Metagenome-assembled genomes of Estonian Microbiome cohort reveal novel species and their links with prevalent diseases

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

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

***Corresponding authors:** Kateryna Pantiukh & Elin Org
Institute of Genomics, Estonian Genome Centre
University of Tartu
Tartu 51010, Estonia
phone: (372) 737 4034; fax: (372) 737 4060
Email: pantiukh@ut.ee, elin.org@ut.ee

Conflict of interest: The authors declare no competing interest.

Keywords: Gut, microbiome, microbiome associations, Estonian population-based cohort, EstMB, metagenome-assembled genome, shotgun metagenomics,

Word count: 6972 words
Number of Figures and Tables: 3 Figures, 1 Table

Summary

Deep metagenomic data from population studies enables genome recovery and construction of population-specific references, including new species and uncovering microbial diversity that global references might miss. We constructed an Estonian population-specific reference of metagenome-assembled genomes (MAGs) from 1,878 stool samples of the EstMB-deep cohort. We assembled 84,762 MAGs, representing 2,257 species, including 353 potentially novel species (15.6%). Additionally, 607 species (26.9%) were not present in the global Unified Human Gastrointestinal Genome (UHGG) reference database and may therefore be population-specific. We further demonstrated the value of *de novo* assembly of bacterial genomes by analysing associations with 33 prevalent diseases and detected 44 significant associations for 15 diseases, including with 10 potentially new species and 5 species absent from UHGG. The correlations, especially with new species, demonstrate that *de novo* bacterial genome assembly from population cohorts can provide significant novel insights linking the microbiome with prevalent diseases and uncovering population-specific differences.

Introduction

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. Using this approach, Foster et al. presented the Human Gastrointestinal Bacteria Genome Collection, which comprised 273 isolates from the human gastrointestinal tract[1]. Similarly, Zou et al. published the Culturable Genome Reference collection, consisting of 1,520 draft genomes of bacteria cultivated from the fecal samples of healthy humans[2]. However, rapid technological advancements have facilitated the generation of vast amounts of metagenomic data and the development of techniques for assembling microbial genomes, consequently improving reference databases. These genomes are known as Metagenome-Assembled Genomes (MAGs).

In 2019, three significant collections of human gut MAGs were unveiled. Almeida et al. presented 92,143 MAGs representing 1,952 microbial species from the human gut[3]. Pasolli et al. revealed 154,723 MAGs representing 4,930 microbial species from diverse body sites, including gut[4], and Nayfach and others reconstructed 60,664 MAGs representing 4,558 microbial species[5]. In 2021, the Unified Human Gastrointestinal Genome (UHGG) collection was launched, combining previous MAG collections into a deduplicated reference with 4,644 genomes[6]. In addition to the gut, reference collections for MAGs from different environments and genomes assembled from cultivated bacteria have been established. The largest collection of bacterial and archaeal genomes currently known is the Genome Taxonomy Database (GTDB; https://gtdb.ecogenomic.org), which currently includes 402,709 genomes and MAGs and is updated annually. GTDB is widely used to establish the novelty of species by analyzing the GTDB annotation of the assembled genome. These global reference collections of MAGs are undoubtedly valuable genetic resources for scientific exploration. Nevertheless, as novel strains may significantly differ in functional capacity from the strains present in the reference database, a global reference can have its limitations. This issue can be addressed by establishing a collection of population-based MAGs that are used as reference in large-scale population biobanks.

In recent years, several population-based biobanks with metagenomic datasets have been established, for example the Dutch Microbiome cohort from the Lifelines biobank[7], Israeli Project 10K cohort[8,9], the FinRisk cohort[10], and the EstMB cohort from the Estonian Biobank[11]. The strength of biobanks is their ability to provide multiple layers of information about the same individuals, including data about the health and well-being of the participants. This comprehensive data has, for instance, enabled to identify the factors influencing the gut microbiome composition, disentangle the effects of host genetics and environment and most

recently, that microbiome can be used to predict disease incidence years before the disease onset[7,11–15]. However, many of these studies rely on reference databases that may have limited resolution and cannot fully account for population-specific differences. Therefore, generating MAGs and developing population-based references could enhance association analyses and lead to novel hypotheses.

Here, we used 1,878 samples from the Estonian Microbiome Cohort (EstMB) sequenced with deep coverage, which allowed us to assemble and annotate 84,762 MAGs representing 2,257 species, forming the EstMB MAGs collection. By comparing the assembled MAGs with the GTDB, we found 353 MAGs (15,6%) of the EstMB MAGs collection to represent potentially novel species. To show the value of the population-based MAG collection, we studied the associations between the species abundance and common diseases. Our findings unveiled associations between bacterial abundance and 15 common diseases, including associations with potentially new species and species absent from the GTDB and the global reference database UHGG.

