Characterizing dysbiosis of gut microbiome in PD: Evidence for overabundance of opportunistic pathogens

Zachary D Wallen, MS,[1] Mary Appah, MS,[1] Marissa N Dean, MD,[1] Cheryl L Sesler, MS[1] Stewart A Factor, DO,[2] Eric Molho, MD,[3] Cyrus P Zabetian, MD, MS,[4] David G Standaert, MD, PhD,[1] and Haydeh Payami*, PhD[1]

[1] Department of Neurology, University of Alabama at Birmingham, Birmingham, AL, 35233, USA
[2] Department of Neurology, Emory University School of Medicine, Atlanta, GA, 30322, USA
[3] Department of Neurology, Albany Medical College, Albany, NY, 12208, USA
[4] VA Puget Sound Health Care System and Department of Neurology, University of Washington, Seattle, WA, 98108, USA

Corresponding author
Haydeh Payami, University of Alabama at Birmingham, Room 429, Civitan International Research Center, 1719 6th Ave S, Birmingham, AL 35233
Phone: 205-934-0371 Email:haydehpayami@uabmc.edu

Competing interests
Authors have no conflict of interest.

Acknowledgement
This work was supported by the National Institute of Neurological Disorders and Stroke grant R01036960 (to HP), The U.S. Army Medical Research Materiel Command endorsed by the U.S. Army through the Parkinson's Research Program Investigator-Initiated Research Award under Award No. W81XWH1810508 (to HP); NIH Udall grant P50 NS062684 (to CPZ) and P50 NS108675 (to DGS), and NIH Training Grant T32NS095775 (to ZDW). Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the U.S. Army or the NIH.

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 represented 3 clusters of co-occurring microorganisms. Cluster 1 was composed of opportunistic pathogens; all were elevated in PD. Cluster 2 were short-chain-fatty-acid producing bacteria; all were reduced in PD. Cluster 3 were carbohydrate-metabolizing probiotics; elevated in PD. Depletion of anti-inflammatory short-chain-fatty-acid producing bacteria and elevated levels of probiotics are confirmatory. Overabundance of opportunistic pathogens is a novel finding and their identity provides a lead to experimentally test their role in PD.

Introduction

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.[3-6] 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 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, life-styles, 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 data sets 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 identified novel 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 (Figure 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 2015-2018. 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 10x 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 (Figure 1), and the difference was statistically significant (P<1E-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, one, for independent validation, and two, to avoid confounding by mixing two clearly different datasets.

Metadata and Confounders

Metadata were collected using two self-administered questionnaires and medical records (Supplementary Table 1). An Environmental and Family History Questionnaire[4,36] was used to collect data relevant to PD. A Gut Microbiome Questionnaire[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, while 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 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=1E-5.

PD microbiomes differed significantly from control microbiomes, in both datasets, with every distance metric measured (P<1E-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 microbiome,[33,34] and with other PD studies in detecting evidence for dysbiosis in PD.[11-30]

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 if results replicate, and also to avoid confounding by the heterogeneity between datasets. Each data set was tested with two methods to test analytic concordance: once using analysis of composition of microbiomes (ANCOM)[40] and again using Kruskal-Wallis rank sum test (KW).[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). All tests were corrected for multiple testing.

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, Figure 2). Five genera had higher abundances in PD than controls: *Porphyromonas, Prevotella, Corynebacterium_1, Bifidobacterium* and *Lactobacillus.* Ten genera had lower abundances in PD than controls: *Faecalibacterium, Agathobacter, Blautia, Roseburia, Fusicatenibacter, Lachnospira, Butyricicoccus, Lachnospiraceae_ND3007_group, Lachnospiraceae_UCG-004,* and *Oscillospira.* Complete MWAS results are in Supplementary Tables 2-5.

