	FALSE Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Devosiaceae	Devosia
27	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Rhodanobacteraceae	Dokdonella
27	FALSE Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Dyadobacter
27	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter
27	FALSE Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Faecalicoccus
27	FALSE Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Fictibacillus

27	FALSE	Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	Flexilinea
27	FALSE	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Fretibacterium
27	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Globicatella
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Glutamicibacter
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Haematobacter
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Halomonas
27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Helcococcus
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Hydrogenophaga
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium
27	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Ignavigranum
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Intrasporangiaceae	Janibacter
27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Johnsonella
27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoanaerobaculum
27	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Lacibacter
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Leucobacter
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Massilia
27	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Merdibacter
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Micropruina
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Moraxella
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Mycobacteriaceae	Mycobacterium
27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	NA
27	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	NA
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	NA	NA
27	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	NA
27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Syntrophomonadaceae	NA
27	FALSE	Bacteria	Chloroflexi	Chloroflexia	Thermomicrobiales	JG30-KF-CM45	NA
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	NA
27	FALSE	Bacteria	Verrucomicrobia	Verrucomicrobiae	NA	NA	NA

27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micromonosporales	Micromonosporaceae	NA
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Streptomycetales	Streptomycetaceae	NA
27	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	NA
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	NA	NA	NA
27	FALSE	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	NA
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	NA	NA	NA
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Micavibrionales	NA	NA
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	NA	NA
27	FALSE	Bacteria	Chloroflexi	Chloroflexia	Kallotenuales	NA	NA
27	FALSE	Bacteria	Tenericutes	Mollicutes	NA	NA	NA
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Rhodocyclaceae	NA
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Neisseriaceae	NA
27	FALSE	Eukaryota	NA	NA	NA	NA	NA
27	FALSE	Bacteria	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	NA	NA
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	NA
27	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	NA
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Neisseriaceae	Neisseria
27	FALSE	Bacteria	Actinobacteria	Actinobacteria	Streptosporangiales	Nocardiopsaceae	Nocardiopsis
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium
27	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	Sphingobacteriaceae	Nubsella
27	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	Sphingobacteriaceae	Pedobacter
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Phenylobacterium
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Phyllobacterium
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Pigmentiphaga
27	FALSE	Bacteria	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	Pirellula
27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Pseudoramibacter
27	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	Psychrobacillus
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Pusillimonas

27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Pygmaiobacter
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Ralstonia
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Reyranellales	Reyranellaceae	Reyranella
27	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	Rummeliibacillus
27	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Selenomonas
27	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	Solibacillus
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas
27	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Spirosoma
27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Stomatobaculum
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Succinivibrionaceae	Succinivibrio
27	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Tetragenococcus
27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Tissierella
27	FALSE	Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	Treponema_2
27	FALSE	Bacteria	Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	Ureaplasma
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Cardiobacteriales	Wohlfahrtiimonadaceae	Wohlfahrtiimonas
27	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	Xanthobacter
27	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	XBB1006
27	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Rhodocyclaceae	Zoogloea
26	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Abiotrophia
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Gammaproteobacteria_Incertae_Sedis	Unknown_Family	Acidibacter
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Streptosporangiales	Thermomonosporaceae	Actinomadura
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Aeromicrobium
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Aggregatibacter
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Alcaligenes
26	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Allofustis
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Altererythrobacter
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Anaerobacillus

26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	Aneurinibacillus
26	FALSE	Bacteria	Epsilonbacteraeota	Campylobacteria	Campylobacterales	Arcobacteraceae	Arcobacter
26	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Atopobium
26	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Weeksellaceae	Bergeyella
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	Bradyrhizobium
26	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Amoebophilaceae	Candidatus_Amoebophilus
26	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Capnocytophaga
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Cardiobacteriales	Cardiobacteriaceae	Cardiobacterium
26	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Catenisphaera
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Cellulomonadaceae	Cellulomonas
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Cellulosilyticum
26	FALSE	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Cetobacterium
26	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	CHKCI002
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_11
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_7
26	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Coriobacteriaceae_UCG-003
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Dietziaceae	Dietzia
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Neisseriaceae	Eikenella
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Ensifer
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	F0332
26	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Spirosomaceae	Flectobacillus
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Geobacillus
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Gracilibacillus
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Hathewaya
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Jeotgalicoccus
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Thermoactinomycetaceae	Kroppenstedtia
26	FALSE	Bacteria	Verrucomicrobia	Verrucomicrobiae	Chthoniobacterales	Chthoniobacteraceae	LD29
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Leifsonia

26	FALSE	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Leptotrichiaceae	Leptotrichia
26	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Libanicoccus
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Macrococcus
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Mesorhizobium
26	FALSE	Bacteria	Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	Mycoplasma
26	FALSE	NA	NA	NA	NA	NA	NA
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	NA
26	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	NA
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Nocardiaceae	NA
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Mitochondria	NA
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	NA
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	NA
26	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Dysgonomonadaceae	NA
26	FALSE	Bacteria	Actinobacteria	NA	NA	NA	NA
26	FALSE	Bacteria	Verrucomicrobia	Verrucomicrobiae	Opitutales	Puniceicoccaceae	NA
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	NA
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	NA
26	FALSE	Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	NA
26	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	NA
26	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	Sphingobacteriaceae	NA
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	NA	NA	NA
26	FALSE	Bacteria	Kiritimatiellaeota	Kiritimatiellae	WCHB1-41	NA	NA
26	FALSE	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	NA
26	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Tannerellaceae	NA
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Neorhizobium
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Papillibacter
26	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Paraeggerthella
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Paucisalibacillus

26	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Pectinatus
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Pelomonas
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Plesiomonas
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Pluralibacter
26	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotellaceae_Ga6A1_group
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Propioniferax
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Pseudochrobactrum
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Pseudoflavonifractor
26	FALSE	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Pseudopropionibacterium
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Pseudorhodoferax
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Robinsoniella
26	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Selenomonas_4
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Shinella
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillales	Azospirillaceae	Skermanella
26	FALSE	Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	Sphaerochaeta
26	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Sporobacterium
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	Sporosarcina
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Succinivibrionaceae	Succinatimonas
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Syntrophomonadaceae	Syntrophomonas
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Tepidimonas
26	FALSE	Bacteria	Deinococcus-Thermus	Deinococci	Thermales	Thermaceae	Thermus
26	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Vagococcus
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Variovorax
26	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Virgibacillus
26	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	W5053
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Xylophilus
26	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Yersinia

25	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinotignum
25	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium
25	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Anaerovorax
25	FALSE	Archaea	Euryarchaeota	Thermoplasmata	Methanomassiliicoccales	Methanomethylophilaceae	Candidatus_ Methanomethylophilus
25	FALSE	Bacteria	Patescibacteria	Saccharimonadia	Saccharimonadales	Saccharimonadaceae	Candidatus_Saccharimonas
25	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_2
25	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Eremococcus
25	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae_UCG-004
25	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae_UCG-006
25	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Flavobacterium
25	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Nocardiaceae	Gordonia
25	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Herbaspirillum
25	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoclostridium_10
25	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Lautropia
25	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Mailhella
25	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Mitsuokella
25	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	NA
25	FALSE	Bacteria	Proteobacteria	NA	NA	NA	NA
25	FALSE	Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	NA	NA
25	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	NA	NA
25	FALSE	Bacteria	Bacteroidetes	NA	NA	NA	NA
25	FALSE	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	NA
25	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	NA
25	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Nosocomiicoccus
25	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Ochrobactrum
25	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Oribacterium
25	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Ornithinibacillus
25	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Paracoccus

25	FALSE Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Propionimicrobium
25	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_V9D2013 _group
25	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Sarcina
25	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Shimwellia
25	FALSE Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Leptotrichiaceae	Sneathia
25	FALSE Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Synergistes
24	FALSE Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinobaculum
24	FALSE Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Aerococcus
24	FALSE Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Alloprevotella
24	FALSE Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Alloscardovia
24	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Anaerofustis
24	FALSE Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Anaeroglobus
24	FALSE Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Muribaculaceae	CAG-873
24	FALSE Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Promicromonosporaceae	Cellulosimicrobium
24	FALSE Bacteria	Verrucomicrobia	Verrucomicrobiae	Opitutales	Puniceicoccaceae	Cerasicoccus
24	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_3
24	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Comamonas
24	FALSE Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	DNF00809
24	FALSE Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Enorma
24	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Herbinix
24	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Cardiobacteriales	Wohlfahrtiimonadaceae	Ignatzschineria
24	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Kosakonia
24	FALSE Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Microbacterium
24	FALSE Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Mogibacterium
24	FALSE Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	NA
24	FALSE Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	Victivallaceae	NA
24	FALSE Bacteria	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	NA
24	FALSE Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Succinivibrionaceae	NA

24	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA
24	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Paeniclostridium
24	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotellaceae_UCG-001
24	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Pseudoglutamicibacter
24	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Rikenella
24	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-007
24	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Sedimentibacter
23	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Allisonella
23	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Dysgonomonadaceae	Dysgonomonas
23	FALSE	Archaea	Euryarchaeota	Thermoplasmata	Methanomassiliicoccales	Methanomassiliicoccaceae	Methanomassiliicoccus
23	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	NA
23	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptostreptococcus
23	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Rikenellaceae_RC9_gut_group
23	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
22	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Dielma
22	FALSE	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium
22	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Holdemanella
22	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	NA
22	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	NA
22	FALSE	Bacteria	NA	NA	NA	NA	NA
22	FALSE	Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	vadinBE97	NA
22	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Sanguibacteroides
22	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella_4
21	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Defluviitaleaceae	Defluviitaleaceae_UCG-011
21	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Enterorhabdus
21	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
21	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	NA
21	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Negativicoccus

