Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Pangenome-resolved deconvolution of gut metatranscriptomes reveals species-specific microbial activity programs in Parkinson's disease

Paul Wilmes

paul.wilmes@uni.lu

University of Luxembourg https://orcid.org/0000-0002-6478-2924

Viacheslav Petrov

University of Luxembourg, Luxembourg Centre for Systems Biomedicine https://orcid.org/0000-0002-5205-9739

Oskar Hickl

Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST)

Remy Villette

University of Luxembourg https://orcid.org/0000-0001-5038-4741

Bérénice Hansen

Luxembourg Centre for Systems Biomedicine, University of Luxembourg

Stella Blumrich

Luxembourg Centre for Systems Biomedicine, University of Luxembourg https://orcid.org/0009-0001-9269-1018

Cedric Laczny

Luxembourg Centre for Systems Biomedicine, University of Luxembourg https://orcid.org/0000-0002-1100-1282

Tuesday Lowndes

Luxembourg Centre for Systems Biomedicine, University of Luxembourg

Zainab Zafar

Luxembourg Centre for Systems Biomedicine, University of Luxembourg

Jordan Caussin

Luxembourg Centre for Systems Biomedicine, University of Luxembourg

Rashi Halder

University of Luxembourg https://orcid.org/0000-0002-1402-1254

Etienne Hanslian

Institute for Social Medicine, Epidemiology and Health Economics, Charité - University Medical Center Berlin

Daniela Koppold

Institute for Social Medicine, Epidemiology and Health Economics, Charité - University Medical Center Berlin

Anika Hartmann Rajput
Department of Dermatology, Venereology and Allergology, Charité - University Medical Center Berlin

Michael Jeitler
Institute for Social Medicine, Epidemiology and Health Economics, Charité - University Medical Center Berlin

Nico Steckhan
Institute for Social Medicine, Epidemiology and Health Economics, Charité - University Medical Center Berlin

Sebastian Schade
Neurosurgery, University Medical Center Göttingen https://orcid.org/0000-0002-6316-6804

Brit Mollenhauer
Department of Neurology, University Medical Center Göttingen

Andreas Michalsen
Institute for Social Medicine, Epidemiology and Health Economics, Charité - University Medical Center Berlin

Jochen Schneider
Luxembourg Centre for Systems Biomedicine, University of Luxembourg

Patrick May
University of Luxembourg https://orcid.org/0000-0001-8698-3770

Article

Keywords:

Posted Date: October 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11117363/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Pangenome-resolved deconvolution of gut metatranscriptomes reveals species-specific microbial activity programs in Parkinson's disease

Viacheslav A. Petrov[1], Oskar Hickl[1,2,3], Rémy Villette[1], Bérénice Hansen[1], Stella Blumrich[1], Cédric C. Laczny[1], Tuesday Lowndes[1], Zainab Zafar[1], Jordan Caussin[1], Rashi Halder[1], Etienne Hanslian[4,5], Daniela A. Koppold[4,5], Anika Hartmann Rajput[8], Michael Jeitler[4,5,6,7], Nico Steckhan[4,9], Sebastian Schade[10], Brit Mollenhauer[10,11], Andreas Michalsen[4,5], Jochen G. Schneider[1,12,13], Patrick May[1*] & Paul Wilmes[1,12*]

1. Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
2. Department of Infection and Immunity, Luxembourg Institute of Health, Luxembourg
3. Systems and Synthetic Biology, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
4. Institute for Social Medicine, Epidemiology and Health Economics, Charité - University Medical Center Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
5. Department of Internal and Integrative Medicine, Immanuel Hospital Berlin-Wannsee Branch, Berlin, Germany
6. Institute of General Practice and Interprofessional Care, University Hospital Tübingen, Tübingen, Germany
7. Robert Bosch Center for Integrative Medicine and Health, Bosch Health Campus, Stuttgart, Germany
8. Department of Dermatology, Venereology and Allergology, Charité - University Medical Center Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
9. Digital Health-Connected Healthcare, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
10. Paracelsus-Elena-Klinik Kassel, Kassel, Germany
11. Department of Neurology, University Medical Center Göttingen, Göttingen, Germany
12. Department of Health, Medicine and Life Sciences, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
13. Department of Internal Medicine II, Saarland University Hospital and Saarland University Faculty of Medicine, Homburg, Germany

*These authors contributed equally to this work and share last authorship

Abstract

Gut microbiome alterations are associated with Parkinson's disease (PD), yet their species-level functional consequences remain poorly understood. Here, we develop a pangenome-resolved framework integrating paired metagenomic and metatranscriptomic data to reconstruct species-level transcriptional profiles. Applied to two independent cohorts comprising 160 individuals and 7,548 profiles across 215 species, this approach revealed reduced genome-wide transcriptional breadth and increased inter-individual heterogeneity in PD. Genomic diversity and transcriptomic variation were significantly associated in 61 of 95 tested species. We identified 2,271 pangenome-resolved features associated with disease status, duration, or severity, with 39% of significant gene-level associations involving

accessory genes. Species-specific responses included reduced metabolic and biosynthetic expression in commensals, including *Coprococcus aceti* and *Streptococcus thermophilus,* alongside increased stress-adaptation, mucin-degradation, and metal-homeostasis functions in other taxa. Cross-cohort classification performance varied substantially between species. These findings establish pangenome-resolved metatranscriptomics as a framework for identifying lineage-specific functional responses within microbiomes.

Introduction

Parkinson's disease (PD) is linked to gastrointestinal dysfunction and gut microbiome alterations. Constipation often precedes motor symptom onset by years, and studies consistently report shifts in microbial composition in PD[1–7], including depletion of short-chain fatty acid (SCFA) producing commensals and enrichment in mucin-associated or opportunistic taxa[8,9]. These observations suggest that microbial functional activity may provide information beyond community composition.

Metatranscriptomics provides a direct readout of microbial gene expression in vivo[8,10] and it has uncovered altered pathways related to organismal motility, bile acid metabolism, glutamate metabolism, and inflammatory interfaces in PD[8,9]. However, conventional analyses quantify transcriptional changes at the community level by collapsing functionally annotated genes, e.g., using KEGG orthology, across taxa. Genes with the same functional annotation but originating from different species are therefore treated as equivalent despite distinct regulation, ecological roles, evolutionary contexts and host interactions, obscuring which microorganisms express disease-associated functions.

Microbial communities, like complex multicellular tissues, consist of heterogeneous cell populations in which signals from individual species are obscured when analyzed in aggregate. In host transcriptomics, deconvolution and pseudo-bulk approaches have transformed the interpretation of complex tissues[11,12] by enabling cell-type-specific analyses from bulk samples. In brief, transcriptomic deconvolution estimates the abundance or expression contribution of different cell types from bulk RNA-sequencing data. Pseudo-bulk approaches aggregate single-cell transcriptomic profiles belonging to the same cell type within a sample to generate cell-type-specific, bulk-like expression profiles. Equivalent strategies remain underdeveloped in microbiome research because microbial communities comprise hundreds to thousands of genetically diverse species with extensive strain-level variation, making species-resolved functional analyses challenging.

To overcome this gap, we developed a taxonomic deconvolution framework that partitions community-wide transcriptional profiles into species-specific signals. By assigning transcripts to reconstructed microbial genomes and aggregating them at the species level within each sample, this approach generates species-level pseudo-bulk transcriptomes, enabling transcriptional responses to be analysed at high taxonomic resolution within complex microbial communities.

Metagenome-assembled genomes (MAGs) yield species-resolved reconstructed genomes directly from metagenomic sequencing data. When integrated with metatranscriptomics, MAGs enable assignment of expressed genes to individual microbial genomes and, thereby, the construction of species-level pseudo-bulk transcriptomes. However, most existing integration approaches rely on reference genomes or dereplicated MAG collections[13], and therefore primarily capture only the conserved core genome. However, in microorganisms, the gene pool is constantly in flux, and a wide range of accessory genes complement the genome[14]. This rapidly evolving accessory gene pool provides an additional instrument of evolutionary flexibility in prokaryotes and serves as a source for adaptation to different environmental conditions, and, importantly, comprises genes linked to pathogenicity and virulence[15].

Together, the core and accessory genomes constitute the pangenome - the complete genetic repertoire of a species[16]. Besides its biological importance and clinical relevance, pangenome-level resolution has rarely been incorporated into metatranscriptomic studies.
We developed a pangenome-resolved taxonomic deconvolution framework that integrates paired metagenomic and metatranscriptomic data to generate species-level pseudo-bulk transcriptomes. We applied this framework to two independent PD cohorts to reveal microbial functional responses hidden in conventional community-level analyses. Specifically, we investigated whether species-resolved pangenome analysis increases functional resolution; identifies lineage-specific transcriptional programs associated with disease status, duration, and symptom severity; and provides new biological insights into microbial functional dysbiosis in PD.

Results

Taxonomic deconvolution framework for gut metatranscriptomes

We first developed a workflow allowing integrated analyses of paired metagenomic (MG) and metatranscriptomic (MT) sequencing data using co-assembly, genome binning, MAG quality control, taxonomic annotation, and transcript remapping to generate species-level deconvoluted expression profiles (Fig. 1). As an input, this framework uses the output of an in-house version of the IMP3 pipeline[17], which governs all preprocessing steps up to generation and initial quality control of MAGs. Specifically, paired MG and MT data from each sample were subjected to hybrid iterative co-assembly into contigs, followed by mapping of the corresponding MG and MT reads back to the assembled contigs (Figure 1A). The resulting contigs were then binned into MAGs using five independent binning algorithms. MAGs were subsequently refined with two metabinners, quality assessed, and taxonomically annotated (Figure 1B).
Critically, unlike conventional bulk MT analyses, where identical functional annotations from multiple taxa are collapsed into single community-level features, our framework retains species context by assigning MT reads to MAGs. Genes with shared annotations within species are subsequently merged into pangenome-resolved features, yielding per-sample species-resolved transcriptomes analogous to cell type-resolved pseudo-bulk profiles used in host single-cell transcriptomics (Figure 1C). This strategy transforms a conventional N × F feature matrix into an expanded (N × M) × F design, where N represents samples, M species, and F features, thereby preserving lineage specificity (Figure 1C).

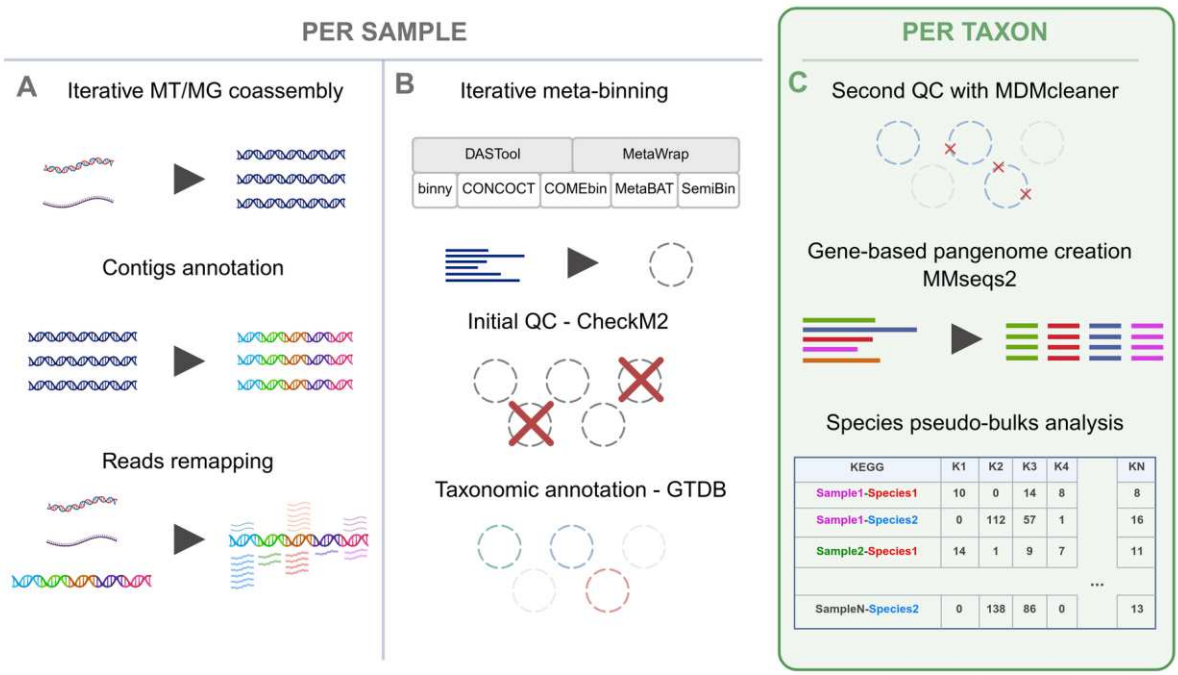


Figure 1. Taxonomic deconvolution framework. Transcriptomic deconvolution partitions metatranscriptomic reads according to reconstructed microbial genomes to generate species-level pseudo-bulk transcriptomes. (A) Metagenomic (MG) and metatranscriptomic (MT) reads are assembled into contigs, annotated, and mapped independently for each sample. (B) Metagenome-assembled genomes (MAGs) are reconstructed by genome binning and subjected to an initial quality assessment. (C) MAGs sharing the same taxonomic assignment across samples are further processed by removing contaminating contigs using MDMcleaner before being combined into species-level gene-based pangenomes by clustering genes into pangenome-resolved features based on sequence similarity. Finally, the read mapping information from the initial mapping step (A) is used to aggregate transcripts assigned to the same microbial species within each sample, generating species-level pseudo-bulk transcriptomes for downstream differential expression analyses.

