Metagenome-assembled genomes from a population-based cohort uncover novel gut species and within-species diversity, revealing prevalent disease associations

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, sub-species 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 Genome Unit Number (GUN), a novel MAG-based metric that informed sub-species analyses. Based on normalized GUN (nGUN), we prioritized *Odoribacter splanchnicus,* a prevalent species with the lowest sub-species heterogeneity, yielding sufficient power for sub-species association study. We identified two dominant

genome units, GU-N1 and GU-N2, with distinct gene repertoires and divergent disease associations. Notably, GU-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 new disease associations on sub-species level obscured at higher taxonomic levels, highlighting the need for genome-resolved approaches in microbiome research.

IMPORTANCE

Microbiome studies increasingly recognize that species-level profiles can mask critical sub-species differences relevant to health and disease. However, our work shows that within-species diversity varies drastically across gut microbes, with some species exhibiting almost as many distinct sub-species clusters as recovered genomes, making association studies at the sub-species level essentially intractable. To address this, we introduce the Genome Unit Number (GUN), a scalable metric for quantifying sub-species structure. Using GUN, we demonstrate that only species with limited within-species diversity, such as *Odoribacter splanchnicus*, currently allow for robust sub-species association testing. These findings emphasize the need to systematically evaluate species structure across the gut microbiome and call for the development of new computational and statistical approaches to enable meaningful sub-species analyses in highly diverse species.

KEYWORDS

Gut microbiome, metagenome assembled genomes, genome unit number, population microbiome, metagenomics, metagenome-wide associations study, sub-species 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],

76 mental health disorders[14] and cancers[15,16]. Furthermore, emerging evidence suggests that the
77 gut microbiome has predictive power, as demonstrated in the context of incident heart failure[17].
78 However, many of these studies still rely solely on reference databases. These databases may
79 lack representatives for many uncultured or underrepresented population-specific microbial
80 species, leading to incomplete or biased interpretations.

81 In the present study, we leveraged deep metagenomic sequencing of a population-based
82 Estonian microbiome-deep (EstMB-deep) cohort to assemble a comprehensive collection of
83 metagenome assembled genomes (MAGs), substantially expanding the reference database of
84 human gut microbes with hundreds of previously uncharacterized species. We integrated these
85 population-specific MAGs with public reference data and conducted association analyses with
86 33 prevalent diseases. To systematically assess within-species diversity, we developed a novel
87 Genome Unit Number (GUN) metric. We demonstrated its utility by identifying specific
88 disease associations at sub-species level in *Odoribacter splanchnicus* that were not apparent at
89 the species level. Our findings demonstrate that genome-resolved microbiome profiling can
90 uncover novel disease-linked microbial signatures beyond that remain hidden using
91 conventional approaches.

92 RESULTS

93 Study design and cohort overview

94 The study aimed to first recover metagenome assembled genomes (MAGs) from a deeply
95 sequenced Estonian gut microbiome cohort (EstMB-deep, N=1,878) and expand the reference
96 database by combining these newly assembled genomes with the existing public Unified
97 Human Gastrointestinal Genome (UHGG) collection and demonstrate the added value of
98 genome-resolved, sub-species 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 (**Figure S1**). In the EstMB-deep subcohort, each sample corresponds to a unique individual, with balanced distributions of age and gender. 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 S2a**). 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 S2b**). A detailed description of the EstMB, including omics and phenotypic data, is provided in Aasmets & Krigul et al. 2022[10].