Correlation network analysis

We questioned if the 15 association signals were independent. We used hypothesis-free correlation network analysis[43] to infer ecological networks of interacting organisms microbiome-

wide (Figure 3, Supplementary Figure 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<3E-4), the highest in the microbiome in our data. The 10 genera that were depleted in PD formed cluster 2, where eight of them clustered at r≥0.4 (P<3E-4), and remaining two (*Oscillospira* and *Lachnospiraceae_UCG-004),* clustered with the others at r=0.25 (P<3E-4) and r=0.35 (P<3E-4). *Lactobacillus* and *Bifidobacterium*, both elevated in PD, were correlated with each other at r=0.33 (P<3E-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 corresponding clusters, we broke the blind by searching PubMed. PubMed results on functional characteristics converged on clusters defined by agnostic network analysis.

Genera in cluster 1: *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. While 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.[44-46]

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 *C. amycolatum* and *C. lactis;* 96% of *Porphyromonas* was composed of ASVs that matched *P. asaccharolytica*, *P. bennonis, P. somerae* or *P. uenonis* with >99%-100% identity; and 98% of *Prevotella* was composed of ASVs that matched *P. bivia, P. buccalis*, *P. disiens*, or *P. 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 10 species, using genus and species name as key word (ex. *Corynebacterium amycolatum*), with search filters: Humans, English, 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.[44-46] *Porphyromonas, Prevotella, Corynebacterium* and other members of cluster 1 are often observed together in these polymicrobial infections.[44-46] With the newly acquired knowledge on the potential biological significance of cluster 1, we questioned if 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 non-significant 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, and in addition, we gained a new significant 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).

Genera in cluster 2: 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 short chain fatty acids (SCFA), mainly butyrate, which help maintain integrity of the gut membrane, and have anti-inflammatory properties.[47,48]

Genera in cluster 3: *Lactobacillus*[49] and *Bifidobacteria*[50] are anaerobic gram positive bacteria. They are among 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. While 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 if 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 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=5E-4, *Roseburia* P=0.02*, Fusicatenibacter* P=0.01*, Lachnospira* P=5E-3*,*

Lachnospiraceae_ND3007_group P=5E-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=5E-3, dataset 2 P=2E-6) and *Lactobacillus* (dataset 2 P=4E-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

Summary We 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. 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.

Alignment with PD literature Reduced levels of SCFA producing bacteria[12,14,16,18,19,21,26,27] and elevated levels of probiotic bacteria in PD[14,16,18,21,25-27] have been reported before, and thus are confirmatory. Overabundance of opportunistic pathogens was a novel finding. We suspect the reason we were able to detect these microorganisms is because they are rare (Figure 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 6 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 while 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. reported at the family level (*Prevotellaceae*), we at genus level (*Prevotella*), and Petrov et al. 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. observed, but the difference was not significant in our datasets (FDR>0.6 in both datasets). Species *P. copri*, which Petrov et al. 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 2 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.

Opportunistic pathogens Overabundance of opportunistic pathogens in PD gut microbiome was a novel and potentially the most exciting finding of this study. Braak and colleagues 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.[54,55] While 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* knock-out 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 is no direct evidence that a pathogen is responsible for the pathology. Here, we present the first 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, for example, compromised membanes.[44-46] 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. While 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.

Anti-inflammatory SCFA producing bacteria 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,[47,48,64-66] each 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,[67-71] and is not specific to PD.

Probiotics We also found elevated levels of *Bifidobacterium* and *Lactobacillus* in PD, which have been noted in some of the prior PD studies, albeit not consistently.[14,16,18,21,25-27] 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] While 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).

Conclusion We uncovered robust and reproducible signals, which reaffirm (SCFA, probiotics) and generate new 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

Subjects (Supplementary Table 1) The study was approved by institutional review boards for ethical conduct of human subject research at all participating institutions. All subjects provided informed consent for their participation. Subjects were enrolled by NGRC investigators, using standardized methods, at four NGRC affiliated movement disorder clinics in 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 2015-2018, 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.

Metadata (Supplementary Table 1) Data were collected using two self-administered questionnaires: an Environmental and Family History Questionnaire (EFQ) and Gut Microbiome Questionnaire (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 Subjects collected stool samples at home using DNA/RNA-free sterile swabs (BD BBL CultureSwab Sterile/Media-free Swabs, Fisher Scientific, Pittsburgh, PA). The sample was collected from excreted stool (the kit is not a rectal swab), thus minimizing contamination by skin microbiota. The stool samples were shipped immediately via standard U.S. 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). 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).