Cohort integration and generation of species-level pseudo-bulk transcriptomes

We applied our newly developed framework to paired MG and MT data from two independent PD microbiome cohorts, MiBiPa[8] (n = 122) and ExpoBiome[18] (n = 180). After participant selection, the combined dataset consisted of 160 individuals, including 66 individuals from the MiBiPa cohort (39 PD, 27 controls) and 94 individuals from the ExpoBiome cohort (49 PD, 45 controls). We balanced the cohorts for diagnosis distribution and included only individuals with coherent metadata available, including sex, age, disease duration, MDS-UPDRS questionnaire parts I - III results, constipation status, and medical treatment information (Supplementary Table 1).

Following MAG reconstruction, quality control, and metadata filtering, we retained an initial set of 9,895 high-quality MAGs (Completeness ≥ 80.0%, Contamination < 5.0%) representing 1,213 different microbial species (Supplementary Table 2). Overall, 55% of the MG and 45% of the MT data could be assigned to these MAGs. We subsequently generated 7,548 high-quality species pseudo-bulk transcriptome profiles representing the transcriptional activity across the pangenomes of the 215 prevalent microbial species Across these profiles, we quantified 17,491,502 unique genes, which were collapsed into 2,406,459

within-species pangenome-resolved features, and functionally annotated them with major annotation systems, including KEGG (12,925 unique KO terms), PFAM (19,760 unique domains), and COG (14,605 unique groups). Together, these data constitute a species-resolved representation of microbial transcriptional activity that captures both core and accessory genomic functions across PD and control microbiomes.

A key assumption of the pseudo-bulk framework is that transcriptional profiles primarily reflect the biological identity of microbial species rather than the sample from which they originate. To evaluate this, we embedded pseudo-bulk transcriptomes using Uniform Manifold Approximation and Projection (UMAP) based on KEGG, PFAM, and COG annotations. Species-resolved pseudo-bulk transcriptomes consistently clustered according to their taxonomic identity (Figure 2A-2C), whereas neither disease status (Figure 2D-2F) nor sample origin (Figure 2G-2I) explained the global structure of the embeddings. These clustering patterns remained robust across multiple UMAP hyperparameter settings (Supplementary Document). Notably, archaeal pseudo-bulk transcriptomes formed a distinct cluster separate from bacterial pseudo-bulk species, further supporting that the pangenome-resolved deconvolution framework preserves biologically meaningful taxonomic relationships while enabling species-level transcriptional analyses (Figure 2A-C).

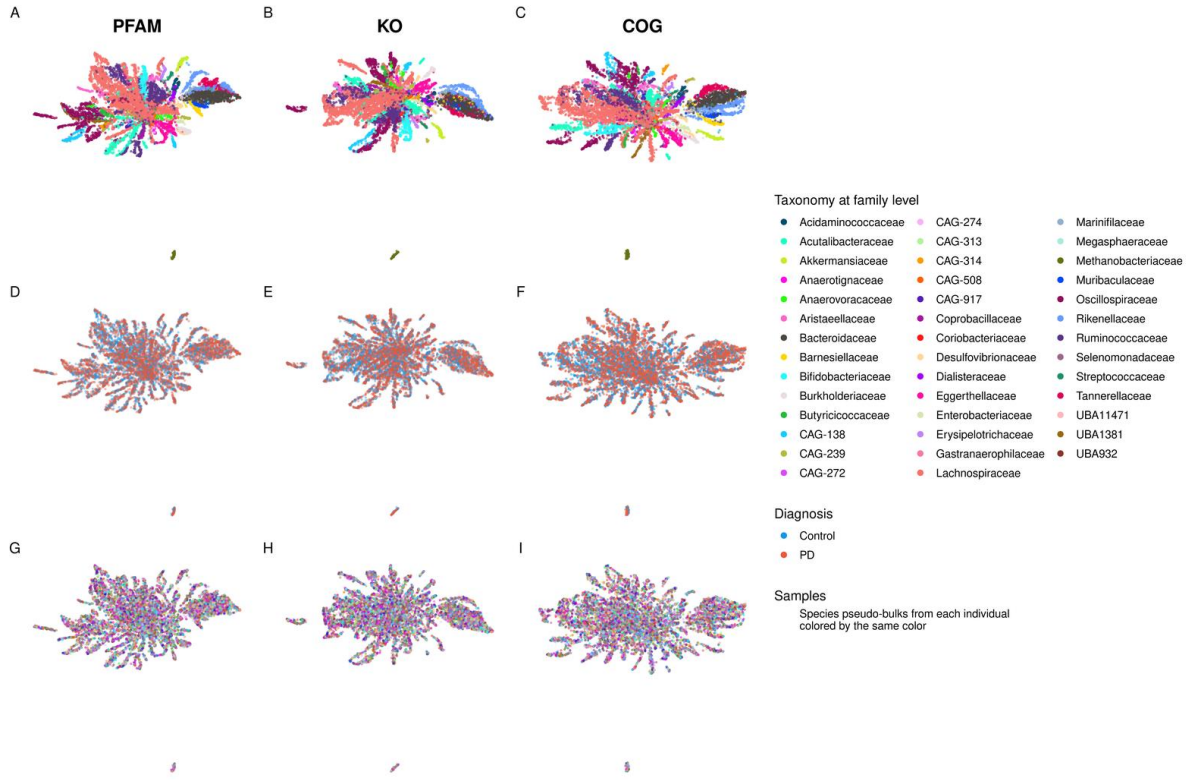


Figure 2. Species-resolved pseudo-bulk transcriptomes organization across taxonomic, functional, and donor annotations. Each point represents one species-level pseudo-bulk transcriptome. UMAP embeddings were generated from transcript counts aggregated at PFAM (A, D, G), KEGG (B, E, H), and COG (C, F, I) annotation levels. Points are colored according to microbial family level (A-C), diagnosis (D-F), or donor (G-I). Across all annotation systems, pseudo-bulk transcriptomes cluster predominantly according to taxonomic identity rather than disease status or sample of origin, demonstrating that the taxonomic deconvolution framework preserves biologically meaningful species-specific transcriptional profiles.

Parkinson's disease is associated with reduced transcriptional breadth and increased functional heterogeneity across microbial pangenomes

Before analysing disease-associated transcriptional changes, we first characterized the reconstructed species pangenomes and evaluated how pangenome size influenced the recovery of species pseudo-bulk transcriptomes and the number of detectable genes across taxa. The size of each species' pangenome (total number of non-redundant genes) strongly correlated with the number of species-resolved pseudo-bulk transcriptomes recovered (R^2= 0.68, p = 2.2×10^{-16}, Figure 3A), indicating that prevalent species with larger pangenomes were consistently represented across samples. For the average number of genes identified per species, the correlation with the number of species-resolved pseudo-bulk transcriptomes was weaker (R^2= 0.03, p = 4.37×10^{-3}, Figure 3B), representing lineage-specific genome size trends. Six bacterial species, mostly of the *Bacteroides* genus, including *B. cellulosilyticus*, *B. fragilis*, *B. ovatus*, *B. thetaiotaomicron*, *Phocaeicola dorei* (previously *B. dorei*), and *Escherichia coli,* demonstrated genome sizes above the average, consistent with the genomic plasticity and large pangenome sizes reported for these taxa (Figure 3C). No significant differences in the proportions of core, accessory, and cloud genes were observed between PD and controls.

We next investigated whether the disease status influences the breadth of transcriptional activity across the microbial genomes identified in individual samples. Transcriptional breadth was defined as the proportion of genes with non-zero expression among all genes in a given species' pseudo-bulk genome or, respectively, across the total pangenome of that species. We applied a Bayesian beta-regression modelling framework accounting for diagnosis, dataset, demographic, and technical covariates, as well as species- and sample-level random effects, for both the mean (μ) and precision (ϕ) parameters of the beta distribution. Genome-level modelling revealed a reduction in genome-wide transcriptional breadth in PD (μ: β=-0.14, 95% CrI -0.25 to -0.02) together with lower precision (ϕ: β=-0.12, 95% CrI -0.22 to -0.02). Thus, PD microbiomes not only expressed a smaller fraction of each genome but also exhibited greater between-sample heterogeneity in the proportion of genes expressed (Figure 3D). Species-specific slope analyses identified 54 taxa with significantly decreased transcriptional breadth, including known species depleted in PD, such as R*oseburia intestinalis*, *R. inulinivorans*, *Agathobacter rectalis*, *Coprococcus aceti*, *Butyribacter intestini*, *Bifidobacterium adolescentis* and *B. longum* (Supplementary Figure 1). In addition, *Eubacterium F sp. 003491505* demonstrated increased gene expression heterogeneity, suggesting niche specialisation in PD.

We then extended this analysis from individual genomes to species pangenomes. On a pangenome level, the mean PD effect was weaker and only marginal, with a negative shift (μ: β=-0.04, 95% CrI -0.09 to 0.00); however, PD showed a stronger association with reduced precision (ϕ: β=-0.33, 95% CrI -0.46 to -0.19), indicating significantly increased heterogeneity across samples (Figure 3E). Consistent with the overall model estimates, the species-specific slopes indicated a reduction in mean transcriptional breadth for only 8 taxa (Supplementary Figure 2), mostly known PD-depleted bacterial species and SCFA producers. At the same time, four taxa showed a positive association (decreased variability of gene expression in PD), and 29 taxa demonstrated a negative association with precision, i.e., increased variability of gene expression in PD (Supplementary Figure 2). Notably, PD-depleted taxa also tended to show lower mean pangenome expression, suggesting not only reduced average functional activity but also a loss of transcriptional breadth across individuals.