21	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	Peptococcus
21	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Pseudoxanthomonas
21	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Salmonella
21	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Senegalimassilia
21	FALSE	Bacteria	Actinobacteria	Actinobacteria	Streptomycetales	Streptomycetaceae	Streptomyces
20	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Dermabacteraceae	Brachybacterium
20	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	Brevibacillus
20	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Eisenbergiella
20	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Moryella
20	FALSE	Archaea	Euryarchaeota	Thermoplasmata	Methanomassiliicoccales	Methanomethylophilaceae	NA
20	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	NA
20	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Oligella
20	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Providencia
20	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_1
20	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Solobacterium
19	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Brevibacteriaceae	Brevibacterium
19	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	CAG-352
19	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Epulopiscium
19	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Hydrogenoanaerobacterium
19	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_6
19	FALSE	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Pyramidobacter
18	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Catenibacterium
18	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Edwardsiella
18	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	GCA-900066225
18	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	Lysinibacillus
18	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Micrococcus
18	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	NA
18	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_2

18	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Sellimonas
17	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	Coprobacter
17	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Megasphaera
17	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	NA
17	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Muribaculaceae	NA
17	FALSE	Bacteria	Firmicutes	Clostridia	DTU014	NA	NA
17	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Pseudogracilibacillus
17	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Terrisporobacter
16	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Acidaminococcus
16	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Butyricimonas
16	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Butyrivibrio
16	FALSE	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Gardnerella
16	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	NA
15	FALSE	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium
15	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Faecalitalea
15	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Kocuria
15	FALSE	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	Methanosphaera
15	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	NA	NA
15	FALSE	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Propionibacterium
15	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-005
14	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Acetanaerobacterium
14	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerofilum
14	FALSE	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Cloacibacillus
14	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Leuconostoc
14	FALSE	Bacteria	Firmicutes	Clostridia	NA	NA	NA
14	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptoclostridium
14	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_5
14	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Succiniclasticum

13	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Brucella
13	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lactonifactor
13	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	NA
13	FALSE	Bacteria	Patescibacteria	Saccharimonadia	Saccharimonadales	NA	NA
13	FALSE	Bacteria	Tenericutes	Mollicutes	Mollicutes_RF39	NA	NA
13	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotellaceae_NK3B31_group
13	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Angelakisella
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Caproiciproducens
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Catabacter
12	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Coprobacillus
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	GCA-900066755
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_NK4B4_group
12	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	NA
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	NA
12	FALSE	Bacteria	Tenericutes	Mollicutes	Izimaplasmatales	NA	NA
12	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Oxalobacter
12	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-004
12	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Weissella
11	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Adlercreutzia
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Christensenellaceae_R-7_group
11	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Gordonibacter
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Harryflintia
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_FCS020_group
11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-010
11	FALSE	Bacteria	Firmicutes	NA	NA	NA	NA
11	FALSE	Bacteria	Bacteroidetes	Bacteroidia	NA	NA	NA
11	FALSE	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Rothia

11	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Shuttleworthia
10	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus
10	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	NA
10	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	Paenibacillus
10	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Slackia
9	FALSE	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Candidatus_Stoquefichus
9	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Family_XIII_AD3011_group
9	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Megamonas
9	FALSE	Bacteria	Patescibacteria	Saccharimonadia	Saccharimonadales	Saccharimonadaceae	NA
9	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriales_Incertae_Sedis	NA
9	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	NA
9	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Oceanobacillus
9	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillibacter
9	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Romboutsia
9	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Sutterella
8	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Eggerthella
8	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Flavonifractor
8	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Morganella
8	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	NA
8	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	NA	NA
8	FALSE	Bacteria	Cyanobacteria	Melainabacteria	Gastranaerophilales	NA	NA
8	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella_3
7	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_1
7	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Marvinbryantia
7	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_7
7	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-010
7	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus
7	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	UC5-1-2E3

6	FALSE	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Bilophila
6	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	CAG-56
6	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_2
6	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiales_vadinBB60_group	NA
6	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides
6	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_9
5	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
5	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Dialister
5	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Dorea
4	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_NK4A136_group
4	FALSE	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Phascolarctobacterium
3	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
3	FALSE	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	NA	NA
3	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Parasutterella
2	FALSE	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella
2	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Fournierella
2	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella
2	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Odoribacter
2	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Paraprevotella
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_1
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_3
1	FALSE	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Intestinibacter
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Negativibacillus
1	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Proteus
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_NK4A214_group
1	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-003
0	FALSE	Bacteria	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Akkermansiaceae	Akkermansia

0	FALSE	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	Barnesiella
0	FALSE	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	NA
0	FALSE	Bacteria	Firmicutes	Bacilli	Bacillales	NA	NA
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-002
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
0	FALSE	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella

Supplementary Table 4. MWAS of dataset 1 conducted using Kruskal-Wallis

Sample size for KW included all samples: N= 201 cases and 132 controls in dataset 1. MRA= mean relative abundance, FC=fold change in patients (PD MRA/control MRA), P= unadjusted significance, FDR (BH)= false discovery rate, adjusted significance. Unclassified genera and genera present in <10% of subjects were excluded from this analysis.

PD MRA	Control MRA	FC	P	FDR (BH)	Kingdom	Phylum	Class	Order	Family	Genus
0.0005	0.0012	0.37	4E-06	2E-04	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_ND3007_group
0.0027	0.0004	6.61	2E-06	2E-04	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
0.0191	0.0362	0.53	7E-06	2E-04	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Agathobacter
0.0153	0.0084	1.83	5E-05	1E-03	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
0.0008	0.0001	13.03	8E-05	1E-03	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Cloacibacillus
0.0353	0.0564	0.63	9E-05	1E-03	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
0.0036	0.0004	8.14	9E-05	1E-03	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Hungatella
0.0029	0.0037	0.80	1E-04	1E-03	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospira
0.0047	0.0012	3.77	1E-04	1E-03	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Megasphaera
0.0034	0.0008	4.20	1E-04	1E-03	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas
0.0140	0.0205	0.68	2E-04	2E-03	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia
0.0017	0.0001	12.72	4E-04	4E-03	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Coprobacillus
0.0077	0.0160	0.48	4E-04	4E-03	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
0.0038	0.0015	2.56	7E-04	6E-03	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella
0.0541	0.0218	2.48	1E-03	7E-03	Bacteria	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Akkermansiaceae	Akkermansia
0.0012	0.0019	0.66	1E-03	7E-03	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Butyricicoccus
0.0045	0.0023	1.95	1E-03	8E-03	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	UBA1819
0.0005	0.0001	4.95	2E-03	0.01	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Varibaculum
0.0020	0.0010	1.96	2E-03	0.01	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium_1
0.0006	0.0003	1.76	3E-03	0.01	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-004
0.0021	0.0038	0.56	3E-03	0.02	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Fusicatenibacter

0.0003	0.0007	0.48	4E-03	0.02	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-004
0.0004	0.0006	0.65	4E-03	0.02	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillospira
0.0027	0.0040	0.69	5E-03	0.02	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerostipes
0.0015	0.0013	1.16	5E-03	0.02	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Desulfovibrio
0.0006	0.0002	2.86	6E-03	0.03	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus
0.0006	0.0004	1.37	7E-03	0.03	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	Methanobrevibacter
0.0071	0.0021	3.32	8E-03	0.03	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Ezakiella
0.0003	0.0017	0.17	9E-03	0.03	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
0.2148	0.2479	0.87	0.01	0.04	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
0.0043	0.0029	1.47	0.01	0.05	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Bilophila
0.0014	0.0018	0.77	0.02	0.07	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_1
0.0008	0.0013	0.64	0.02	0.07	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_3
0.0006	0.0005	1.13	0.03	0.09	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Family_XIII_AD3011_group
0.0012	0.0009	1.32	0.03	0.10	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-010
0.0013	0.0014	0.95	0.03	0.10	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-013
0.0021	0.0014	1.42	0.04	0.10	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Eisenbergiella
0.0023	0.0009	2.58	0.04	0.10	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Peptoniphilus
0.0008	0.0002	3.81	0.04	0.10	Bacteria	Epsilonbacteraeota	Campylobacteria	Campylobacterales	Campylobacteraceae	Campylobacter
0.0003	0.0001	2.12	0.04	0.11	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Murdochiella
0.0052	0.0079	0.66	0.04	0.12	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Phascolarctobacterium
0.0060	0.0092	0.65	0.04	0.12	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoclostridium
0.0020	0.0005	4.09	0.05	0.13	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_6
0.0009	0.0011	0.86	0.06	0.15	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Intestinimonas
0.0302	0.0233	1.29	0.07	0.15	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides
0.0070	0.0067	1.05	0.06	0.15	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-005
0.0006	0.0001	5.47	0.07	0.16	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Sellimonas

0.0015	0.0004	3.31	0.08	0.17	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Anaerococcus
0.0020	0.0031	0.67	0.08	0.17	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae_UCG-003
0.0036	0.0032	1.14	0.08	0.17	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillibacter
0.0015	0.0022	0.69	0.09	0.20	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-003
0.0002	0.0002	0.80	0.11	0.23	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Eggerthella
0.0002	0.0001	3.08	0.11	0.23	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Hydrogenoanaerobacterium
0.0020	0.0016	1.28	0.12	0.24	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Butyricimonas
0.0032	0.0014	2.30	0.13	0.25	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium
0.0027	0.0014	1.96	0.13	0.25	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Holdemanella
0.0038	0.0024	1.60	0.14	0.27	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_5
0.0039	0.0022	1.74	0.15	0.27	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Acidaminococcus
0.0058	0.0034	1.69	0.15	0.27	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella_4
0.0386	0.0337	1.15	0.16	0.28	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
0.0015	0.0004	3.79	0.16	0.28	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Faecalitalea
0.0003	0.0001	4.63	0.16	0.28	Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	Victivallaceae	Victivallis
0.0003	0.0005	0.57	0.17	0.29	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	CAG-56
0.0006	0.0009	0.64	0.18	0.30	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-010
0.0161	0.0109	1.48	0.19	0.32	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-002
0.0010	0.0007	1.45	0.24	0.39	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Finegoldia
0.0002	0.0003	0.65	0.25	0.40	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	GCA-900066225
0.0065	0.0046	1.41	0.26	0.42	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Christensenellaceae_R-7_group
0.0004	0.0001	3.97	0.27	0.42	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Dielma
0.0072	0.0071	1.02	0.28	0.44	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus_1
0.0024	0.0029	0.83	0.29	0.45	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Dorea
0.0002	0.0003	0.75	0.31	0.48	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	DTU089
0.0015	0.0013	1.15	0.32	0.48	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Intestinibacter

