METAGENOME-ASSEMBLED GENOMES FROM A POPULATION-BASED COHORT UNCOVER NOVEL GUT SPECIES AND STRAIN DIVERSITY, REVEALING PREVALENT DISEASE ASSOCIATIONS

A short running title: Utility of MAGs in large cohort study

Kateryna Pantiukh[1]*, Kertu Liis Krigul[1], Oliver Aasmets[1], Elin Org[1]*

[1] Institute of Genomics, Estonian Genome Centre, University of Tartu, Tartu, Estonia

*Address correspondence to Kateryna Pantiukh and Elin Org pantiukh@ut.ee, elin.org@ut.ee

Present address: Institute of Genomics, Estonian Genome Centre, University of Tartu, Tartu 51010, Estonia, phone: (372) 737 4034; fax: (372) 737 4060

ABSTRACT

Metagenomic profiling has advanced understanding of microbe-host interactions. However, widely used read-based approaches are limited by incomplete reference databases and the inability to resolve strain-level variation. Here, we present a scalable, genome-resolved framework that integrates population-specific metagenome-assembled genomes (MAGs) to discover novel species, strain diversity, and disease associations. From 1,878 deeply sequenced samples in the Estonian microbiome cohort (EstMB-deep), we reconstructed 84,762 MAGs representing 2,257 species, including 353 (15.6%) previously uncharacterized species reaching up to 30% relative abundances in some individuals. We integrated these MAGs with the Unified Human Gastrointestinal Genome (UHGG) collection to create an expanded reference (GUTrep), enabling profiling of 2,509 EstMB individuals and testing associations with 33 prevalent diseases. Of 25 diseases with significant associations, 8 involved newly identified species, underscoring the value of population-specific MAGs. To quantify within-species diversity, we developed the Strain Richness Index (SRI), a novel MAG-based metric that informed strain-level analyses. Based on SRI, we prioritized *Odoribacter splanchnicus,* a prevalent species with the lowest strain heterogeneity, yielding sufficient power for strain-level analysis. We identified two dominant strains, N1 and N2, with distinct gene repertoires and divergent disease associations.

Notably, strain N1 was negatively associated with gastritis and duodenitis and hypertensive heart disease, associations undetected at the species level. Our study expands the human gut reference landscape, demonstrates the importance of population-specific MAGs for uncovering novel microbial diversity, and reveals strain-level disease associations obscured at higher taxonomic levels, highlighting the need for genome-resolved approaches in microbiome research.

KEYWORDS

Gut microbiome, Metagenome-assembled genomes, strain richness index, population microbiome, metagenomics, metagenome-wide associations study, strain-level diversity.

INTRODUCTION

The human gut microbiome exhibits remarkable diversity across individuals and populations, necessitating comprehensive global reference databases to enable accurate taxonomic and functional profiling of microbial communities. In recent years, considerable research effort has been directed towards establishing collections of global reference genomes of the human gut microbiome. Initially, the focus was on sequencing bacteria that could be isolated and cultured[1,2]. However, rapid technological advancements have facilitated the generation of vast amounts of metagenomic data and the development of techniques for assembling genomes from unculturable species, consequently improving reference databases. These Metagenome-Assembled Genomes (MAGs) substantially expand the number of gut microbial species, as 81% of the species in the current version of the Unified Human Gastrointestinal Genome (UHGG) collection were identified by MAGs while having no corresponding representative in any human gut culture database[3]. Moreover, MAG assembly enables genome-centric analyses, such as identifying strains of species present in a population and conducting strain-level association studies[4,5]. Therefore, MAGs enable us to significantly improve our understanding of the ecosystem under study.

The importance of MAGs recovery is exemplified in population biobanks that include deeply phenotyped individuals and their microbiome samples. In this case, it becomes possible to identify correlations between known, newly reconstructed species and specific genome structure and various environmental, dietary, or health-related factors. In recent years, several population-based biobanks with metagenomic datasets have been established, for example, the Dutch Microbiome cohort from the Lifelines biobank[6], the Israeli Project 10K cohort[7,8], the FinRisk cohort[9], and the EstMB cohort from the Estonian Biobank[10]. Analyses of these datasets have demonstrated that gut microbiome composition is associated with a range of environmental and lifestyle factors, particularly diet and medication use[10,11], and that variation in the microbiome is associated with several diseases, such as cardiovascular diseases[12,13], mental health disorders[14] and cancers[15,16]. Furthermore, emerging evidence suggests that the gut microbiome has predictive power, as demonstrated in the context of incident heart failure[17]. However, many of these studies still rely solely on reference databases. These databases may lack representatives for many uncultured or underrepresented population-specific microbial species, leading to incomplete or biased interpretations.

In the present study, we leveraged deep metagenomic sequencing of a population⍰based Estonian microbiome-deep (EstMB-deep) cohort to assemble a comprehensive collection of metagenome⍰assembled genomes (MAGs), substantially expanding the reference database of human gut microbes with hundreds of previously uncharacterized species. We integrated these population⍰specific MAGs with public reference data and conducted association analyses with 33 prevalent diseases. To systematically assess within-species diversity, we developed a novel Strain Richness Index (SRI). We demonstrated its utility by identifying strain⍰specific disease associations in *Odoribacter splanchnicus* that were not apparent at the species level. Our findings demonstrate that genome⍰resolved, strain⍰aware microbiome profiling can uncover novel disease⍰linked microbial signatures beyond that remain hidden using conventional approaches.

RESULTS

Study design and cohort overview

The study aimed to first recover metagenome-assembled genomes (MAGs) from a deeply sequenced Estonian gut microbiome cohort (EstMB-deep, N=1,878) and expand the reference database by combining these newly assembled genomes with the existing public Unified Human Gastrointestinal Genome (UHGG) collection and demonstrate the added value of genome-resolved, strain-level analysis in identifying disease associations[3] (**Figure 1a**).

The EstMB-deep subcohort used in the study is a subset of the volunteer-based Estonian Microbiome cohort that is resequenced with much deeper coverage than the initial sample set. In brief, the EstMB cohort included 1,764 women (70.31%) and 745 men (29.69%), and the EstMB-deep subcohort consists of 1,308 women (69.65%) and 570 men (30.35%), with both cohorts representing individuals aged 23 to 89 (**Figure 1b**). Compared to the EstMB average sequencing depth (30.63 ± 3.12 million reads per sample), EstMB-deep achieved over threefold higher coverage (106.70 ± 42.1 million, **Figure 1c**). A detailed description of the EstMB, including omics and phenotypic data, is provided in Aasmets & Krigul et al. 2022[10].

Creating a representative MAG pool of gut bacteria in the Estonian population

To characterize population-specific microbes and expand publicly available human gut microbiome databases with microbial genomes from the Estonian population, we performed *de novo* Metagenome Assembled Genomes (MAGs) reconstruction from all 1878 samples in the EstMB-deep cohort. The MAG reconstruction pipeline is summarized in **Figure 1d**. We successfully reconstructed 84,762 MAGs from EstMB-deep, with an average of 45.13 MAGs per sample. Among these, 42,049 (49.61%) were high quality (HQ) MAGs, i.e. MAGs with completeness > 90% and contamination < 5%; 26,806 (31.63%) were medium quality (MQ) MAGs, i.e. MAGs with completeness > 50% and contamination < 10%; and all others 15,907 (18.77%) were low quality (LQ) MAGs according to CheckM **(Figure S1)**. To describe the

Estonian population species pool, we clustered all MAGs with dRep using a 95.0% ANI threshold, ensuring that the final clusters represent distinct species.

The species-level clustering procedure yielded 2,257 clusters (**Table S1**). For each cluster, the representative MAG was selected based on genome completeness, minimal contamination, strain heterogeneity, and N50 (a parameter reflecting assembly fragmentation level). We refer to these 2,257 species representative MAGs as the "ESTrep" collection. The majority of ESTrep MAGs from the ESTrep collection (72.97%, n=1647 MAGs) were >90% complete and <5% contaminated (**Figure 1e-g**). Additionally, 475 of them (21.04%) contained the 5S, 16S, and 23S rRNA genes along with at least 18 tRNAs, meeting the 'high quality' criteria defined by the Genomic Standards Consortium[18], and we refer to these as HQ-mimag MAGs (**Figure 1e**).

MAG assembly remains essential for detecting novel population-specific species

Next, we identified previously uncharacterized species within the ESTrep MAGs collection. MAGs were categorized as novel species if their taxonomic classification at the species level or higher couldn't be assigned using the GTDB-Tk[19], a common approach for evaluating whether a newly reconstructed MAG represents a new species[20,21]. Of the 2,257 representative MAGs, 353 (15.64%) were classified as novel. Among these, 231 MAGs (65.44%) had > 90% and < 5% contamination, and 57 (16.15%) also contained rRNA and tRNA genes, meeting the MIMAG guidelines for high-quality MAGs[18]. We observed a strong correlation between the number of novel species discovered and the number of samples analyzed ($R^2 = 0.97$). Specifically, for every 500 samples, approximately 102 novel species were identified (**Figure 2a**). As we have not observed any indication of a plateau with the current sample size, we expect that analysing more samples will reveal additional species.