Results

Cohort overview and study design

In our research, we aimed to build an Estonian population-specific Metagenome-Assembled Genomes (MAGs) reference to enhance association studies (**Figure 1a**). This study included gut metagenomic data from the Estonian Microbiome Deep (EstMB-deep) cohort (N= 1878), which is a subset of the volunteer-based Estonian Microbiome cohort (EstMB). A detailed overview of the EstMB, including omics and phenotypic data availability, is described in Aasmets & Krigul et al. 2022[11]. The EstMB-deep cohort includes 1,308 women (69.64 %) and 570 men (30.35 %), representing ages 23 to 89 (**Figure 1b**) and encompassing various geographical regions across Estonia. The EstMB-deep samples were resequenced with deep coverage, yielding an average of 11.73 +/- 3.9 Gb data per sample and 56.13 +/- 19.37 million reads per sample for metagenome assembly (**Figure 1c**).

Population-specific Metagenome-Assembled Genomes (MAGs) reference

Firstly, we aimed to recover gut bacterial genomes from the 1878 EstMB-deep samples. The pipeline for MAG recovery is depicted in **Figure 1d**. We used the MEGAHIT assembler for contigs assembly, prioritizing speed due to the large sample size. Binning involved a multi-binning approach using Metabat, Maxbin, and Vamb with Das Tools for bin selection, enhancing final bin quality. We successfully reconstructed 84,762 metagenome-assembled genomes (MAGs), an average of 45 MAGs per sample. Among these, 42,048 MAGs (49.6%) have completeness > 90% and contamination < 5%.

In order to define the species structure within the Estonian population, all 84,762 MAGs were clustered at the species level with an average nucleotide identity (ANI) threshold of 95%. The clustering procedure resulted in a total of 2,257 species-level clusters (**Table S1**). The representative MAG was selected from all the MAGs belonging to the same species-level cluster based on genome completeness, minimal contamination, strain heterogeneity, and assembly N50. The majority of the 2,257 species-level representative genomes (72.97% or 1647 MAGs) were >90% complete and <5% contaminated, and 475 of them (21.05%) had the 5S, 16S, and 23S rRNA genes together with at least 18 of the tRNAs. These 475 MAGs (21%) satisfy the 'high quality' criteria set for MAGs by the Genomic Standards Consortium[16] (**Figure 1e**). The median completeness of reconstructed 2,257 representative MAGs was 95.55%, interquartile range, IQR = 90.82–98.64% (**Figure 1f**) and median contamination was 0.57%, IQR = 0.14–1.81% (**Figure 1g**).

A genome taxonomy was assigned for each cluster representative MAG using the Genome Taxonomy Database Tool kit GTDB-Tk v2.3.02, release 214.0 (**Table S1**). Representative MAGs belonged to 20 different phyla, while the majority of 2,257 representative MAGs (1762 MAGs,

78%) belonged to three phyla – Bacillota_A (1213 MAGs, 54%), Bacteroidota (337 MAGs, 15%) and Bacillota (212 MAGs, 9%) (**Figure S1**). The bacterial species most represented in our MAGs collection were *Odoribacter splanchnicus* (MAG recovered from 70.93% samples), *Barnesiella intestinihominis* (62.83%), *Parabacteroides distasonis* (60,38%), *Alistipes putredinis* (54,53%) and *Agathobacter rectalis* (51.92%) (**Figure S2, Table S2**).

Novel species and comparison of the population-specific reference with global reference UHGG

We considered MAGs to belong to a new species if their taxonomic classification at the species level or higher couldn't be assigned using the GTDB-Tk[17,18], a common approach for evaluating MAG novelty[3–5]. From the 2,257 representative MAGs, 353 MAGs (15,64%) represent a new species according to the GTDB criteria. Of all 353 new species MAGs, 231 MAGs (65.43%) exhibit completeness greater than 90% and contamination less than 5%, including 57 MAGs (16.14%) where rRNA and tRNA were detected, meeting the quality standards outlined in the MIMAG guidelines for high-quality genomes[16]. We also compared our collection of species-representative MAGs with the latest and most comprehensive publicly available global gut reference dataset, the Unified Human Gastrointestinal Genome (UHGG)[6]. From 2,257 species from the EstMB collection, 1650 species were also found at USGG (USGG+), and 607 species were unique for EstMB compared to USGG (USGG-). We refer to these MAGs as EstMB-specific MAGs representing potentially EstMB-specific species. Among the EstMB-specific species, 328 were found in the GTDB (GTDB+), suggesting they had probably been previously identified from environments other than the human gut. The remaining 279 EstMB-specific species were not present in the GTDB (GDTB-) and were first assembled in this study. Interestingly, out of the 353 new species identified according to the GTDB criteria, 74 MAGs were found in USGG (USGG+/GDTB-) (**Figure 2a**). This was unexpected, as we anticipated that all MAGs from published global references would also be included in GTDB. Possible explanations for this may be a time lag until GTDB updates, quality criteria restrictions for including in GTDB, and other technical issues. The newly discovered species are distributed across 182 different genera, typically with only a few species per genus, with one notable exception - 27 new species were found within the genus Collinsella (**Figure S3**).