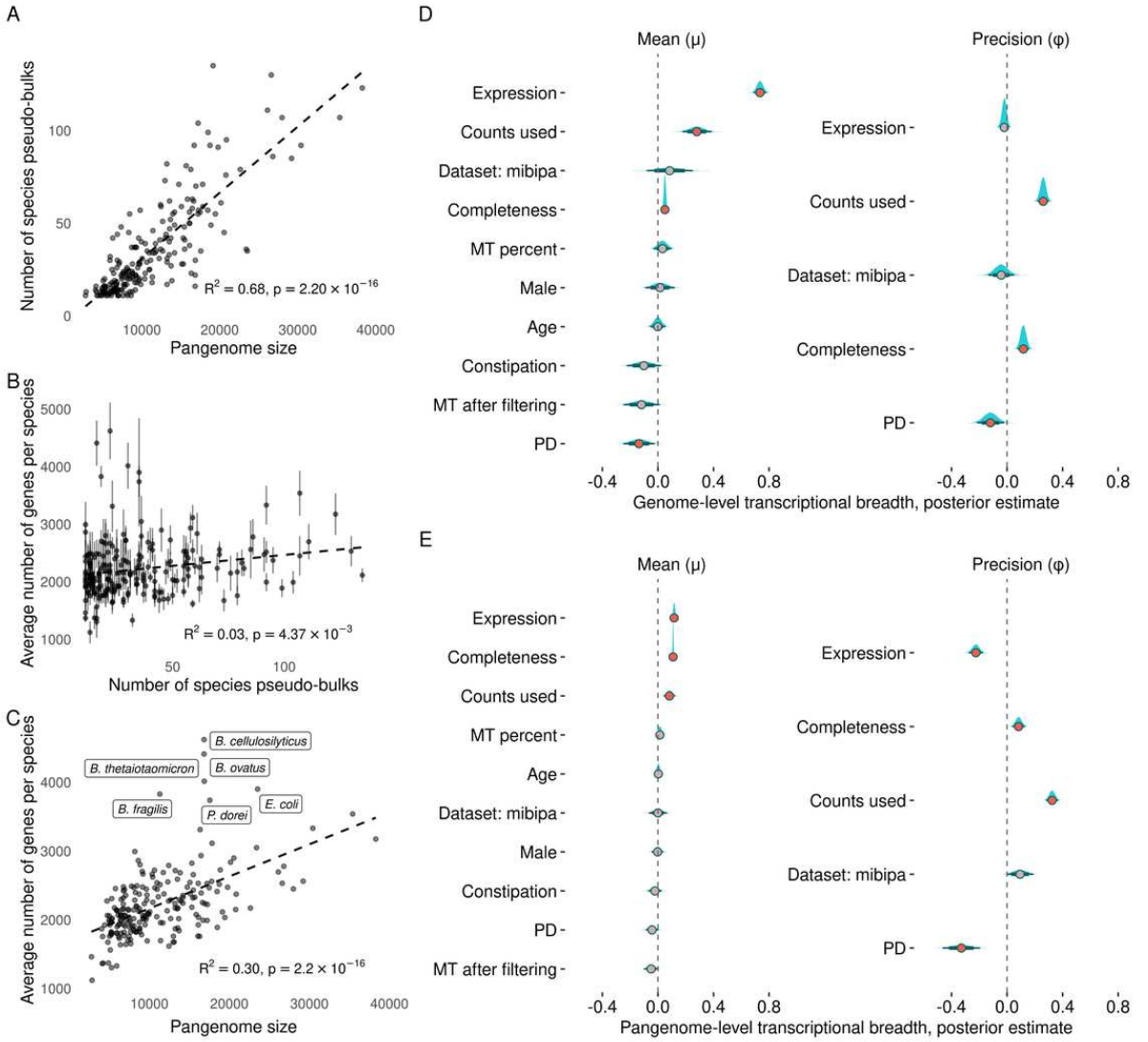


Figure 3. Pangenome characteristics and transcriptional breadth. (A) Relationship between species pangenome size and the number of recovered species pseudo-bulk transcriptomes. (B) Relationship between the average number of expressed genes per species pseudo-bulk transcriptome and the number of recovered species pseudo-bulk transcriptomes. Whisker plots show the distribution of expressed genes per pseudo-bulk transcriptome for each species. (C) Relationship between the average number of expressed genes per species pseudo-bulk transcriptome and species pangenome sizes. Species that deviated from the general tendency are signed. (D-E) Bayesian beta-regression estimates for the mean (μ) and precision (ϕ) of genome- and pangenome-level transcriptional breadth. Dots represent posterior estimates of parameters. Parameters colored in red have credible intervals excluding zero. Posterior densities are marked in blue. Models included diagnosis, cohort, demographic, technical, genome quality, and sequencing covariates together with species- and sample-level random effects.

Strain-level genomic diversity is associated with Parkinson's disease status and clinical severity

We next examined whether within-species genomic variation contributed to PD-associated heterogeneity. Among species present in at least 30 samples, 2,405 strains were identified across 95 species using 99% average nucleotide identity (ANI), with extensive variation in

strain diversity between taxa (Supplementary Fig. 3). Because many species displayed highly uneven strain distributions, conventional strain-based association analyses of disease status and severity lacked statistical power. To address this, we developed an ANI distance-based approach that quantifies within-species genomic variation using pairwise distances between genomes. The utilisation of distance allows the use of an established framework for β-diversity analysis, including PERMANOVA, while adjusting for demographic, clinical, and technical covariates.

Across species, disease status, disease duration, and all clinical severity measures explained comparable proportions of genomic variation (mean $R^2 \approx 2–3\%$), with motor symptom severity (UDPRS II and III) showing the largest average effects (Figure 4A). Biological covariates, such as sex, age, and constipation, explained a comparable proportion of variance in the disease status (Figure 4A). At nominal $P < 0.05$, we detected 33 exploratory species-variable associations involving 27 species (Figure 4B), including *Akkermansia muciniphila* with constipation and MDS-UPDRS II/III, and *Dorea formicigenerans* and *Eubacterium_F sp003491505* with disease status and constipation. Genomic dispersion also differed nominally between PD and controls for *Eubacterium_F sp003491505*, *Dorea_A longicatena,* and *Alistipes communis* (Figures 4C–E). Additionally, some species showed pronounced genome clustering, as illustrated by *Akkermansia muciniphila* (Figure 4F). This clustering was not explained by the measured covariates, including cohort origin, and may reflect latent population structure within these microbial species.

Because genetic variation is a major determinant of gene expression, as evident from large-scale expression quantitative trait loci analyses[19], we next asked whether within-species genomic diversity predicts transcriptional diversity. To test whether genomic diversity corresponded to transcriptional organization, we compared within-species genomic and transcriptomic distance matrices using Mantel tests. Significant associations were observed for 61 of 95 species (FDR<0.05; mean R=0.340, range 0.156–0.604; Supplementary Fig. 4), including PD-associated *Evtepia gabavorous*[20] (R=0.589), *A. muciniphila* (R=0.580), *PeH17 sp000435055* (R=0.408) and *Methanocatella* (*Methanobrevibacter*) *smithii* (R=0.531). These findings link strain-level genomic architecture to microbial expression states and support the use of cohort-derived genomic references for functional MT analysis.

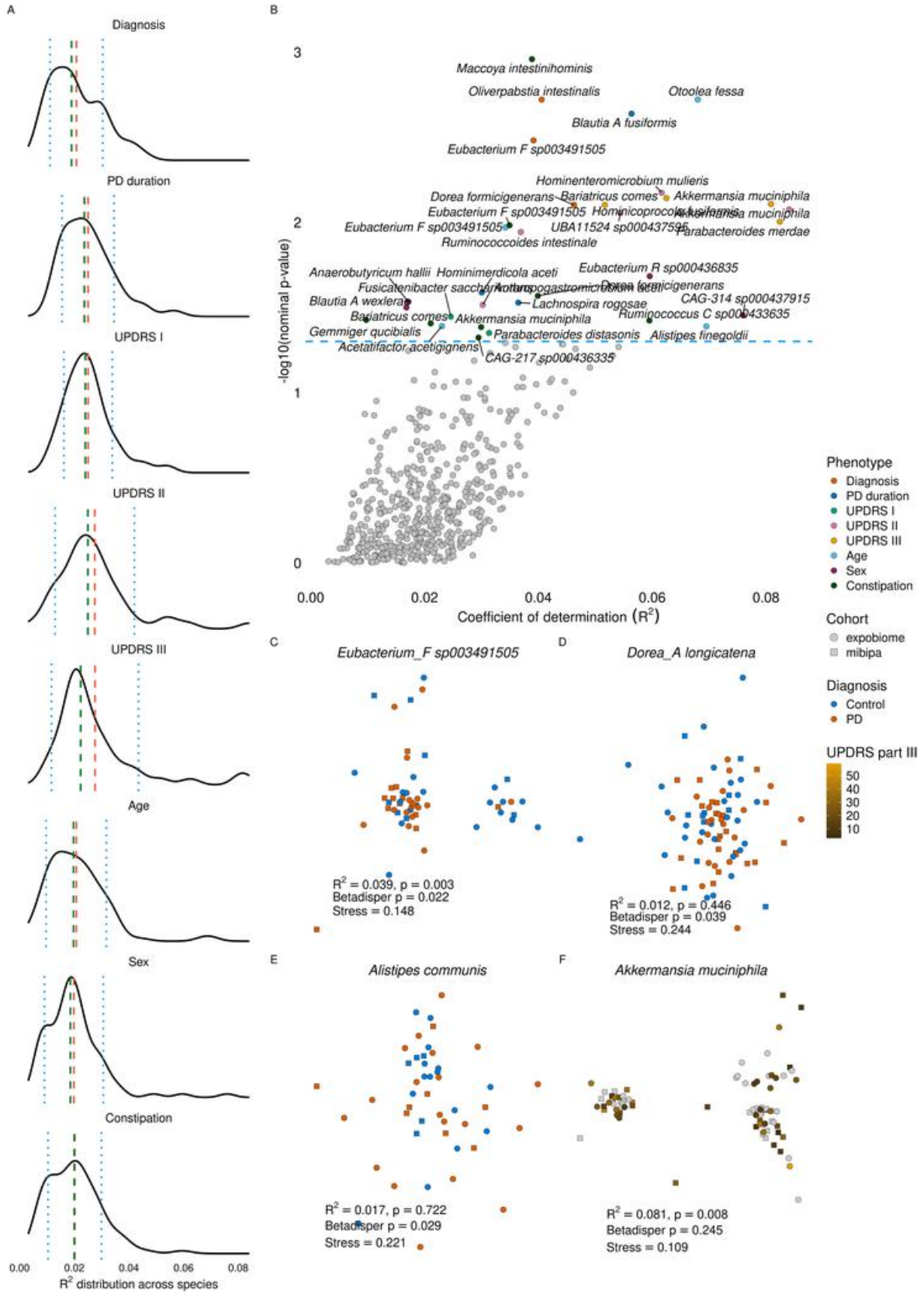


Figure 4. Strain-level metagenomic variation in Parkinson's disease. (A) Distribution of PERMANOVA R^2 values quantifying the proportion of within-species genomic variation explained by different clinical covariates. The red dashed line represents the mean, the green dashed line

represents the median, and the blue dotted lines represent the standard deviation. (B) Summary of associations between within-species genomic dissimilarity variation explained by clinical covariates. Each point represents one species–covariate association, colored according to the tested covariate type. Grey dots represent non-significant associations($P \geq 0.05$), and the blue dashed line represents the nominal significance ($P=0.05$). (C–F) Two-dimensional non-metric multidimensional scaling (nMDS) of ANI-derived within-species genomic distance matrices illustrating representative strain-level genomic variation. Each dot represents one reconstructed genome (MAG) of a species. Panels C–E show examples of species with significant differences in genomic dispersion between PD and HC groups, whereas Panel F shows genomic clustering in *Akkermansia muciniphila*. In Panel F, individuals without measured UPDRS Part III scores are colored in grey.

Per-species pan-transcriptomic signatures of the gut microbiome in Parkinson's disease

To identify disease-associated microbial functions, we modeled species pseudo-bulk transcript counts using negative binomial regression, accounting for biological and technical covariates. Analyses were performed for genes and KEGG, PFAM, and COG annotations in species detected in at least 30 individuals; disease duration and MDS-UPDRS analyses were restricted to individuals with PD.

