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.

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

67 In the present study, we leveraged deep metagenomic sequencing of a population-based
68 Estonian microbiome-deep (EstMB-deep) cohort to assemble a comprehensive collection of
69 metagenome-assembled genomes (MAGs), substantially expanding the reference database of
70 human gut microbes with hundreds of previously uncharacterized species. We integrated these
71 population-specific MAGs with public reference data and conducted association analyses with
72 33 prevalent diseases. To systematically assess within-species diversity, we developed a novel
73 Strain Richness Index (SRI). We demonstrated its utility by identifying strain-specific disease
74 associations in *Odoribacter splanchnicus* that were not apparent at the species level. Our
75 findings demonstrate that genome-resolved, strain-aware microbiome profiling can uncover
76 novel disease-linked microbial signatures beyond that remain hidden using conventional
77 approaches.

78 **RESULTS**

79 **Study design and cohort overview**

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

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

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

94 To characterize population-specific microbes and expand publicly available human gut
95 microbiome databases with microbial genomes from the Estonian population, we performed *de*
96 *novo* Metagenome Assembled Genomes (MAGs) reconstruction from all 1878 samples in the
97 EstMB-deep cohort. The MAG reconstruction pipeline is summarized in **Figure 1d**. We
98 successfully reconstructed 84,762 MAGs from EstMB-deep, with an average of 45.13 MAGs per
99 sample. Among these, 42,049 (49.61%) were high quality (HQ) MAGs, i.e. MAGs with
100 completeness > 90% and contamination < 5%; 26,806 (31.63%) were medium quality (MQ)
101 MAGs, i.e. MAGs with completeness > 50% and contamination < 10%; and all others 15,907
102 (18.77%) were low quality (LQ) MAGs according to CheckM **(Figure S1)**. To describe the

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

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

113 **MAG assembly remains essential for detecting novel population-specific species**

114 Next, we identified previously uncharacterized species within the ESTrep MAGs collection.
115 MAGs were categorized as novel species if their taxonomic classification at the species level or
116 higher couldn't be assigned using the GTDB-Tk[19], a common approach for evaluating whether a
117 newly reconstructed MAG represents a new species[20,21]. Of the 2,257 representative MAGs, 353
118 (15.64%) were classified as novel. Among these, 231 MAGs (65.44%) had > 90% and < 5%
119 contamination, and 57 (16.15%) also contained rRNA and tRNA genes, meeting the MIMAG
120 guidelines for high-quality MAGs[18]. We observed a strong correlation between the number of
121 novel species discovered and the number of samples analyzed (R^2 = 0.97). Specifically, for every
122 500 samples, approximately 102 novel species were identified (**Figure 2a**). As we have not
123 observed any indication of a plateau with the current sample size, we expect that analysing
124 more samples will reveal additional species.

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

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

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

134 As the success of metagenome assembly and genome reconstruction depends on multiple
135 technical and analytical factors, we did not expect to recover all microbial genomes present in
136 the gut. Therefore, we constructed an integrated species-level reference by combining newly
137 reconstructed MAGs from the Estonian population with publicly available human gut-associated
138 species. This integrated reference, called the GUTrep collection (**Figure 1a**), was generated by
139 deduplicating the ESTrep MAGs collection and UHGG MAGs collection[22] at a 95.0% ANI
140 threshold, retaining the highest-quality MAG for each species. When two MAGs from one species
141 were present (one from ESTrep and one from UHGG), the highest-quality MAG was selected for
142 the final collection. The final GUTrep database comprises 4,792 species, of which 3,285
143 (68.55%) originated from UHGG and 1,507 (31.45%) from ESTrep, thereby substantially
144 improving the UHGG dataset. Notably, the ESTrep contribution includes 353 novel species, 607
145 known species absent from UHGG and 927 higher-quality MAGs already represented in UHGG.

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

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

165 **Newly assembled species provide valuable input for association studies**

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

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

181 3.13×10^{-6}) (**Figure 2g, Table S4**). These findings emphasize the importance of population-
182 specific reference databases for detecting disease-associated microbiome changes. However,
183 further studies are needed to confirm whether these associations generalize beyond the
184 Estonian cohort.

185 **MAG data enables strain-level diversity analysis across species**

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

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

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

201 The SRI values varied widely, ranging from 0.4 to 100 (**Figure 3a, Table S6**), indicating
202 substantial differences in strain diversity across species in the population. *Odoribacter*
203 *splanchnicus* exhibited the lowest SRI (0.4), with one strain per ~250 MAGs, reflecting low
204 diversity despite high prevalence. In contrast, *Prevotella copri* had one of the highest SRIs (94.0),
205 consistent with its well-documented heterogeneity, where nearly every MAG represents a

206 unique strain, making it difficult to conduct strain-level association analysis in the population.
207 Interestingly, *Alistipes_A* genus appeared in both the lowest and highest SRI groups.