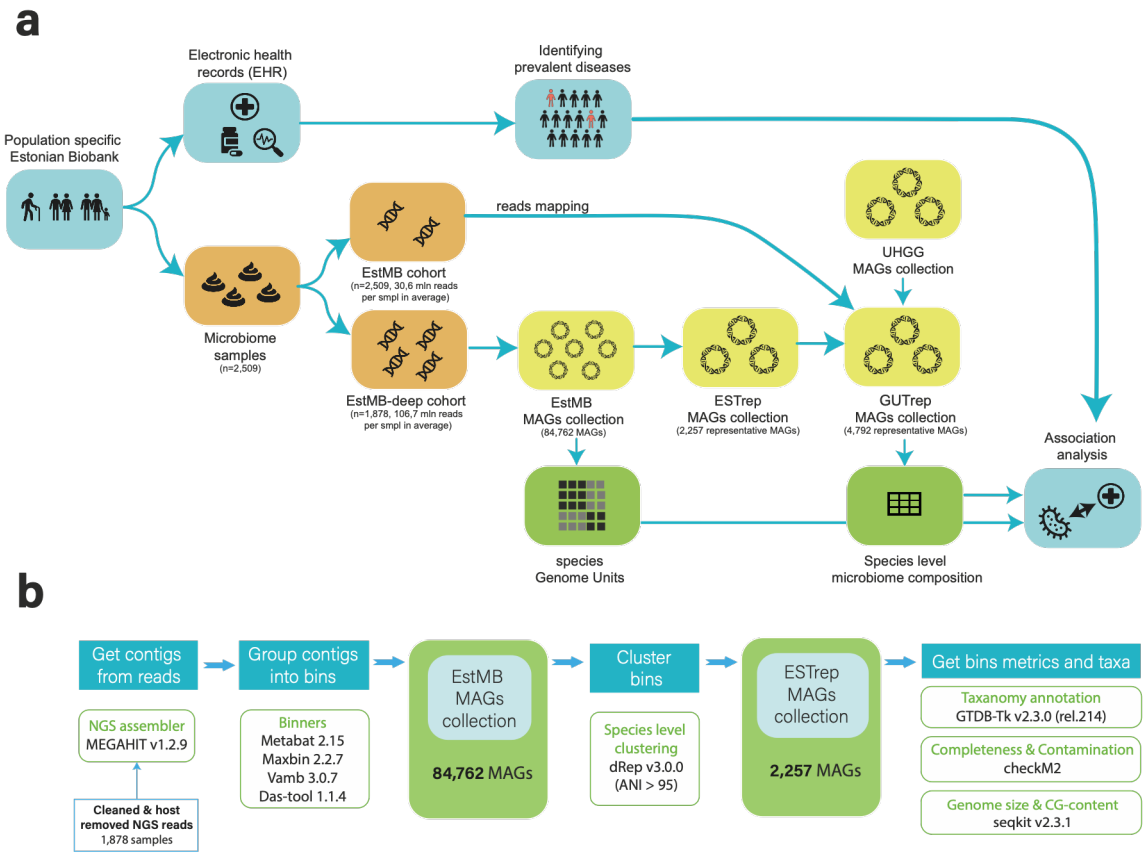


FIG 1. Study overview and cohort description. **a.** Study workflow overview. **b.** Overview of the metagenome assembled genomes (MAGs) recovery pipeline.

112 **Creating a representative MAGs pool of gut bacteria in the Estonian population**

113 To characterize population-specific microbes and expand publicly available human gut
114 microbiome databases with microbial genomes from the Estonian population, we performed
115 *de novo* metagenome assembled genomes (MAGs) reconstruction from all 1878 samples in the
116 EstMB-deep cohort. The MAG reconstruction pipeline is summarized in **Figure 1b**. We
117 successfully reconstructed 84,762 MAGs from EstMB-deep, with an average of 45.13 MAGs
118 per sample. Among these, 42,049 (49.61%) were almost complete MAGs, i.e. MAGs with
119 completeness > 90% and contamination < 5% according to CheckM. To describe the Estonian
120 population species pool, we clustered all MAGs with dRep using a 95.0% ANI threshold,
121 ensuring that the final clusters represent distinct species.

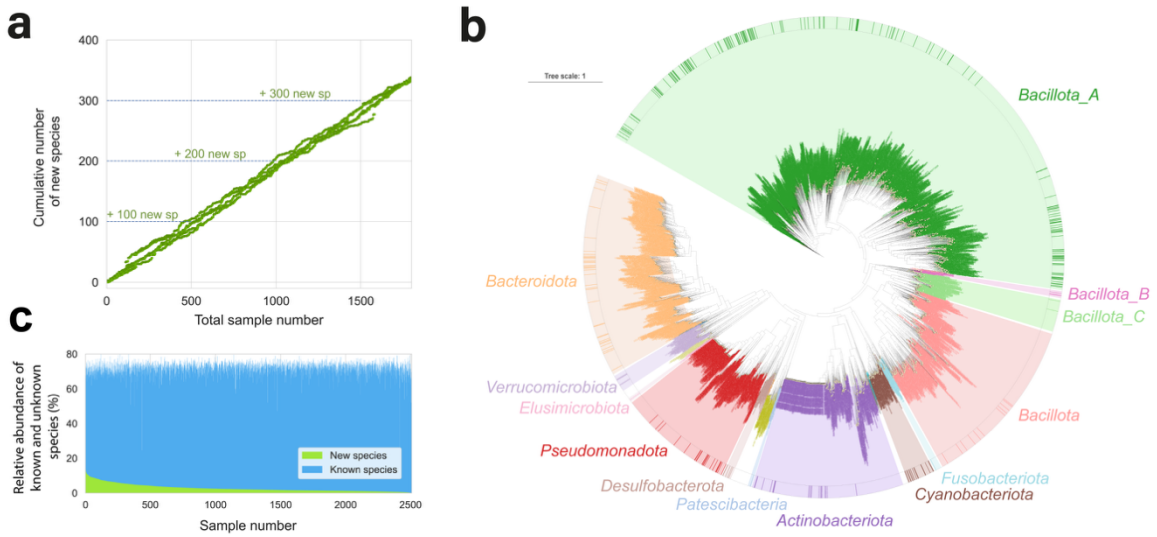
122 The species-level clustering procedure yielded 2,257 clusters (**Table S1**). For each cluster, the
123 representative MAG was selected based on genome completeness, minimal contamination,
124 strain heterogeneity, and N50 (a parameter reflecting assembly fragmentation level). We refer
125 to these 2,257 species representative MAGs as the "ESTrep" collection. The majority of
126 ESTrep MAGs from the ESTrep collection (72.97%, n=1647 MAGs) were almost complete
127 MAGs (MAGs >90% complete and <5% contaminated). Under the MIMAG standard[18], 475
128 representative MAGs (21.05%) qualified as high quality, exhibiting >90% completeness, <5%
129 contamination, and the presence of the 5S, 16S, and 23S rRNA genes together with at least 18
130 tRNAs. The remaining genomes included 1,714 MAGs (75,94%) of medium quality (MQ) and
131 68 MAGs (3,01%) of low quality (LQ) (Figure S3).