Across phenotypes, 2,271 pangenome-resolved features were significant after multiple-testing correction ($q < 0.05$, Supplementary Table 3). Among these, 535 functions across 67 microbial species were associated with disease status, and 479 functions across 38 species were associated with disease duration (Figure 5A). We also detected 611 functions from 38 species associated with non-motor symptoms of PD assessed based on the UPDRS scale part I, 293 functions from 33 species associated with motor symptoms (UPDRS scale part II), and 353 functions from 36 species associated with motor examination scores (UPDRS scale part III; Figure 5A). Importantly, among 353 associations identified for coding gene-level pangenome-resolved features, 139 (39%) were identified exclusively in the accessory gene pool and would thus be missed using a standard single-reference-based approach (Supplementary Table 3). Disease status was enriched for increased expression, whereas greater non-motor burden was predominantly associated with reduced expression and greater motor impairment more often with increased expression; duration showed no consistent direction.

To assess whether different clinical phenotypes were associated with shared microbial transcriptional programs, we correlated signed $-\log_{10}$ p-values across pan-transcriptomic features. The strongest concordance was observed between motor symptoms and results of motor examination ($R = 0.56$, $p < 2e\text{-}16$), indicating shared changes in the gut metatranscriptome between patient-reported motor disabilities and examiner-rated motor symptoms. A moderate overlap was detected between non-motor symptom burden and motor experiences of daily living, followed by non-motor symptoms and motor symptoms ($R = 0.29$, $p < 2e\text{-}16$). This may reflect shared systemic or gut-brain-axis processes, including autonomic dysfunction, constipation, inflammation, or overall clinical burden, which can influence both non-motor symptoms and perceived motor disability. The correlation in the pan-transcriptomic response between results of the motor examinations and the non-motor symptoms severity was

low ($R = 0.06$, $p < 2e\text{-}16$). Disease duration showed only weak similarity to transcriptional signatures with non-motor symptoms ($R = 0.17$, $p < 2e\text{-}16$) and almost no overlap with both UPDRS parts II and III (UPDRS II, $R = 0.05$, $p < 2e\text{-}16$; UPDRS III, $R = 0.09$, $p < 2e\text{-}16$), suggesting that the gut microbial transcriptional changes relate more strongly to current clinical manifestations than to time since diagnosis.

We identified several pan-transcriptomic signatures of the gut microbiome in response to PD, its duration, and severity, described in detail in Supplementary notes. Specifically, in response to PD, microorganisms demonstrate two major transcriptomic trends (Figure 5B, Supplementary Table 3). Some taxa, particularly gut commensals such as *Coprococcus aceti* and *Streptococcus thermophilus*, exhibited transcriptional signatures of reduced cellular fitness, including downregulation of genes involved in metabolism, protein synthesis, DNA repair, cell-wall biogenesis, and methylglyoxal defence, together with increased expression of oxidative stress-related genes. Another group of microorganisms, exemplified by *Phocaeicola massiliensis*, *Hominimerdicola aceti*, and *Akkermansia muciniphila*, in contrast, showed stress-adaptation responses, including increased acid-stress resistance, antimicrobial resistance, and cell-envelope remodelling. Several taxa also showed increased mucin-degrading activity, including *Akkermansia muciniphila*, *Phocaeicola massiliensis*, and *Ruminococcoides intestinale*.

Members of the Lachnospiraceae family demonstrated the most consistent transcriptomic response to PD severity and duration (Figure 6A, Supplementary Table 3). The strongest effect of disease duration was identified in *Lachnospira rogosae*, which showed increased stress adaptation and DNA maintenance, alongside reduced vitamin, cofactor, and amino acid biosynthesis. It also exhibited altered carbohydrate and acetoin metabolism, membrane remodeling, and changes in c-di-GMP, flagellar, and quorum-sensing pathways, suggesting a shift in motility–biofilm behavior during disease progression.

With increasing PD severity, particularly greater non-motor symptom burden, *Lachnospira eligens_A* showed broad suppression of genes involved in transcription, translation, energy metabolism, and B-vitamin and amino-acid biosynthesis, consistent with reduced microbial metabolic activity. Greater motor impairment was additionally associated with decreased expression of acetate-related and cell-wall functions, together with alterations in lipid metabolism and chemotaxis pathways, further suggesting a loss of potentially beneficial *Lachnospira* activity in more severe disease. In contrast, another Lachnospiraceae species, *Hominisplanchenecus faecis*, exhibited increased transcriptional activity in nitrogen and amino-acid metabolism, particularly pathways involving glutamate, glutamine, tryptophan, and arginine. Together, these opposing transcriptional patterns suggest functional divergence and metabolic remodelling within Lachnospiraceae as motor and non-motor symptoms worsen.

In addition to that, in PD, multiple gut microbes showed coordinated changes in metal sensing, transport, and detoxification, with *Otoolea fessa* showing the strongest metal-responsive signature. Similar increases in iron, copper, cobalt, manganese, zinc, and toxic-metal handling pathways across PD status, severity, and duration suggest altered metal availability or metal-associated stress in the gut environment.

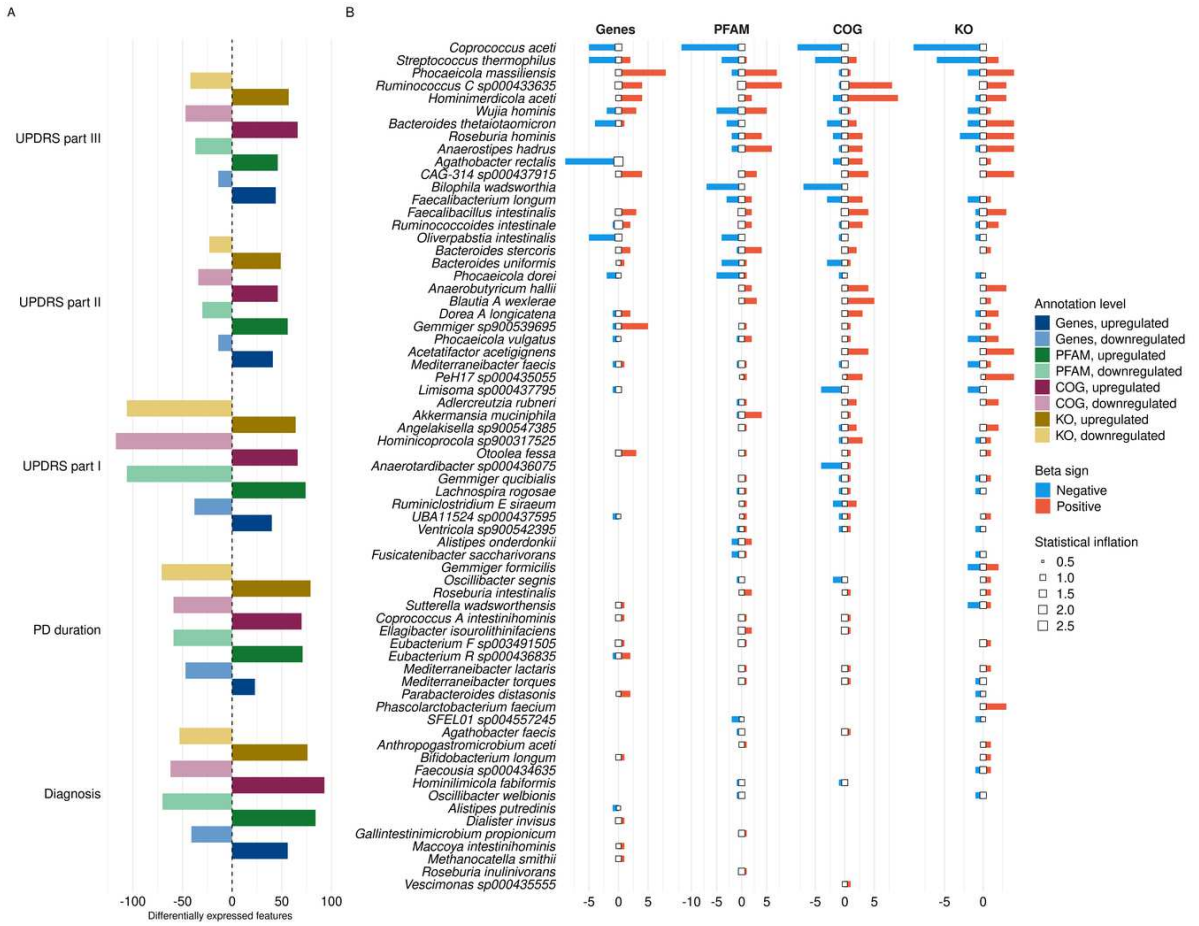


Figure 5. Species-resolved pangenome features associated with Parkinson's disease and clinical phenotypes. (A) Number of significant differentially expressed pangenome-resolved features with PD status, disease duration, and clinical severity (MDS-UDPRS I-III), stratified by functional annotation level (genes, KEGG, PFAM, and COG). (B) Species-specific distribution of pangenome-resolved features associated with PD status. Bars show the total number of significantly upregulated (red) and downregulated (blue) features for each microbial species across all annotation levels. Only features significant after correction for multiple tests (FDR<0.05) are shown.

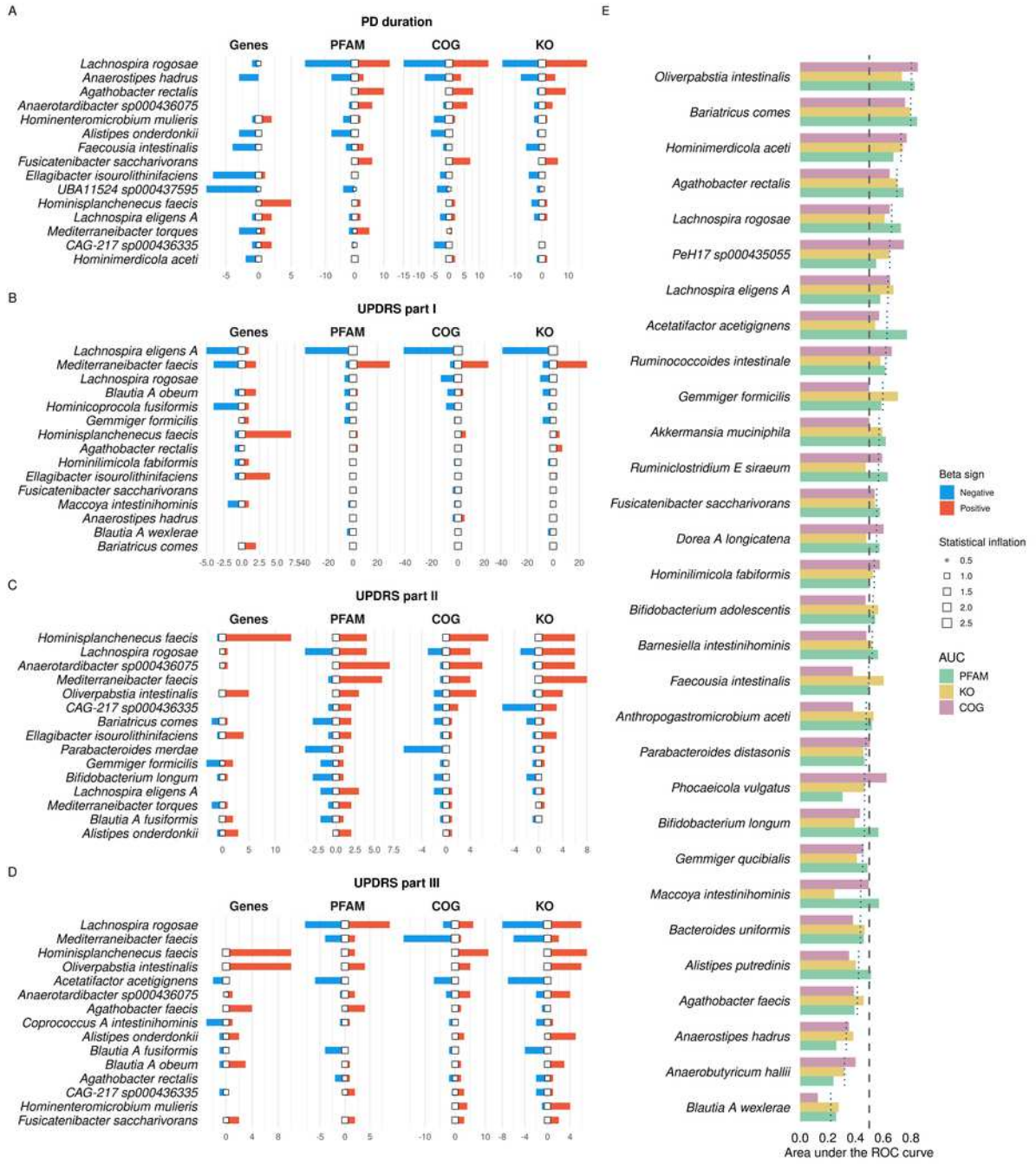


Figure 6. Species-specific pangenome-resolved features associated with disease duration and clinical severity. (A-D) Number of significantly upregulated (red) and downregulated (blue) pangenome-resolved features with disease duration (A), UPDRS part I (B), UPDRS part II (C), and UPDRS part III (D), split by microbial taxa and functional annotation level. Species are ranked according to the total number of significant associated features summed across all annotation levels. Only features significant after multiple testing correction ($q<0.05$) and only the top 15 species for each phenotype are shown. (E) Species-specific classification performance in the independent validation cohort. Dotted lines represent the per-species area under the receiver operating characteristic (ROC) curve (AUC), averaged across functional annotation levels. The dashed line represents random classification performance (AUC = 0.5).