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 6 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

Sequence QC 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.

Amplicon sequence variant (ASV) inference and ASV tables 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 1 ASV table for dataset 1 and 6 ASV tables for dataset 2. The 6 ASV tables of dataset 2 were merged using the mergeSequenceTables function. Chimeras were detected and removed using the removeBimeraDenovo function.

Data transformation The following 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 1.4.2 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 assignment 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 (RDP) naïve 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+/).

Phylogenetic trees 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]

Phyloseq Object For each dataset, a phyloseq object was created 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

Principal component analysis 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 (Figure 1).

Confounders We interrogated 47 variables (extracted from metadata) 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 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). Since 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 (ß 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 were present only in PD cases and could not be included as covariates in PERMANOVA or MWAS. To gauge the effect of PD on ß 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 implemented in the cor.test function in R.

Global composition of microbiome (ß diversity) PERMANOVA was used to identify variables that had a significant effect on ß 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.

(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 10lbs in past year + gastrointestinal discomfort on day of stool collection + constipation in past three months + alcohol use + fruits or vegetables daily + stool sample travel time]

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

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.

(Model C) Subset of PD cases not on a given PD medication vs controls: [Distance ~ case/control]

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 alpha to be 0.5. To calculate Canberra distances, data were transformed to relative abundances and distances were calculated using the vegdist function in vegan.

MWAS 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 three 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 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 (Figure 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 (Figure 3, Supplementary Figure 1) For each dataset, and for cases and controls separately, 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 3,000 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<3E-4, which is the maximum significance attainable with 3,000 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 Data will be publicly available at NCBI Sequence Read Archive (SRA).

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

Contributions ZDW and HP were responsible for the design and execution of the study and wrote the first draft of the paper. All co-authors reviewed and critiqued the paper. SAF, EM, CPZ, DGS and MND were responsible for the clinical aspects of the study. ZDW and MA performed bioinformatics and statistical analysis. CLS assisted with blasting and literature searches.

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, doi:10.1038/ng.3955 (2017).

2 Tanner, C. M. Advances in environmental epidemiology. *Mov Disord* **25 Suppl 1**, S58-62 (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, doi:10.1016/j.nbd.2012.06.025 (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, doi:10.1038/tpj.2012.38 (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, doi:10.1016/j.parkreldis.2016.08.002 (2016).

7 Schmidt, T. S. B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association to Modulation. *Cell* **172**, 1198-1215, doi:10.1016/j.cell.2018.02.044 (2018).

8 Chen, H. *et al.* Meta-analyses on prevalence of selected Parkinson's nonmotor symptoms before and after diagnosis. *Transl Neurodegener* **4**, 1, doi:10.1186/2047-9158-4-1 (2015).

9 Houser, M. C. *et al.* Stool Immune Profiles Evince Gastrointestinal Inflammation in Parkinson's Disease. *Mov Disord* **33**, 793-804, doi:10.1002/mds.27326 (2018).

10 Forsyth, C. B. *et al.* Increased intestinal permeability correlates with sigmoid mucosa alpha-synuclein staining and endotoxin exposure markers in early Parkinson's disease. *PLoS One* **6**, e28032, doi:10.1371/journal.pone.0028032 (2011).

11 Scheperjans, F. *et al.* Gut microbiota are related to Parkinson's disease and clinical phenotype. *Mov Disord* **30**, 350-358, doi:10.1002/mds.26069 (2015).

12 Keshavarzian, A. *et al.* Colonic bacterial composition in Parkinson's disease. *Mov Disord* **30**, 1351-1360, doi:10.1002/mds.26307 (2015).

13 Hasegawa, S. *et al.* Intestinal Dysbiosis and Lowered Serum Lipopolysaccharide-Binding Protein in Parkinson's Disease. *PLoS One* **10**, e0142164, doi:10.1371/journal.pone.0142164 (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*, doi: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, doi:10.1016/j.cell.2016.11.018 (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, doi:10.1002/mds.26942 (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, doi:10.1186/s13073-017-0428-y (2017).