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

133 Next, we identified previously uncharacterized species within the ESTrep MAGs collection.
134 MAGs were categorized as novel species if their taxonomic classification at the species level
135 or higher couldn't be assigned using the GTDB-Tk[19], a common approach for evaluating

136 whether a newly reconstructed MAG represents a new species[20,21]. Of the 2,257 representative
137 MAGs, 353 (15.64%) were classified as novel. Among these, 231 MAGs (65.44%) had > 90%
138 and < 5% contamination, and 57 (16.15%) also contained rRNA and tRNA genes, meeting the
139 MIMAG guidelines for high-quality MAGs[18]. We observed a strong correlation between the
140 number of novel species discovered and the number of samples analyzed (R^2 = 0.97).
141 Specifically, for every 500 samples, approximately 102 novel species were identified (**Figure**
142 **2a**). As we have not observed any indication of a plateau with the current sample size, we
143 expect that analysing more samples will reveal additional species.

144 Although Estonia is considered a Westernised population, novel species still make up a
145 significant proportion of the microbiome community. These newly identified species were
146 distributed across multiple phyla (**Figure 2b**). On average, 2.82% of the total reads per sample
147 were assigned to these novel species, even reaching a maximum relative abundance of 32.34%
148 in some samples (**Figure 2c**). Since these species are absent from public databases and may be
149 population-specific, microbiome studies that rely solely on existing references may
150 substantially underestimate microbial diversity.



151

152 **FIG 2.** Overview of novel species from the EstMB MAGs collection**. a.** The relationship
153 between the number of samples analyzed and the cumulative number of novel species

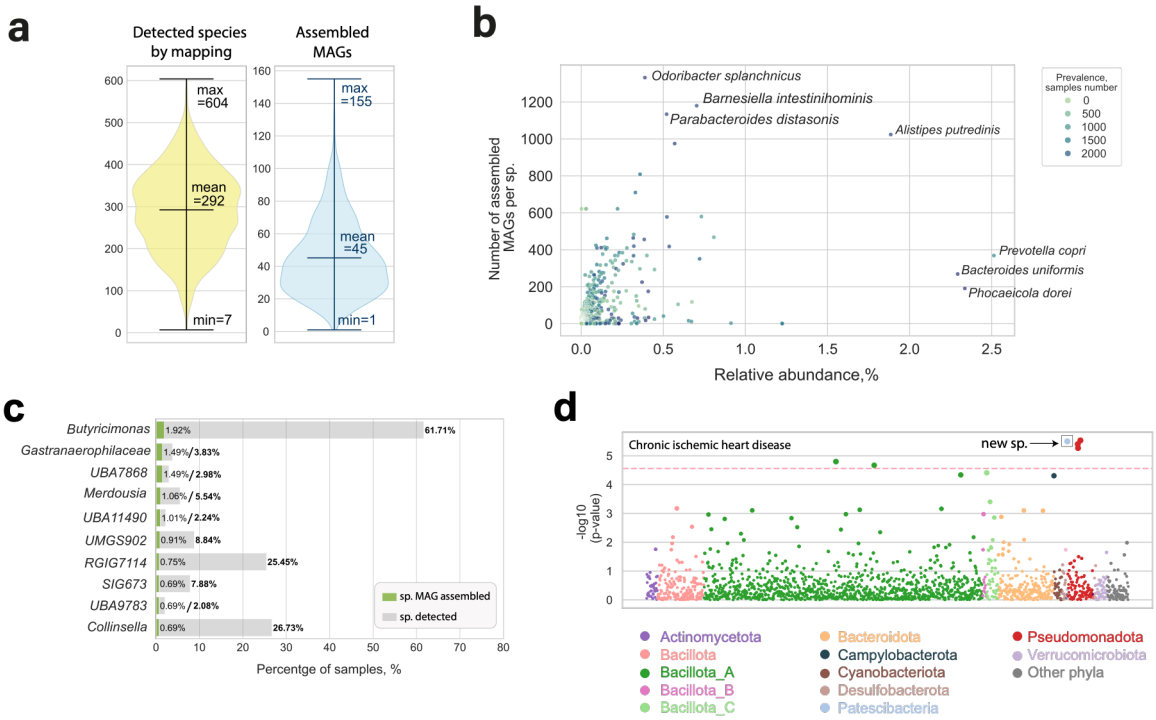
identified. **b**. Phylogenetic tree of the ESTrep species. The inner circle displays a phylogenetic tree of species, with branches colored by phylum (according to the Genome Taxonomy Database (GTDB-Tk v2.3.0), the outer ring highlights novel species assembled in this study. **c.** Relative abundances of known and novel species.

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 3a**). The most prevalent (>95%) species detected by read mapping were all well-known gut microbes: *Phocaeicola dorei*, *Bacteroides species* (*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 S4)**. We observe that samples with more species detected by mapping also tended to have more MAGs recovered (**Figure S5**). However, the number of assembled genomes per species did not clearly correlate with species prevalence or mean abundance (**Figure 3b, 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 3c, 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 S6**), illustrating the challenges of comprehensive genome reconstruction.