Although Estonia is considered a Westernised population, novel species still make up a significant proportion of the microbiome community. These newly identified species were distributed across multiple phyla (**Figure 2b)**. On average, 2.82% of the total reads per sample were assigned to these novel species, even reaching a maximum relative abundance of 32.34%

in some samples (**Figure 2c**). Since these species are absent from public databases and may be population-specific, microbiome studies that rely solely on existing references may substantially underestimate microbial diversity.

Integrating population-specific and global MAGs improves reference quality and uncovers assembly biases

As the success of metagenome assembly and genome reconstruction depends on multiple technical and analytical factors, we did not expect to recover all microbial genomes present in the gut. Therefore, we constructed an integrated species-level reference by combining newly reconstructed MAGs from the Estonian population with publicly available human gut-associated species. This integrated reference, called the GUTrep collection (**Figure 1a**), was generated by deduplicating the ESTrep MAGs collection and UHGG MAGs collection[22] at a 95.0% ANI threshold, retaining the highest-quality MAG for each species. When two MAGs from one species were present (one from ESTrep and one from UHGG), the highest-quality MAG was selected for the final collection. The final GUTrep database comprises 4,792 species, of which 3,285 (68.55%) originated from UHGG and 1,507 (31.45%) from ESTrep, thereby substantially improving the UHGG dataset. Notably, the ESTrep contribution includes 353 novel species, 607 known species absent from UHGG and 927 higher-quality MAGs already represented in UHGG.

To estimate microbiome composition across the EstMB dataset (n = 2,509), we mapped all reads against the GUTrep collection. This approach, which does not require deep sequencing, identified 3,423 species in total. On average, each sample contained 292 species, whereas MAG assembly yielded an average of 45 MAGs per sample (**Figure 2d**). The most prevalent (>95%) species detected by read mapping were all well-known gut microbes: *Phocaeicola dorei*, *Bacteroides spp* (*B. uniformis*, *B. xylanisolvens*, *B. ovatus*), *Faecalibacterium prausnitzii*, and *Odoribacter splanchnicus, P. dorei and B. uniformis* also being among the most abundant species in addition to *Prevotella copri* (>2% on average) **(Figure S2)**. We observe that samples with more species detected by mapping also tended to have more MAGs recovered (**Figure S3**).

However, the number of assembled genomes per species did not clearly correlate with species prevalence or mean abundance (**Figure 2e, Table S2**). Moreover, the difference between these values can range from minimal to substantial. For example, despite one of the most prevalent species, *Bacteroides xylanisolvens,* being detected in 97.13% of samples and having a mean relative abundance of 0.39%, only 18 MAGs were assembled for this species. This pattern, common among newly identified species, highlights that many species detected by mapping are represented only by a few MAGs, complicating genome-centric analysis (**Figure 2f, Table S3**). Among 3,423 species detected, only 199 were represented by more than 100 assembled MAGs and just 19 species had over 500 recovered MAGs (**Figure S4**), illustrating the challenges of comprehensive genome reconstruction.

Newly assembled species provide valuable input for association studies

Next, we utilized the comprehensive electronic health records (EHR) data from the Estonian population to perform a microbiome-wide association study (MWAS) of common diseases, using the population-based GUTrep reference. We included 33 prevalent diseases (≥ 100 cases each; **Table S4**), spanning various categories, such as the respiratory system (7 diseases), circulatory system (7 diseases), and digestive system (4 diseases) disorders. Associations between species abundance and diseases were assessed using linear regression models adjusted for BMI, gender, and age. To reduce multiple testing, we limited the analysis to species present in ≥1% of the samples, resulting in 1,595 species.

We identified 105 significant associations (Bonferroni-adjusted $p < 2.71 \times 10^{-5}$) between 96 bacterial species and 25 diseases (**Table S4, Table S5**). Notably, newly assembled species were associated with 8 out of the 33 diseases, including asthma, chronic ischemic heart disease, chronic rhinitis, nasopharyngitis and pharyngitis, female infertility, heart failure, haemorrhoids, iron deficiency anaemia, and vitamin D deficiency. For example, one of the strongest associations was observed for chronic ischemic heart disease, involving a newly assembled species from the *Nanosynbacter* genus (species ID: H2144_Nanosynbacter_undS, adjusted p =

3.13×10^{-6}) (**Figure 2g, Table S4**). These findings emphasize the importance of population-specific reference databases for detecting disease-associated microbiome changes. However, further studies are needed to confirm whether these associations generalize beyond the Estonian cohort.

MAG data enables strain-level diversity analysis across species

Most large-scale microbe-disease or MWAS studies are conducted at the species level, although strain-level analysis is often recommended for understanding the functional insights[23–25]. However, not all species exhibit a strain structure that allows sufficient case numbers for robust strain-level association testing. Metagenome assembly provides the opportunity to characterize this diversity within species and describe strain structures and prevalences in the population. Here, we introduce the Strain Richness Index (SRI), a metric that quantifies genetic variation within species, i.e. how many strains per individual species can be detected in the population. Specifically, the number of strain clusters detected per species is normalized by the number of MAGS:

Strain Richness Index (SRI) = (Number of strains / Number of MAGs) x * 100%

This normalization allows comparison of within-species diversity and allows for systematic assessment of strain structure across bacterial species in the human gut microbiome. We focused on species with >10 reconstructed MAGs, yielding 376 species across diverse phyla. Notably, none of the newly identified species were included due to an insufficient number of MAGs. Strain clusters were defined at 99% Average Nucleotide Identity (ANI).

The SRI values varied widely, ranging from 0.4 to 100 (**Figure 3a, Table S6**), indicating substantial differences in strain diversity across species in the population. *Odoribacter splanchnicus* exhibited the lowest SRI (0.4), with one strain per ~250 MAGs, reflecting low diversity despite high prevalence. In contrast, *Prevotella copri* had one of the highest SRIs (94.0), consistent with its well-documented heterogeneity, where nearly every MAG represents a unique strain, making it difficult to conduct strain-level association analysis in the population. Interestingly, *Alistipes_A* genus appeared in both the lowest and highest SRI groups.

We also examined SRI distribution across six phyla with ≥10 species present in each: *Bacillota, Bacteroidota, Verrucomicrobiota, Bacillota_A*, *Proteobacteria,* and *Cyanobacteroidota* - all of which exhibited a broad range in SRI distributions (**Figure 3b**). *Bacillota* species tended to have higher SRI values, indicating that this phylum generally tends to have a higher number of strains per species. In contrast, *Verrucomicrobiota*, *Cyanobacteroidota*, and *Proteobacteria* exhibited lower SRI values, suggesting that species in these phyla typically have fewer strains per species. However, due to the small sample sizes in some phyla, further studies are needed to confirm whether these differences represent true phylum-level trends.

Strain-level analysis reveals novel phenotype associations undetected at the species level

In order to demonstrate the value of strain-level MWAS analysis, we selected *O. splanchnicus* due to its low SRI (SRI = 0.4) and high prevalence (detected in 96.14% of samples, assembled in 72.68%, **Figure 3c**). Among its MAGs, we identified four distinct strain clusters, two of which were rare (found in 2 and 19 samples, respectively). Therefore, we focused on the two major clusters with high case numbers: strain N1(n=974 samples, original strain ID: 1_2.3.4.6.9) and strain N2 (n=335 samples, original strain ID: 1_1) (**Figure 3d**).

Logistic regression models adjusted for BMI, gender, and age were used to assess the association between the presence or absence of *O. splanchnicus* strains N1 and N2 and the same 33 diseases previously analyzed at the species-level MWAS. Our analysis identified a significant association between the presence of strain N1 and two different diseases - gastritis and duodenitis, and hypertensive heart disease (**Figure 3e**). The odds ratio for strain N1 was less than 1 in both diseases (gastritis and duodenitis OR = 0,56, hypertensive heart disease OR = 0,63), indicating that its presence is associated with a reduced likelihood of having the disease. Notably, these associations were not detected at the species level, highlighting the added resolution of strain-level analysis.

To explore functional differences, we performed a pan-genome analysis of strains N1 and N2. We carried out Principal coordinate analysis of predicted gene cluster presence/absence, which showed clear separation between the strains (**Figure 3f**). We identified that the two strains formed distinct clusters, indicating clear genomic differentiation based on gene content (**Figure 3g**), with 40 gene clusters unique to one of the two (**Figure 3h, Table S7**). While most encoded hypothetical or uncharacterized proteins, some were annotated with putative functions based on the Clusters of Orthologous Genes (COG20)[26]. Strain N2 harboured a broader repertoire of genes associated with stress response, iron acquisition, and antimicrobial resistance—traits consistent with enhanced survival in inflammatory gastrointestinal environments. These included elevated copy numbers of the extracytoplasmic stress sigma factor RpoE (σ^E), iron transport components FecR and CirA, and multidrug resistance elements such as AcrR and an ABC-type efflux pump (YadH). In contrast, strain N1 was enriched for redox maintenance proteins such as YyaL/DsbD, suggesting a distinct strategy centered on oxidative stress mitigation.