We observed a strong correlation between the number of new species discovered and the sample size analyzed (R^2 = 0.97). Specifically, for every 500 samples, we identified approximately 102 additional novel species (**Figure 2b**). The linear trend and the absence of any apparent plateau in the graph imply that expanding our sample size, provided the samples maintain similar sequencing characteristics and microbial diversity, could further augment the discovery of new species.

Estimation of species prevalence using population-specific reference

MAGs representing new species are often reconstructed from a limited number of samples. In fact, MAGs for only five new species were recovered in more than 1% of samples. This reflects a general trend where only 511 out of 2257 species were found to have MAGs in more than 1% of samples. Therefore, to estimate species prevalence in the Estonian population, we mapped the reads against our population-specific reference. As the reference, we use a collection of all representative MAGs recovered from EstMB-deep cohort samples, which includes EstMB-specific species as well as new species reconstructed for the first time during this study.

After mapping the reads against the representative MAGs in the EstMB-deep cohort, we detected an average of 389 species per sample, including an average of 30 new species per sample. In comparison, the average number of recovered genomes per sample was 45, which highlights the complexity of metagenome assembly (**Figure 2c**). Among 2,257 species from the population-specific reference, 1,595 species had a prevalence of more than 1%. The prevalence of species, grouped by their presence in the GTDB and USGG, is shown in **Figure 2d**. Even species which have been recovered from a large number of samples can be found in significantly more samples after mapping (**Figure 2e, Table S2**). MAGs representing EstMB-specific species

(USGG-), which are usually extracted from a small number of samples, show a prevalence exceeding 80% for some species. For example, *Bacteroides faecalis* has a prevalence of 97.23%, although only 1 MAG was assembled, and *Bacteroides intestinigallinarum* has a prevalence of 95.85% although only 2 MAGs were assembled. As for the new species, the gap between the number of recovered genomes and species prevalence is even more pronounced (**Figure 2f, Table S3**). For example, one unidentified species belonging to the genus *Butyricimonas* (species ID: H0366_Butyricimonas_undS) was recovered in only 36 samples, but it was found in 1056 samples, i.e. more than half of the total number of samples (**Table S3**). On average, the combined relative abundance of all new species per sample is 3.99% (**Figure S4**), the minimum combined relative abundance for a sample is 0.54% and the maximum 34.76%.

In general, species prevalence estimated by read mapping is always higher than the number of samples where the MAGs belonging to those species were reconstructed. Interestingly, no correlation was observed between the number of samples in which species-level MAGs were detected by mapping (species prevalence) and the number of samples in which species-level MAGs were recovered (**Figure S5**).

Association of gut species with prevalent diseases in the Estonian Microbiome cohort

To explore the relationship between common diseases and microbiome composition in the Estonian population, we took advantage of the rich information about participants' health available from the electronic health records (EHR). Linear regression models, adjusted for BMI, gender, and age, were constructed to evaluate the association between the diseases and species abundance. A prevalence threshold of 1% for the bacterial species was used to limit the number of tests carried out, resulting in 1,595 species for the analysis. A total of 33 prevalent diseases with more than 100 cases were included in the analysis (**Table S4**). These diseases belonged to various groups, with the largest being Diseases of the Respiratory System (7 diseases), Diseases of the Circulatory System (7 diseases), and Diseases of the Digestive System (4 diseases) (**Table S4**).

Overall, we identified 44 associations between 41 bacterial species and 15 diseases (with significance level after Bonferroni correction 3.13e-05) (**Figure 3, Table 1**). Among these 41 species, 10 (24%) represent new species. Associations with these new species were sometimes among the strongest observed. For example, the most significant association with heart failure was an association with a new species from the CACTXU01 genus (species ID: H1293_CAKTXU01_undS, adjusted P value 5,7 x 10^{-10}). Detected associations also included associations with 5 known EstBM-specific species absent from USGG. While most species displayed only significant association with only a single disease, two known species (*Ventrimonas sp900540335* and *Succiniclasticum sp900544275*) and one new species from the Alistipes genus (species ID: H2126_Alistipes_undS) exhibited associations with two distinct conditions. *Ventrimonas sp900540335* was associated with essential hypertension and vitamin D deficiency. *Succiniclasticum sp900544275* was associated with chronic ischemic heart disease and heart failure. New species belonging to the Alistipes genus were associated with glaucoma and major depressive disorder. Further research is needed to clarify the identified relationships and determine if they are reproducible and universal across other populations, or if they are specific to the Estonian population. Nevertheless, our results highlight the value of creating a population-specific reference database for improving association studies.