Species-resolved transcriptional signatures generalize across independent Parkinson's disease cohorts

Next, we evaluated the ability of species-resolved transcriptional signatures to distinguish PD from controls in an independent validation dataset. Given the available sample size, our objective was not to develop a comprehensive diagnostic classifier but to assess the generalizability of species-specific transcriptional programs. Species present in at least 60 individuals were modelled separately using boosted decision-tree classifiers trained in ExpoBiome and evaluated while minimizing information leakage in MiBiPa. Ninety models were generated from KO, PFAM, and COG features. Cross-cohort performance was assessed using the area under the receiver operating characteristic curve (AUC).
For PFAM-based models, the AUC range was 0.24 - 0.85 across species, while KO- and COG-based models demonstrated AUC ranges of 0.25 - 0.80 and 0.13 - 0.85, respectively (Figure 6E, Supplementary Table 4). When averaged across annotation types within each species, the AUC range was 0.22 - 0.80. In total, 17 species achieved an average AUC greater than 0.5, including nine species with AUCs above 0.6, and three species with AUCs above 0.7, indicating that a subset of microbial species exhibits transcriptional signatures that reproducibly distinguish PD across independent cohorts.

Discussion

Here, we developed a species-resolved, pangenome-based metatranscriptomic framework that preserves taxonomic context while capturing within-species functional diversity. Applying this approach to PD revealed transcriptional alterations extending beyond changes in microbial composition. Notably, 39% of statistically significant gene-level associations involved accessory genes, highlighting functional information potentially missed by analyses restricted to a single representative genome. Furthermore, genomic and transcriptomic variation were significantly associated in 61 of 95 species, indicating that within-species genomic diversity is closely linked to microbial transcriptional variation. Together, these findings demonstrate the value of incorporating cohort-derived pangenomes into functional microbiome analyses.
Our study shows that PD is associated both with reduced transcriptional breadth within individual species genomes and with increasingly heterogeneous recruitment of accessory genomic functions across individuals. At the pangenome level, increased heterogeneity was more pronounced than the reduction in mean transcriptional breadth, suggesting that disease-associated differences involve variable expression of species' broader functional repertoires rather than a uniform loss of transcriptional activity. Such variation may reflect real differences in strain composition, accessory gene content or responses to the intestinal environment. These findings provide a functional perspective on the substantial inter-individual variability observed in PD microbiome studies and emphasize that taxonomically similar communities need not exhibit equivalent transcriptional states.
PD is associated with reduced gut SCFA concentrations[21] and depletion of SCFA-producing commensals such as *Coprococcus* spp.[22,23]. This decrease, in principle, could result either from reduced SCFA production by resident bacteria or from loss of SCFA-producing taxa. Our species-resolved pantranscriptomic data from *Coprococcus aceti* and *Streptococcus thermophilus* support the latter explanation: both species showed broad reductions in metabolic activity and cellular fitness, while expression of SCFA-production pathways remained largely unchanged. These findings suggest that reduced SCFA availability in PD may be driven more by depletion of SCFA-producing commensals than by suppression of SCFA biosynthesis within surviving populations.

The taxonomic resolution also revealed changes in functions potentially involved in interactions with the intestinal environment. Increased expression of mucin- and glycan-associated features was observed in several species, extending previous reports of enrichments in mucolytic taxa in PD[6,8,23]. Together with altered cell-envelope- and LPS-associated functions, these observations suggest remodeling of microbial processes relevant to the host interface. Such changes could influence interactions with the intestinal mucus barrier and mucosal immune system[24] and extend previous MG-based observations of increased abundance of microbial envelope-related pathways in PD[23] although their physiological consequences cannot be inferred directly from transcriptional data.

Particularly pronounced phenotype-associated remodeling occurred within members of the Lachnospiraceae family shown to be depleted in multiple studies[6,25]. Increasing disease duration was associated with stress adaptation and cellular maintenance in *Lachnospira rogosae*, accompanied by reduced biosynthetic expression and altered carbohydrate and membrane-related functions. In contrast, greater non-motor symptom burden was associated with broad transcriptional suppression in *Lachnospira eligens_A*, whereas *Hominisplanchenecus faecis* showed increased expression of nitrogen and amino-acid-related functions with greater symptom severity. These divergent patterns indicate that related microbial species exhibit distinct transcriptional associations with clinical phenotypes.

Additional species-specific signatures involved microbial carbonyl metabolism and environmental stress responses. Reduced expression of a bacterial DJ-1 family glyoxalase in *Streptococcus thermophilus* and increased expression of acetoin-utilization transport components in *PeH17 sp000435055* suggest altered microbial metabolic responses[26–31]. However, the relationship between bacterial DJ-1 homologues, host PARK7 biology, and acetoin availability remains unresolved, raising the hypothesis that altered DJ-1-associated functions may be linked to increased acetoin production and subsequent changes in the PD gut microbiome (Supplementary Notes). Similarly, changes in microbial metal sensing, transport, and detoxification suggest altered responses to metal availability or metal-associated stress. Because metal concentrations and exposure histories were not measured, these signatures cannot establish increased environmental exposure, but they provide a plausible hypothesis linking heavy-metal exposure, gut microbial physiology, and disease progression. Together, these observations motivate future studies integrating microbial transcriptional profiles with metabolite measurements and exposome data.

Several studies have constructed and evaluated machine-learning classifiers for PD based either on taxonomic or functional compositions of gut microbiomes[7,32–35]. The taxonomic deconvolution coupled with pantranscriptomic resolution of gene expression allowed us to go deeper and, for the first time, provide per-species prediction potential for PD, while the use of two independent cohorts increased the robustness of our estimates. Species showing the strongest cross-cohort performance were taxa previously implicated in PD both in the literature and the present study. *Oliverpabstia intestinalis*, which achieved the highest mean AUC of 0.80, was associated with both disease status and motor symptom severity in our analyses. This species has previously been reported to be enriched in the gut microbiome of individuals with PD, and it has been negatively correlated with caffeine metabolism, a pathway known to be lower in PD[36]. *Bariatricus* (*Coprococcus*) *comes*, which also achieved a mean AUC of 0.80, was associated with both motor and non-motor symptom severity. Both the genera *Coprococcus* and the species *Bariatricus comes* have been reported to be depleted in PD[2,6,37]. *Agathobacter rectalis* was associated with disease status, disease duration, and both motor and non-motor symptom severity. This species has been shown to attenuate microglia-mediated neuroinflammation in an Alzheimer's disease model[38] and has been reported to be depleted in PD[9,39].

In contrast, several species, including *Blautia_A wexlerae*, *Anaerobutyricum hallii*, and *Anaerostipes hadrus*, demonstrated mean AUC values of 0.22, 0.32, and 0.33, respectively. AUC values substantially below 0.5 indicate class inversion. In cross-cohort validation, reversal of the direction of discrimination relative to the training cohort may indicate poor feature transportability and cohort-dependent effects and was previously shown in PD[7,40,41]. Prioritizing taxa with reproducible, directionally concordant associations may reduce cohort-specific effects and improve the robustness and generalizability of microbiome-based PD classifiers.

Several limitations of our study should be mentioned. Microbial pangenomes and pan-transcriptomes were reconstructed from short-read metagenomic and metatranscriptomic data without isolate-based sequencing; as a result, the true genetic diversity of the microbiome may be underestimated. To mitigate this and to maximise reliability, we included only high-quality MAGs in downstream analyses; however, this filtering resulted in information loss, as reads mapping to lower-quality MAGs, unbinned contigs, and reads that could not be assigned to assembled contigs were excluded. The extent of this information loss was quantified and is reported in the study. Variation in taxon prevalence resulted in unequal statistical power across species. Therefore, differential expression analyses and machine-learning models were more powered for prevalent taxa. This issue was particularly relevant for analyses of disease duration and symptom severity, for which the effective sample size was further constrained by the inclusion of only individuals with PD in the analysis and led to the exclusion of such taxa as *Coprococcus aceti* and *Streptococcus thermophilus*, underrepresented in the case of PD. Larger cohorts will be required to improve statistical power for less prevalent taxa and to more robustly assess associations with disease, its duration and symptom severity.

Together, our findings show that PD-associated microbiome alterations involve heterogeneous, lineage-specific transcriptional responses linked to within-species genomic diversity. By preserving species identity and incorporating accessory gene repertoires, pangenome-resolved metatranscriptomics provides a framework for identifying microbial functional changes that may be obscured by community-level or single-reference analyses. This approach offers a basis for investigating microbial functional heterogeneity across diseases and prioritizing species, strains and pathways for experimental validation.

Methods

Study design, cohorts, and participants' selection

The study was conducted on two independent cohorts of individuals with Parkinson's disease (PD) and healthy controls (HC). The first cohort, MiBiPa, included 46 PD, 49 HCs, and 27 patients with idiopathic REM sleep behavior disorder[8]. The second cohort, ExpoBiome, included 60 PD, 60 HCs, and 60 patients with rheumatoid arthritis. The study protocol is described in detail elsewhere[18]. In both cohorts, PD was diagnosed in accordance with the UK Brain Bank Diagnostic Criteria (UKBBC). From both cohorts, we included only individuals with PD and HC in the current study. Among them, we selected participants with paired gut metagenomic (MG) and metatranscriptomic (MT) data and available clinical data on age, sex, constipation status, PD duration, and disease severity. Severity was measured with the Movement Disorder Society – Unified PD Rating Scale (MDS-UPDRS) parts I (non-motor symptoms), II (motor symptoms), and III (motor examination)[42]. Further, we excluded four individuals from the ExpoBiome cohort due to diarrhea, and nine other

individuals: three from the ExpoBiome cohort and six from the MiBiPa cohort, with less than 15% of the MT reads retained after quality check (see Sequencing data processing and analysis). As a result, 94 individuals from the ExpoBiome cohort and 66 individuals from the MiBiPa cohort were studied herein, for a total of 160 participants.

Fecal sample collection

Fecal sample collection conformed to the Declaration of Helsinki and was approved by the ethics committee of the Physician's Board Hessen, Germany (FF 38/2016), the institutional review board of the Charité-Universitätsmedizin Berlin (EA1/204/19), and the Ethics Review Panel (ERP) of the University of Luxembourg (ERP 21-001 A ExpoBiome). Fecal samples were collected at the different clinical care units and immediately put on dry ice and stored at -80°C to guarantee optimal sample integrity and quality. The collected frozen stool samples were homogenized (6875D Freezer/Mill Spex - Instrument Solutions Benelux BV) and separated into 150 mg aliquots for subsequent biomolecular extractions. Each sample was weighed to allow relative corrections of the fecal mass quantifications.