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

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

217 In order to demonstrate the value of strain-level MWAS analysis, we selected *O. splanchnicus*
218 due to its low SRI (SRI = 0.4) and high prevalence (detected in 96.14% of samples, assembled in
219 72.68%, **Figure 3c**). Among its MAGs, we identified four distinct strain clusters, two of which
220 were rare (found in 2 and 19 samples, respectively). Therefore, we focused on the two major
221 clusters with high case numbers: strain N1(n=974 samples, original strain ID: 1_2.3.4.6.9) and
222 strain N2 (n=335 samples, original strain ID: 1_1) (**Figure 3d**).

223 Logistic regression models adjusted for BMI, gender, and age were used to assess the
224 association between the presence or absence of *O. splanchnicus* strains N1 and N2 and the same
225 33 diseases previously analyzed at the species-level MWAS. Our analysis identified a significant
226 association between the presence of strain N1 and two different diseases - gastritis and
227 duodenitis, and hypertensive heart disease (**Figure 3e**). The odds ratio for strain N1 was less
228 than 1 in both diseases (gastritis and duodenitis OR = 0,56, hypertensive heart disease OR =
229 0,63), indicating that its presence is associated with a reduced likelihood of having the disease.
230 Notably, these associations were not detected at the species level, highlighting the added
231 resolution of strain-level analysis.

232 To explore functional differences, we performed a pan-genome analysis of strains N1 and N2.
233 We carried out Principal coordinate analysis of predicted gene cluster presence/absence, which
234 showed clear separation between the strains (**Figure 3f**). We identified that the two strains
235 formed distinct clusters, indicating clear genomic differentiation based on gene content (**Figure**
236 **3g**), with 40 gene clusters unique to one of the two (**Figure 3h, Table S7**). While most encoded
237 hypothetical or uncharacterized proteins, some were annotated with putative functions based
238 on the Clusters of Orthologous Genes (COG20)[26]. Strain N2 harboured a broader repertoire of
239 genes associated with stress response, iron acquisition, and antimicrobial resistance—traits
240 consistent with enhanced survival in inflammatory gastrointestinal environments. These
241 included elevated copy numbers of the extracytoplasmic stress sigma factor RpoE (σ^E), iron
242 transport components FecR and CirA, and multidrug resistance elements such as AcrR and an
243 ABC-type efflux pump (YadH). In contrast, strain N1 was enriched for redox maintenance
244 proteins such as YyaL/DsbD, suggesting a distinct strategy centered on oxidative stress
245 mitigation.

246 **DISCUSSION**

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

258 at the species level and provided functional genomic insights that may explain these
259 associations.

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

281 A common argument against investing in resource-intensive *de novo* assembly is that the
282 rapidly growing and regularly updated public gut genome catalogues increasingly capture
283 known microbial diversity, suggesting that most gut species will soon be represented, and
284 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.

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

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

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

359 Our study also has some limitations that should be considered. First, the reliance on short-read
360 sequencing, may reduce assembly contiguity and strain resolution compared to long-read
361 approaches[44]. Although long read technologies offer higher genomic completeness , their
362 current cost limits their use in large-scale population studies. A practical compromise could
363 involve using short-read sequencing for most samples and applying long-read sequencing to key
364 or novel taxa. Second, the observed associations are correlative, and further validation in

365 longitudinal and experimental studies is needed to assess causality. Third, our strain-level
366 analysis focused on one species due to sample representation and analytical feasibility, and
367 broader application across species remains a key next step. Finally, while functional differences
368 between strains were identified, interpretation was limited by the prevalence of unannotated
369 genes. Future improvements in genome annotation and cross-cohort replication will be
370 essential to build on these findings. Despite these challenges, our findings demonstrate the
371 value of population-scale metagenomics in uncovering novel microbial diversity and strain-level
372 functional signatures relevant to human health.

373 **CONCLUSION**

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

378 **METHODS**

379 **Estonian Microbiome cohort description**

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

390 subset of stool samples from the EstMB cohort that have been resequenced with over three
391 times deeper coverage (N = 1878).

392 **Microbiome sample collection and DNA extraction**

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

405 **Shotgun metagenomic sequencing**

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

415 **EstMB MAGs metagenome assembly and binning**

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

423 **ESTrep MAGs collection**

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

436 MAGs were classified into three quality tiers. High-quality (HQ) MAGs were defined as those
437 with >90% completeness and <5% contamination. A subset of HQ MAGs meeting the Minimum
438 Information about a Metagenome-Assembled Genome (MIMAG) standards—defined by the
439 presence of ≥21 tRNA genes and a full complement of rRNA genes (5S, 16S, and 23S)—were
440 named as HQ-mimag[18]. Medium-quality (MQ) MAGs were defined as those with >50%
441 completeness and <10% contamination. MAGs not meeting HQ or MQ thresholds were classified

442 as low-quality (LQ). Assembly statistics, including total assembly size, number of contigs, N50,
443 and GC content, were calculated using SeqKit (v2.3.1)[58].