18 Petrov, V. A. *et al.* Analysis of Gut Microbiota in Patients with Parkinson's Disease. *Bull Exp Biol Med* **162**, 734-737, doi:10.1007/s10517-017-3700-7 (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, doi:10.1007/s11427-016-9001-4 (2017).

20 Hopfner, F. *et al.* Gut microbiota in Parkinson disease in a northern German cohort. *Brain Res* **1667**, 41-45, doi:10.1016/j.brainres.2017.04.019 (2017).

21 Lin, A. *et al.* Gut microbiota in patients with Parkinson's disease in southern China. *Parkinsonism Relat Disord* **53**, 82-88, doi:10.1016/j.parkreldis.2018.05.007 (2018).

22 Qian, Y. *et al.* Alteration of the fecal microbiota in Chinese patients with Parkinson's disease. *Brain Behav Immun* **70**, 194-202, doi:10.1016/j.bbi.2018.02.016 (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, doi:10.1002/mds.27105 (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, doi:10.1038/s41531-019-0100-x (2019).

25 Barichella, M. *et al.* Unraveling gut microbiota in Parkinson's disease and atypical parkinsonism. *Mov Disord* **34**, 396-405, doi:10.1002/mds.27581 (2019).

26 Pietrucci, D. *et al.* Dysbiosis of gut microbiota in a selected population of Parkinson's patients. *Parkinsonism Relat Disord*, doi: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, doi:10.1016/j.ebiom.2019.05.064 (2019).

28 Lin, C. H. *et al.* Altered gut microbiota and inflammatory cytokine responses in patients with Parkinson's disease. *J Neuroinflammation* **16**, 129, doi:10.1186/s12974-019-1528-y (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, doi:10.1016/j.neulet.2019.134297 (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, doi:10.3389/fnmol.2019.00171 (2019).

31 Gerhardt, S. & Mohajeri, M. H. Changes of Colonic Bacterial Composition in Parkinson's Disease and Other Neurodegenerative Diseases. *Nutrients* **10**, doi: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, doi:10.3233/JPD-191711 (2019).

33 Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560-564, doi:10.1126/science.aad3503 (2016).

34 Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565-569, doi:10.1126/science.aad3369 (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, doi:10.3389/fmicb.2017.02224 (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, doi:10.1093/bioinformatics/bts342 (2012).

38 Lance, G. N. & Williams, W. T. Computer programs for hierarchical polythetic classification (“similarity analyses”). *The Computer Journal* **9**, 60-64 (1966).

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

40 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).

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, doi:10.1186/s40168-017-0237-y (2017).

43 Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**, e1002687, doi:10.1371/journal.pcbi.1002687 (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, doi:10.1128/JCM.00551-07 (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, doi:10.1111/1462-2920.13632 (2017).

46 Choi, Y. *et al.* Co-occurrence of Anaerobes in Human Chronic Wounds. *Microb Ecol* **77**, 808-820, doi:10.1007/s00248-018-1231-z (2019).

47 Hamer, H. M. *et al.* Review article: the role of butyrate on colonic function. *Aliment Pharmacol Ther* **27**, 104-119, doi:10.1111/j.1365-2036.2007.03562.x (2008).

48 Canani, R. B. *et al.* Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. *World J Gastroenterol* **17**, 1519-1528, doi:10.3748/wjg.v17.i12. 1519 10.3748/wjg.v17.i12.1519 (2011).

49 O'Callaghan, J. & O'Toole, P. W. Lactobacillus: host-microbe relationships. *Curr Top Microbiol Immunol* **358**, 119-154, doi:10.1007/82_2011_187 (2013).

50 O'Callaghan, A. & van Sinderen, D. Bifidobacteria and Their Role as Members of the Human Gut Microbiota. *Front Microbiol* **7**, 925, doi:10.3389/fmicb.2016.00925 (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, doi:10.1038/s41591-019-0439-x (2019).