Nucleic acids extraction

Nucleic acid extraction from fecal samples and sequencing procedure for the MiBiPa cohort are described elsewhere[8]. Samples for bio-macromolecular extractions were incubated in RNAlater-ice at -20°C overnight before RNA and DNA extractions. After overnight incubation on 1.5mL RNALater-ice, three cold autoclaved 4 mm stainless steel milling balls (Retsch) were added to each 150 mg stool aliquot, which were then homogenized by shaking for 2 min at 10 Hz in an Oscillating Mill (Geno/Grinder 2025, Spex - Instrument Solutions Benelux BV), followed by centrifugation at 700 g for 2 min at 4°C. The supernatant was collected into a 2 mL Eppendorf tube and was subsequently centrifuged at 14 000 g for 15 min at 4°C. After removal of the supernatant, cooled stainless steel milling balls (5×(2 mm) + 2×(5 mm)) together with 600 µl of cold RLT buffer (containing 10 µl/ml of 2-mercaptoethanol) were added to the resulting interphase pellet, before re-suspension of the pellet by quick vortexing. Bead beating was performed for 30 seconds at 25 Hz in cold racks (4°C; Oscillating Mill Geno/Grinder 2025). The resulting lysate was then transferred to a 96-deep well plate before completion of the pre-processing.
To the lysate, 1µl RiboGuard (Biosearch Technologies - LGC) was added before each sample was loaded onto a QIAshredder column and centrifuged for 2 min at 12 000 g. The flow-through was then processed via the AllPrep DNA/RNA/Protein mini kit (QIAGEN), following the manufacturer's instructions. First, the flow-through was added to the AllPrep DNA spin column (carefully, not to transfer any of the debris pelleted from the previous step), which was placed in a 2 ml collection tube (supplied) and then centrifuged for 2 min at 12 000 g. The flow-through was kept, and 400 µl of pure ethanol was added and mixed by pipetting. For each sample, 700 µl was loaded onto an RNeasy spin column, which was placed in a 2 ml collection tube (supplied), and centrifuged for 1 min at 12 000 g. The spin column membrane was washed by first adding 700 µl Buffer RW1 and centrifuging for 1 min at 12 000 g. Then, two washing steps of 500 µl Buffer RPE were added to the RNeasy spin column and centrifuged for 1 min at 12 000 g. To dry the column, a final centrifugation of 2 min at 12 000 g was used. The long centrifugation dries the spin column membrane, ensuring that no ethanol is carried over during RNA elution, as residual ethanol may interfere with downstream reactions. After centrifugation, the RNeasy spin column was carefully removed from the collection tube, avoiding contact with the flow-through. The collection tube containing the flow-through was discarded. To elute the RNA, the RNeasy spin column was placed in a new 1.5 ml collection

tube, 50 µl RNase-free water was added directly to the spin column membrane, and centrifuged for 1 min at 12 000 g. This step was repeated once to maximise RNA elution.
Purification of the genomic DNA was performed after full purification and collection of the RNA. In the interim, the DNA spin columns were placed in the fridge to prevent degradation but were removed and left at room temperature for at least 10 minutes before processing, to ensure all components would be processed at room temperature. The rest of the extraction procedure was performed according to the manufacturer's instructions: 500 µl of Buffer AW1 was added to each column before centrifugation for 1 min at 12 000 g. Next, 500 µl of Buffer AW2 was added to each column before centrifugation for 2 min (dry spin). After placing the spin column in a new 1.5 ml collection tube, 50 µl Buffer EB (preheated to 70°C) was directly added to the spin column membrane and incubated at room temperature (15–25°C) for 2 min, and then centrifuged for 1 min at 12 000 g to elute the DNA. This step was repeated once (including incubation) to maximise DNA elution.
Both DNA and RNA elutions were subjected to enzyme clean-up and concentration via either the DNA or RNA Clean-up & Concentration kit (Zymo Research), followed by either a DNase I or RNase A treatment, respectively. RNA elutions were DNase-treated by adding them to the spin column of the RNA Clean-up & Concentration kit (Zymo Research), following the manufacturer's instructions. First 200µl of RNA Binding buffer was added per sample, followed by another 300µl of ≥95% Ethanol. The solution was then mixed well (via vortex or pipette action) and up to 700µl transferred to a prepared Zymo-Spin IC column (in a 2mL collection tube). The samples were centrifuged at 13 000 g for 30 seconds, and the flow-through was discarded. If necessary, this step was repeated until all of the solution was transferred and run through the column. Next, a vial of the DNase I reaction master mix was generated by combining 5µl of both the DNase I (final concentration: 1U/µl) and DNA digestion buffer (provided) for every sample, and mixing gently via pipette action. 10µl was added to the centre of each Zymo-Spin IC column membrane and incubated at room temperature for 15 mins. After incubation, 400µl RNA Prep buffer was added to the column and centrifuged at 13 000 g for 30 seconds, and the flow-through was discarded. Next, the column was subjected to two washing steps with RNA Washing buffer, one with 700µl and a following centrifugation at 13 000 g for 30 seconds; the other with 400µl and centrifugation at 13 000 g for 1 min (dry spin). The final flow-through was discarded and the column transferred to a DNA LoBind 1.5 mL tube (Eppendorf). 30 – 60 µl of Nuclease-free water was added to the centre of the column membrane and incubated at room temperature for 1 minute, followed by a final centrifugation at 16 000 g for 1 minute. The final elution was subjected to several quantification and quality checks via a spectrophotometer (NanoDrop One C – Thermo Fisher Scientific), fluorometer (Qubit 4 – Thermo Fisher Scientific), and fragment analyzer (Bioanalyzer 2100 – Agilent).

DNA elutions were RNase-treated by first combining and gently mixing with 5µl of RNase A (final concentration: 20 µg/ml) and incubating at 65°C for 60 mins. Afterwards, the elutes were immediately placed on ice before being run through the DNA Clean-up & concentration kit (Zymo Research) according to the manufacturer's instructions. Samples were combined with 200 µl DNA Binding buffer and mixed well via vortexing; up to 700µl was then transferred to a prepared Zymo-Spin column (in a 2mL collection tube). The samples were centrifuged at 13 000 g for 30 seconds, and the flow-through was discarded. The column was washed twice using 200 µl DNA Wash buffer, followed by centrifugation at 13 000 g for 30 seconds and 1 min (dry spin), respectively. Finally, the Zymo-Spin column was transferred to a DNA LoBind 1.5 mL tube (Eppendorf), 30 - 60 µl of Nuclease-free water was added to the centre of the column membrane and incubated at room temperature for 1 minute, followed by a final centrifugation at 16 000 g for 1 minute. The final elution was subjected to several quantification and quality checks via a spectrophotometer (NanoDrop One C – Thermo Fisher Scientific), fluorometer (Qubit 4 – Thermo Fisher Scientific), and fragment analyzer (Bioanalyzer 2100 - Agilent).
Samples with low yield or yield below the limit of detection were re-extracted. Biomolecular extracts of samples that remained low in yield after re-extraction were pooled with the original extraction and concentrated via Refrigerated CentreVap Concentrator (Labconic). This was done by transferring the full volume from each extraction into a single DNA LoBind 1.5mL tube, then placing them (open) inside the CentriVap at either room temperature or 4°C, for DNA or RNA, respectively. The samples were run until the volume was reduced to 20 – 30 µl. The samples were then re-quantified via pectrophotometer (Nanodrop One C – Thermo Fisher Scientific), fluorometer (Qubit 4 – Thermo Fisher Scientific) and fragment analyzer (Bioanalyzer 2100 - Agilent).

Library preparation and sequencing

Library preparation for metagenomic DNA was done using the xGen DNA library preparation kit (cat. no. 10009822, Integrated DNA Technologies) using xGen UDI-UMI adapters (cat. no. 10005903) according to the manufacturer's protocol. The metagenomic DNA was enzymatically fragmented for 10 min, and DNA libraries were prepared without PCR amplification. The average insert size of the libraries was 600bp. Prepared libraries were quantified using Qubit (DNA HS kit, ThermoFischer) and quality-checked with Fragment Analyzer (Agilent). Library preparation for metatranscriptomic RNA was performed using Illumina stranded total RNA prep with Ribo-Zero Plus microbiome kit (Illumina, 20072063) according to the provided protocol. Prepared libraries were quantified using Qubit 4.0 (DNA HS kit, ThermoFischer) and quality-checked with Fragment Analyzer (Agilent). Sequencing for all libraries was performed at the LCSB genomics platform (RRID: SCR_021931) on a NextSeq2000 (Illumina) instrument using 2x150 bp read length. The fastq files were generated using the onboard DRAGEN Bio-IT server version 4.2.7.

Sequencing data processing and analysis

Paired MG and MT reads were processed and coassembled per sample using the Integrated Meta-omic Pipeline 3 (IMP3)[17] as follows. The reads were trimmed, quality-filtered, and then mapped against the human reference genome and transcriptome (T2TCHM13v2) to remove human host sequences. Remaining MG and MT reads were iteratively coassembled using MEGAHIT v1.2.9[43]. Assembled contigs were annotated using Bakta[44]. Functional annotation of Bakta predictions was done with Mantis[45] using KEGG[46], PFAM[47], and COG[48] databases. Feature quantification for MG and MT reads was done using featureCounts[49]. Based on the sequencing and quality control results, we derived several additional technical covariates,

including the raw number of MG and MT reads sequenced, the number of reads retained after quality control, and the percentage of MG and MT reads passing quality control.

MAG reconstruction, taxonomic annotation, and quality check

Metagenome-assembled genomes (MAGs) reconstruction was done per sample with IMP3. Assembled contigs underwent iterative meta-binning using five different binners: binny[50], CONCOCT[51], COMEbin[52], MetaBAT[53], and SemiBin[54]. The resulting bins were evaluated with two meta-binners, DASTool[55] and MetaWrap[56], and the best bins were retained for further analyses. For the taxonomic annotation, gtdb-tk v2.6.1[57] with GTDB v226[58] was used. All MAGs underwent an initial quality check with CheckM2 v1.1.0[59], and MAGs with completeness ≥ 80.0 and contamination < 5.0 were kept. After this step, the remaining MAGs were checked with MDMcleaner v0.8.7[60], and all contigs without the flag "keep" were removed from MAGs as contaminants. Final MAGs were again checked with CheckM2, with completeness ≥ 80 and contamination < 5 as thresholds, and the MAGs that passed were used in the downstream analyses. For each sample, we derived four technical covariates: the number of MG and MT reads mapped to MAGs that passed quality control, and the proportion of quality-controlled MG and MT reads mapping to these MAGs. Furthermore, for each MAG, we estimated expression as the total number of reads mapped to the MAG divided by its total length.

Pangenomes construction and species pseudo-bulks generation

Pangenomes were constructed with an in-house Nextflow-based pipeline. MAGs for each species identified in the dataset were collected across all samples, and nucleic acid sequences of all annotated features were extracted. Sequences were then clustered with MMseqs2 v17.b804f[61] using "easy-cluster" mode with "--min-seq-id 0.95", "-c 0.90", and "--cov-mode 0" to produce per-species gene-based pangenomes. Feature annotations and corresponding MG and MT count estimates were integrated from the feature quantification results (see Sequencing data processing and analysis) through matching gene, sample, and contig identifiers. In pangenomes, core genes were defined as those present in >90% of genomes, accessory genes as those present in 10–90% of genomes, and cloud genes as those present in <10% of genomes. Genes clustered within taxa according to a predefined similarity threshold were aggregated into pangenome-resolved features, yielding species-level pseudo-bulk count matrices. In these matrices, each row represents a high-quality MAG recovered from an individual's metagenome ("Sample X - Species Y"), while columns correspond to pangenome-aware feature units and cell values contain the associated MT counts.
The percentage of expressed genes per genome was computed for each species pseudo-bulk by dividing the number of genes with at least one MT count by the total number of genes in the species pseudo-bulk. The percentage of expressed genes per pangenome was computed for each species pseudo-bulk by dividing the number of genes with at least one MT count by the species pangenome size (total number of pangenome-aware feature units in the species pangenome). Only species with at least 20 MAGs were used for this analysis.