DISCUSSION

Our study presents a scalable, genome-resolution framework for population-scale microbiome analysis, enabling improved species and strain-level characterization and discovery of disease associations. By expanding the gut microbial reference database with thousands of metagenome-assembled genomes (MAGs), including novel bacterial species and diverse strains of known taxa, we address a major limitation in current reference datasets, which often underrepresent global microbiome diversity. We demonstrate that genome-resolution microbiome analysis, coupled with population-specific MAG catalogues, enables more comprehensive species- and strain-level association studies. Furthermore, we introduce the Strain Richness Index, a novel quantitative metric of within-species genetic diversity, which we applied across 378 gut species to guide candidate selection for strain-level analysis. Using this framework, we uncovered strain-specific disease signals for *O. splanchnicus* that were invisible at the species level and provided functional genomic insights that may explain these associations.

Read-based taxonomic profiling remains the most widely adopted approach to characterize microbial communities, particularly in association studies[27,28]. Alternatively, metagenome-assembled genomes (MAGs) offer a culture-independent, reference-free approach to recover community structure[29]. Despite deep sequencing (an average of 106 million read pairs per sample), metagenome assembly reconstructed only ~45 genomes per sample on average, compared to 292 species detected by read mapping, highlighting that even abundant taxa are not always fully recoverable by assembly and *de novo* assembly alone fails to capture the full microbial diversity. This finding challenges the common assumption that high-abundance taxa can be reliably assembled given sufficient sequencing depth[30–32]. We observed multiple cases where prevalent and relatively abundant species yielded few MAGs. For instance, *Bacteroides xylanisolvens,* present in 97.13% of samples at 0.39% mean abundance, yielded only 18 MAGs. Similar trends were observed for novel species, e.g., an undefined sp from the *Butyricimonas* genus (ID: H0366) was assembled from just 36 samples but detected in >55% samples by mapping. These findings support earlier observations that low-abundance but genetically distinct species may assemble more readily than abundant, genetically diverse taxa[33] and underscore the need for a hybrid strategy combining MAG assembly and high-resolution read mapping against population-specific reference. Our GUTrep database exemplifies such a strategy, integrating local MAGs with the Unified Human Gastrointestinal Genome (UHGG) collection to improve reference coverage. Notably, 31% of dereplicated GUTrep species originated from our Estonian-specific MAGs, illustrating the added value of local assembly efforts.

A common argument against investing in resource-intensive *de novo* assembly is that the rapidly growing and regularly updated public gut genome catalogues increasingly capture known microbial diversity, suggesting that most gut species will soon be represented, and further assembly may become redundant. However, our findings challenge this assumption, and support continued *de novo* assembly in new population studies. Early efforts reported high proportions of novel taxa, with 77% of MAGs classified as novel in Pasolli et al[34] and 66% in Almeida et al[35]. More recent studies, such as Leviatan et al., still report 310 novel species out of 3,594 assembled (8.6%)[36]. In our cohort, we recovered 353 novel species from 2,257 MAGs (15.6%), with ~102 additional novel species per 500 samples, and no indication of a discovery plateau. Moreover, we confirm a previously reported finding that many novel species assembled from a few samples were nonetheless widespread by mapping, suggesting that assembly remains essential even in well-characterized industrialized populations[33]. These results reinforce the idea that local assembly efforts complement global references and remain critical for uncovering the full spectrum of microbial diversity.

Another advantage of *de novo* assembly is its ability to uncover strain-level variation[5,24]. Strains of the same species can differ significantly in function and disease associations[37]. As a classic example, well-known gut microbe Escherichia coli species includes strains which can be pathogenic (e.g., enterohaemorrhagic O157:H7), probiotic (Nissle 1917), or commensal (K-12), and this demonstrates how it can be insufficient to study the microbe at the species level[38]. While strain-level taxonomic profilers such as MetaPhlAn 4.0[39] offer efficient resolution, they lack the genomic content necessary for detailed functional analyses.

In contrast, reconstructing MAGs directly from samples linked to host metadata allows for in-depth investigation of within-species genomic variation in relation to specific phenotypes. Nevertheless, this approach is not feasible for all species in the population. Many taxa, especially newly discovered ones, are only recovered in a small number of samples, limiting their use in association analysis. In our dataset, no newly identified species was assembled in more than 36 samples; the most prevalent was a novel species from the *Butyricimonas* genus (MAG ID: H0366). For species that are well represented in the MAG dataset, high intraspecies genomic diversity can further complicate analysis. Thus, within-species genomic variability becomes a critical consideration when selecting candidate species for strain-level analysis.

Our strain richness index, or SRI, helps to assess whether strain-level analysis is feasible by quantifying within-species diversity. High SRI value of the species might indicate that each individual harbours a unique strain, complicating population-level associations. In this study, we analysed within-species diversity across 378 gut species, expanding upon previous work that showed a strain richness of 92 gut species based on pure isolates from at least three different individuals[40]. Our results significantly expand on this by including a broader range of species, each represented by more than 10 MAGs. Consistent with earlier findings, we observe substantial variability in strain richness across species[40]. However, our data provides a more detailed and comprehensive picture due to the larger number of species and genomes analysed, allowing for a broader estimation range of the SRI estimation. Among the 14 overlapping species between the two studies, some show similar patterns of genomic diversity - for example, *Odoribacter splanchnicus* and *Barnesiella intestinihominis* exhibit consistently low diversity, while others like *Bifidobacterium longum* remain highly diverse. Other overlapping examples, such as *Fusicatenibacter saccharivorans* and *Coprococcus eutactus*, display divergent diversity estimates, likely reflecting methodological differences (culture isolate vs metagenome-based) and highlighting the need for further comparative research. Highly diverse species such as *Prevotella copri* well known from other studies[41,42], are absent from the Chen-Liaw dataset, but prominent in ours. Our larger dataset also allowed the investigation of phylum-level patterns, suggesting that the common phyla, such as *Bacteroidota and Bacillota_A* species, tend to have higher SRI values, while *Verrucomicrobiota, Cyanobacteroidota, and Proteobacteria* exhibit lower diversity. These phylum-level differences in strain richness suggest possible evolutionary or ecological constraints. However, further studies are needed, particularly for the less common phyla, to validate and understand the underlying mechanisms.

Understanding microbiome diversity at both species and strain levels enhances resolution in metagenome-wide association studies (MWAS). At the species level, we identified 96 bacterial species significantly associated with 25 common diseases, of which 8 diseases involved previously uncharacterized species, highlighting the limits of relying solely on global references.

These differences may reflect population-specific variation driven by local differences in diet, genetics, and lifestyle[11]. At the strain level, we identified associations undetectable at the species level, such as *Odoribacter splanchnicus* strain N1 negative association with gastritis, duodenitis and hypertensive heart disease. Previous research has shown that the abundance of the genus *Odoribacter* is negatively correlated with systolic blood pressure in overweight and obese pregnant women, suggesting that SCFA-producing taxa may influence host blood pressure[43]. Comparative genomic analysis of the MAGs from two *O. splanchnicus* strains helped us to identify a set of gene clusters that differed between the two groups. These genomic features suggest that strain N1 is functionally better adapted to conditions characteristic of gastritis and duodenitis, such as oxidative stress, nutrient limitation, and host antimicrobial pressure. Together, our findings highlight the value of strain-resolved metagenomic approaches in revealing disease-relevant microbial functions that would remain hidden in broader taxonomic analysis. While our study focused on a single, ethnically homogeneous Northern European population, several key findings, such as the detection of widespread yet previously uncharacterized species and low-diversity strain structures within common gut taxa, are likely to extend beyond the Estonian population. However, microbiome composition is influenced by genetic, dietary and environmental factors that vary between populations. Future studies in diverse cohorts will be essential to evaluate the generalizability of our results and validate the species - and strain-level disease associations uncovered here. The genome-resolved analytical framework we present is scalable and readily applicable to other population-scale microbiome datasets, enabling cross-cohort comparison and discovery.

Our study also has some limitations that should be considered. First, the reliance on short-read sequencing, may reduce assembly contiguity and strain resolution compared to long-read approaches[44]. Although long read technologies offer higher genomic completeness , their current cost limits their use in large-scale population studies. A practical compromise could involve using short-read sequencing for most samples and applying long-read sequencing to key or novel taxa. Second, the observed associations are correlative, and further validation in longitudinal and experimental studies is needed to assess causality. Third, our strain-level analysis focused on one species due to sample representation and analytical feasibility, and broader application across species remains a key next step. Finally, while functional differences between strains were identified, interpretation was limited by the prevalence of unannotated genes. Future improvements in genome annotation and cross-cohort replication will be essential to build on these findings. Despite these challenges, our findings demonstrate the value of population-scale metagenomics in uncovering novel microbial diversity and strain-level functional signatures relevant to human health.

CONCLUSION

In conclusion, this study expands the human gut genomic reference, underscores the importance of population-specific MAGs in uncovering novel microbial diversity, and reveals strain-level disease associations obscured at higher taxonomic levels, thereby highlighting the critical need for genome-resolved approaches in microbiome research.

METHODS

Estonian Microbiome cohort description

The Estonian Microbiome Cohort (EstMB) was established in 2017, when stool, oral, and blood samples were collected from 2509 Estonian Biobank (EstBB) participants[10]. The EstBB is a volunteer-based population cohort initiated in 1999 that currently includes over 212,000 adults of European ancestry (≥ 18 years old) across Estonia[45]. Extensive information is available for the EstMB participants, including data from self-reported questionnaires and electronic health records (EHRs) (completed by medical professionals) covering diseases, medication use and medical procedures both before and after sample collection. In addition to the questionnaire and EHR data, the participants' anthropometric measurements (e.g., height, weight, blood pressure, and waist and hip circumferences) were taken during a pre-registered visit upon delivering the stool sample. The Estonian Microbiome Deep cohort (EstMB-deep) includes a subset of stool samples from the EstMB cohort that have been resequenced with over three times deeper coverage (N = 1878).