53 Doron, S. & Snydman, D. R. Risk and safety of probiotics. *Clin Infect Dis* **60 Suppl 2**, S129-134, doi:10.1093/cid/civ085 (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 neuroinvasion by an unknown pathogen. *J Neural Transm (Vienna)* **110**, 517-536, doi:10.1007/s00702-002-0808-2 (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, doi:10.1002/mds.27556 (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, doi:10.1016/S1474-4422(18)30162-5 (2018).

58 Svensson, E. *et al.* Vagotomy and subsequent risk of Parkinson's disease. *Ann Neurol* **78**, 522-529, doi:10.1002/ana.24448 (2015).

59 Liu, B. *et al.* Vagotomy and Parkinson disease: A Swedish register-based matched-cohort study. *Neurology* **88**, 1996-2002, doi:10.1212/WNL.0000000000003961 (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, doi:10.1016/j.neuron.2019.05.035 (2019).

61 Stolzenberg, E. *et al.* A Role for Neuronal Alpha-Synuclein in Gastrointestinal Immunity. *J Innate Immun*, doi: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, doi:10.1007/s00702-017-1726-7 (2017).

63 Matheoud, D. *et al.* Intestinal infection triggers Parkinson's disease-like symptoms in Pink1(-/-) mice. *Nature* **571**, 565-569, doi:10.1038/s41586-019-1405-y (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, doi:10.1038/s41598-019-45311-y (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, doi:10.1111/imm.12933 (2018).

66 Furusawa, Y. *et al.* Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446-450, doi:10.1038/nature12721 (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**, doi: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. *Medicine (Baltimore)* **98**, e14513, doi:10.1097/MD.0000000000014513 (2019).

69 Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55-60, doi:10.1038/nature11450 (2012).

70 Guo, Z. *et al.* Intestinal Microbiota Distinguish Gout Patients from Healthy Humans. *Sci Rep* **6**, 20602, doi:10.1038/srep20602 (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 Enteral Nutr* **39**, 569-577, doi:10.1177/0148607114529596 (2015).

72 Zhang, K. & Ni, Y. Tyrosine decarboxylase from Lactobacillus brevis: soluble expression and characterization. *Protein Expr Purif* **94**, 33-39, doi:10.1016/j.pep.2013.10.018 (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**, doi: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, doi:10.1212/WNL.0000000000003127 (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, doi:10.1038/s41587-019-0209-9 (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, doi:10.1038/nmeth.3869 (2016).

79 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).

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. *Applied and environmental microbiology* **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, doi:10.1089/10665270050081478 (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 community analyses. *F1000Res* **5**, 1492, doi:10.12688/f1000research.8986.2 (2016).

83 Wright, E. S. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* **16**, 322, doi:10.1186/s12859-015-0749-z (2015).

84 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (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, doi:10.1371/journal.pone.0061217 (2013).

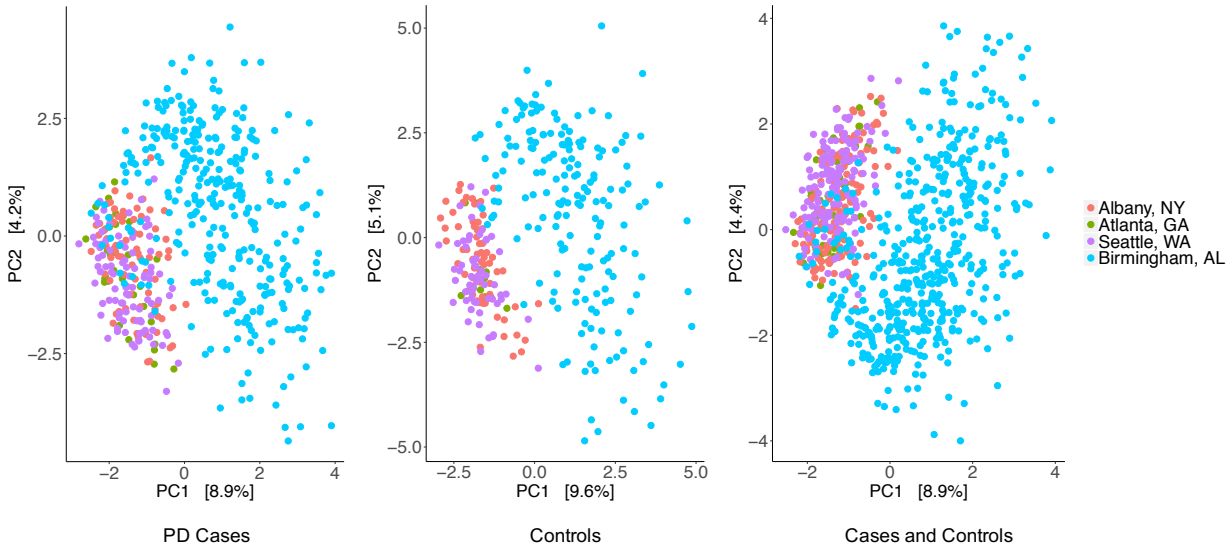
86 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290, doi:10.1093/bioinformatics/btg412 (2004).