0.0038	0.0032	1.19	0.32	0.48	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_NK4A214_group
0.0033	0.0040	0.82	0.34	0.50	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Sutterella
0.0077	0.0082	0.94	0.36	0.52	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-014
0.0023	0.0023	0.99	0.39	0.55	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Odoribacter
0.0004	0.0004	0.94	0.43	0.59	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Holdemania
0.0015	0.0012	1.26	0.43	0.60	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Romboutsia
0.0006	0.0005	1.19	0.44	0.60	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_1
0.0087	0.0126	0.69	0.45	0.60	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter
0.0005	0.0008	0.66	0.45	0.60	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelatoclostridium
0.0003	0.0005	0.51	0.47	0.61	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	Coprobacter
0.0020	0.0018	1.10	0.47	0.61	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Negativibacillus
0.0010	0.0029	0.35	0.46	0.61	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_7
0.0042	0.0054	0.79	0.49	0.62	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_NK4A136_group
0.0040	0.0023	1.72	0.50	0.63	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
0.0064	0.0058	1.12	0.52	0.64	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	Barnesiella
0.0121	0.0133	0.90	0.52	0.64	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
0.0007	0.0005	1.44	0.56	0.68	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Turicibacter
0.0024	0.0026	0.90	0.58	0.69	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Paraprevotella
0.0010	0.0009	1.10	0.59	0.70	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella
0.0026	0.0024	1.08	0.60	0.70	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_9
0.0012	0.0006	1.95	0.61	0.70	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Veillonella
0.0008	0.0005	1.45	0.62	0.71	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
0.0001	0.0002	0.82	0.63	0.72	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Angelakisella
0.1131	0.1351	0.84	0.64	0.72	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
0.0021	0.0014	1.47	0.64	0.72	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella

0.0017	0.0010	1.74	0.74	0.81	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	CAG-352
0.0211	0.0155	1.37	0.78	0.85	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas
0.0079	0.0090	0.89	0.81	0.86	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Dialister
0.0024	0.0026	0.92	0.81	0.86	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Flavonifractor
0.0039	0.0047	0.83	0.82	0.86	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Parasutterella
0.0115	0.0080	1.45	0.81	0.86	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus_2
0.0040	0.0052	0.76	0.85	0.88	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_6
0.0006	0.0007	0.81	0.88	0.90	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-001
0.0106	0.0155	0.68	0.88	0.90	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_9
0.0002	0.0002	0.94	0.91	0.92	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	GCA-900066575
0.0020	0.0015	1.36	1.00	1.00	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_2

Supplementary Table 5. MWAS of dataset 2 conducted using Kruskal-Wallis

Sample size for KW included all samples: N= 323 cases and 184 controls in dataset 2. MRA= mean relative abundance, FC=fold change in patients (PD MRA/control MRA), P= unadjusted significance, FDR (BH)= false discovery rate, adjusted significance. Unclassified genera and genera present in <10% of subjects were excluded from this analysis. MRA values of 0.0000 correspond to MRAs that were <0.0001.

PD MRA	Control MRA	FC	P	FDR (BH)	Kingdom	Phylum	Class	Order	Family	Genus
0.0239	0.0088	2.72	4E-09	6E-07	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
0.0004	0.0011	0.38	2E-07	1E-05	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-004
0.0097	0.0172	0.56	1E-06	6E-05	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Agathobacter
0.0047	0.0078	0.60	7E-06	3E-04	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
0.0002	0.0001	2.19	8E-06	3E-04	Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Eubacterium
0.0007	0.0011	0.59	2E-05	6E-04	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_ND3007_group
0.0013	0.0018	0.73	3E-05	7E-04	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-013
0.0039	0.0050	0.78	7E-05	1E-03	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Anaerostipes
0.0003	0.0001	4.43	2E-04	3E-03	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Lawsonella
0.0276	0.0416	0.66	2E-04	3E-03	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
0.0006	0.0003	2.11	4E-04	0.01	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Turicibacter
0.0208	0.0258	0.81	6E-04	0.01	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas
0.0036	0.0026	1.38	7E-04	0.01	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	UBA1819
0.0039	0.0015	2.53	7E-04	0.01	Bacteria	Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium_1
0.0001	0.0002	0.58	8E-04	0.01	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae_UCG-003
0.0017	0.0007	2.56	1E-03	0.01	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Anaerococcus
0.0003	0.0007	0.39	1E-03	0.01	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-001
0.0015	0.0008	1.81	1E-03	0.01	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Desulfovibrio
0.0056	0.0036	1.57	1E-03	0.01	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
0.0036	0.0053	0.68	2E-03	0.01	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospira
0.0003	0.0005	0.64	2E-03	0.01	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillospira
0.0005	0.0003	1.79	2E-03	0.01	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Varibaculum
0.0027	0.0012	2.14	2E-03	0.01	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Peptoniphilus
0.0015	0.0008	1.77	2E-03	0.02	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Hungatella
0.0045	0.0028	1.61	3E-03	0.02	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	Methanobrevibacter
0.0022	0.0000	220.2	3E-03	0.02	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Delftia
0.0085	0.0086	0.99	3E-03	0.02	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus

0.0026	0.0009	2.94	3E-03	0.02	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas
0.0025	0.0006	4.39	3E-03	0.02	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella
0.0031	0.0046	0.69	0.01	0.03	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Fusicatenibacter
0.0008	0.0005	1.69	0.01	0.03	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus
0.0079	0.0101	0.78	0.01	0.03	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus_2
0.0001	0.0000	34.45	0.01	0.03	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Parvimonas
0.0003	0.0002	1.72	0.01	0.04	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Mobiluncus
0.0005	0.0003	1.45	0.01	0.04	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces
0.0012	0.0005	2.60	0.01	0.04	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Finegoldia
0.0003	0.0001	2.69	0.01	0.04	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	S5-A14a
0.0187	0.0237	0.79	0.01	0.04	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia
0.0005	0.0002	2.25	0.01	0.04	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Murdochiella
0.0048	0.0015	3.18	0.01	0.04	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Ezakiella
0.0004	0.0005	0.78	0.01	0.04	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	DTU089
0.0000	0.0000	2.06	0.01	0.05	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Cutibacterium
0.0013	0.0019	0.68	0.02	0.06	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Butyricicoccus
0.0000	0.0000	2.91	0.02	0.07	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Scardovia
0.0034	0.0053	0.64	0.02	0.07	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_6
0.0034	0.0044	0.79	0.02	0.07	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus_1
0.0000	0.0000	0.72	0.02	0.07	Bacteria	Firmicutes	Bacilli	Bacillales	Family_XI	Gemella
0.0004	0.0001	2.58	0.03	0.09	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Fastidiosipila
0.0038	0.0063	0.61	0.03	0.09	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-014
0.0037	0.0018	2.09	0.03	0.09	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Eisenbergiella
0.0008	0.0007	1.23	0.03	0.09	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-004
0.0007	0.0002	4.09	0.03	0.09	Bacteria	Epsilonbacteraeota	Campylobacteria	Campylobacterales	Campylobacteraceae	Campylobacter
0.0027	0.0045	0.61	0.03	0.09	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Veillonella
0.0002	0.0000	11.43	0.03	0.10	Bacteria	Lentisphaerae	Lentisphaeria	Victivallales	Victivallaceae	Victivallis
0.0000	0.0001	0.59	0.04	0.12	Bacteria	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Granulicatella
0.0000	0.0000	0.95	0.04	0.13	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Cuneatibacter
0.0001	0.0001	1.45	0.04	0.13	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-009
0.0000	0.0000	1.58	0.05	0.13	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-011
0.0001	0.0001	0.86	0.05	0.14	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Candidatus_Soleaferrea
0.0001	0.0002	0.73	0.05	0.14	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
0.0000	0.0000	0.51	0.05	0.14	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-008