Dimensionality reduction

Before the dimensionality reduction step, pangenome-aware feature units of species pseudo-bulk samples were separately aggregated with KEGG, PFAM, and COG annotation systems, allowing for placing pseudo-bulk samples from different species in the same annotation space.

This resulted in three separate tables with per-species pseudo-bulk samples as rows and KEGG, PFAM, or COG annotations in columns, with MT counts as cell values. From the tables, we excluded all per-species pseudo-bulk samples with fewer than 100 genes with non-zero expression. These tables were then filtered to retain only species with at least 20 pseudo-bulk samples, resulting in 215 species included in the analysis. To reduce the data size before the dimension reduction, we excluded all features present in less than 25% of pseudo-bulk samples within each species.
Dimensionality reduction was performed separately for each filtered count matrix using the R implementation of the Uniform Manifold Approximation and Projection (UMAP) algorithm[62]. To assess the robustness of the resulting embeddings to parameter selection, we evaluated a grid of hyperparameter combinations, varying "n_neighbors" (15, 20, 30, 40, and 50) and "min_dist" (0.1, 0.3, 0.5, 0.7, and 0.9).

Strain inference

The strain inference was done within each species on quality-checked MAGs present in at least 30 samples. In the initial step, Galah[63] was used with the following parameters: "--ani 99.0", "--precluster-ani 95". To compute the Average Nucleotide Identity (ANI)-based distances between MAGs, dRep[64] was used with the following parameters: "--S_algorithm fastANI", "-pa 0.95", and "-sa 0.99". Pairwise distances between genomes were further transformed into R "dist" objects to make within-species genomic dissimilarity matrices. Projections of genomic dissimilarity matrices onto two-dimensional space were done with the monoMDS function of the vegan R package[65].

Machine learning

Machine learning was conducted in the R programming language v4.4.3 "Trophy Case"[66]. Machine learning analyses were performed on species-resolved pseudo-bulk profiles, separately for each species. The modelling feature set consisted of metatranscriptomic levels of pangenome-resolved features annotated using the PFAM, KO, and COG databases, while disease status (HC versus PD) was used as the response variable. Separate models were built for each annotation type. Given the sample size of the total cohort, model training was restricted to species detected in at least 60 individuals. To reduce data leakage between training and validation sets and to improve the robustness of model evaluation, we used one cohort for training and the other for validation. Because the ExpoBiome cohort included a larger number of samples, it was used as the training cohort, whereas the MiBiPa cohort was used for validation.
Before model training, pangenome-resolved features for each species in the training set were processed by adding a pseudocount of one, followed by total-sum scaling and log transformation. Features with prevalence below 15% were removed. For the remaining features, variance was calculated, and only features within the top 75% of the variance distribution were retained. The validation data were processed using a similar normalization procedure, including pseudocount addition, total-sum scaling, and log transformation. Further, the validation set was restricted to the features retained from the corresponding training set to ensure that feature selection was performed exclusively within the training data.
Model training was done with the XGBoost algorithm[67] via its R interface (xgboost R package). All models were run in 100 iterations with the gbtree booster and regularisation lambda equal to one, minimising the cross-entropy loss function; other parameters were left at default. After training, each model was evaluated on the validation set, and the area under the receiver operating characteristic curve (AUC) was computed for the validation set using the pROC R package[68].

Statistical analysis

Statistical analysis was conducted in the R programming language v4.4.3 "Trophy Case"[66]. Genomic and pangenomic expression breadth analysis was conducted using Bayesian beta-regression implemented in the R brms package[69], an interface to the Stan probabilistic programming language[70], and further handled by the R posterior package[71]. Only species with at least 20 pseudo-bulk samples were included in this analysis. Expression breadth was modelled as a beta-distributed proportion with a logit-linked mean (μ) and log-linked precision (ϕ), enabling simultaneous estimation of average expression breadth and its dispersion.
The mean model (1) included diagnosis, cohort, age, sex, constipation, the number (*mt_after_filtering*) and fraction (*mt_percent*) of MT reads that passed quality control, the number of per-sample MT reads mapped to MAGs that passed quality control (*counts_used_MT*), MAG completeness, and overall expression, with species-level varying intercepts and diagnosis slopes and a sample-level varying intercept.

(1) *Expression breadth ~ Diagnosis + Cohort + Age + Sex + Constipation + mt_after_filtering + mt_percent + counts_used_MT + Completeness + Expression + (1 + Diagnosis|Species) + (1|Sample)*

The precision model (2) included diagnosis, completeness, the number of per-sample MT reads mapped to MAGs that passed quality control (*counts_used_MT*), overall expression, cohort, and species-level varying intercepts and diagnosis slopes.

(2) *Expression breadth ~ Diagnosis + Completeness + counts_used_MT + Expression + Cohort + (1+Diagnosis|Species)*

Continuous covariates were z-scaled. Priors were Normal(0, 1) for fixed effects, Normal(0, 1.5) for the intercept, Exponential(1) for random-effect standard deviations, LKJ(2) for random-effect correlations, Normal(0, 0.5) for fixed effects in the precision submodel, and Normal(0, 1) for the precision intercept.
The model for genome expression breadth was fit with 10 chains of 8,000 iterations (4,000 warm-up; adapt_delta = 0.97, max_treedepth = 15). The model for pangenome expression breadth was fit with 12 chains of 14,000 iterations (7,000 warm-up; adapt_delta = 0.97, max_treedepth = 15). All models showed good convergence ($\hat{R}$=1.00 for all reported parameters) and adequate effective sample sizes.
The within-species genomic dissimilarity PERMANOVA analysis was done with the adonis2 model of the vegan R package. The model (3) included a distance matrix (DM) as a dependent variable, and a set of covariates: (i) main variable - either disease status, disease duration, or disease severity, based on the UPDRS parts I-III scale; (ii) biological covariates - sex, constipation status, and age; (iii) technical covariates - cohort origin, MAGs completeness, contamination, abundance and expression. Continuous covariates were z-scaled.

(3) *DM ~ Main variable + Sex + Constipation + Age + Cohort + Completeness + Contamination + Abundance + Expression*

The model was run in marginal mode with 9,999 permutations. Disease duration and severity were modelled only on data from PD patients. The homogeneity of within-group dispersions for disease status was assessed with the betadisper function of the vegan R package using "type = median" and "bias.adjust = TRUE". For both PERMANOVA and dispersion analysis, only nominally significant p-values were reported.
Correlation between within-species genomic and transcriptomic dissimilarity matrices was assessed with the Mantel test based on Pearson's product-moment correlation implemented in

the vegan R package. To compute transcriptomic dissimilarity matrices, the expression counts of individual pangenome-aware feature units were extracted for each species and organised in sample-by-feature matrices. These matrices were filtered to retain features present in at least 20% of samples per each species. A pseudocount of one was added to each matrix, followed by total sum scaling and computation of composition-aware Aitchison distance with the vegdist function of the vegan R package to produce transcriptomic dissimilarity matrices. To estimate the significance of correlation, genomic dissimilarity matrices were permuted against transcriptomic dissimilarity matrices 99,999 times, and p-values were corrected with FDR[72], with q-value < 0.05 as a significance threshold.

For each species, differential expression analysis was conducted on pseudo-bulk samples at four levels of resolution: individual pangenome-aware features (coding and non-coding genes, including rRNA, tmRNA, tRNA, CRISPR, and ncRNA) and functional categories derived from KEGG, PFAM, and COG annotations. Specifically, we modeled associations between expression and diagnosis, disease duration, and different aspects of disease severity, based on the UPDRS parts I-III scale. For the analysis, we used species present in at least 30 individuals. The analysis of disease duration and severity was done only on samples from PD patients. During the analysis, we conducted a systematic evaluation of the model specifications (covariates included) and types (linear model with total-sum-scaled, clr-transformed gene expression values and negative binomial model with raw counts) based on statistical inflation and model behavior. The concept of statistical inflation is derived from genomic inflation[73], which is used in genome-wide association studies to assess the behavior of regression models. Statistical inflation was estimated by first converting each p-value to a chi-square statistic with one degree of freedom using the inverse chi-square distribution $\chi^2 = \text{qchisq}(1 - p, df = 1)$. The inflation factor was then calculated as the median of these observed chi-square statistics, excluding missing values, divided by the theoretical median of a chi-square distribution with one degree of freedom. As a result, the optimal model (4) showed the best performance based on the negative binomial regression with sex, age, constipation, cohort origin, and sequencing library size (*mt_after_filtering*) as covariates, and a total sum of species pseudo-bulk transcript counts (*tot_sum*) as an offset.

(4) Counts ~ Main variable + Sex + Age + Constipation + Cohort + mt_after_filtering + offset(log(tot_sum))

The modelling was done with the glmmTMB R package[74] using quadratic parameterization for the negative binomial distribution[75]. Continuous covariates were z-scaled. Before the analysis, for diagnosis, feature abundances were filtered to retain features present in at least 20% of samples in each group or in 50% of samples in any one group; for the disease severity and duration, we filtered features present in less than 20% of samples. Unidentified features were set to zero. Correction for multiple testing was done with FDR separately for each species and level of resolution (genes, KEGGs, PFAMs, and COGs), with a q-value < 0.05 as a significance threshold. Visualisation was done with R packages ggplot2[76], Polychrome[77], patchwork[78], ggdist[79], and ggrepel[80]. Data were handled with the R package data.table[81].

Data availability

Metagenomic and metatranscriptomic data for the MiBiPa cohort are available on SRA, under the BioProject: PRJNA782492. Metagenomic and metatranscriptomic data for the ExpoBiome cohort are available on ENA, under: in prep. Metadata for the MiBiPa cohort is available via the ELIXIR Luxembourg Data Catalogue: https://datacatalog.elixir-luxembourg.org/e/dataset/ELU-2-81D473-1. Metadata for the ExpoBiome cohort is available via: in prep.

Code availability

The IMP pipeline, which was used for analysis of metagenomic and metatranscriptomic data, is available at https://gitlab.lcsb.uni.lu/IMP/imp3. The code used in the paper is available at: https://gitlab.com/uniluxembourg/Personalfolders/v.a.petrov/pangenome_paper.

References

1. Scheperjans, F. *et al.* Gut microbiota are related to Parkinson's disease and clinical phenotype. *Mov. Disord. Off. J. Mov. Disord. Soc.* **30**, 350–358 (2015).

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

3. Petrov, V. A. *et al.* Analysis of Gut Microbiota in Patients with Parkinson's Disease. *Bull. Exp. Biol. Med.* **162**, 734–737 (2017).

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

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

6. Romano, S. *et al.* Meta-analysis of the Parkinson's disease gut microbiome suggests alterations linked to intestinal inflammation. *Npj Park. Dis.* **7**, 1–13 (2021).

7. Romano, S. *et al.* Machine learning-based meta-analysis reveals gut microbiome alterations associated with Parkinson's disease. *Nat. Commun.* **16**, 4227 (2025).

8. Villette, R. *et al.* Integrated multi-omics highlights alterations of gut microbiome functions in prodromal and idiopathic Parkinson's disease. 2024.12.13.628341 Preprint at https://doi.org/10.1101/2024.12.13.628341 (2024).

9. Villette, R. *et al.* Human gut microbiome gene co-expression network reveals a loss in taxonomic and functional diversity in Parkinson's disease. *Npj Biofilms Microbiomes* **11**, 142 (2025).

10. Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2**, 16180 (2016).

11. Hickey, J. W. *et al.* Organization of the human intestine at single-cell resolution. *Nature* **619**, 572–584 (2023).

12. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).

13. Vannier, N. *et al.* Genome-resolved metatranscriptomics reveals conserved root colonization determinants in a synthetic microbiota. *Nat. Commun.* **14**, 8274 (2023).