444 **Population-based reference GUTrep MAGs collection**

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

454 **Species relative abundance and prevalence estimation**

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

461 **Species-level association study**

462 For the association study, we used species-level relative abundance data from the EstMB cohort
463 as previously described. We tested associations between centered log-ratio (CLR) transformed
464 species abundances and participants' health status for common diseases in the Estonian
465 population. We selected 33 diseases based on ICD10 codes from the electronic health records,
466 each with at least 100 prevalent cases within the EstMB cohort. The remaining samples were
467 considered as controls for each studied disease. From the 4,792 bacterial species in the GUTrep

468 reference, we included 1,842 species with a prevalence >1% for the association analysis. Linear
469 regression models, adjusted for BMI, gender, and age, were constructed to evaluate the
470 association between the selected diseases and CLR-transformed species abundance. A stringent
471 Bonferroni correction was applied to the significance level, adjusting for the number of analyzed
472 species, resulting in a corrected alpha of 2.71 × 10^-5 (from an original alpha of 0.05).

473 **Strain richness index estimation**

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

480
$$SRI = \frac{Number\ of\ strain\ clusters}{Total\ number\ of\ MAGs} * 100\%$$

481 **Strain level association study**

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

492 1_2.3.4.6.9) and strain N2 (n = 335; original strain ID: 1_1). Logistic regression analyses
493 adjusted for sex, age, and BMI were performed for the same 33 diseases previously examined at
494 the species level. To account for multiple testing, a stringent Bonferroni correction was applied,
495 resulting in a corrected significance threshold of $\alpha = 1.5 \times 10^{-3}$ (original $\alpha = 0.05$).

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

499 **ACKNOWLEDGEMENTS**

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

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

511 **AUTHOR CONTRIBUTIONS**

512 Conceptualization, K.P., K.L.K. and E.O.; Methodology, K.P. and O.A.; Data analysis, K.P.;
513 Visualization, K.P.; prepared the first draft of the manuscript, which all authors reviewed and
514 edited, K.P. All authors agreed to submit the manuscript, read and approved the final draft, and
515 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.

542 3. Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D.,
543 and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*,
544 499–504. https://doi.org/10.1038/s41586-019-0965-1.

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

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

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

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

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

560 9. Salosensaari, A., Laitinen, V., Havulinna, A.S., Meric, G., Cheng, S., Perola, M., Valsta, L., Alfthan,
561 G., Inouye, M., Watrous, J.D., et al. (2021). Taxonomic signatures of cause-specific mortality
562 risk in human gut microbiome. Nat. Commun. *12*, 2671. https://doi.org/10.1038/s41467-
563 021-22962-y.

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

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

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

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

578 14. Brushett, S., Gacesa, R., Vich Vila, A., Brandao Gois, M.F., Andreu-Sánchez, S., Swarte, J.C.,
579 Klaassen, M.A.Y., Collij, V., Sinha, T., Bolte, L.A., et al. (2023). Gut feelings: the relations
580 between depression, anxiety, psychotropic drugs and the gut microbiome. Gut Microbes *15*,
581 2281360. https://doi.org/10.1080/19490976.2023.2281360.

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

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

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

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

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

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

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

607 21. Rodríguez-Cruz, U.E., Castelán-Sánchez, H.G., Madrigal-Trejo, D., Eguiarte, L.E., and Souza, V.
608 (2024). Uncovering novel bacterial and archaeal diversity: genomic insights from
609 metagenome-assembled genomes in Cuatro Cienegas, Coahuila. Front. Microbiol. *15*,
610 1369263. https://doi.org/10.3389/fmicb.2024.1369263.

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

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

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

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

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

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

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

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

636 30. Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., and Nielsen, P.H.
637 (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage
638 binning of multiple metagenomes. Nat. Biotechnol. *31*, 533–538.
639 https://doi.org/10.1038/nbt.2579.

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

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

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

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

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

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

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

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

669 39. Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P.,
670 Dubois, L., Huang, K.D., Thomas, A.M., et al. (2023). Extending and improving metagenomic
671 taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat. Biotechnol. *41*,
672 1633–1644. https://doi.org/10.1038/s41587-023-01688-w.

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

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

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

682 43. Gomez-Arango, L.F., Barrett, H.L., McIntyre, H.D., Callaway, L.K., Morrison, M., and Dekker
683 Nitert, M. (2016). Increased Systolic and Diastolic Blood Pressure Is Associated With Altered
684 Gut Microbiota Composition and Butyrate Production in Early Pregnancy. Hypertension *68*,
685 974–981. https://doi.org/10.1161/HYPERTENSIONAHA.116.07910.

686 44. Olson, N.D., Treangen, T.J., Hill, C.M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., and Pop, M.
687 (2019). Metagenomic assembly through the lens of validation: recent advances in assessing
688 and improving the quality of genomes assembled from metagenomes. Brief. Bioinform. *20*,
689 1140–1150. https://doi.org/10.1093/bib/bbx098.