Microbiome sample collection and DNA extraction

The participants collected a fresh stool sample immediately after defecation with a sterile Pasteur pipette and placed it inside a polypropylene conical 15 mL tube. The participants were instructed to time their sample collection as close as possible to the visiting time in the study center. The samples were stored at −80 °C until DNA extraction. The median time between sampling and arrival at the freezer in the core facility was 3 h 25 min (mean 4 h 34 min), and the transport time was not significantly associated with alpha (Spearman correlation, p-value 0.949 for observed richness and 0.464 for Shannon index) nor beta diversity (p-value 0.061, R-squared 0.0005). Microbial DNA extraction was performed after all samples were collected using a QIAamp DNA Stool Mini Kit (Qiagen, Germany). For the extraction, approximately 200 mg of stool was used as a starting material for the DNA extraction kit, according to the manufacturer's instructions. DNA was quantified from all samples using a Qubit 2.0 Fluorometer with a dsDNA Assay Kit (Thermo Fisher Scientific).

Shotgun metagenomic sequencing

Sequencing for the main EstMB cohort was done using shotgun metagenomic paired-end sequencing on the Illumina NovaSeq 6000 platform and described in detail in[10]. The EstMB-deep cohort samples were selected based on DNA quality and resequenced at higher depth using paired-end shotgun metagenomic sequencing on the MGISEQ-2000 platform. Sequencing reads' quality control (QC) was performed using FastQC (v0.12.1)[46], and human reads were filtered using Bowtie2 (v0.6.5)[47] against the GRCh38.p14 human genome reference. While following the QC, the EstMB cohort had an average of 30.63 ± 3.12 million reads per sample, the EstMB-deep cohort resulted in 106.70 ± 42.1 million reads per sample, indicating over three times deeper sequencing coverage.

EstMB MAGs metagenome assembly and binning

EstMB MAGs collection refers to all MAGs recovered from the EstMB-deep cohort, which comprises 1,878 samples sequenced at deep coverage. Reads were assembled into contigs with MEGAHIT (v1.2.9)[48]. Binning was performed separately for each sample from the EstMB-deep cohort. Contigs were binned using binners: MetaBAT (v2.15)[49], MaxBin (v2.2.7)[50], and VAMB (v3.0.7)[51], with further refining with DAS Tool (v1.1.4)[52]. MAGs resulting from this process form the EstMB MAGs collection. MAG quality, including completeness and contamination, was estimated using CheckM (v2.3.1)[53].

ESTrep MAGs collection

ESTrep MAGs collection refers to representative MAGs from the EstMB MAGs collection described earlier. Representative MAGs were selected from the EstMB MAGs collection by clustering MAGs from the EstMB MAGs collection on the species level (Average Nucleotide Identity, ANI index = 95) with dRep[54]. Taxonomy of all representative MAGs was assigned using GTDB-Tk (v2.3.0)[19], a software toolkit for assigning objective taxonomic classifications to bacterial and archaeal genomes based on the Genome Taxonomy Database (GTDB)[19,55]. If a MAG could not be taxonomically classified at the species level or higher using GTDB-Tk, this indicates that the genome does not closely match any existing entries in the GTDB reference database. Therefore, it was treated as a novel species. This criterion is widely used in studies involving MAG assembly[20,21]. MAG completeness and contamination were estimated using CheckM (v2.3.1)[53]. Ribosomal RNA genes were identified with Barrnap v0.8[56], and tRNA genes were predicted using tRNAscan-SE v2.0.0[57].

MAGs were classified into three quality tiers. High-quality (HQ) MAGs were defined as those with >90% completeness and <5% contamination. A subset of HQ MAGs meeting the Minimum Information about a Metagenome-Assembled Genome (MIMAG) standards—defined by the presence of ≥21 tRNA genes and a full complement of rRNA genes (5S, 16S, and 23S)—were named as HQ-mimag[18]. Medium-quality (MQ) MAGs were defined as those with >50% completeness and <10% contamination. MAGs not meeting HQ or MQ thresholds were classified as low-quality (LQ). Assembly statistics, including total assembly size, number of contigs, N50, and GC content, were calculated using SeqKit (v2.3.1)[58].

Population-based reference GUTrep MAGs collection

The GUTrep MAG collection is a non-redundant set of representative MAGs, created by combining MAGs from the current study (ESTrep MAGs) with those from the Unified Human Gastrointestinal Genome (UHGG) collection[35]. This integrated reference includes both population-specific taxa identified in our cohort and globally distributed species that, while detected in our samples, could not be completely assembled but are present in public databases. To remove redundancy, MAGs from both collections were clustered at the species level using an average nucleotide identity (ANI) threshold of 95% with dRep[54]. For each species cluster containing MAGs from both sources, the higher-quality MAG — based on completeness, contamination, and assembly statistics, was retained as the representative.

Species relative abundance and prevalence estimation

To evaluate species-level relative abundance and prevalence, we used all samples from the EstMB cohort, as it includes more samples than the EstMB-deep cohort. Deep sequencing is less critical for read profiling against an established reference database, whereas the larger sample size of the EstMB cohort is crucial for the subsequent association analyses. Reads were mapped against the GUTrep MAGs collection using CoverM[59] and aggregated into a relative abundance table with a custom Python script.

Species-level association study

For the association study, we used species-level relative abundance data from the EstMB cohort as previously described. We tested associations between centered log-ratio (CLR) transformed species abundances and participants' health status for common diseases in the Estonian population. We selected 33 diseases based on ICD10 codes from the electronic health records, each with at least 100 prevalent cases within the EstMB cohort. The remaining samples were considered as controls for each studied disease. From the 4,792 bacterial species in the GUTrep reference, we included 1,842 species with a prevalence >1% for the association analysis. Linear regression models, adjusted for BMI, gender, and age, were constructed to evaluate the association between the selected diseases and CLR-transformed species abundance. A stringent Bonferroni correction was applied to the significance level, adjusting for the number of analyzed species, resulting in a corrected alpha of 2.71 × 10^-5 (from an original alpha of 0.05).

Strain richness index estimation

The strain richness index quantifies the normalized number of strains observed for a given species per 100 assembled MAGs. To calculate this value, strain clusters were identified for each species using dRep[54] with a 99% ANI index threshold, considering only those species with more than 100 assembled MAGs. The number of strain clusters was then divided by the total number of MAGs for that species and multiplied by 100 to express the result as a percentage. The corresponding formula is shown below:

$$SRI = \frac{Number\ of\ strain\ clusters}{Total\ number\ of\ MAGs} * 100\%$$

Strain level association study

Candidate species for strain-level association analysis were selected based on two criteria: (1) a high number of reconstructed MAGs per species, and (2) the lowest strain richness index (SRI), indicating fewer strain clusters per species. These criteria were established to ensure sufficiently large sample sizes for robust microbiome-wide association studies (MWAS). Strain clusters were defined using dRep[54]. Based on these criteria, *O. splanchnicus* was selected as the candidate species for strain-level analysis. We examined *O. splanchnicus* strain-level population structure and focused on two out of the five most prevalent identified *O. splanchnicus* strain clusters. The remaining three clusters were excluded due to their presence in only a small subset of samples. For association analyses, we used presence/absence data from the two selected clusters. These clusters were designated as strain N1 (n = 974; original strain ID:

1_2.3.4.6.9) and strain N2 (n = 335; original strain ID: 1_1). Logistic regression analyses adjusted for sex, age, and BMI were performed for the same 33 diseases previously examined at the species level. To account for multiple testing, a stringent Bonferroni correction was applied, resulting in a corrected significance threshold of $\alpha = 1.5 \times 10^{-3}$ (original $\alpha = 0.05$).

For species cluster structure visualisation, we used ANIclustermap[60]. Pangenome analysis and pangenome visualisation were performed using the Anvi'o workflow with standard parameters[61].

ACKNOWLEDGEMENTS

We express our gratitude to all individuals who made valuable contributions to the Estonian Microbiome cohort, as well as to those who developed the software and databases utilized in this study. We thank Mait Metspalu, Andres Metspalu, Lili Milani and Tõnu Esko from the Estonian Biobank research team for the Estonian Biobank health data collection.

Data analysis was carried out in part in the High-Performance Computing Centre of the University of Tartu, and we thank the HPC Support Team of the Institute of Computer Science at the University of Tartu for delivering exceptional service and assistance in installing the necessary programs on the cluster. Special thanks to the Writing Retreat organized by the Institute of Genomics, University of Tartu, for providing a conducive atmosphere for writing this paper. We also thank MMHP (Million Metagenomes of Human Project) for providing the sequencing for the EstMB-deep cohort.

AUTHOR CONTRIBUTIONS

Conceptualization, K.P., K.L.K. and E.O.; Methodology, K.P. and O.A.; Data analysis, K.P.; Visualization, K.P.; prepared the first draft of the manuscript, which all authors reviewed and edited, K.P. All authors agreed to submit the manuscript, read and approved the final draft, and assumed full responsibility for its content, including the accuracy of the data.