0.0001	0.0004	0.40	0.06	0.14	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella_3
0.0004	0.0003	1.40	0.06	0.15	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	GCA-900066225
0.0001	0.0002	0.58	0.06	0.16	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Brevibacteriaceae	Brevibacterium
0.0002	0.0001	1.42	0.06	0.16	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Howardella
0.0023	0.0017	1.37	0.07	0.16	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Butyricimonas
0.0002	0.0001	2.00	0.07	0.17	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Mogibacterium
0.0000	0.0000	1.46	0.07	0.17	Bacteria	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Anaerofustis
0.0001	0.0001	0.75	0.07	0.17	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Family_XIII_UCG-001
0.0096	0.0047	2.06	0.07	0.17	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Acidaminococcus
0.0003	0.0002	1.33	0.07	0.17	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Anaeroglobus
0.0001	0.0001	0.82	0.08	0.17	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Phocea
0.0001	0.0002	0.76	0.08	0.18	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	Peptococcus
0.0373	0.0460	0.81	0.08	0.18	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Phascolarctobacterium
0.0000	0.0000	0.72	0.09	0.19	Bacteria	Firmicutes	Clostridia	Clostridiales	Defluviitaleaceae	Defluviitaleaceae_UCG-011
0.0007	0.0004	1.68	0.10	0.20	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Sellimonas
0.0041	0.0032	1.31	0.10	0.21	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-005
0.0000	0.0000	0.65	0.10	0.21	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-007
0.0001	0.0001	0.79	0.10	0.21	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelatoclostridium
0.0036	0.0008	4.50	0.10	0.21	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella_4
0.0055	0.0040	1.36	0.10	0.21	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella
0.1610	0.1887	0.85	0.12	0.23	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
0.0000	0.0000	0.98	0.12	0.23	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Holdemania
0.0028	0.0028	0.99	0.14	0.26	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Dorea
0.0003	0.0003	0.90	0.15	0.28	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	CAG-56
0.0004	0.0009	0.40	0.17	0.32	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_2
0.0051	0.0040	1.26	0.17	0.33	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_5
0.0002	0.0003	0.86	0.18	0.33	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	GCA-900066575
0.0032	0.0034	0.93	0.19	0.35	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_NK4A136_group
0.0001	0.0001	1.63	0.20	0.37	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Faecalitalea
0.0032	0.0033	0.99	0.23	0.41	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Christensenellaceae_R-7_group
0.0199	0.0149	1.34	0.24	0.42	Bacteria	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Akkermansiaceae	Akkermansia
0.0018	0.0080	0.22	0.24	0.43	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella
0.0031	0.0032	0.98	0.26	0.46	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_9
0.0001	0.0001	1.04	0.27	0.47	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Shuttleworthia

0.0001	0.0000	3.41	0.28	0.48	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Atopobium
0.0104	0.0114	0.91	0.30	0.50	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoclostridium
0.0058	0.0051	1.14	0.32	0.53	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillibacter
0.0001	0.0001	1.54	0.32	0.53	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus
0.1960	0.2043	0.96	0.33	0.53	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
0.0006	0.0006	0.93	0.34	0.54	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Eggerthella
0.0044	0.0033	1.35	0.34	0.55	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Parasutterella
0.0020	0.0021	0.93	0.35	0.55	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-003
0.0000	0.0001	0.37	0.35	0.55	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium
0.0008	0.0007	1.16	0.35	0.55	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_UCG-010
0.0001	0.0001	1.53	0.36	0.55	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Negativicoccus
0.0042	0.0015	2.80	0.36	0.55	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Cloacibacillus
0.0018	0.0017	1.08	0.37	0.55	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Bilophila
0.0026	0.0027	0.98	0.37	0.55	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_NK4A214_group
0.0001	0.0001	0.67	0.37	0.55	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_FCS020_group
0.0053	0.0037	1.45	0.38	0.56	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Dialister
0.0006	0.0003	1.97	0.40	0.58	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium
0.0002	0.0001	2.54	0.41	0.59	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Terrisporobacter
0.0015	0.0009	1.67	0.41	0.59	Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Sutterella
0.0222	0.0199	1.11	0.41	0.59	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides
0.0020	0.0024	0.83	0.42	0.59	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Intestinimonas
0.0006	0.0002	2.78	0.43	0.60	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_6
0.0018	0.0020	0.92	0.43	0.60	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Negativibacillus
0.0086	0.0035	2.48	0.43	0.60	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Megasphaera
0.0001	0.0001	0.41	0.45	0.61	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae_NK4B4_group
0.0000	0.0001	0.76	0.45	0.61	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Coprobacillus
0.0138	0.0112	1.24	0.46	0.62	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-002
0.0076	0.0131	0.58	0.49	0.64	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Proteus
0.0027	0.0020	1.39	0.49	0.64	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Romboutsia
0.0001	0.0000	1.32	0.49	0.64	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerofilum
0.0008	0.0004	2.10	0.50	0.64	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	Coprobacter
0.0006	0.0005	1.29	0.52	0.67	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus
0.0000	0.0000	1.08	0.52	0.67	Bacteria	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Catabacter
0.0000	0.0000	1.19	0.53	0.67	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	GCA-900066755

0.0004	0.0003	1.56	0.54	0.68	Bacteria	Synergistetes	Synergistia	Synergistales	Synergistaceae	Pyramidobacter
0.0001	0.0000	2.06	0.55	0.68	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminiclostridium_1
0.0001	0.0000	1.58	0.55	0.68	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Candidatus_Stoquefichus
0.0004	0.0003	1.20	0.57	0.69	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Angelakisella
0.0007	0.0006	1.26	0.57	0.70	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XIII	Family_XIII_AD3011_group
0.0009	0.0010	0.89	0.58	0.70	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_3
0.0032	0.0033	0.97	0.59	0.71	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Flavonifractor
0.0000	0.0000	1.12	0.62	0.73	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Acetanaerobacterium
0.0002	0.0001	1.31	0.62	0.73	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Slackia
0.0001	0.0001	0.63	0.62	0.73	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Rothia
0.0000	0.0000	3.20	0.63	0.73	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Dielma
0.0041	0.0055	0.75	0.64	0.74	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_9
0.0005	0.0002	2.14	0.65	0.75	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Marvinbryantia
0.0004	0.0005	0.92	0.66	0.75	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Fournierella
0.0004	0.0003	1.30	0.67	0.75	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus_1
0.0007	0.0007	0.93	0.68	0.77	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Intestinibacter
0.0001	0.0001	2.34	0.70	0.78	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Hydrogenoanaerobacterium
0.0052	0.0044	1.18	0.72	0.80	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Barnesiellaceae	Barnesiella
0.0000	0.0000	1.48	0.73	0.80	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lactonifactor
0.0001	0.0001	1.16	0.74	0.80	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Gordonibacter
0.0016	0.0017	0.98	0.74	0.80	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
0.0003	0.0002	1.25	0.75	0.80	Bacteria	Actinobacteria	Coriobacteriia	Coriobacteriales	Eggerthellaceae	Adlercreutzia
0.0002	0.0002	0.79	0.75	0.80	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	UC5-1-2E3
0.0170	0.0159	1.07	0.79	0.84	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
0.0028	0.0015	1.82	0.80	0.84	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae_1	Clostridium_sensu_stricto_1
0.0015	0.0011	1.33	0.88	0.92	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_7
0.0202	0.0206	0.98	0.88	0.92	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
0.0006	0.0005	1.05	0.91	0.95	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcaceae_UCG-010
0.0000	0.0000	1.34	0.95	0.98	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Caproiciproducens
0.0002	0.0002	1.32	0.95	0.98	Bacteria	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Weissella
0.0014	0.0009	1.51	0.96	0.98	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Tyzzerella
0.0000	0.0000	1.23	0.98	0.99	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Harryflintia
0.0053	0.0049	1.09	0.98	0.99	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Marinifilaceae	Odoribacter
0.0012	0.0015	0.80	0.99	0.99	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Paraprevotella

Supplementary Table 6. PubMed search results for *Porphyromonas, Prevotella*, or *Corynebacterium_1*

Species that comprised each genus (and accounted for at least 80% of the ASVs in the genus) were identified based on 100% sequence identity using DADA2-SILVA reference database or 100% or >99% identity and high statistical confidence using NCBI 16S rRNA database. Then each species was searched in PubMed using "genus species" as search term. Search filters: Humans, English, Title/abstract. The citations were tabulated for articles that addressed function, characteristics or relevance to human health; method papers were omitted. All infections are in human samples, except *C. lactis* (newly discovered) was found in an abscess in a companion dog.

Search term	PubMed Return	Subject matter
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/10342655	septic arthritis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/10482033	surgical or catheter related infection, pilonidal cyst
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/11749760	endocarditis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/12235925	blood cultures
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/12439810	mastitis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/12565065	infective endocarditis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/15315020	opportunistic infections
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/15786829	Peritonitis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/17284316	endocarditis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/18174873	infections in pediatric oncology
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/18809563	endocarditis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/19153032	response to antibiotic tigecycline
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/19876565	predominant species in infections in cancer patients
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/20624090	resistance to antibiotic macrolide
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/22361761	clinical diphtheroid samples
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/23806703	surgical site infection
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/26324578	vaginosis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/28011352	blood stream infection
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/28264610	breast abscess
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/28700261	infection of orbital implant
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/29793964	respiratory infection after lung transplant
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/30102894	Bloodstream and venous catheter-related infections
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/30248572	cystic neutrophilic granulomatous mastitis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/30803027	bacteremia
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/8727888	clinical isolates, multiple sources