193 **FIG 3.** Assembly–mapping comparison and MWAS associations **a.** Average number of species
194 detected by read mapping (yellow) versus number of recovered MAGs per sample (blue). **b.**
195 Relationship between species prevalence, mean relative abundance, and number of assembled
196 MAGs per species. **c.** Prevalence of the top 10 novel species with the highest number of
197 recovered MAGs, comparing recovery by MAG assembly (green bars) and detection by read
198 mapping (grey bars). **d.** Metagenome-wide association results between GUTrep species
199 abundances and chronic ischemic heart disease. Each data point corresponds to a single species,
200 with vertical position reflecting the log-transformed *p*-value from linear regression; significant
201 associations for newly reconstructed species are highlighted with a box.

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

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

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

217 assembled species from the *Nanosynbacter* genus (species ID: H2144_Nanosynbacter_undS,
218 adjusted $p = 3.13 \times 10^{-6}$) (**Figure 3d, Table S4**). These findings emphasize the importance of
219 population-specific reference databases for detecting disease-associated microbiome changes.
220 However, further studies are needed to confirm whether these associations generalize beyond
221 the Estonian cohort.

222 **MAGs data enables within species genetic diversity analysis across species**

223 Most large-scale microbe-disease or MWAS studies are conducted at the species level,
224 although strain-level analysis is often recommended for understanding the functional
225 insights[23–25]. However, strain definitions vary substantially across studies, and many species
226 do not exhibit genomic clustering patterns that provide sufficient sample sizes for robust sub-
227 species-level association testing. Metagenome assembly enables high-resolution reconstruction
228 of species-level genomic diversity and allows the identification of taxa for which such analyses
229 are feasible, thereby broadening the range of detectable biological and clinical associations. To
230 capture this variation systematically and to select species suitable for sub-species-level
231 MWAS, we introduce the **Genome Unit Number (GUN)** metric. GUN quantifies genomic
232 diversity within a species by counting the number of highly similar genome units per each
233 species (Average Nucleotide Identity, ANI > 99%). We further define the **normalized GUN**
234 **(nGUN)**, which scales the GUN by the total number of MAGs recovered for the analyzed
235 species, allowing comparison across species with differing MAG counts.

236 $$\text{nGUN} = (\text{Number of genome units} / \text{Number of MAGs}) \times 100\%$$

237 We focused on species with >10 reconstructed MAGs, yielding 376 species across diverse
238 phyla. Notably, none of the newly identified species were included due to an insufficient
239 number of MAGs.

240 The nGUN values varied widely, ranging from 0.4 to 100 (**Figure 4a, Table S6**), indicating
241 substantial differences in species population genetic diversity. *Odoribacter splanchnicus*
242 exhibited the lowest nGUN (0.4), with one genome unit per ~250 MAGs, reflecting low
243 diversity despite high prevalence. In contrast, *Prevotella copri* had one of the highest nGUN
244 (94.0), consistent with its well-documented heterogeneity, where nearly every MAG represents
245 a unique genome unit, making it difficult to conduct sub-species association analysis in the
246 population. Interestingly, *Alistipes_A* genus appeared in both the lowest and highest nGUN
247 groups.

248 We also examined nGUN distribution across six phyla with ≥10 species present in each:
249 *Bacillota, Bacteroidota, Verrucomicrobiota,* Bacillota_A, *Pseudomonadota,* and
250 *Cyanobacteriota* all of which exhibited a broad range in nGUN distributions (**Figure 4b**).
251 *Bacillota* species tended to have higher nGUN values, indicating that this phylum generally
252 tends to have a higher number of genome units per species. In contrast, *Verrucomicrobiota*,
253 *Cyanobacteriota*, and *Pseudomonadota* exhibited lower nGUN values, suggesting that species
254 in these phyla typically have fewer genome units per species. However, due to the small sample
255 sizes in some phyla, further studies are needed to confirm whether these differences represent
256 true phylum-level trends.