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* (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, doi:10.1371/journal.pone.0098679 (2014).

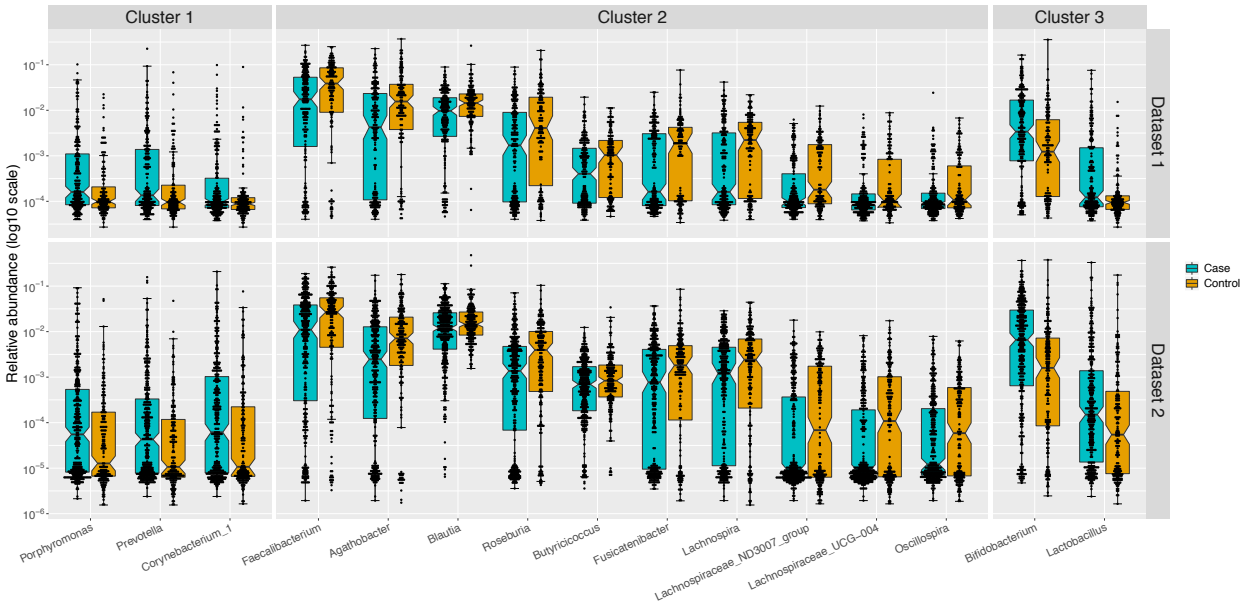
89 Blondel, V., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. P10008 (2008).