Discussion

In this study, we created a collection of Estonian human gut metagenome-assembled genomes (MAGs), uncovering previously unidentified bacterial species. Utilizing this collection as a reference, we analyzed the microbiome composition of our samples and conducted a metagenome association study for prevalent diseases. Our findings reveal significant associations between both known and novel species, highlighting the need for population-specific references.

The important question for large population cohort holders is whether *de novo* Metagenome-Assembled Genome (MAGs) assembly will yield new species or species not present at publicly available human gut reference collections like USGG[6]. Early studies on *de novo* assembly of gut microbial species often yielded a high percentage of new species (77% MAGs in the Pasolli article[4], and 66% in the Almeida articles[6]). However, more recent works show a much lower rate of potentially new species discoveries due to the increasing comprehensiveness of the databases. For instance, Leviatan et al. identified 310 novel species among 3,594 species assembled (8.6%)[9]. In our study, we successfully assembled 353 new species out of 2,257 assembled species, resulting in a rate of 15.6%. Although the rate of newly discovered species has decreased in recent years, at least in well-studied westernized populations, our findings indicate that discovering new species is not only possible but expected. We managed to identify approximately 102 additional novel species in every 500 samples. Our results indicate that even in Estonia, which represents an industrialized European population, the number of novel species is expected to grow with larger sample sizes. Therefore, with the growing number of biobanks, particularly those representing understudied populations, we expect a significant number of novel species that could be fundamental for understanding the microbiome-host interactions.

Additionally, we show that based on our population-specific reference database, 44 bacteria were associated with the prevalence of at least one common disease. Moreover, out of the 44 bacteria, 9 were novel and had not been previously identified in any database. This highlights the limitations of relying solely on publicly available global reference databases, as population-specific variations in the microbiome composition can arise from a variety of factors such as diet, genetics, and lifestyle[19,20]. For instance, research indicates significant differences in the gut microbiomes of the Japanese compared to other populations, characterized by the higher Actinobacteria abundance, increased carbohydrate genes, and depletion of genes for methanogenesis[21]. These differences occur not only at the species level, but also at the strain level, where specific strains within the same species can predominate in different populations, potentially influencing disease associations[22]. Thus, it is essential to identify population-specific microbiome compositions to conduct robust disease association studies effectively.

On average, we assembled 43 genomes per sample, indicating that despite deep sequencing (with an average of 53.04 million read pairs per sample), we're unable to recover all the species present in a sample as the average number of detected species was 389. The considerable gap between the number of genomes recovered and their actual prevalence could be attributed to several factors. One of the most important factors influencing the recovery of MAGs is sequencing depth and the choice of the assembly pipeline. For example, including the DAS tool in a pipeline has been shown to recover a higher number of species[23]. However, some factors remain beyond the control of researchers and cannot be enhanced to maximize MAG recovery yield. These include the unequal representation of members of the microbial community within the sample, the complexity of specific genomes and the presence of closely related microorganisms with similar genomes. All of these factors complicate the assembly process[24,25]. However, the key takeaway is that the number of recovered MAGs does not represent species prevalence.

In addition to having a population-specific reference for bacterial species, the recovered genomes allow us to study genomic variability in greater detail, particularly focusing on bacterial strains. Since the number of samples containing specific strains is usually much smaller than those containing specific species, detailed strain-level analysis may require a larger number of samples. Another limitation of our study arises from the length of reads utilized. The use of short reads in de novo assembly can result in lower quality MAGs compared to what could be achieved with long-read sequencing[26]. However, the expense associated with long-read sequencing remains considerably higher than that of short-read sequencing, requiring a balance between cost and quality, particularly when analyzing a large number of samples for population cohorts. One potential solution to this challenge could involve using short-read sequencing for

all samples and utilizing long-read sequencing for a smaller subset of particularly intriguing samples, such as those where novel species have been identified. Assessing the quality of the *de novo* assembled genomes presents challenges, particularly when dealing with newly discovered bacteria belonging to rare groups with limited similar genomes in databases. In such cases, evaluating completeness and contamination of the assembled genome can deviate significantly from reality due to the presence of unexpected genome fragments or the absence of expected ones[Philip Hugenholtz, 2008]. While CheckM v2 handles this issue more effectively[27], it remains imperfect, underscoring the need for cautious validation of results.