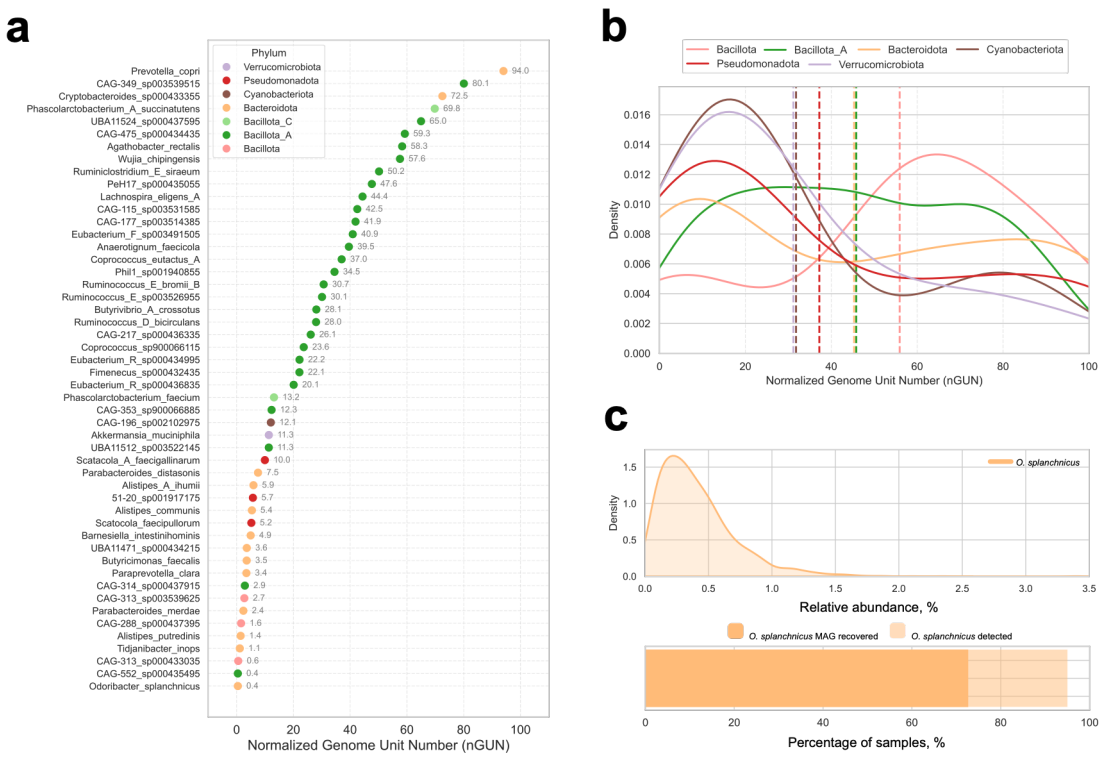


FIG4. Within-species diversity and sub-species analysis of *Odoribacter splanchnicus*. a. Normalized genome unit number (nGUN) values for the top 50 species with the highest number of metagenome assembled genomes. **b.** Distribution of nGUN values across major gut bacteria phyla. **c.** *Odoribacter splanchnicus* relative abundance, number of recovered MAGs and prevalence across samples.

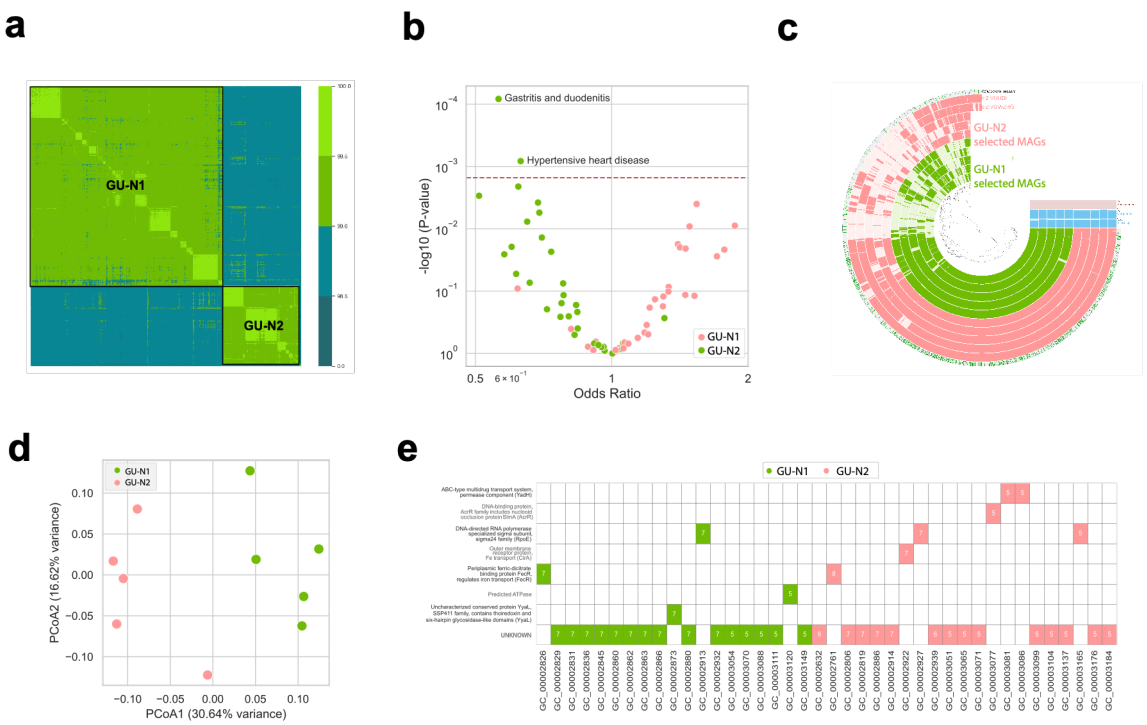
nGUN–guided sub-species analysis reveals phenotype associations missed at the species level

In order to demonstrate the value of nGUN–guided sub-species MWAS analysis, we selected *O. splanchnicus* due to its low nGUN (0.4) and high prevalence (detected in 96.14% of samples, assembled in 72.68%, **Figure 4c**). Among its MAGs, we identified four distinct genome units, two of which were rare (found in 2 and 19 samples, respectively). Therefore, we focused on the two major genome units with high case numbers: genome unit N1 (GU-N1, n =

270 974 samples, original cluster ID: 1_2.3.4.6.9) and genome unit N2 (GU-N2, n=335 samples,
271 original cluster ID: 1_1) (**Figure 5a**).

272 Logistic regression models adjusted for BMI, gender, and age were used to assess the
273 association between the presence or absence of *O. splanchnicus* GU-N1 and GU-N2 and the
274 same 33 diseases previously analyzed at the species-level MWAS. Our analysis identified a
275 significant association between the presence of GU-N1 and two different diseases - gastritis
276 and duodenitis, and hypertensive heart disease (**Figure 5b**). The odds ratio for GU-N1 was less
277 than 1 in both diseases (gastritis and duodenitis OR = 0.56, hypertensive heart disease OR =
278 0.63), indicating that its presence is associated with a reduced likelihood of having the disease.
279 Notably, these associations were not detected at the species level, highlighting the added
280 resolution of sub-species-level analysis.