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

693 46. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. Version
694 0121 Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

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

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

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

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

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

710 52. Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., and Banfield, J.F.
711 (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring
712 strategy. Nat. Microbiol. *3*, 836–843. https://doi.org/10.1038/s41564-018-0171-1.

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

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

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

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

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

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

731 59. Aroney, S.T.N., Newell, R.J.P., Nissen, J.N., Camargo, A.P., Tyson, G.W., and Woodcroft, B.J.
732 (2025). CoverM: read alignment statistics for metagenomics. Bioinformatics *41*, btaf147.
733 https://doi.org/10.1093/bioinformatics/btaf147.

734 60. Musiał, K., Petruńko, L., and Gmiter, D. (2024). Simple approach to bacterial genomes
735 comparison based on Average Nucleotide Identity (ANI) using fastANI and ANIclustermap.
736 Acta Univ. Lodz. Folia Biol. Oecologica *18*, 66–71. https://doi.org/10.18778/1730-
737 2366.18.10.

738 61. Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O.
739 (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ *3*,
740 e1319. https://doi.org/10.7717/peerj.1319.

741

742

743 **SUPPLEMENTAL MATERIAL**

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

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

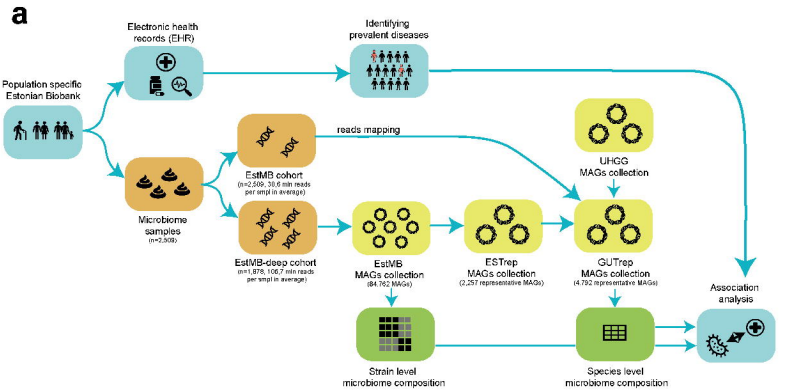
746 **FIGURE LEGENDS**

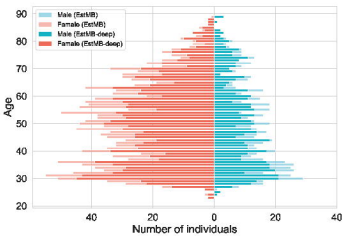
747 **Figure 1.** Study overview and cohort description. **a.** Study workflow overview. **b.** Distribution of
748 age and gender across the Estonian Microbiome cohort (EstMB) and Estonian Microbiome deep
749 sequencing cohort (EstMB-deep). **c.** Distribution of the number of reads across different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

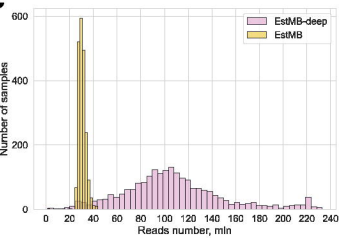
754 **Figure 2. Overview of species from the EstMB MAG collection. a.** The relationship between
755 the number of samples analyzed and the cumulative number of novel species identified. **b**.
756 Phylogenetic tree of the ESTrep species. The inner circle displays a phylogenetic tree of species,
757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

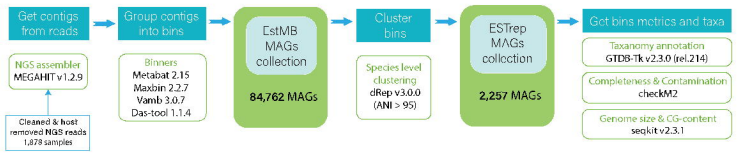
768 **Figure 3. Within-species diversity and strain level analysis of *Odoribacter splanchnicus*. a.**
769 Strain richness index (SRI) for the top 50 species with the highest number of metagenome-
770 assembled genomes. **b.** Distribution of SRI values across major gut bacteria phyla. **c.** *Odoribacter*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

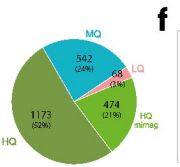
779

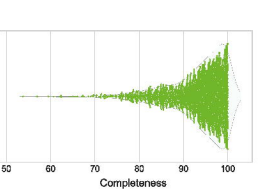

a
Electronic health records (EHR)
Identifying prevalent diseases
Population specific Estonian Biobank
reads mapping
UHGG MAGs collection
EstMB cohort
(n=2,509, 30.6 mln reads per smpl in average)
Microbiome samples
(n=2,509)
EstMB-deep cohort
(n=1,878, 106.7 mln reads per smpl in average)
EstMB MAGs 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