Thanks to the depth of our sequencing, we successfully recovered a large number of MAGs, with the majority of the 2,257 representative MAGs being of good quality. This deep sequencing data enabled the discovery of a new species, potentially specific to the Estonian population. Expanding the sample size will likely enable us to recover even more novel species and uncover previously undetected microbiome-diseases associations. Due to our cohort being population-based rather than clinical, we were able to capture a diverse microbiome composition, which is reflective of the broader population. This collection of Estonian population MAGs serves as a fundamental resource for further research.

Acknowledgements

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

Data analysis was carried out in part in the High-Performance Computing Centre of University of Tartu, and we thank 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 EstMB-deep cohort.

This work was funded by Estonian Research Council grant (PRG1414 to E.O) and an EMBO Installation grant (No. 3573 to E.O.). EO was supported by European Regional Development Fund Project No. 15-0012 GENTRANSMED and Estonian Center of Genomics/Roadmap II project No 16-0125.

Declaration of interests.

The authors declare no competing interests.

Author contributions.

Conceptualization, K.P. 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. and E.O.; 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.

Supplemental information titles and legends

Supplementary_Figures_S1-S5. Figures S1–S5.

Supplementary_Table_S1-S4. Excel file containing supplementary tables S1-S4.

Figure S1. Phylum-level annotation of the representative MAGs.

Figure S2. Top 20 bacterial species with the highest number of recovered species-level MAGs from all samples.
Figure S3. The distribution of novel species per genus.
Figure S4. Relative abundance of new species per sample.
Figure S5. Relationship between the number of samples where species-level MAGs were detected by mapping versus the number of samples where species-level MAGs were recovered.
Table S1. Quality parameters, taxonomic annotation, and cluster information for representative genomes of species-level clusters (n=2,257).
Table S2. Species Prevalence and number of recovery MAGs of the top species clusters.
Table S3. Species Prevalence and number of recovery MAGs of the top new species clusters.
Table S4. List of analyzed diseases.