281 To explore functional differences, we performed a pangenome analysis of GU-N1 and GU-N2.
282 We carried out Principal coordinate analysis of predicted gene cluster presence/absence, which
283 showed clear separation between the genome units (**Figure 5c**). We identified that the two
284 genome units formed distinct clusters, indicating clear genomic differentiation based on gene
285 content (**Figure 5d**), with 40 gene clusters unique to one of the two (**Figure 5e, Table S7**).
286 While most encoded hypothetical or uncharacterized proteins, some were annotated with
287 putative functions based on the Clusters of Orthologous Genes (COG20)[26]. GU-N22 harboured
288 a broader repertoire of genes associated with stress response, iron acquisition, and
289 antimicrobial resistance - traits consistent with enhanced survival in inflammatory
290 gastrointestinal environments. These included elevated copy numbers of the extracytoplasmic
291 stress sigma factor RpoE (σ^E), iron transport components FecR and CirA, and multidrug
292 resistance elements such as AcrR and an ABC-type efflux pump (YadH). In contrast, GU-N1
293 was enriched for redox maintenance proteins such as YyaL/DsbD, suggesting a distinct strategy
294 centered on oxidative stress mitigation.



295

296 **FIG 5.** Genomic structure and disease-associated variation across two major *O. splanchnicus*
297 genome units **a.** Heatmap of Average Nucleotide Identity (ANI) values among *O. splanchnicus*
298 MAGs, revealing two distinct genome units. **b.** Volcano plot of associations between the two
299 major *O. splanchnicus* genome units and 33 disease phenotypes. The red line indicates the
300 Bonferroni-corrected significance threshold. **c.** Pangenome analysis of five representative
301 MAGs from each *O. splanchnicus* genome unit (GU-N1 and GU-N2). **d.** Principal coordinates
302 analysis (PCoA) of *O. splanchnicus* representative MAGs based on predicted gene cluster
303 presence/absence profiles. **e.** Gene clusters uniquely present in only one of the two major *O.*
304 *splanchnicus* genome units.

305 **DISCUSSION**

306 Our study presents a scalable, genome-resolution framework for population-scale microbiome
307 analysis, enabling improved species and sub-species characterization and discovery of disease
308 associations. By expanding the gut microbial reference database with thousands of
309 metagenome assembled genomes (MAGs), including novel bacterial species, as well as sub-
310 species genome units of known taxa, we address a major limitation in current reference
311 datasets, which often underrepresent global microbiome diversity. We demonstrate that

genome-resolution microbiome analysis, coupled with population-specific MAGs catalogues, enables more comprehensive species- and sub-species-level association studies. Furthermore, we introduce the Genome Unit Number (GUN), a novel quantitative metric of within-species genetic diversity, which we applied across 378 gut species to guide candidate selection for sub-species-level analysis. Using this framework, we uncovered genome unit-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 and binning 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 MAGs reconstruction and *de novo* MAG assembly and binning 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 species 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

337 with the Unified Human Gastrointestinal Genome (UHGG) collection to improve reference
338 coverage. Notably, 31% of dereplicated GUTrep species originated from our Estonian-specific
339 MAGs, illustrating the added value of local assembly efforts.

340 A common argument against investing in resource-intensive *de novo* assembly is that the
341 rapidly growing and regularly updated public gut genome catalogues increasingly capture
342 known microbial diversity, suggesting that most gut species will soon be represented, and
343 further assembly may become redundant. However, our findings challenge this assumption,
344 and support continued *de novo* assembly in new population studies. Early efforts reported high
345 proportions of novel taxa, with 77% of MAGs classified as novel in Pasolli et al[34] and 66% in
346 Almeida et al[35]. More recent studies, such as Leviatan et al., still report 310 novel species out
347 of 3,594 assembled (8.6%)[36]. In our cohort, we recovered 353 novel species from 2,257 MAGs
348 (15.6%), with ~102 additional novel species per 500 samples, and no indication of a discovery
349 plateau. Moreover, we confirm a previously reported finding that many novel species
350 assembled from a few samples were nonetheless widespread by mapping, suggesting that
351 assembly remains essential even in well-characterized industrialized populations[33]. These
352 results reinforce the idea that local assembly efforts complement global references and remain
353 critical for uncovering the full spectrum of microbial diversity.

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

362 level and can report presents of different strains even among strains coexisting within the same
363 sample[62].