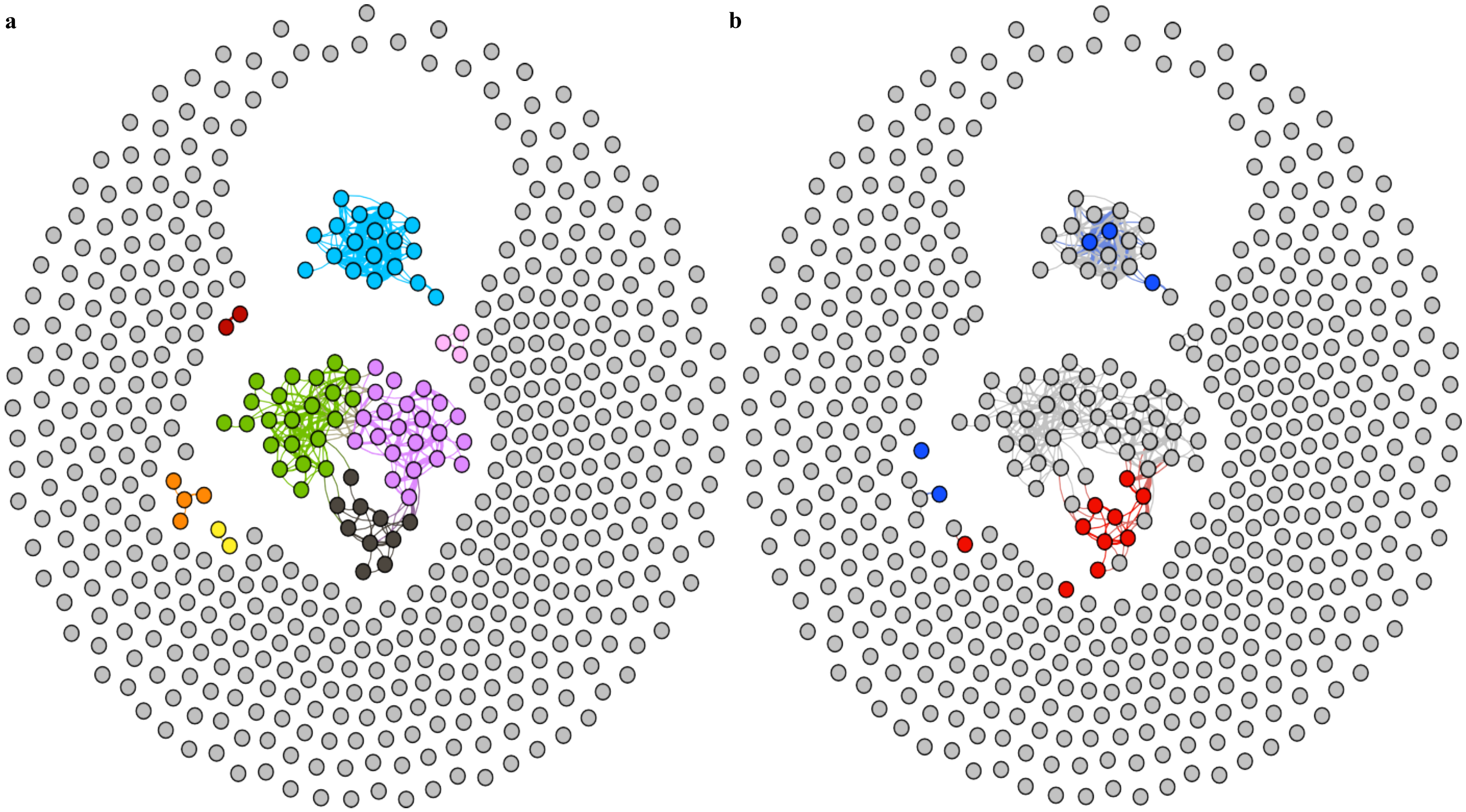
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/8874085	sepsis
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/9157120	neonatal sepsis fatal in premature infant
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/9488824	wound, bloodstream, and urinary tract infections
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/9505178	infection after orthopedic surgery
Corynebacterium amycolatum	https://www.ncbi.nlm.nih.gov/pubmed/9868692	Cardioverter-Lead Electrode Infection
Corynebacterium lactis	https://www.ncbi.nlm.nih.gov/pubmed/25937144	Infection in companion dog
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5896039/	colorectal cancer
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6247719/	causes Lemierre's syndrome
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/15528728	clinical isolates
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/15722627	Clinical isolates, multiple sources
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/15888469	predominant in polymicrobial flora in 48 inflamed sinuses
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/15897651	Lemierre's syndrome
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/16887693	liver abscess
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/19390440	causes Lemierre's syndrome
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/21407153	tubo-ovarian abscess
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/23435719	causes Lemierre's syndrome (acute otopjaryngeal infection)
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/23474186	pleural empyema in immunocompetent diabetic patient
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/24679105	polymicrobial foot infection
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/7548548	extraoral infections
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/7752213	418 children with infection, found in infections across body sites
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/7857230	cause chest wall abscess in one woman
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/8126176	bacterial vaginosis
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/8518760	male and female genital ulcers
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/8907604	female genital tract infection
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/9200028	intravenous catheter related bacteremia in child with cancer
Porphyromonas asaccharolytica	https://www.ncbi.nlm.nih.gov/pubmed/9772922	infected cardiac myxoma
Porphyromonas bennonis	https://www.ncbi.nlm.nih.gov/pubmed/19542133	identification and characterization in clinical specimen from various body sites
Porphyromonas somerae	https://www.ncbi.nlm.nih.gov/pubmed/16145091	chronic skin, soft tissue and bone infections
Porphyromonas somerae	https://www.ncbi.nlm.nih.gov/pubmed/30541687	abscesses, biopsies, wounds
Porphyromonas uenonis	https://www.ncbi.nlm.nih.gov/pubmed/15528728	identification as pathogen
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/10823756	enhanced HIV expression

Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/10875323	septic arthritis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/11368254	bacterial vaginosis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/11707013	septic arthritis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/14532256	Paronychia
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/15722627	Clinical specimens
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/16192439	abdominal cutaneous ulcer
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/16316686	Lemierre's syndrome
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/17367470	virulence
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/1747864	bacterial vaginosis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/17982605	penile abscess
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/18237241	Chorionic plate inflammation
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/19053926	Oral lichen planus
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/19271076	septic arthritis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/19283879	chest wall abscess
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/20711427	bacterial vaginosis in HIV infected women
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/21214658	amniotic fluid infection
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/21376823	Skin and soft tissue infection
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/22375046	inguinal bubo
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/23001520	Abdominal wall phlebitis following renal transplant
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/24452170	empyema
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/24787738	Pelvic inflammatory disease
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/25114266	Necrotizing fasciitis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/28008411	Proctitis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/28903767	bacterial vaginosis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/28931859	bacterial vaginosis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/29772525	bacterial vaginosis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/29860038	multi-center survey of multi-drug resistant isolates
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/8013486	endocarditis
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/8205934	obstetrics gynecology specimen
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/8270797	association with cervical cancer
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/8324131	bacterial vaginosis in pregnant women

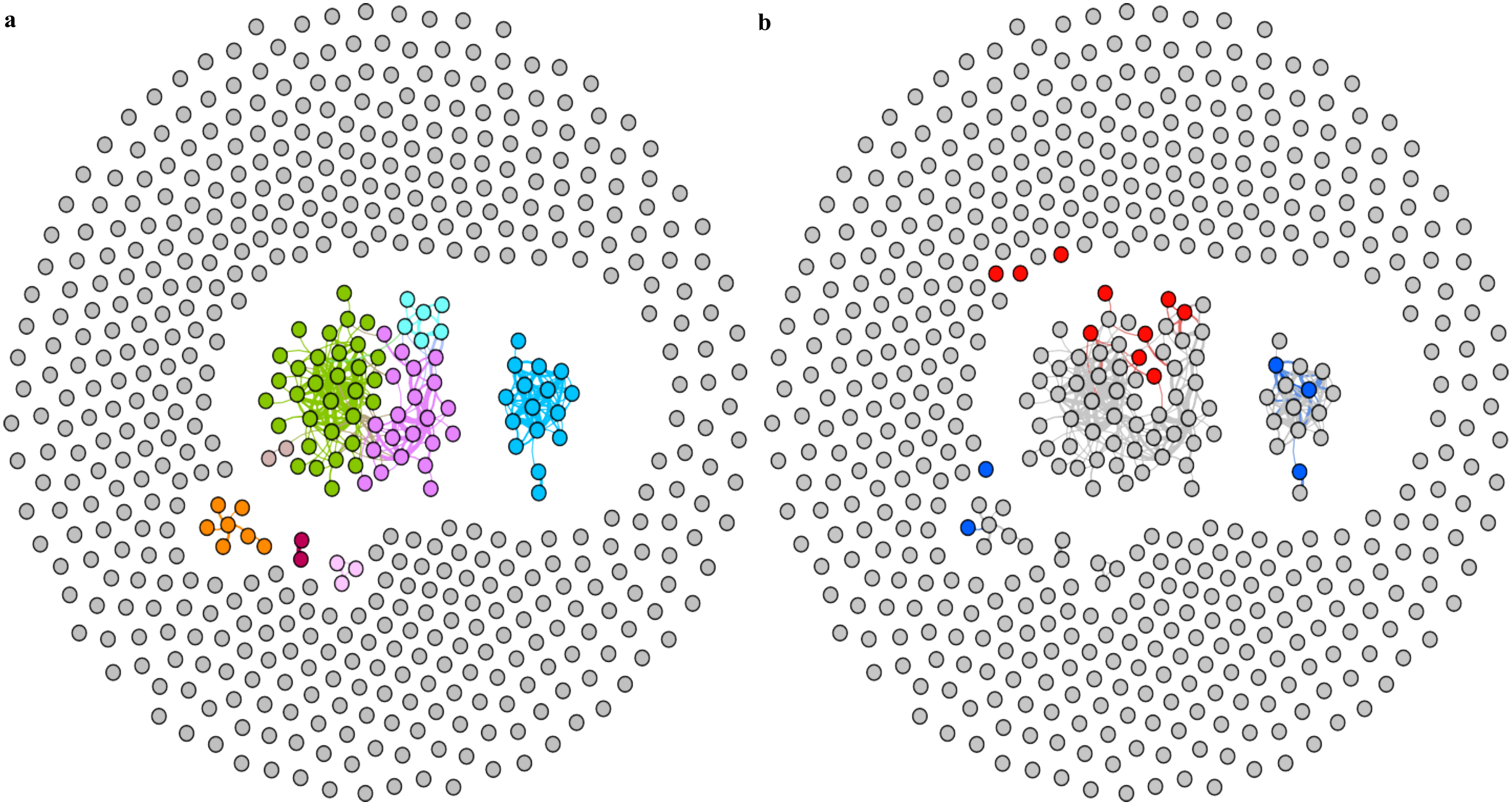
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/8677085	bacteremia after C-section
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/8907604	female genital tract infection
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/9003606	dog/cat bite wound
Prevotella bivia	https://www.ncbi.nlm.nih.gov/pubmed/9745330	Periodontal abscesses
Prevotella buccalis	https://www.ncbi.nlm.nih.gov/pubmed/14662931	urinary tract infection after renal transplant
Prevotella buccalis	https://www.ncbi.nlm.nih.gov/pubmed/24565649	Endodontic infections
Prevotella buccalis	https://www.ncbi.nlm.nih.gov/pubmed/9266340	Periodontitis
Prevotella disiens	https://www.ncbi.nlm.nih.gov/pubmed/15508748	Periodontitis
Prevotella disiens	https://www.ncbi.nlm.nih.gov/pubmed/1747864	bacterial vaginosis
Prevotella disiens	https://www.ncbi.nlm.nih.gov/pubmed/19161595	bacterial vaginosis
Prevotella disiens	https://www.ncbi.nlm.nih.gov/pubmed/24565649	Endodontic infections
Prevotella disiens	https://www.ncbi.nlm.nih.gov/pubmed/26183701	cranioplasty infection
Prevotella disiens	https://www.ncbi.nlm.nih.gov/pubmed/8205934	obstetrics gynecology specimen
Prevotella disiens	https://www.ncbi.nlm.nih.gov/pubmed/8324131	bacterial vaginosis in pregnant women
Prevotella disiens	https://www.ncbi.nlm.nih.gov/pubmed/8907604	female genital tract infection
Prevotella timonensis	https://www.ncbi.nlm.nih.gov/pubmed/17392225	breast abscess
Prevotella timonensis	https://www.ncbi.nlm.nih.gov/pubmed/29307650	various sites mostly genital and wound

Supplementary Figure 1. Correlation Network Analysis.