Figures

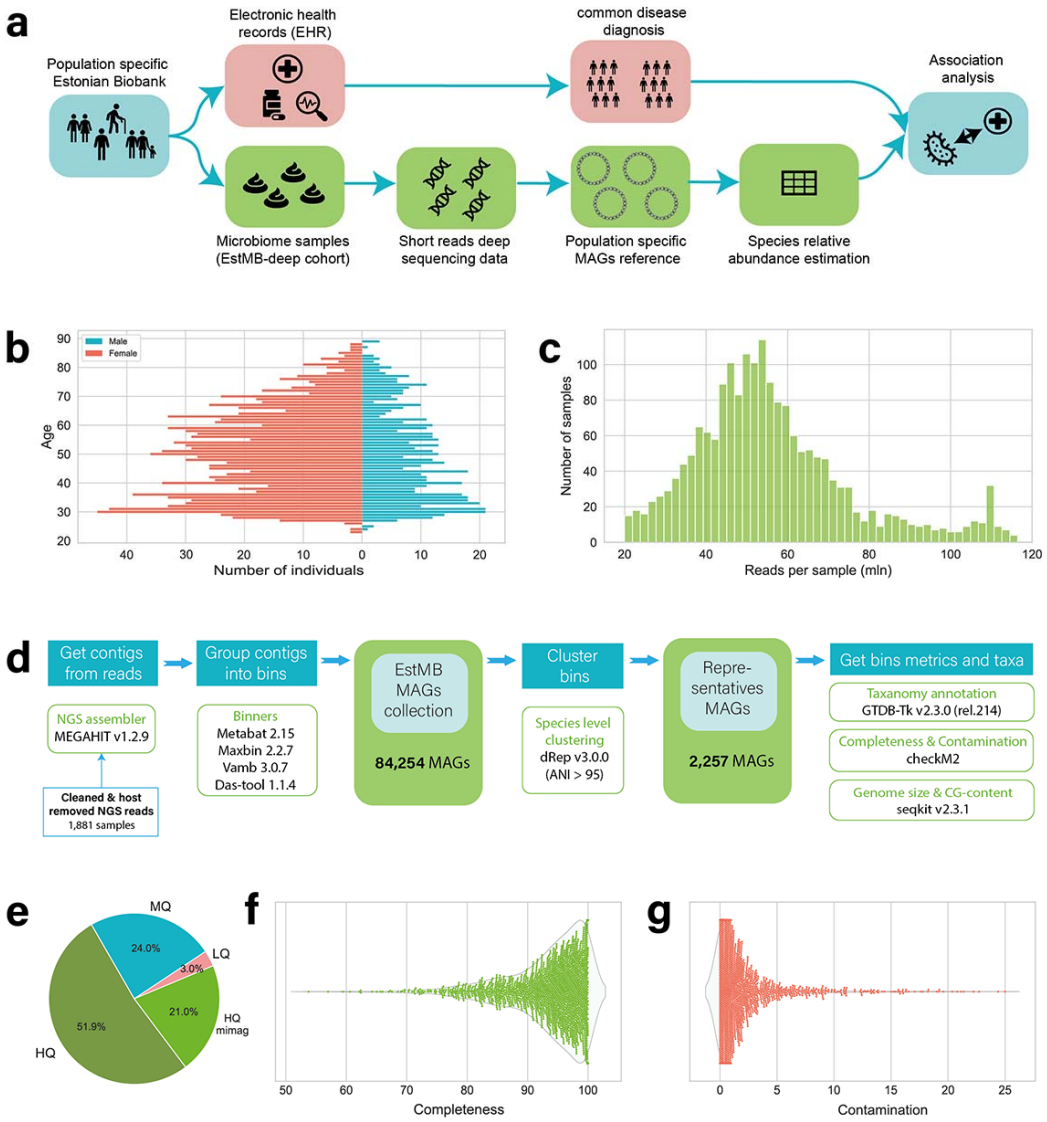

Figure 1. Workflow for MAG Reconstruction and Species-Level Clustering a. Study overview **b.** Distribution of age and gender across Estonian Microbiome cohort **c.** Distribution of

the number of paired-end (2 x 150 bp) reads **d**. Metagenome-Assembled Genome (MAG) reconstruction pipeline **e.** Quality estimation for the MAGs representing final species-level clusters. High-quality MAGs according to the MIMAG standards (HQ mimag) – 475; High-quality MAGs (HQ) – 1172; Medium-quality MAGs (MQ) – 542; Low-quality MAGs (LQ) – 68 **f.** Completeness of the representative MAGs (%) **g.** Contamination of the representative MAGs (%).