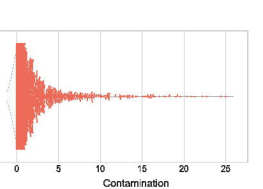

b
Male (EstMB)
Female (EstMB)
Male (EstMB-deep)
Female (EstMB-deep)
Age
Number of individuals

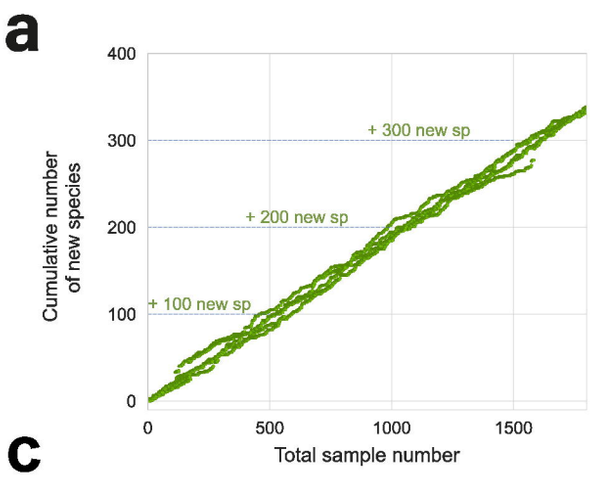

c
EstMB-deep
EstMB
Number of samples
Reads number, mln

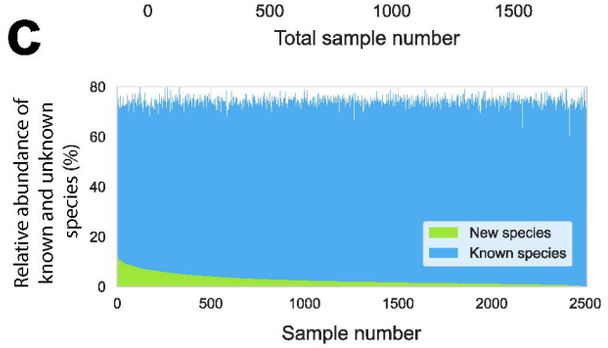

d
Get contigs from reads
NGS assembler
MEGAHIT v1.2.9
Cleaned & host removed NGS reads
1,878 samples
Group contigs into bins
Binners
Metabat 2.15
Maxbin 2.2.7
Vamb 3.0.7
Das-tool 1.1.4
EstMB MAGs collection
84,762 MAGs
Cluster bins
Species level clustering
dRep v3.0.0
(ANI > 95)
ESTrep MAGs collection
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