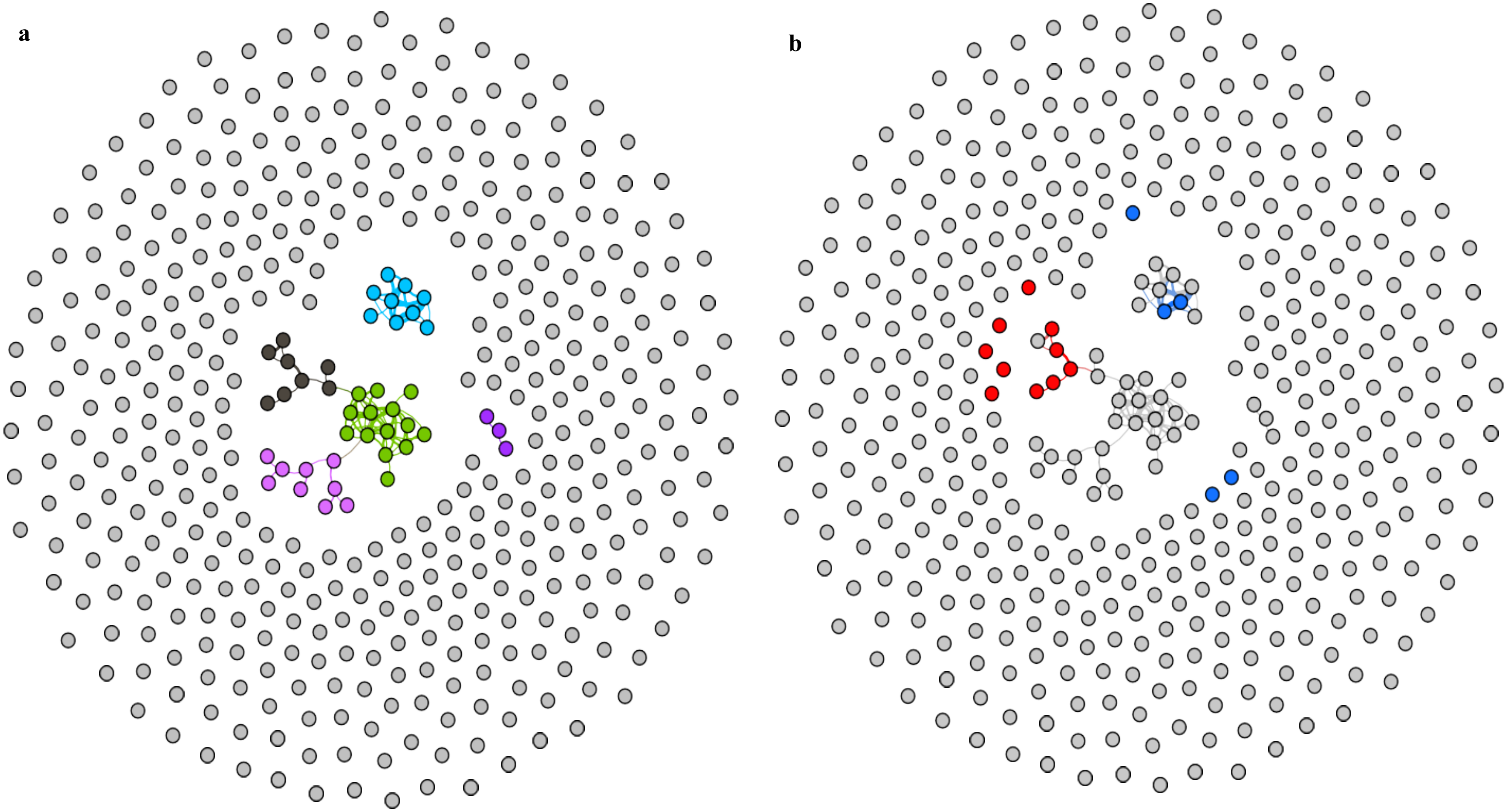
Dataset 2 Cases



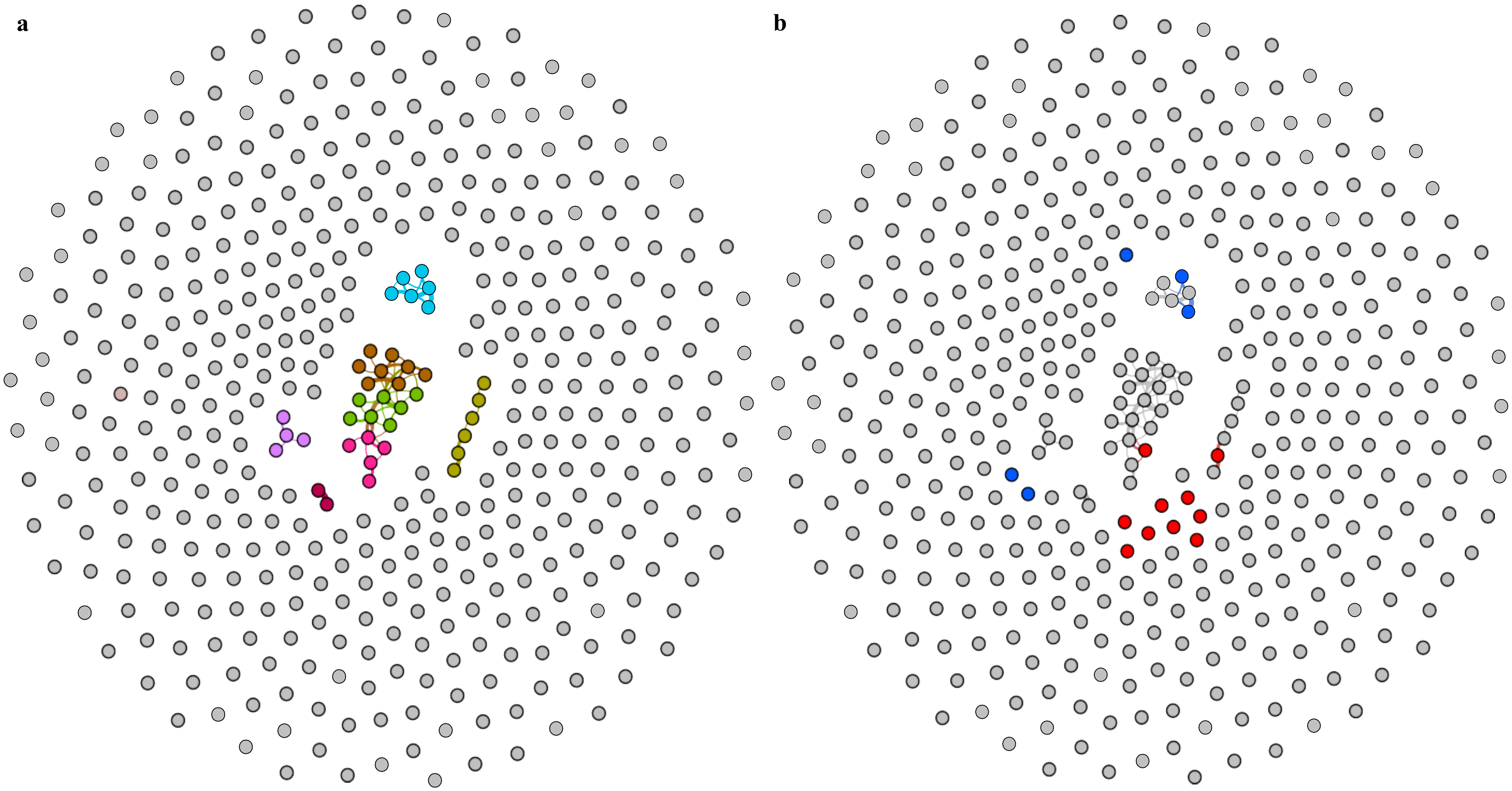
Dataset 2 Controls



Dataset 1 Cases



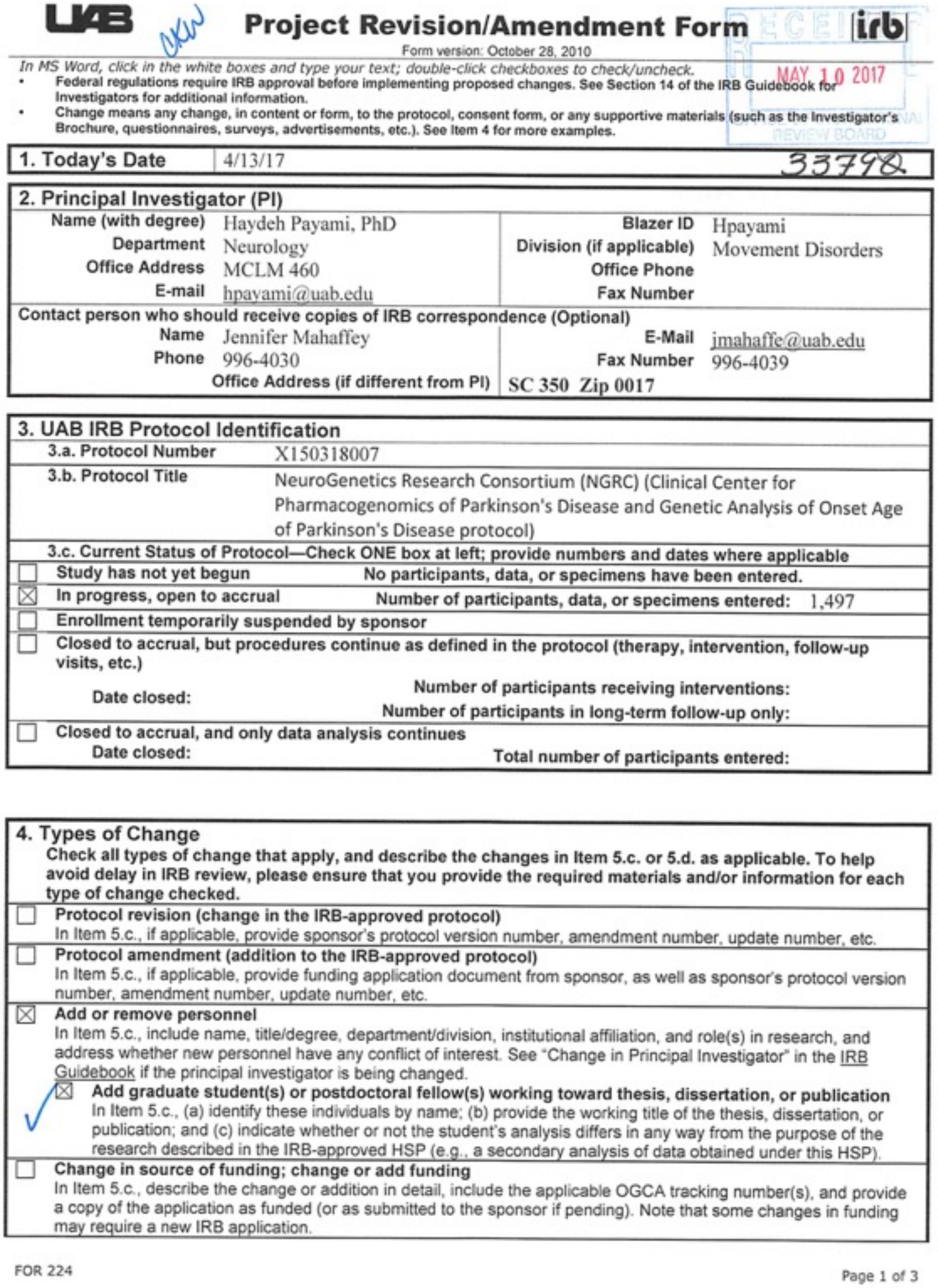
Dataset 1 Controls



We calculated pairwise correlations in relative abundances for all genera microbiome-wide, for each dataset and in cases and controls separately. Sample size: dataset 1 cases= 201, dataset 1 controls=132, dataset 2 cases= 323, dataset controls =184. To display, we set an arbitrary threshold of correlation coefficient at r≥|0.4| to connect genera that were correlated. At r≥|0.4| all correlations were significant at P<3E-4 (the limit for 3,000 permutations). The graphics denoted by "a" display the algorithm-predicted clusters in different colors. Graphics denoted by "b" are identical to their "a" counterpart except algorithm generated colors are now shown in grey and PD-associated taxa are highlighted in blue (if increased in PD) or red (if decreased in PD). Dataset 2 has more power due to larger sample size, and has greater resolution due to deeper sequencing, nonetheless, the general patterns are similar in the two datasets. Generally, the 15 PD-associated genera fall in 3 clusters. As best seen in dataset 2 cases, which has the largest sample size and power, *Porphyromonas*, *Prevotella*, and *Corynebacterium_1* co-occur in cluster 1. Eight of the 10 in cluster 2 also connect at r≥|0.4|, the other two, *Oscillospira* connects to cluster 2 at r=0.25 (P<3E-4) and *Lachnospiraceae_UCG-004* connects at r=0.35 (P<3E-4). *Lactobacillus* and *Bifidobacterium* connect to each other (cluster 3) at r=0.33 (P<3E-4).

APPENDIX C

IRB DOCUMENTATION



UAB okw

Project Revision/Amendment Form

irb

Form version: October 28, 2010

In MS Word, click in the white boxes and type your text; double-click checkboxes to check/uncheck.

- **Federal regulations require IRB approval before implementing proposed changes. See Section 14 of the IRB Guidebook for Investigators for additional information.**