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



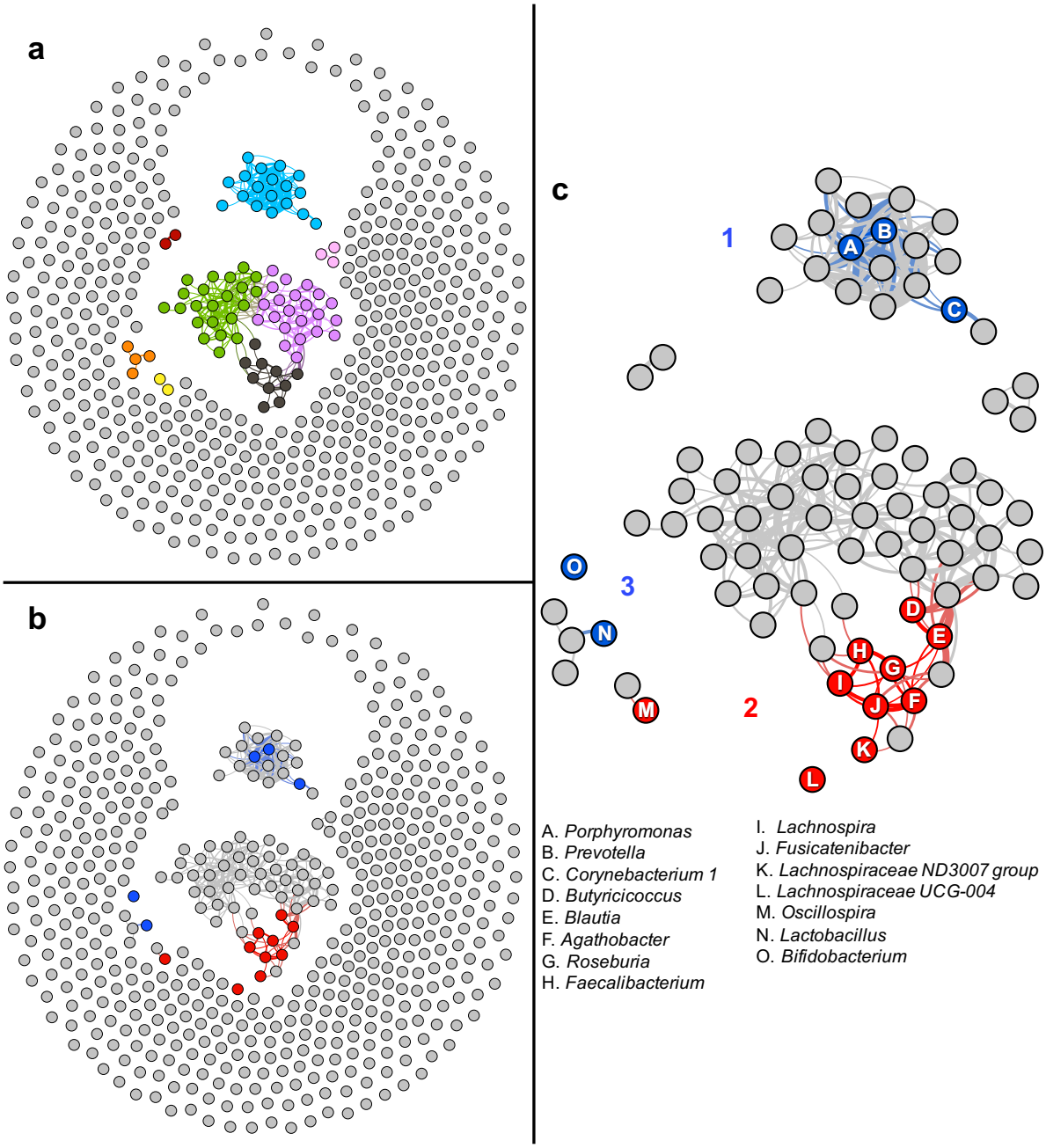
Principal Component (PC) Analysis was used to generate the graphs for PD cases (left), controls (middle), and cases and controls combined (right) where each point represents the composition of the gut microbiome of one individual and distances indicate degree of similarity to other individuals. Percentages 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-5). Dataset 1: red (Albany, NY), purple (Seattle, WA) and green (Atlanta, GA). Dataset 2: blue (Birmingham, AL).

Figure 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. 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.