DATA AND CODE AVAILABILITY

The source code for the analyses is available at GitHub:

https://github.com/Chartiza/EstMB_MAGs_db_paper.

Representative MAGs from the EstMB-deep cohort samples have been deposited in the European Nucleotide Archive under study accession PRJEB76860. The phenotype data contain sensitive information from healthcare registers, and they are available under restricted access through the Estonian biobank upon submission of a research plan and signing a data transfer agreement. All data access to the Estonian Biobank must follow the informed consent regulations of the Estonian Committee on Bioethics and Human Research, which are clearly described in the Data Access section at https://genomics.ut.ee/en/content/estonian-biobank. A preliminary request for raw metagenome and phenotype data must first be submitted via the email address releases@ut.ee.

FUNDING

This work was funded by the Estonian Research Council grant (PRG1414 to E.O.) and an EMBO Installation grant (No. 3573 to E.O.). EstMB sample collection was supported by the Estonian Center of Genomics/Roadmap II project No. 16-0125.

CONFLICTS OF INTEREST

The authors declare no competing interests.

REFERENCES

1. Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D., Dunn, M., Mkandawire, T.T., Zhu, A., Shao, Y., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat. Biotechnol. *37*, 186–192. https://doi.org/10.1038/s41587-018-0009-7.

2. Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. *37*, 179–185. https://doi.org/10.1038/s41587-018-0008-8.

3. Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504. https://doi.org/10.1038/s41586-019-0965-1.

4. Hildebrand, F. (2021). Ultra-resolution Metagenomics: When Enough Is Not Enough. mSystems *6*, e00881-21. https://doi.org/10.1128/mSystems.00881-21.

5. Chen, L., Wang, D., Garmaeva, S., Kurilshikov, A., Vich Vila, A., Gacesa, R., Sinha, T., Segal, E., Weersma, R.K., Wijmenga, C., et al. (2021). The long-term genetic stability and individual specificity of the human gut microbiome. Cell *184*, 2302-2315.e12. https://doi.org/10.1016/j.cell.2021.03.024.

6. Gacesa, R., Kurilshikov, A., Vich Vila, A., Sinha, T., Klaassen, M.A.Y., Bolte, L.A., Andreu-Sánchez, S., Chen, L., Collij, V., Hu, S., et al. (2022). Environmental factors shaping the gut microbiome in a Dutch population. Nature *604*, 732–739. https://doi.org/10.1038/s41586-022-04567-7.

7. Shilo, S., Bar, N., Keshet, A., Talmor-Barkan, Y., Rossman, H., Godneva, A., Aviv, Y., Edlitz, Y., Reicher, L., Kolobkov, D., et al. (2021). 10 K: a large-scale prospective longitudinal study in Israel. Eur. J. Epidemiol. *36*, 1187–1194. https://doi.org/10.1007/s10654-021-00753-5.

8. Reicher, L., Shilo, S., Godneva, A., Lutsker, G., Zahavi, L., Shoer, S., Krongauz, D., Rein, M., Kohn, S., Segev, T., et al. (2025). Deep phenotyping of health–disease continuum in the Human Phenotype Project. Nat. Med. https://doi.org/10.1038/s41591-025-03790-9.

9. Salosensaari, A., Laitinen, V., Havulinna, A.S., Meric, G., Cheng, S., Perola, M., Valsta, L., Alfthan, G., Inouye, M., Watrous, J.D., et al. (2021). Taxonomic signatures of cause-specific mortality risk in human gut microbiome. Nat. Commun. *12*, 2671. https://doi.org/10.1038/s41467-021-22962-y.

10. Aasmets, O., Krigul, K.L., Lüll, K., Metspalu, A., and Org, E. (2022). Gut metagenome associations with extensive digital health data in a volunteer-based Estonian microbiome cohort. Nat. Commun. *13*, 869. https://doi.org/10.1038/s41467-022-28464-9.

11. Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. Nature *555*, 210–215. https://doi.org/10.1038/nature25973.

12. Palmu, J., Salosensaari, A., Havulinna, A.S., Cheng, S., Inouye, M., Jain, M., Salido, R.A., Sanders, K., Brennan, C., Humphrey, G.C., et al. (2020). Association Between the Gut Microbiota and Blood Pressure in a Population Cohort of 6953 Individuals. J. Am. Heart Assoc. *9*, e016641. https://doi.org/10.1161/JAHA.120.016641.

13. Lin, Y.-T., Sayols-Baixeras, S., Baldanzi, G., Dekkers, K.F., Hammar, U., Nguyen, D., Nielsen, N., Eklund, A.C., Varotsis, G., Holm, J.B., et al. (2025). The association between the gut microbiome and 24-h blood pressure measurements in the SCAPIS study. Commun. Med. *5*, 276. https://doi.org/10.1038/s43856-025-00980-x.

14. Brushett, S., Gacesa, R., Vich Vila, A., Brandao Gois, M.F., Andreu-Sánchez, S., Swarte, J.C., Klaassen, M.A.Y., Collij, V., Sinha, T., Bolte, L.A., et al. (2023). Gut feelings: the relations between depression, anxiety, psychotropic drugs and the gut microbiome. Gut Microbes *15*, 2281360. https://doi.org/10.1080/19490976.2023.2281360.

15. Kartal, E., Schmidt, T.S.B., Molina-Montes, E., Rodríguez-Perales, S., Wirbel, J., Maistrenko, O.M., Akanni, W.A., Alashkar Alhamwe, B., Alves, R.J., Carrato, A., et al. (2022). A faecal

microbiota signature with high specificity for pancreatic cancer. Gut *71*, 1359–1372. https://doi.org/10.1136/gutjnl-2021-324755.

16. Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J.S., Voigt, A.Y., Palleja, A., Ponnudurai, R., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat. Med. *25*, 679–689. https://doi.org/10.1038/s41591-019-0406-6.

17. Erawijantari, P.P., Kartal, E., Liñares-Blanco, J., Laajala, T.D., Feldman, L.E., The FINRISK Microbiome DREAM Challenge and ML4Microbiome Communities, Carmona-Saez, P., Shigdel, R., Claesson, M.J., Bertelsen, R.J., et al. (2023). Microbiome-based risk prediction in incident heart failure: a community challenge. Preprint, https://doi.org/10.1101/2023.10.12.23296829 https://doi.org/10.1101/2023.10.12.23296829.

18. The Genome Standards Consortium, Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731. https://doi.org/10.1038/nbt.3893.

19. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics *36*, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.

20. Duru, I.C., Lecomte, A., Shishido, T.K., Laine, P., Suppula, J., Paulin, L., Scheperjans, F., Pereira, P.A.B., and Auvinen, P. (2024). Metagenome-assembled microbial genomes from Parkinson's disease fecal samples. Sci. Rep. *14*, 18906. https://doi.org/10.1038/s41598-024-69742-4.

21. Rodríguez-Cruz, U.E., Castelán-Sánchez, H.G., Madrigal-Trejo, D., Eguiarte, L.E., and Souza, V. (2024). Uncovering novel bacterial and archaeal diversity: genomic insights from metagenome-assembled genomes in Cuatro Cienegas, Coahuila. Front. Microbiol. *15*, 1369263. https://doi.org/10.3389/fmicb.2024.1369263.

22. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.

23. Van Rossum, T., Ferretti, P., Maistrenko, O.M., and Bork, P. (2020). Diversity within species: interpreting strains in microbiomes. Nat. Rev. Microbiol. *18*, 491–506. https://doi.org/10.1038/s41579-020-0368-1.

24. Hildebrand, F. (2021). Ultra-resolution Metagenomics: When Enough Is Not Enough. mSystems *6*, 10.1128/msystems.00881-21. https://doi.org/10.1128/msystems.00881-21.

25. Carrow, H.C., Batachari, L.E., and Chu, H. (2020). Strain diversity in the microbiome: Lessons from Bacteroides fragilis. PLOS Pathog. *16*, e1009056. https://doi.org/10.1371/journal.ppat.1009056.

26. Galperin, M.Y., Vera Alvarez, R., Karamycheva, S., Makarova, K.S., Wolf, Y.I., Landsman, D., and Koonin, E.V. (2025). COG database update 2024. Nucleic Acids Res. *53*, D356–D363. https://doi.org/10.1093/nar/gkae983.

27. Jin, D.-M., Morton, J.T., and Bonneau, R. (2024). Meta-analysis of the human gut microbiome uncovers shared and distinct microbial signatures between diseases. mSystems *9*, e00295-24. https://doi.org/10.1128/msystems.00295-24.

28. Maghini, D.G., Oduaran, O.H., Olubayo, L.A.I., Cook, J.A., Smyth, N., Mathema, T., Belger, C.W., Agongo, G., Boua, P.R., Choma, S.S.R., et al. (2025). Expanding the human gut microbiome atlas of Africa. Nature *638*, 718–728. https://doi.org/10.1038/s41586-024-08485-8.

29. Yang, C., Chowdhury, D., Zhang, Z., Cheung, W.K., Lu, A., Bian, Z., and Zhang, L. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. Comput. Struct. Biotechnol. J. *19*, 6301–6314. https://doi.org/10.1016/j.csbj.2021.11.028.

30. Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., and Nielsen, P.H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat. Biotechnol. *31*, 533–538. https://doi.org/10.1038/nbt.2579.

31. Vicedomini, R., Quince, C., Darling, A.E., and Chikhi, R. (2021). Strainberry: automated strain separation in low-complexity metagenomes using long reads. Nat. Commun. *12*, 4485. https://doi.org/10.1038/s41467-021-24515-9.

32. Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome. Nature *568*, 505–510. https://doi.org/10.1038/s41586-019-1058-x.

33. Feng, X., and Li, H. (2024). Evaluating and improving the representation of bacterial contents in long-read metagenome assemblies. Genome Biol. *25*, 92. https://doi.org/10.1186/s13059-024-03234-6.

34. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649-662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

35. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.

36. Leviatan, S., Shoer, S., Rothschild, D., Gorodetski, M., and Segal, E. (2022). An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species. Nat. Commun. *13*, 3863. https://doi.org/10.1038/s41467-022-31502-1.

37. Ravichandar, J.D., Rutherford, E., Chow, C.-E.T., Han, A., Yamamoto, M.L., Narayan, N., Kaplan, G.G., Beck, P.L., Claesson, M.J., Dabbagh, K., et al. (2022). Strain level and comprehensive microbiome analysis in inflammatory bowel disease via multi-technology meta-analysis identifies key bacterial influencers of disease. Front. Microbiol. *13*, 961020. https://doi.org/10.3389/fmicb.2022.961020.

38. Leimbach, A., Hacker, J., and Dobrindt, U. (2013). E. coli as an All-Rounder: The Thin Line Between Commensalism and Pathogenicity. In Between Pathogenicity and Commensalism Current Topics in Microbiology and Immunology., U. Dobrindt, J. H. Hacker, and C. Svanborg, eds. (Springer Berlin Heidelberg), pp. 3–32. https://doi.org/10.1007/82_2012_303.

39. Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat. Biotechnol. *41*, 1633–1644. https://doi.org/10.1038/s41587-023-01688-w.

40. Chen-Liaw, A., Aggarwala, V., Mogno, I., Haifer, C., Li, Z., Eggers, J., Helmus, D., Hart, A., Wehkamp, J., Lamousé-Smith, E.S.N., et al. (2024). Gut microbiota strain richness is species specific and affects engraftment. Nature. https://doi.org/10.1038/s41586-024-08242-x.

41. Metwaly, A., and Haller, D. (2019). Strain-Level Diversity in the Gut: The P. copri Case. Cell Host Microbe *25*, 349–350. https://doi.org/10.1016/j.chom.2019.02.006.

42. Fehlner-Peach, H., Magnabosco, C., Raghavan, V., Scher, J.U., Tett, A., Cox, L.M., Gottsegen, C., Watters, A., Wiltshire-Gordon, J.D., Segata, N., et al. (2019). Distinct Polysaccharide Utilization Profiles of Human Intestinal Prevotella copri Isolates. Cell Host Microbe *26*, 680-690.e5. https://doi.org/10.1016/j.chom.2019.10.013.

43. Gomez-Arango, L.F., Barrett, H.L., McIntyre, H.D., Callaway, L.K., Morrison, M., and Dekker Nitert, M. (2016). Increased Systolic and Diastolic Blood Pressure Is Associated With Altered Gut Microbiota Composition and Butyrate Production in Early Pregnancy. Hypertension *68*, 974–981. https://doi.org/10.1161/HYPERTENSIONAHA.116.07910.

44. Olson, N.D., Treangen, T.J., Hill, C.M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., and Pop, M. (2019). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief. Bioinform. *20*, 1140–1150. https://doi.org/10.1093/bib/bbx098.

45. Milani, L., Alver, M., Laur, S., Reisberg, S., Haller, T., Aasmets, O., Abner, E., Alavere, H., Allik, A., Annilo, T., et al. (2025). The Estonian Biobank's journey from biobanking to personalized medicine. Nat. Commun. *16*, 3270. https://doi.org/10.1038/s41467-025-58465-3.

46. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. Version 0121 Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

47. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

48. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

49. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

50. Wu, Y.-W., Simmons, B.A., and Singer, S.W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics *32*, 605–607. https://doi.org/10.1093/bioinformatics/btv638.

51. Nissen, J.N., Johansen, J., Allesøe, R.L., Sønderby, C.K., Armenteros, J.J.A., Grønbech, C.H., Jensen, L.J., Nielsen, H.B., Petersen, T.N., Winther, O., et al. (2021). Improved metagenome binning and assembly using deep variational autoencoders. Nat. Biotechnol. *39*, 555–560. https://doi.org/10.1038/s41587-020-00777-4.

52. Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., and Banfield, J.F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. *3*, 836–843. https://doi.org/10.1038/s41564-018-0171-1.

53. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

54. Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868. https://doi.org/10.1038/ismej.2017.126.

55. Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004. https://doi.org/10.1038/nbt.4229.

56. Seemann, T. (2013). Barrnap 0.8: Basic Rapid Ribosomal RNA Predictor. Available at. https://github.com/tseemann/barrnap

57. Chan, P.P., Lin, B.Y., Mak, A.J., and Lowe, T.M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. *49*, 9077–9096. https://doi.org/10.1093/nar/gkab688.

58. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

59. Aroney, S.T.N., Newell, R.J.P., Nissen, J.N., Camargo, A.P., Tyson, G.W., and Woodcroft, B.J. (2025). CoverM: read alignment statistics for metagenomics. Bioinformatics *41*, btaf147. https://doi.org/10.1093/bioinformatics/btaf147.

60. Musiał, K., Petruńko, L., and Gmiter, D. (2024). Simple approach to bacterial genomes comparison based on Average Nucleotide Identity (ANI) using fastANI and ANIclustermap. Acta Univ. Lodz. Folia Biol. Oecologica *18*, 66–71. https://doi.org/10.18778/1730-2366.18.10.

61. Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ *3*, e1319. https://doi.org/10.7717/peerj.1319.

SUPPLEMENTAL MATERIAL

Supplemental Figures (pantiukh_Suppl_Figures.docx). Fig. S1 to S4.

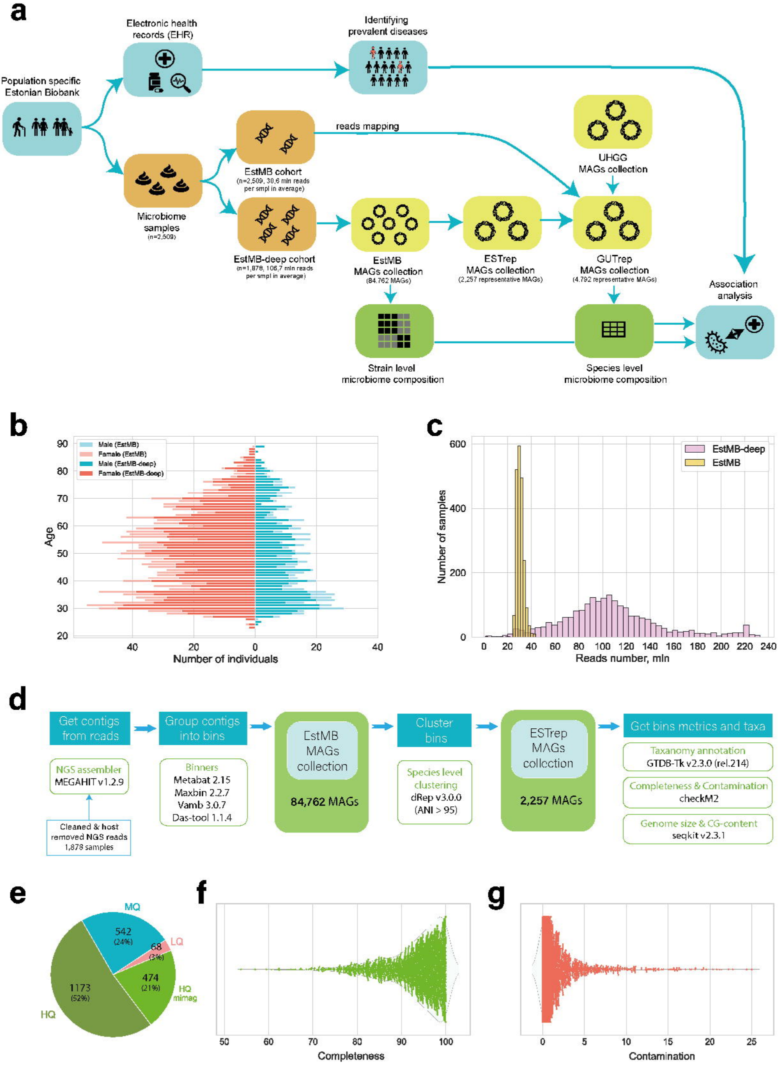
Supplemental Tables (pantiukh_Supplementary_Tables_S1-S7.xlsx). Tables S1 to S7.