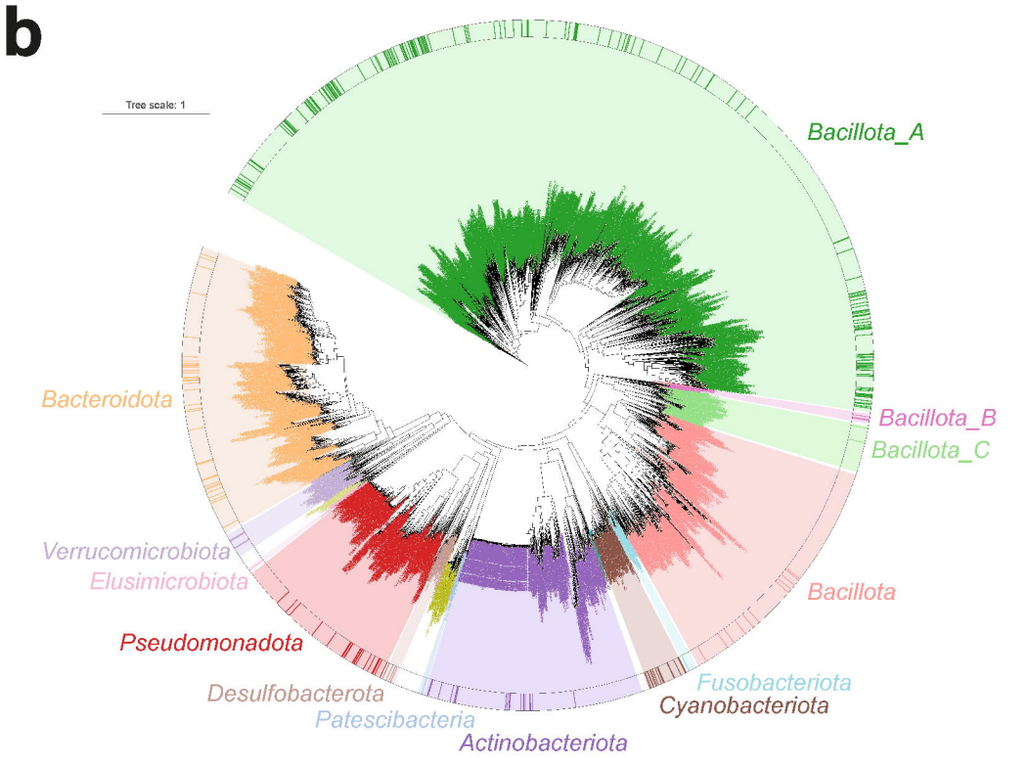

e
MQ
542
(24%)
LQ
68
(3%)
474
(21%)
HQ mimag
1173
(52%)
HQ

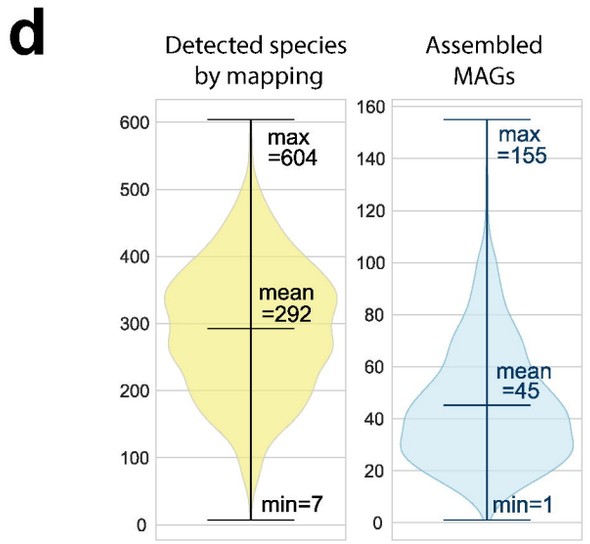

f
Completeness

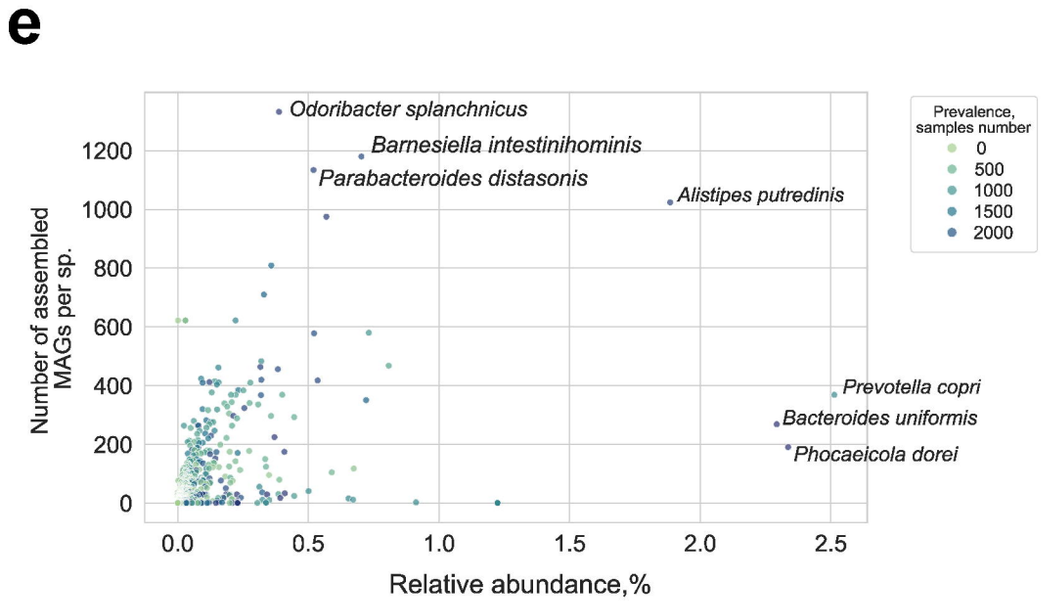

g
Contamination

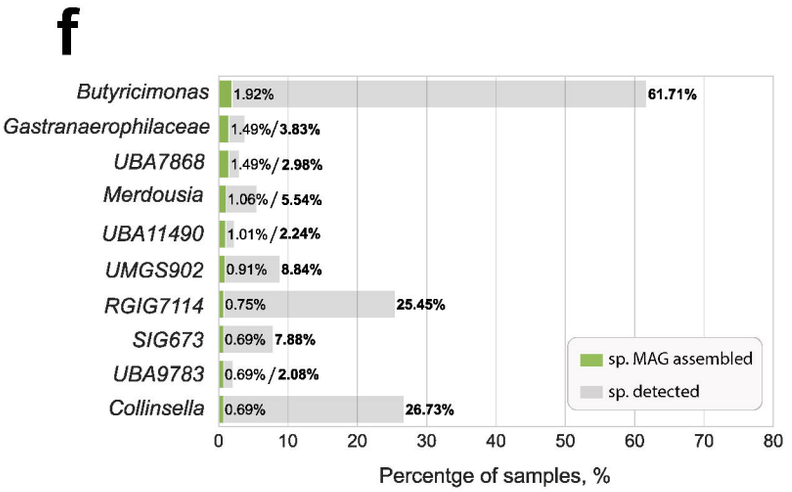

a
Cumulative number of new species
Total sample number
+ 100 new sp
+ 200 new sp
+ 300 new sp