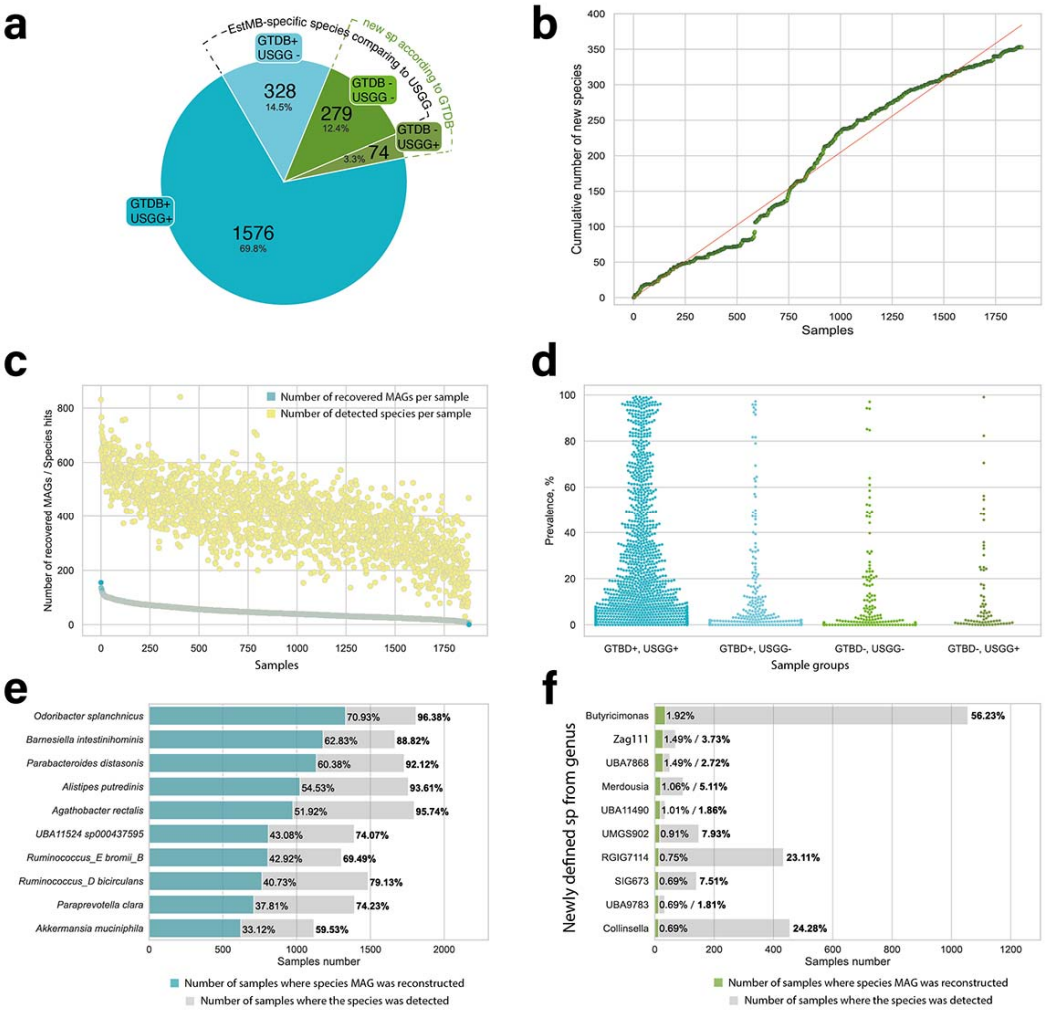

Figure 2. Overview of species from the EstMB MAG collection a. Presence of EstMB Species in GTDB and UHGG Databases: Dark green (UHGG+, GTDB-) and light green (UHGG-, GTDB-) colors indicate new species based on GTDB criteria of novelty. Light blue (UHGG-, GTDB+) and light green (UHGG-, GTDB-) colors denote EstMB-specific species compared to the UHGG database **b.** The relationship between the number of samples analyzed and the cumulative number of new species identified **c.** Number of recovered species and species detected by mapping the reads against the EstMB MAG Collection. Samples are sorted from those with the highest to the lowest number of recovered MAGs **d.** The prevalence of species in EstMB grouped by the presence in the GTDB and UHGG databases **e.** The prevalence and number of recovered MAGs per species. The top 10 species with the highest number of recovered MAGs are shown. Blue bars represent the number of samples where MAG of the species were recovered, while gray bars show the species prevalence in EstMB **f.** The prevalence and number of recovered MAGs per new species. The top 10 new species with the highest number of recovered MAGs are

shown. Blue bars represent the number of samples where MAG of the new species were recovered, while gray bars show the new species prevalence.