- **Change means any change, in content or form, to the protocol, consent form, or any supportive materials (such as the Investigator's Brochure, questionnaires, surveys, advertisements, etc.). See Item 4 for more examples.**

1. Today's Date

4/13/17 33798

2. Principal Investigator (PI)

Name (with degree)	Haydeh Payami, PhD	**Blazer ID**	Hpayami
Department	Neurology	**Division (if applicable)**	Movement Disorders
Office Address	MCLM 460	**Office Phone**	
E-mail	hpayami@uab.edu	**Fax Number**	

Contact person who should receive copies of IRB correspondence (Optional)

Name	Jennifer Mahaffey	**E-Mail**	jmahaffe@uab.edu
Phone	996-4030	**Fax Number**	996-4039
Office Address (if different from PI)	SC 350 Zip 0017		

3. UAB IRB Protocol Identification

3.a. Protocol Number X150318007

3.b. Protocol Title NeuroGenetics Research Consortium (NGRC) (Clinical Center for Pharmacogenomics of Parkinson's Disease and Genetic Analysis of Onset Age of Parkinson's Disease protocol)

3.c. Current Status of Protocol—Check ONE box at left; provide numbers and dates where applicable

□ **Study has not yet begun** — **No participants, data, or specimens have been entered.**

☒ **In progress, open to accrual** — **Number of participants, data, or specimens entered:** 1,497

□ **Enrollment temporarily suspended by sponsor**

□ **Closed to accrual, but procedures continue as defined in the protocol (therapy, intervention, follow-up visits, etc.)**
Date closed:
Number of participants receiving interventions:
Number of participants in long-term follow-up only:

□ **Closed to accrual, and only data analysis continues**
Date closed:
Total number of participants entered:

4. Types of Change

Check all types of change that apply, and describe the changes in Item 5.c. or 5.d. as applicable. To help avoid delay in IRB review, please ensure that you provide the required materials and/or information for each type of change checked.

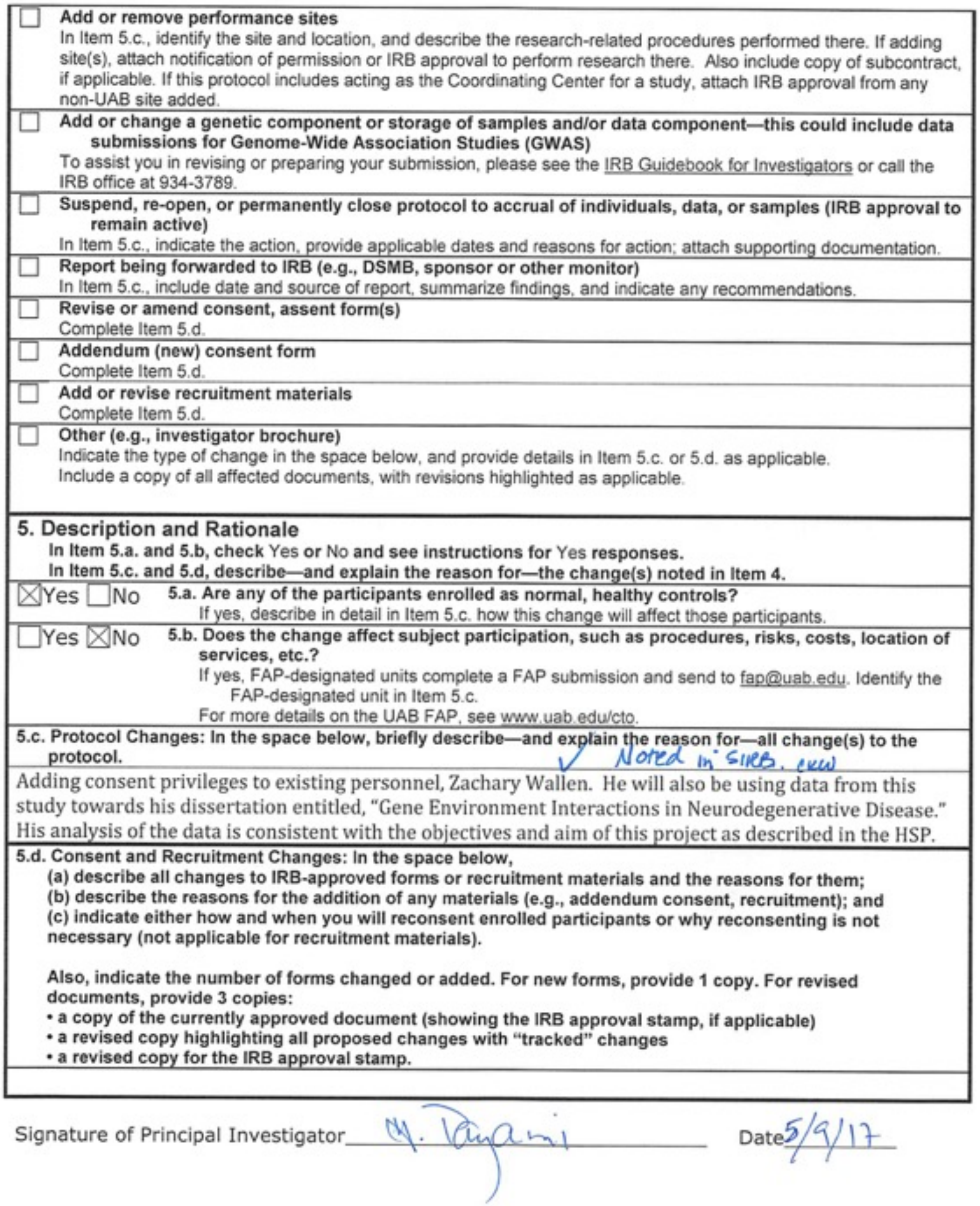
□ **Protocol revision (change in the IRB-approved protocol)**
In Item 5.c., if applicable, provide sponsor's protocol version number, amendment number, update number, etc.