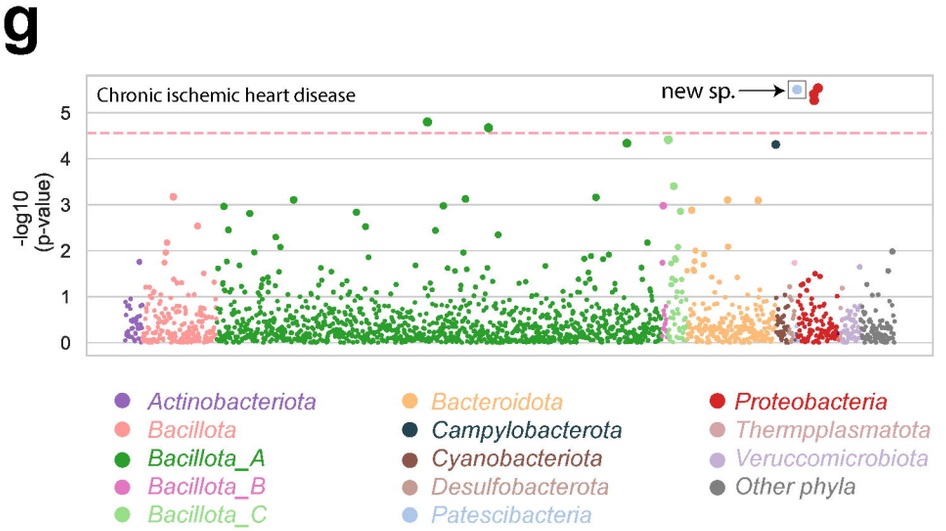

c
Relative abundance of known and unknown species (%)
Sample number
New species
Known species