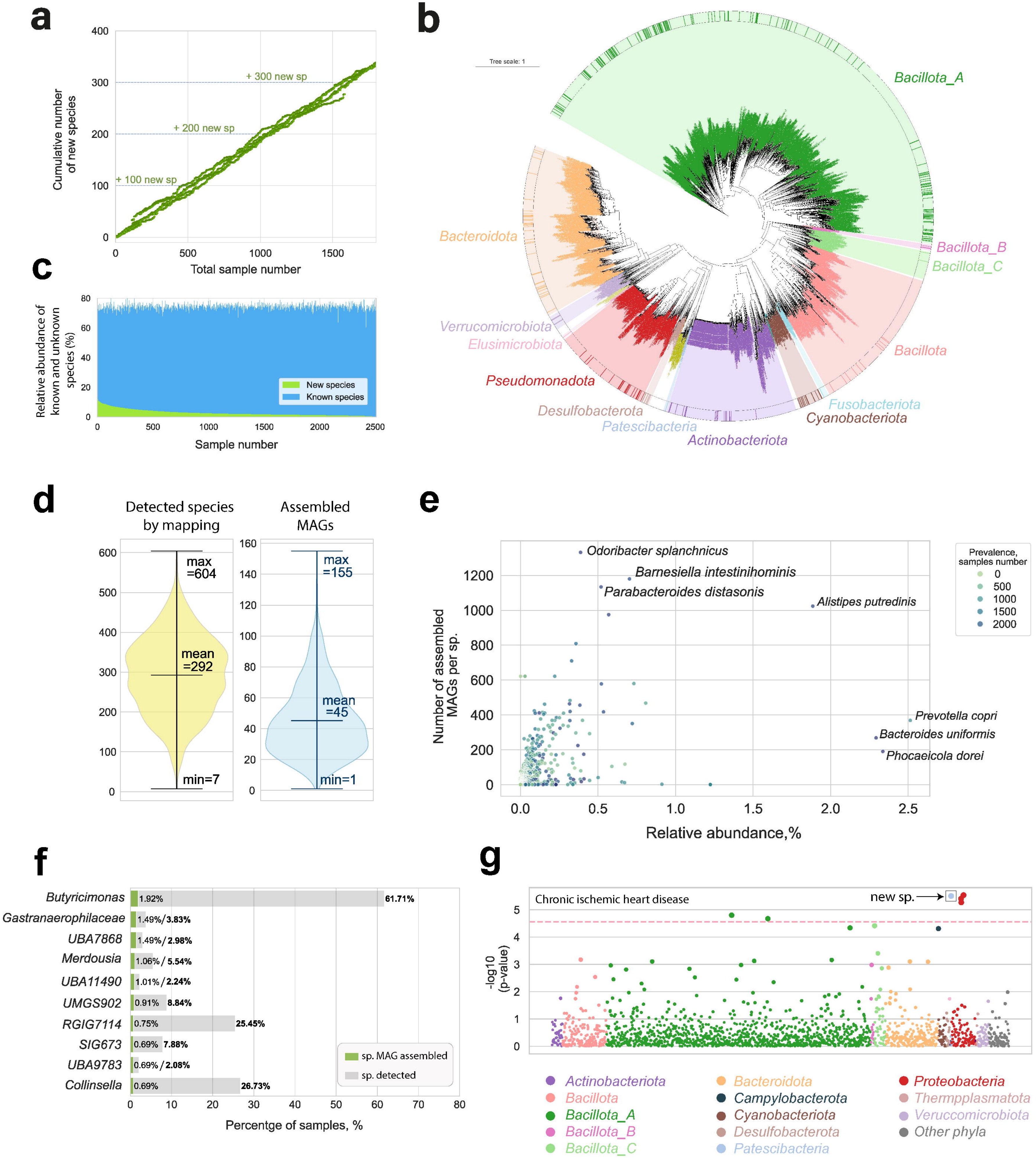
FIGURE LEGENDS

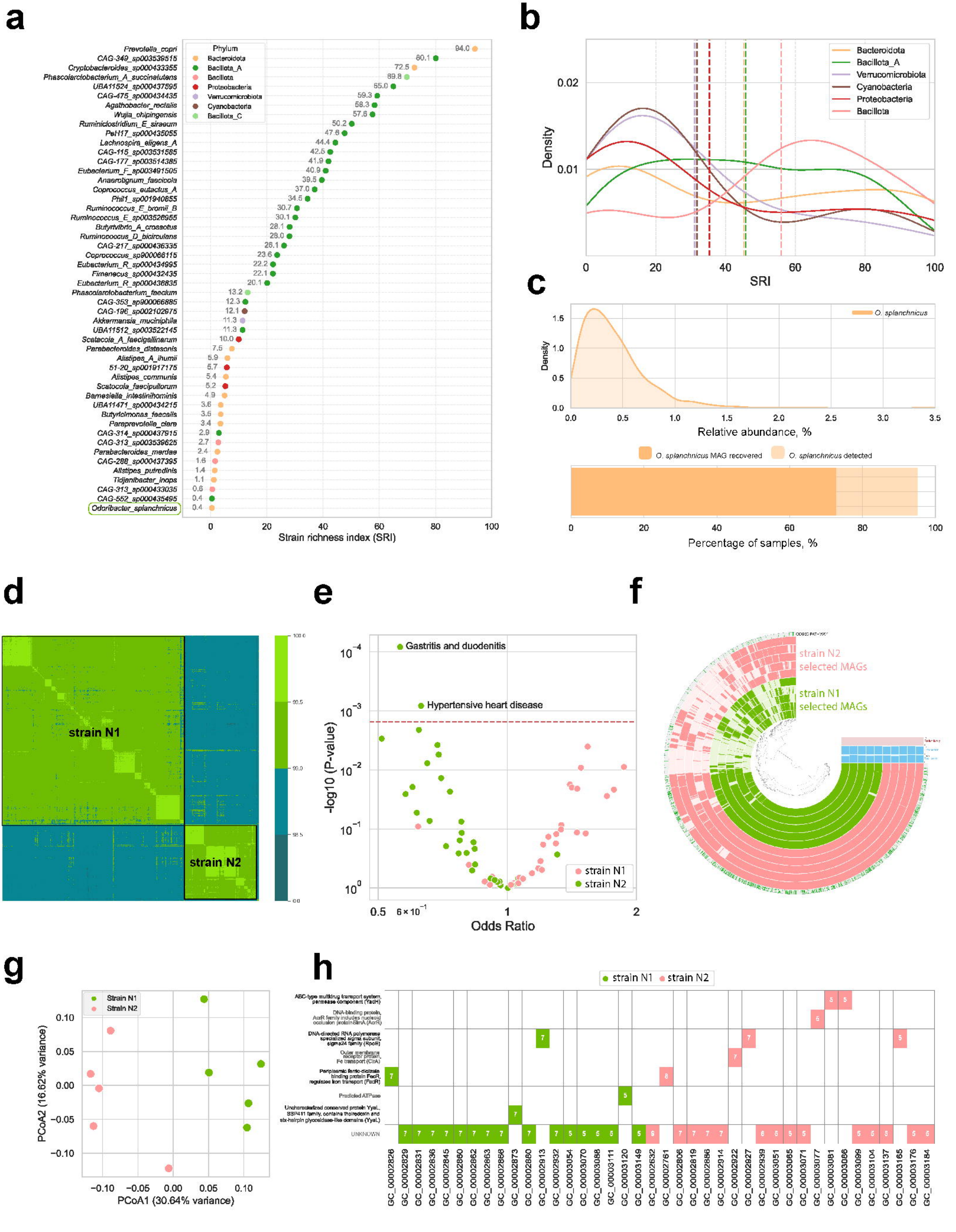
Figure 1. Study overview and cohort description. **a.** Study workflow overview. **b.** Distribution of age and gender across the Estonian Microbiome cohort (EstMB) and Estonian Microbiome deep sequencing cohort (EstMB-deep). **c.** Distribution of the number of reads across different genders of EstMB and EstMB-deep cohorts. **d.** Overview of the metagenome assembled genomes (MAGs) recovery pipeline. **e.** Quality distribution of ESTrep MAGs (HQ, high quality; MQ, medium quality; and LQ, low quality). **f.** Completeness (%) of MAGs in the ESTrep Collection. **g.** Contamination (%) of MAGs in the ESTrep collection.

Figure 2. Overview of species from the EstMB MAG collection. a. The relationship between the number of samples analyzed and the cumulative number of novel species identified. **b.** Phylogenetic tree of the ESTrep species. The inner circle displays a phylogenetic tree of species, with branches colored by phylum (according to the Genome Taxonomy Database (GTDB-Tk v2.3.0), the outer ring highlights novel species assembled in this study. **c.** Relative abundances of known and novel species. **d.** Average number of species detected by read mapping (yellow) versus number of recovered MAGs per sample (blue). **e.** Relationship between species prevalence, mean relative abundance, and number of assembled MAGs per species. **f.** Prevalence of the top 10 novel species with the highest number of recovered MAGs, comparing recovery by MAG assembly (green bars) and detection by read mapping (grey bars). **g.** Metagenome-wide association results between GUTrep species abundances and chronic ischemic heart disease. Each data point corresponds to a single species, with vertical position reflecting the log-transformed *p*-value from linear regression; significant associations for newly reconstructed species are highlighted with a box.