14. Guerra, A. The pangenome: a statistical model, not a fixed biological property. *Bioinforma. Adv.* **6**, vbag069 (2026).

15. Jackson, R. W., Vinatzer, B., Arnold, D. L., Dorus, S. & Murillo, J. The influence of the accessory genome on bacterial pathogen evolution. *Mob. Genet. Elem.* **1**, 55–65 (2011).

16. Matthews, C. A., Watson-Haigh, N. S., Burton, R. A. & Sheppard, A. E. A gentle introduction to pangenomics. *Brief. Bioinform.* **25**, bbae588 (2024).

17. Narayanasamy, S. *et al.* IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).

18. Hansen, B. *et al.* Protocol for a multicentre cross-sectional, longitudinal ambulatory clinical trial in rheumatoid arthritis and Parkinson's disease patients analysing the relation between the gut microbiome, fasting and immune status in Germany (ExpoBiome). *BMJ Open* **13**, e071380 (2023).

19. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

20. Marzouk, N. H., Rashwan, H. H., El-Hadidi, M., Ramadan, R. & Mysara, M. Proinflammatory and GABA eating bacteria in Parkinson's disease gut microbiome from a meta-analysis perspective. *Npj Park. Dis.* **11**, 145 (2025).

21. Aho, V. T. E. *et al.* Relationships of gut microbiota, short-chain fatty acids, inflammation, and the gut barrier in Parkinson's disease. *Mol. Neurodegener.* **16**, 6 (2021).

22. Notting, F., Pirovano, W., Sybesma, W. & Kort, R. The butyrate-producing and spore-forming bacterial genus Coprococcus as a potential biomarker for neurological disorders. *Gut Microbiome* **4**, e16 (2023).

23. Wallen, Z. D. *et al.* Metagenomics of Parkinson's disease implicates the gut microbiome in multiple disease mechanisms. *Nat. Commun.* **13**, 6958 (2022).

24. Vilela Rodrigues, T. *et al.* Bacterial cell envelope components of gut commensals: effects in host-microbe interaction. *Ann. Med.* **58**, 2618429.

25. Shen, T. *et al.* The Association Between the Gut Microbiota and Parkinson's Disease, a Meta-Analysis. *Front. Aging Neurosci.* **13**, 636545 (2021).

26. Huh, E. *et al.* Methylglyoxal Aggregates α-Synuclein by Inhibiting SIRT1-Hif-1α in Intestinal Enteroendocrine Cells. *Int. J. Toxicol.* **45**, 22–28 (2026).

27. Heremans, I. P. *et al.* Parkinson's disease protein PARK7 prevents metabolite and protein damage caused by a glycolytic metabolite. *Proc. Natl. Acad. Sci.* **119**, e2111338119 (2022).

28. Skou, L. D., Johansen, S. K., Okarmus, J. & Meyer, M. Pathogenesis of DJ-1/PARK7-Mediated Parkinson's Disease. *Cells* **13**, 296 (2024).

29. Kumari, S. *et al.* Quantitative metabolomics of saliva using proton NMR spectroscopy in patients with Parkinson's disease and healthy controls. *Neurol. Sci.* **41**, 1201–1210 (2020).

30. Singh, Y. *et al.* DJ-1 (Park7) affects the gut microbiome, metabolites and the development of innate lymphoid cells (ILCs). *Sci. Rep.* **10**, 16131 (2020).

31. Gibson, R., Dalvi, S. P. & Dalvi, P. S. DJ-1 and Parkinson's disease. *Brain Disord.* **3**, 100020 (2021).

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

33. Qian, Y. *et al.* Gut metagenomics-derived genes as potential biomarkers of Parkinson's disease. *Brain* **143**, 2474–2489 (2020).

34. Zhao, Z. *et al.* Microbial biomarker discovery in Parkinson's disease through a network-based approach. *Npj Park. Dis.* **10**, 203 (2024).

35. Pietrucci, D. *et al.* Can Gut Microbiota Be a Good Predictor for Parkinson's Disease? A Machine Learning Approach. *Brain Sci.* **10**, 242 (2020).

36. Chen, C.-C. *et al.* Investigating Plasma Metabolomics and Gut Microbiota Changes Associated With Parkinson Disease. *Neurology* **104**, e213592 (2025).

37. Nishiwaki, H. *et al.* Meta-Analysis of Gut Dysbiosis in Parkinson's Disease. *Mov. Disord.* **35**, 1626–1635 (2020).

38. Lv, X. *et al.* Gut commensal *Agathobacter rectalis* alleviates microglia-mediated neuroinflammation against pathogenesis of Alzheimer disease. *iScience* **27**, 111116 (2024).

39. De Jaegher, S. *et al.* Identifying microbial biomarkers of neurodegeneration: a comparative study in Alzheimer's and Parkinson's disease. *Front. Microbiomes* **5**, (2026).

40. Li, M. *et al.* Performance of Gut Microbiome as an Independent Diagnostic Tool for 20 Diseases: Cross-Cohort Validation of Machine-Learning Classifiers. *Gut Microbes* **15**, 2205386.

41. Kleine Bardenhorst, S. *et al.* Gut microbiota dysbiosis in Parkinson disease: A systematic review and pooled analysis. *Eur. J. Neurol.* **30**, 3581–3594 (2023).

42. Goetz, C. G. *et al.* Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov. Disord. Off. J. Mov. Disord. Soc.* **23**, 2129–2170 (2008).

43. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinforma. Oxf. Engl.* **31**, 1674–1676 (2015).

44. Schwengers, O. *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genomics* **7**, 000685 (2021).

45. Queirós, P., Delogu, F., Hickl, O., May, P. & Wilmes, P. Mantis: flexible and consensus-driven genome annotation. *GigaScience* **10**, giab042 (2021).

46. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

47. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

48. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).

49. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* **30**, 923–930 (2014).

50. Hickl, O., Queirós, P., Wilmes, P., May, P. & Heintz-Buschart, A. binny: an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets. *Brief. Bioinform.* **23**, bbac431 (2022).

51. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

52. Wang, Z. *et al.* Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nat. Commun.* **15**, 585 (2024).

53. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

54. Pan, S., Zhu, C., Zhao, X.-M. & Coelho, L. P. A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat. Commun.* **13**, 2326 (2022).

55. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).

56. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).

57. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

58. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).

59. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

60. Vollmers, J., Wiegand, S., Lenk, F. & Kaster, A.-K. How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner. *Nucleic Acids Res.* **50**, e76 (2022).

61. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

62. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/ARXIV.1802.03426 (2018).

63. Aroney, S. T. N., Camargo, A. P., Tyson, G. W. & Woodcroft, B. J. Galah: More scalable dereplication for metagenome assembled genomes. https://doi.org/10.5281/zenodo.13637856 (2024) doi:10.5281/zenodo.13637856.

64. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

65. Oksanen, J. *et al.* vegan: Community Ecology Package. (2026).

66. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2025).

67. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. https://doi.org/10.48550/ARXIV.1603.02754 (2016) doi:10.48550/ARXIV.1603.02754.

68. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

69. Bürkner, P.-C. Bayesian Item Response Modeling in R with brms and Stan. *J. Stat. Softw.* **100**, 1–54 (2021).

70. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *J. Stat. Softw.* **76**, 1–32 (2017).

71. Bürkner, P.-C. *et al.* posterior: Tools for Working with Posterior Distributions. (2026).

72. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

73. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).

74. Brooks, M. *et al.* glmmTMB: Generalized Linear Mixed Models using Template Model Builder. 1.1.14 https://doi.org/10.32614/CRAN.package.glmmTMB (2017).

75. Hardin, J. W. & Hilbe, J. M. *Generalized Linear Models and Extensions*. (Stata Press, College Station, Texas, 2018).

76. Wickham, H. *et al.* ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. (2025).

77. Brock, G. & Coombes, K. R. Polychrome: Qualitative Palettes with Many Colors. 1.6.1 https://doi.org/10.32614/CRAN.package.Polychrome (2017).

78. Pedersen, T. L. patchwork: The Composer of Plots. (2024).

79. Kay, M. & Wiernik, B. M. ggdist: Visualizations of Distributions and Uncertainty. (2025).

80. Slowikowski, K. *et al.* ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. (2024).

81. Barrett, T. *et al.* data.table: Extension of 'data.frame'. (2026).

Acknowledgements

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 863664), and was further supported by the Luxembourg National Research Fund (FNR) CORE/16/BM/11333923 (MiBiPa), the Institute for Advanced Studies of the University of Luxembourg (AUDACITY grant MCI-BIOME_2019); Microbiomes in One Health Doctoral Training Unit – MICROH-DTU (FNR PRIDE/18/11823097); Metaproteome-based Leveraged Microbiome Management in the Context of One Health– (METAMIC3, HORIZON–MSCA–2024–DN 101225682); Investigating the potential of microbiome-derived amyloidogenic proteins as clinical biomarkers for the early detection of Parkinson's disease (AMY-Dx, ERC Proof of Concept Grant 101247183) to P.W.; Michael J. Fox Foundation (#025802) to P.M.

Author information

These authors contributed equally: Patrick May, Paul Wilmes

Authors and Affiliations

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Viacheslav A. Petrov, Oskar Hickl, Rémy Villette, Bérénice Hansen, Stella Blumrich, Cédric C. Laczny, Tuesday Lowndes, Zainab Zafar, Jordan Caussin, Rashi Halder, Jochen G. Schneider, Patrick May & Paul Wilmes

Department of Infection and Immunity, Luxembourg Institute of Health, Luxembourg

Oskar Hickl

Systems and Synthetic Biology, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

Oskar Hickl

Institute for Social Medicine, Epidemiology and Health Economics, Charité - University Medical Center Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

Etienne Hanslian, Daniela A. Koppold, Michael Jeitler, Nico Steckhan & Andreas Michalsen

Department of Internal and Integrative Medicine, Immanuel Hospital Berlin-Wannsee Branch, Berlin, Germany

Etienne Hanslian, Daniela A. Koppold, Michael Jeitler & Andreas Michalsen

Institute of General Practice and Interprofessional Care, University Hospital Tübingen, Tübingen, Germany

Michael Jeitler

Robert Bosch Center for Integrative Medicine and Health, Bosch Health Campus, Stuttgart, Germany

Michael Jeitler

Department of Dermatology, Venereology and Allergology, Charité - University Medical Center Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

Anika Hartmann Rajput

Digital Health-Connected Healthcare, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany

Nico Steckhan

Paracelsus-Elena-Klinik Kassel, Kassel, Germany

Sebastian Schade & Brit Mollenhauer

Department of Neurology, University Medical Center Göttingen, Göttingen, Germany

Brit Mollenhauer

Department of Health, Medicine and Life Sciences, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Jochen G. Schneider & Paul Wilmes

Department of Internal Medicine II, Saarland University Hospital and Saarland University Faculty of Medicine, Homburg, Germany

Jochen G. Schneider

Contributions

Supervision: P.W. Conceptualization: V.A.P., P.M. and P.W. Patient recruitment, clinical coordination, and sampling: P.W., B.H., R.V., C.C.L., E.H., D.A.K., A.R.K., M.J., N.S., S.S., B.M., A.M., J.G.S. Multi-omic data generation: T.L., Z.Z., J.C., R.H. Bioinformatics: O.H., V.A.P., P.M., R.V., S.B. Statistics and data visualization: V.A.P. Initial manuscript draft: V.A.P. and P.M. Review and editing: O.H., P.M., C.C.L., and P.W. Funding acquisition: C.C.L. and P.W. All authors read and approved the submitted version.

Corresponding authors

Correspondence to: Viacheslav A. Petrov or Paul Wilmes

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarynotes.docx
- SupplementaryFigure4.jpg
- SupplementaryFigure2.jpg
- SupplementaryDocument.pdf
- SupplementaryFigure3.jpg
- SupplementaryFigure1.jpg
- Supplementarytables.xlsx