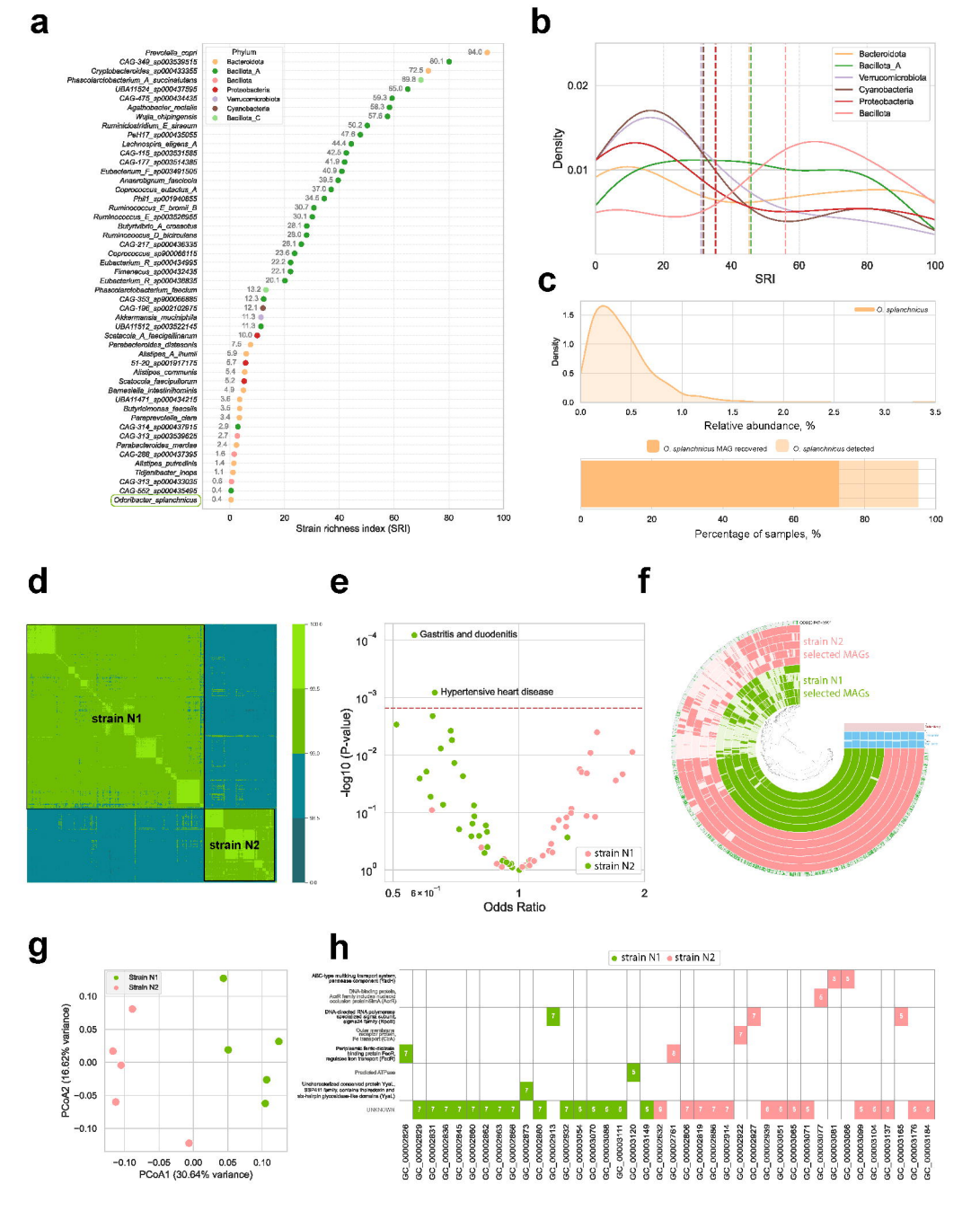

b
Tree scale: 1
Bacillota_A
Bacillota_B
Bacillota_C
Bacteroidota
Verrucomicrobiota
Elusimicrobiota
Pseudomonadota
Desulfobacterota
Patescibacteria
Actinobacteriota
Cyanobacteriota
Fusobacteriota
Bacillota


d
Detected species by mapping
Assembled MAGs
max =604
mean =292
min=7
max =155
mean =45
min=1


e
Odoribacter splanchnicus
Barnesiella intestinihominis
Parabacteroides distasonis
Alistipes putredinis
Prevotella copri
Bacteroides uniformis
Phocaeicola dorei
Number of assembled MAGs per sp.
Relative abundance,%
Prevalence, samples number
0
500
1000
1500
2000


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, %


g
Chronic ischemic heart disease
new sp.
-log10 (p-value)
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
CAG-349_sp000393615
Cryptobacteroides_sp000433355
Phascolarctobacterium_A_succinatutans
UBA11524_sp000437595
CAG-476_sp000434435
Agathobacter_rectalis
Wujia_chipingensis
Ruminiclostridium_E_siraeum
Pel117_sp000435055
Lachnospira_eligens_A
CAG-115_sp000351585
CAG-177_sp000314385
Eubacterium_F_sp003491505
Anaerotignum_faecicola
Coprococcus_eutactus_A
Phil1_sp001940655
Ruminococcus_E_bromii_B
Ruminococcus_E_sp003526955
Butyrivibrio_A_crossotus
Ruminococcus_D_bicirculens
CAG-217_sp000436335
Coprococcus_sp900066115
Eubacterium_R_sp000436995
Finenecus_sp000432435
Eubacterium_R_sp000436835
Phascolarctobacterium_faecium
CAG-353_sp900066985
CAG-196_sp002102675
Akkermansia_muciniphila
UBA11512_sp003522145
Scatocola_A_faecipallilarum
Parabacteroides_distasonis
Alistipes_A_jhumi
51-20_sp001917175
Alistipes_communis
Scatocola_fascipullorum
Barnesiella_intestinihominis
UBA11471_sp000434215
Butyricimonas_faecalis
Paraprevotella_clara
CAG-314_sp000437615
CAG-313_sp003539625
Parabacteroides_merdae
CAG-288_sp000437395
Alistipes_putredinis
Tidjanibacter_inopa
CAG-313_sp000433035
CAG-552_sp000435495
Odoribacter_splanchnicus
94.0
80.1
72.5
69.8
65.0
59.3
58.3
57.6
50.2
47.6
44.4
42.6
41.9
40.9
39.5
37.0
34.6
30.7
30.1
28.1
28.0
26.1
23.6
22.2
22.1
20.1
13.2
12.3
12.1
11.3
11.3
10.0
7.5
5.9
5.7
5.4
5.2
4.9
3.6
3.5
3.4
2.9
2.7
2.4
1.6
1.4
1.1
0.6
0.4
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
strain N2
selected MAGs
strain N1
selected MAGs
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 (TkoH)
DNA-binding protein, AsnR family inducible nucleoid-associated protein (HlyW)
DNA-directed RNA polymerase specialized sigma subunit, sigma24 family (RpoE)
Outer membrane receptor protein, Fe transport (CirA)
Periplasmic ferric-chelate binding protein FeuA, regulates iron transport (FeuA)
Predicted ATPase
Uncharacterized conserved protein YpsL, SSF141 family contains transposon and stx-hairpin glycosidase-like domains (YpsL)
UNKNOWN
GC_00002826
GC_00002829
GC_00002831
GC_00002836
GC_00002845
GC_00002850
GC_00002852
GC_00002853
GC_00002866
GC_00002873
GC_00002880
GC_00002913
GC_00002922
GC_00003054
GC_00003070
GC_00003088
GC_00003111
GC_00003120
GC_00003149
GC_00002632
GC_00002761
GC_00002806
GC_00002819
GC_00002866
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