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 used to detect clusters of co-occurring microorganisms. To display, we used an arbitrary

correlation coefficient threshold at r≥ |0.4| to connect the genera that were correlated. All correlations noted were significant at P<3E-4 (the limit for 3,000 permutations). Here we show the result for PD cases in dataset 2 because it had larger sample size and greater sequencing depth than dataset 1. (See Supplementary Figure 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 panel a but shown in grey, and PD-associated genera highlighted in blue (if increased in PD) or red (if decreased in PD). **(c)** Zoomed in version of panel b. The 15 PD-associated genera fell in 3 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 10 genera that were reduced in PD, eight of which are shown connected at r≥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, for example, *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).

Table 1. Effect of PD and other key variables on the global composition of microbiome

	Dataset 1 (201 cases, 132 controls)						Dataset 2 (323 cases, 184 controls)					
	Aitchison		GUniFrac		Canberra		Aitchison		GUniFrac		Canberra	
	%variation	P	%variation	P	%variation	P	%variation	P	%variation	P	%variation	P
Model A. All PD vs Control	0.71	<1E-5	1.38	<1E-5	0.57	<1E-5	0.56	<1E-5	0.89	<1E-5	0.38	<1E-5
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-5	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-5	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-5	0.56	<1E-5	0.55	<1E-5	0.88	<1E-5	0.38	<1E-5
PD not on anticholinergics vs control	0.73	<1E-5	1.31	<1E-5	0.58	<1E-5	0.57	<1E-5	0.92	2E-05	0.39	<1E-5
PD not on MAO-B inhibitors vs control	0.81	<1E-5	1.50	3E-05	0.66	<1E-5	0.71	<1E-5	1.07	<1E-5	0.45	<1E-5
PD not on dopamine agonists vs control	0.81	2E-04	1.51	3E-05	0.70	<1E-5	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-5	0.60	<1E-5	0.48	3E-05	0.74	3E-05	0.37	<1E-5
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

Model A tested PD vs. control without any other variable in the model. Model B included 11 variables (including case/control) and each was tested while adjusting for the other 10, without priority. For Model C patients were stratified by the PD medication they were taking at the time of stool collection; those not on medication were tested against controls

All analyses were repeated with 3 different distance measures: Aitchison, Canberra, and GUniFrac (generalized UniFrac). % variation was the inter-individual variation explained by each variable. P value was calculated using 999,999 permutations, setting the highest achievable significance at P=1E-5.

* Analysis included subset of samples that had complete data on all 11 variables, Dataset 1 N= 160 cases, 111 controls, Dataset 2 N= 283 cases, 167 controls.

** Low power due to N<50 not on L-dopa, and <20 not on any PD medication. For other medications, N of patients not on medication was 88 to 179 in dataset 1 and 153 to 312 in dataset 2.

Table 2. PD-associated genera.

PD-associated genera					MWAS significant In Dataset 1				MWAS significant In Dataset 2				Cluster	Pub-Med
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	pathogen
Bacteroidetes	*Bacteroidia*	*Bacteroidales*	*Prevotellaceae*	*Prevotella*	0.002	2.56	400	6E-03	0.001	4.39	463	2E-02	1	pathogen
Actinobacteria	*Actinobacteria*	*Corynebacteriales*	*Corynebacteriaceae*	*Corynebacterium_1*	0.001	1.96	360	1E-02	0.002	2.53	465	8E-03	1	pathogen
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	SCFA
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*	0.0006	0.65	367	2E-02	0.0005	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*	0.0004	6.61	407	2E-04	0.004	1.57	458	1E-02	3	probiotic

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). Clusters were identified hypothesis-free using correlation network analysis (Figure 3). PubMed search was conducted after analyses were completed using genus and species name as search term (Supplementary Table 6). Function (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, 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).

MWAS=microbiome-wide association study

MRA= mean relative abundance in controls

FC = fold change in patients (MRA in patients/MRA in controls)

ANCOM= analysis of composition of microbiomes

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). Threshold 0.8 was used for significance. All shown genera were above significance threshold.

KW= Kruskal-Wallis

FDR=false discovery rate

NC=uncultured, not characterized