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

374 Our normalized genome unit number, or nGUN, helps to assess whether sub-species analysis
375 is feasible by quantifying within-species diversity. High nGUN value of the species might
376 indicate that each individual harbours a unique genome unit, complicating population-level
377 associations. In this study, we analysed within-species diversity across 378 gut species,
378 expanding upon previous work that showed a strain richness of 92 gut species based on pure
379 isolates from at least three different individuals[40]. Our results significantly expand on this by
380 including a broader range of species, each represented by more than 10 MAGs. Consistent with
381 earlier findings, we observe substantial variability in sub-species diversity across species[40].
382 However, our data provides a more detailed and comprehensive picture due to the larger
383 number of species. Among the 14 overlapping species between the two studies, some show
384 similar patterns of genomic diversity - for example, *Odoribacter splanchnicus* and *Barnesiella*
385 *intestinihominis* exhibit consistently low diversity, while others like *Bifidobacterium longum*
386 remain highly diverse. Other overlapping examples, such as *Fusicatenibacter saccharivorans*

387 and *Coprococcus eutactus*, display divergent diversity estimates, likely reflecting
388 methodological differences (culture isolate vs metagenome-based) and highlighting the need
389 for further comparative research. Highly diverse species such as *Prevotella copri,* well known
390 from other studies[41,42], are absent from the Chen-Liaw dataset, but prominent in ours. Our
391 larger dataset also allowed the investigation of phylum-level patterns, suggesting that the
392 common phyla, such as *Bacteroidota and Bacillota_A* species, tend to have higher nGUN
393 values, while *Verrucomicrobiota, Cyanobacteriota,* and *Pseudomonadota* exhibit lower
394 diversity. These phylum-level differences in within-species diversity suggest possible
395 evolutionary or ecological constraints. However, further studies are needed, particularly for the
396 less common phyla, to validate and understand the underlying mechanisms.

397 Understanding microbiome diversity at both species and sub-species levels enhances resolution
398 in metagenome-wide association studies (MWAS). At the species level, we identified 96
399 bacterial species significantly associated with 25 common diseases, of which 8 diseases
400 involved previously uncharacterized species, highlighting the limits of relying solely on global
401 references. These differences may reflect population-specific variation driven by local
402 differences in diet, genetics, and lifestyle[11]. At the sub-species level, we identified associations
403 undetectable at the species level, such as *Odoribacter splanchnicus* genome unit GU-N1
404 negative association with gastritis, duodenitis and hypertensive heart disease. Previous research
405 has shown that the abundance of the genus *Odoribacter* is negatively correlated with systolic
406 blood pressure in overweight and obese pregnant women, suggesting that SCFA-producing
407 taxa may influence host blood pressure[43]. Comparative genomic analysis of the MAGs from
408 two *O. splanchnicus* genome units helped us to identify a set of gene clusters that differed
409 between the two groups. These genomic features suggest that GU-N1 is functionally better
410 adapted to conditions characteristic of gastritis and duodenitis, such as oxidative stress, nutrient
411 limitation, and host antimicrobial pressure. Together, our findings highlight the value of

412 genome-resolved metagenomic approaches in revealing disease-relevant microbial functions
413 that would remain hidden in broader taxonomic analysis. While our study focused on a single,
414 ethnically homogeneous Northern European population, several key findings, such as the
415 detection of widespread yet previously uncharacterized species and species structures within
416 common gut taxa, are likely to extend beyond the Estonian population. However, microbiome
417 composition is influenced by genetic, dietary and environmental factors that vary between
418 populations. Future studies in diverse cohorts will be essential to evaluate the generalizability
419 of our results and validate the species - and sub-species disease associations uncovered here.
420 The genome-resolved analytical framework we present is scalable and readily applicable to
421 other population-scale microbiome datasets, enabling cross-cohort comparison and discovery.

422 Our study also has some limitations that should be considered. First, the reliance on short-read
423 sequencing may reduce assembly contiguity and sub-species resolution compared to long read
424 approaches[44]. Although long read technologies offer higher genomic completeness, their
425 current cost limits their use in large-scale population studies. A practical compromise could
426 involve using short-read sequencing for most samples and applying long read sequencing to
427 key or novel taxa. Second, the observed associations are correlative, and further validation in
428 longitudinal and experimental studies is needed to assess causality. Third, our association study
429 on sub-species level is focused on one species due to sample representation and analytical
430 feasibility, and broader application across species remains a key next step. Finally, the genome
431 unit number (GUN) is influenced by multiple biological and technical factors that may bias its
432 ability to accurately reflect within-species microdiversity. These include pangenome size,
433 genome length, the propensity for horizontal gene transfer, and the level of homologous
434 recombination. In addition, when a species is present within a sample as several highly similar
435 genomic variants that may belong to different genome units, metagenomic assembly becomes
436 more challenging. In such cases, MAGs may fail to be recovered or may represent composite

437 genomes spanning multiple genome variants, which can lead to underestimation or conflation
438 of true within-species diversity. Moreover, while functional differences between genome units
439 of one species were identified, interpretation was limited by the prevalence of unannotated
440 genes. Future improvements in genome annotation and cross-cohort replication will be
441 essential to build on these findings. Despite these challenges, our findings demonstrate the
442 value of population-scale metagenomics in uncovering novel microbial diversity and functional
443 signatures relevant to human health.