Figure 3. Within-species diversity and strain level analysis of *Odoribacter splanchnicus*. a. Strain richness index (SRI) for the top 50 species with the highest number of metagenome-assembled genomes. **b.** Distribution of SRI values across major gut bacteria phyla. **c.** *Odoribacter splanchnicus* relative abundance, number of recovered MAGs and prevalence across samples. **d.** Heatmap of Average Nucleotide Identity (ANI) values among *O. splanchnicus* MAGs, revealing two distinct strain clusters. **e.** Volcano plot of associations between the two major *O. splanchnicus* strains and 33 disease phenotypes. The red line indicates the Bonferroni-corrected significance threshold. **f.** Pan-genome analysis of five representative MAGs from each *O. splanchnicus* strain (N1 and N2). **g.** Principal coordinates analysis (PCoA) of *O. splanchnicus* representative MAGs based on predicted gene cluster presence/absence profiles. **h.** Gene clusters uniquely present in only one of the two major *O. splanchnicus* strains.


a
Population specific
Estonian Biobank
Electronic health
records (EHR)
Identifying
prevalent diseases
reads mapping
UHGG
MAGs collection
EstMB cohort
(n=2,509, 30,6 mln reads
per smpl in average)
Microbiome
samples
(n=2,509)
EstMB-deep cohort
(n=1,878, 106,7 mln reads
per smpl in average)
EstMB
MAGs collection
(84.762 MAGs)
ESTrep
MAGs collection
(2,257 representative MAGs)
GUTrep
MAGs collection
(4.792 representative MAGs)
Association
analysis
Strain level
microbiome composition
Species level
microbiome composition
b
Male (EstMB)
Famale (EstMB)
Male (EstMB-deep)
Famale (EstMB-deep)
Age
90
80
70
60
50
40
30
20
40
20
0
20
40
Number of individuals
c
EstMB-deep
EstMB
Number of samples
600
400
200
0
0
20
40
60
80
100
120
140
160
180
200
220
240
Reads number, mln
d
Get contigs
from reads
NGS assembler
MEGAHIT v1.2.9
Cleaned & host
removed NGS reads
1,878 samples
Group contigs
into bins
Binners
Metabat 2.15
Maxbin 2.2.7
Vamb 3.0.7
Das-tool 1.1.4
EstMB
MAGs
collection
84,762 MAGs
Cluster
bins
Species level
clustering
dRep v3.0.0
(ANI > 95)
ESTrep
MAGs
collection
2,257 MAGs
Get bins metrics and taxa
Taxanomy annotation
GTDB-Tk v2.3.0 (rel.214)
Completeness & Contamination
checkM2
Genome size & CG-content
seqkit v2.3.1
e
MQ
542
(24%)
LQ
68
(3%)
474
(21%)
HQ
mimag
1173
(52%)
HQ
f
50
60
70
80
90
100
Completeness
g
0
5
10
15
20
25
Contamination


a
Cumulative number of new species
Total sample number
+ 100 new sp
+ 200 new sp
+ 300 new sp
0
100
200
300
400
0
500
1000
1500
b
Tree scale: 1
Bacillota_A
Bacillota_B
Bacillota_C
Bacillota
Fusobacteriota
Cyanobacteriota
Actinobacteriota
Patescibacteria
Desulfobacterota
Pseudomonadota
Elusimicrobiota
Verrucomicrobiota
Bacteroidota
c
Relative abundance of known and unknown species (%)
Sample number
New species
Known species
0
20
40
60
80
0
500
1000
1500
2000
2500
d
Detected species by mapping
Assembled MAGs
max =604
mean =292
min=7
max =155
mean =45
min=1
0
100
200
300
400
500
600
0
20
40
60
80
100
120
140
160
e
Number of assembled MAGs per sp.
Relative abundance,%
Odoribacter splanchnicus
Barnesiella intestinihominis
Parabacteroides distasonis
Alistipes putredinis
Prevotella copri
Bacteroides uniformis
Phocaeicola dorei
Prevalence, samples number
0
500
1000
1500
2000
0
200
400
600
800
1000
1200
0.0
0.5
1.0
1.5
2.0
2.5
f
Butyricimonas 1.92% 61.71%
Gastranaerophilaceae 1.49%/3.83%
UBA7868 1.49%/2.98%
Merdousia 1.06%/5.54%
UBA11490 1.01%/2.24%
UMGS902 0.91% 8.84%
RGIG7114 0.75% 25.45%
SIG673 0.69% 7.88%
UBA9783 0.69%/2.08%
Collinsella 0.69% 26.73%
sp. MAG assembled
sp. detected
Percentge of samples, %
0
10
20
30
40
50
60
70
80
g
Chronic ischemic heart disease
new sp.
-log10 (p-value)
0
1
2
3
4
5
Actinobacteriota
Bacillota
Bacillota_A
Bacillota_B
Bacillota_C
Bacteroidota
Campylobacterota
Cyanobacteriota
Desulfobacterota
Patescibacteria
Proteobacteria
Thermpplasmatota
Veruccomicrobiota
Other phyla


a
Phylum
Bacteroidota
Bacillota_A
Bacillota
Proteobacteria
Verrucomicrobiota
Cyanobacteria
Bacillota_C
Prevotella_copri 94.0
CAG-349_sp003539515 80.1
Cryptobacteroides_sp000433355 72.5
Phascolarctobacterium_A_succinatutens 69.8
UBA11524_sp000437595 65.0
CAG-475_sp000434435 59.3
Agathobacter_rectalis 58.3
Wujia_chipingensis 57.6
Ruminiclostridium_E_siraeum 50.2
Pelt17_sp000435055 47.6
Lachnospira_eligens_A 44.4
CAG-115_sp003531585 42.5
CAG-177_sp003514385 41.9
Eubacterium_F_sp003491505 40.9
Anaerotignum_faecicola 39.5
Coprococcus_eutactus_A 37.0
Phil1_sp001940855 34.5
Ruminococcus_E_bromii_B 30.7
Ruminococcus_E_sp003526955 30.1
Butyrivibrio_A_crossotus 28.1
Ruminococcus_D_bicirculans 28.0
CAG-217_sp000436335 26.1
Coprococcus_sp900066115 23.6
Eubacterium_R_sp000434995 22.2
Firmicus_sp000432435 22.1
Eubacterium_R_sp000436835 20.1
Phascolarctobacterium_faecium 13.2
CAG-353_sp900066885 12.3
CAG-196_sp002102975 12.1
Akkermansia_muciniphila 11.3
UBA11512_sp003522145 11.3
Scatocola_A_faecigallinarum 10.0
Parabacteroides_distasonis 7.5
Alistipes_A_ihumii 5.9
51-20_sp001917175 5.7
Alistipes_communis 5.4
Scatocola_faecipullorum 5.2
Barnesiella_intestinihominis 4.9
UBA11471_sp000434215 3.6
Butyricimonas_faecalis 3.5
Paraprevotella_clara 3.4
CAG-314_sp000437915 2.9
CAG-313_sp003539625 2.7
Parabacteroides_merdae 2.4
CAG-288_sp000437395 1.6
Alistipes_putredinis 1.4
Tidjanibacter_inops 1.1
CAG-313_sp000433035 0.6
CAG-552_sp000435495 0.4
Odoribacter_splanchnicus 0.4
0
20
40
60
80
100
Strain richness index (SRI)
b
Bacteroidota
Bacillota_A
Verrucomicrobiota
Cyanobacteria
Proteobacteria
Bacillota
Density
0.02
0.01
0
20
40
60
80
100
SRI
c
O. splanchnicus
Density
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Relative abundance, %
O. splanchnicus MAG recovered
O. splanchnicus detected
0
20
40
60
80
100
Percentage of samples, %
d
strain N1
strain N2
100.0
99.5
99.0
98.5
0.0
e
Gastritis and duodenitis
Hypertensive heart disease
-log10 (P-value)
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
strain N1
strain N2
0.5
6×10^{-1}
1
2
Odds Ratio
f
COG20 PATHWAY
strain N2
selected MAGs
strain N1
selected MAGs
Redundancy
Completion
Total length
g
Strain N1
Strain N2
PCoA2 (16.62% variance)
0.10
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10
PCoA1 (30.64% variance)
h
strain N1
strain N2
ABC-type multidrug transport system, permease component (YadH)
DNA-binding protein, AcrR family includes nucleoid occlusion protein SlmA (AcrR)
DNA-directed RNA polymerase specialized sigma subunit, sigma24 family (RpoE)
Outer membrane receptor protein, Fe transport (CirA)
Periplasmic ferric-dicitrate binding protein FecR, regulates iron transport (FecR)
Predicted ATPase
Uncharacterized conserved protein YyaL, SSP411 family, contains thoiredoxin and six-hairpin glycosidase-like domains (YyaL)
UNKNOWN
GC_00002826
GC_00002829
GC_00002831
GC_00002836
GC_00002845
GC_00002860
GC_00002862
GC_00002863
GC_00002868
GC_00002873
GC_00002880
GC_00002913
GC_00002932
GC_00003054
GC_00003070
GC_00003088
GC_00003111
GC_00003120
GC_00003149
GC_00002632
GC_00002761
GC_00002806
GC_00002819
GC_00002886
GC_00002914
GC_00002922
GC_00002927
GC_00002939
GC_00003051
GC_00003065
GC_00003071
GC_00003077
GC_00003081
GC_00003086
GC_00003099
GC_00003104
GC_00003137
GC_00003165
GC_00003176
GC_00003184
7
5
6