□ **Protocol amendment (addition to the IRB-approved protocol)**
In Item 5.c., if applicable, provide funding application document from sponsor, as well as sponsor's protocol version number, amendment number, update number, etc.

☒ **Add or remove personnel**
In Item 5.c., include name, title/degree, department/division, institutional affiliation, and role(s) in research, and address whether new personnel have any conflict of interest. See "Change in Principal Investigator" in the IRB Guidebook if the principal investigator is being changed.

✓ ☒ **Add graduate student(s) or postdoctoral fellow(s) working toward thesis, dissertation, or publication**
In Item 5.c., (a) identify these individuals by name; (b) provide the working title of the thesis, dissertation, or publication; and (c) indicate whether or not the student's analysis differs in any way from the purpose of the research described in the IRB-approved HSP (e.g., a secondary analysis of data obtained under this HSP).

□ **Change in source of funding; change or add funding**
In Item 5.c., describe the change or addition in detail, include the applicable OGCA tracking number(s), and provide a copy of the application as funded (or as submitted to the sponsor if pending). Note that some changes in funding may require a new IRB application.



☐ **Add or remove performance sites**
In Item 5.c., identify the site and location, and describe the research-related procedures performed there. If adding site(s), attach notification of permission or IRB approval to perform research there. Also include copy of subcontract, if applicable. If this protocol includes acting as the Coordinating Center for a study, attach IRB approval from any non-UAB site added.

☐ **Add or change a genetic component or storage of samples and/or data component—this could include data submissions for Genome-Wide Association Studies (GWAS)**
To assist you in revising or preparing your submission, please see the IRB Guidebook for Investigators or call the IRB office at 934-3789.

☐ **Suspend, re-open, or permanently close protocol to accrual of individuals, data, or samples (IRB approval to remain active)**
In Item 5.c., indicate the action, provide applicable dates and reasons for action; attach supporting documentation.

☐ **Report being forwarded to IRB (e.g., DSMB, sponsor or other monitor)**
In Item 5.c., include date and source of report, summarize findings, and indicate any recommendations.

☐ **Revise or amend consent, assent form(s)**
Complete Item 5.d.

☐ **Addendum (new) consent form**
Complete Item 5.d.

☐ **Add or revise recruitment materials**
Complete Item 5.d.

☐ **Other (e.g., investigator brochure)**
Indicate the type of change in the space below, and provide details in Item 5.c. or 5.d. as applicable.
Include a copy of all affected documents, with revisions highlighted as applicable.

5. Description and Rationale

In Item 5.a. and 5.b, check Yes **or** No **and see instructions for** Yes **responses.**
In Item 5.c. and 5.d, describe—and explain the reason for—the change(s) noted in Item 4.

☒Yes ☐No **5.a. Are any of the participants enrolled as normal, healthy controls?**
If yes, describe in detail in Item 5.c. how this change will affect those participants.

☐Yes ☒No **5.b. Does the change affect subject participation, such as procedures, risks, costs, location of services, etc.?**
If yes, FAP-designated units complete a FAP submission and send to fap@uab.edu. Identify the FAP-designated unit in Item 5.c.
For more details on the UAB FAP, see www.uab.edu/cto.

5.c. Protocol Changes: In the space below, briefly describe—and explain the reason for—all change(s) to the protocol. ✓ Noted in SIRB. cew

Adding consent privileges to existing personnel, Zachary Wallen. He will also be using data from this study towards his dissertation entitled, "Gene Environment Interactions in Neurodegenerative Disease." His analysis of the data is consistent with the objectives and aim of this project as described in the HSP.

5.d. Consent and Recruitment Changes: In the space below,
(a) describe all changes to IRB-approved forms or recruitment materials and the reasons for them;
(b) describe the reasons for the addition of any materials (e.g., addendum consent, recruitment); and
(c) indicate either how and when you will reconsent enrolled participants or why reconsenting is not necessary (not applicable for recruitment materials).

Also, indicate the number of forms changed or added. For new forms, provide 1 copy. For revised documents, provide 3 copies:
- **a copy of the currently approved document (showing the IRB approval stamp, if applicable)**
- **a revised copy highlighting all proposed changes with "tracked" changes**
- **a revised copy for the IRB approval stamp.**

Signature of Principal Investigator H. Payami Date 5/9/17

FOR 224 Page 2 of 3

FOR IRB USE ONLY

□ Received & Noted ☑ Approved Expedited* □ To Convened IRB

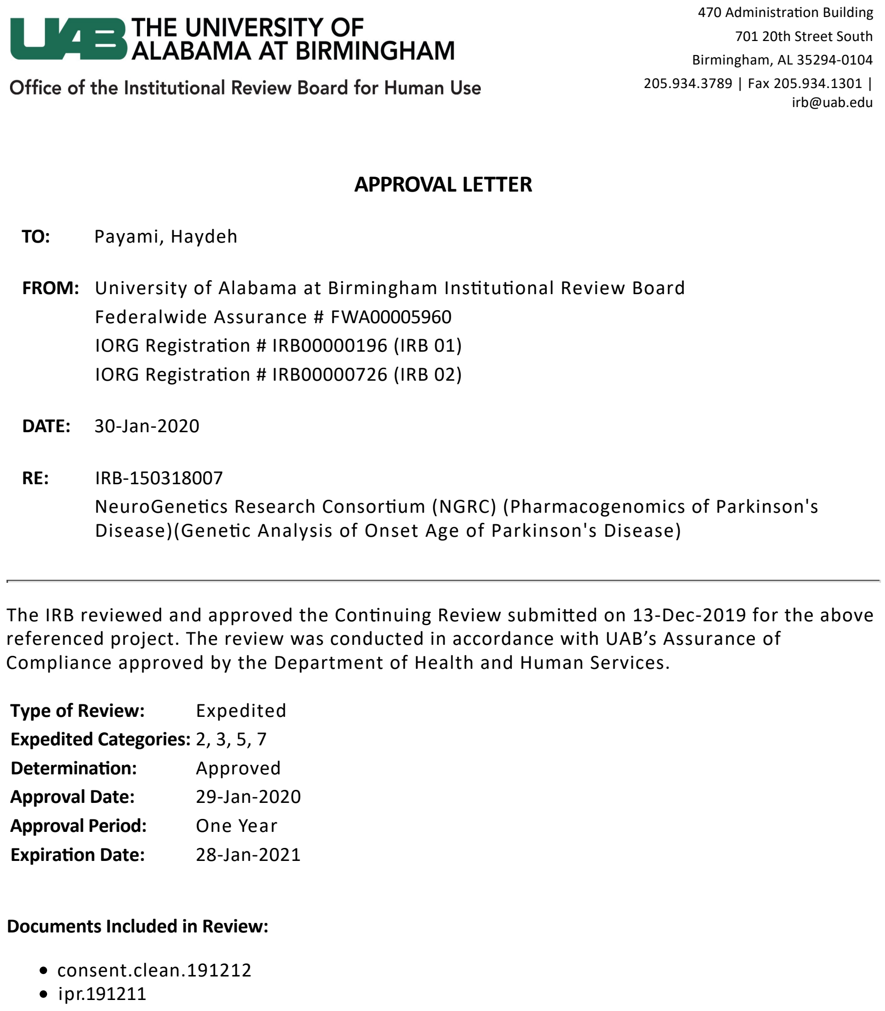
Margie Lawson, CIP 5/12/17

Signature (Chair, Vice-Chair, Designee) Date

DOLA 4/4/17

Change to Expedited Category Y / [✓ N] / NA

*No change to IRB's previous determination of approval criteria at 45 CFR 46.111 or 21 CFR 56.111



THE UNIVERSITY OF
ALABAMA AT BIRMINGHAM

Office of the Institutional Review Board for Human Use

470 Administration Building
701 20th Street South
Birmingham, AL 35294-0104
205.934.3789 | Fax 205.934.1301 |
irb@uab.edu

APPROVAL LETTER

TO: Payami, Haydeh

FROM: University of Alabama at Birmingham Institutional Review Board
Federalwide Assurance # FWA00005960
IORG Registration # IRB00000196 (IRB 01)
IORG Registration # IRB00000726 (IRB 02)

DATE: 30-Jan-2020

RE: IRB-150318007
NeuroGenetics Research Consortium (NGRC) (Pharmacogenomics of Parkinson's Disease)(Genetic Analysis of Onset Age of Parkinson's Disease)

The IRB reviewed and approved the Continuing Review submitted on 13-Dec-2019 for the above referenced project. The review was conducted in accordance with UAB's Assurance of Compliance approved by the Department of Health and Human Services.

Type of Review:	Expedited
Expedited Categories:	2, 3, 5, 7
Determination:	Approved
Approval Date:	29-Jan-2020
Approval Period:	One Year
Expiration Date:	28-Jan-2021

Documents Included in Review:

- consent.clean.191212
- ipr.191211