444 CONCLUSION

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

449 METHODS

450 Estonian Microbiome cohort description

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

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

463 **Microbiome sample collection and DNA extraction**

464 The participants collected a fresh stool sample immediately after defecation with a sterile
465 Pasteur pipette and placed it inside a polypropylene conical 15 mL tube. The participants were
466 instructed to time their sample collection as close as possible to the visiting time in the study
467 center. The samples were stored at −80 °C until DNA extraction. The median time between
468 sampling and arrival at the freezer in the core facility was 3 h 25 min (mean 4 h 34 min), and
469 the transport time was not significantly associated with alpha (Spearman correlation, p-value
470 0.949 for observed richness and 0.464 for Shannon index) nor beta diversity (Bray–Curtis, p-
471 value 0.061, R-squared 0.0005). Microbial DNA extraction was performed after all samples
472 were collected using a QIAamp DNA Stool Mini Kit (Qiagen, Germany). For the extraction,
473 approximately 200 mg of stool was used as a starting material for the DNA extraction kit,
474 according to the manufacturer's instructions. DNA was quantified from all samples using a
475 Qubit 2.0 Fluorometer with a dsDNA Assay Kit (Thermo Fisher Scientific).

476 **Shotgun metagenomic sequencing**

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

485 over three times deeper sequencing coverage. Sequencing depth was evaluated with Nonpareil
486 v3.5.5[63].

487 **EstMB MAGs metagenome assembly and binning**

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

495 **ESTrep MAGs collection**

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

508 MAGs were classified into three quality tiers. High-quality (HQ) MAGs were defined as those
509 with >90% completeness and <5% contamination and by the presence of ≥ 21 tRNA genes and
510 a full complement of rRNA genes (5S, 16S, and 23S)[18]. Medium-quality (MQ) MAGs were
511 defined as those with >50% completeness and <10% contamination. MAGs not meeting HQ
512 or MQ thresholds were classified as low-quality (LQ). Assembly statistics, including total
513 assembly size, number of contigs, N50, and GC content, were calculated using SeqKit
514 (v2.3.1)[58].

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

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

525 **Species relative abundance and prevalence estimation**

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

532 **Species-level association study**

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

544 **Genome Unit Number estimation**

545 The Genome Unit Number (GUN) quantifies the number of genome units (GUs) identified for
546 a given species. Genome units are defined as clusters of genomes from the same species sharing
547 an average nucleotide identity (ANI) greater than 99% (analysis is restricted to species
548 represented by more than 100 assembled MAGs). The normalized GUN (nGUN) represents
549 the number of genome units per 100 MAGs and is calculated by dividing the species GUN by
550 the total number of MAGs for that species and multiplying by 100 to express the value as a
551 percentage.

552 The corresponding formula is shown below:

$$nGUN = \frac{species\ GUN}{Total\ number\ of\ MAGs} * 100\%$$

553

554 **Sub-species association study**

555 Candidate species for sub-species association analysis were selected based on two criteria: (1)
556 a high number of reconstructed MAGs per species, and (2) the lowest normalized GUN
557 (nGUN), indicating fewer sub-species clusters per species. These criteria were established to
558 ensure sufficiently large sample sizes for robust microbiome-wide association studies
559 (MWAS). Genome units were defined using dRep[54]. Based on these criteria, *O. splanchnicus*
560 was selected as the candidate species for sub-species analysis. We examined *O. splanchnicus*
561 sub-species population structure and focused on two out of the five most prevalent identified
562 *O. splanchnicus* genome units. The remaining three genome units were excluded due to their
563 presence in only a small subset of samples. For association analyses, we used presence/absence
564 data from the two selected genome units. These genome units were designated as GU-N1 (n =
565 974; original cluster ID: 1_2.3.4.6.9) and GU-N2 (n = 335; original cluster ID: 1_1). Logistic
566 regression analyses adjusted for sex, age, and BMI were performed for the same 33 diseases
567 previously examined at the species level. To account for multiple testing, a stringent Bonferroni
568 correction was applied, resulting in a corrected significance threshold of $\alpha = 1.5 \times 10^{-3}$ (original
569 $\alpha = 0.05$).

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

573 **ACKNOWLEDGEMENTS**

574 We express our gratitude to all individuals who made valuable contributions to the Estonian
575 Microbiome cohort, as well as to those who developed the software and databases utilized in
576 this study. We thank Mait Metspalu, Andres Metspalu, Lili Milani and Tõnu Esko from the
577 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.

720 31. Vicedomini, R., Quince, C., Darling, A.E., and Chikhi, R. (2021). Strainberry: automated
721 strain separation in low-complexity metagenomes using long reads. Nat. Commun. *12*,
722 4485. https://doi.org/10.1038/s41467-021-24515-9.

723 32. Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019). New
724 insights from uncultivated genomes of the global human gut microbiome. Nature *568*,
725 505–510. https://doi.org/10.1038/s41586-019-1058-x.

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

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

734 35. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard,
735 K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of
736 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–
737 114. https://doi.org/10.1038/s41587-020-0603-3.

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

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

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

751 39. Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M.,
752 Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., et al. (2023). Extending and
753 improving metagenomic taxonomic profiling with uncharacterized species using
754 MetaPhlAn 4. Nat. Biotechnol. *41*, 1633–1644. https://doi.org/10.1038/s41587-023-
755 01688-w.

756 40. Chen-Liaw, A., Aggarwala, V., Mogno, I., Haifer, C., Li, Z., Eggers, J., Helmus, D.,
757 Hart, A., Wehkamp, J., Lamousé-Smith, E.S.N., et al. (2024). Gut microbiota strain
758 richness is species specific and affects engraftment. Nature.
759 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.

62. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat. Biotechnol. 39, 727–736 (2021).

63. Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R. & Konstantinidis, K. T. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. mSystems 3, e00039-18 (2018).

SUPPLEMENTAL MATERIAL

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

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