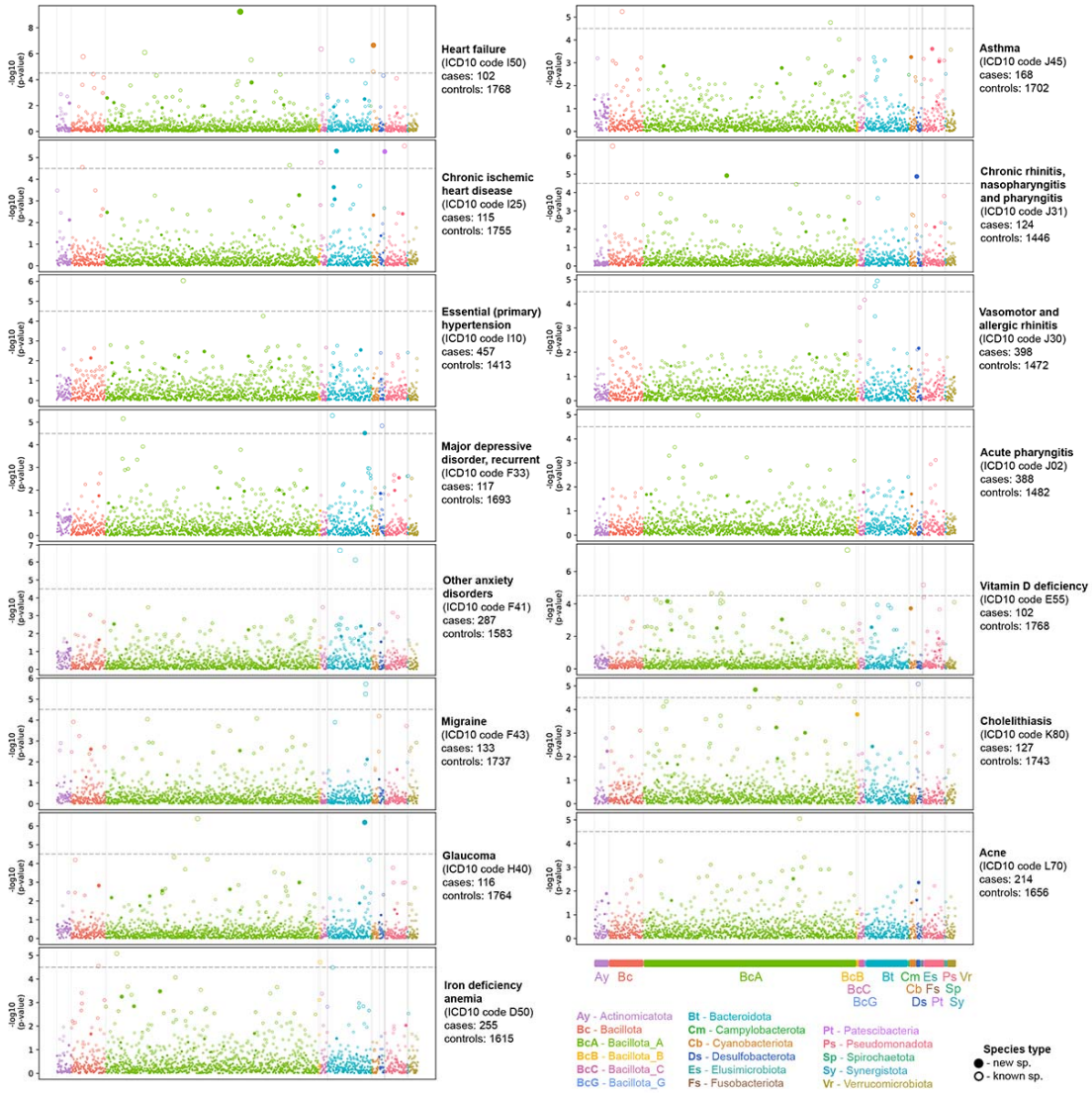

Figure 3. Metagenome-wide associations between the abundance of gut bacterial species and disease. Each data point corresponds to a single species, with its vertical position reflecting the logarithmic p-value obtained from linear regression models. The plot showcases results for 15 diseases, where statistically significant associations were identified. Species are arranged based on their phylogenetic relationships, with known species denoted by empty dots and new species represented by solid dots.

Tables

Table 1. Associations between prevalent diseases and bacterial species. Potentially novel species are highlighted in green. Species not overlapping with the Unified Human Gastrointestinal Genome (UHGG) reference database are highlighted in blue.

n	ICD10 code	disease	bacterial species	p-value

1	L70	Acne	*Blautia_A massiliensis*	0,000008915983941
2	J02	Acute pharyngitis	*UMGS1470 sp900552105*	0,00001054774178
3	J45	Asthma	*Turicibacter bilis*	0,000005749297645
			Coproplasma sp905207385	0,00001721266946
4	K80	Cholelithiasis	***UBA1259* sp.**	0,00001432202061
			Evtepia gabavorous	0,000009760191658
			Bilophila faecipullorum	0,000008196822562
5	I25	Chronic ischemic heart disease	*Escherichia coli*	0,000002892356114
			***Nanosynbacter* sp.**	0,000005235151474
			Merdibacter merdavium	0,00002774592404
			***JADFUS01* sp.**	0,000004977293113
			Copromonas sp900541255	0,00002242419627
			Succiniclasticum sp900544275	0,00001693128266
6	J31	Chronic rhinitis, nasopharyngitis and pharyngitis	***Bilophila* sp.**	0,00001340086411
			***SIG427* sp.**	0,00001204590597
			Leuconostoc falkenbergense	0,0000002986438041
7	I10	Essential (primary) hypertension	*Ventrimonas sp900540335*	0,0000009373027266
8	H40	Glaucoma	*Ruminiclostridium_E sp900539195*	0,0000004064963786
			***Alistipes* sp.**	0,0000006397283819
9	I50	Heart failure	*Scatousia sp900554385*	0,00002455821243
			Streptococcus parasanguinis_E	0,000001705269539
			Lachnospira sp900316325	0,0000008099232244
			***CAKTXU01* sp.**	0,0000000005783512104
			***Stercorousia* sp.**	0,0000002232363058
			CAKNOI01 sp930990125	0,000003004294387
			Prevotella sp022769285	0,000003320167463
			Succiniclasticum sp900544275	0,0000004401380839
10	D50	Iron deficiency anemia	***CALXSV01 sp944390945***	0,00001947728143
			Phil1 sp004558525	0,000008346134583
			Absicoccus sp000434355	0,00002861282809
11	F33	Major depressive disorder, recurrent	*Hungatella effluvii*	0,000006997386775
			Desulfovibrio faecigallinarum	0,00001453214507
			Bacteroides faecis	0,000005161550417
			***Alistipes* sp.**	0,00003053997865
12	G43	Migraine	***Merdimorpha sp002314265***	0,000001888145566
13	F41	Other anxiety	*Paraprevotella sp900548345*	0,0000007644509752

		disorders	*Bacteroides uniformis*	0,0000001983804184
14	J30	Vasomotor and allergic rhinitis	*Bacteroides ovatus*	0,00001853318449
			Bacteroides intestinigallinarum	0,0000113423063
15	E55	Vitamin D deficiency	*Ventrimonas sp900540335*	0,00002326338047
			Borkfalkia excrementipullorum	0,000006283389628
			Enterocloster clostridioformis	0,0000000416377185
			Aphodousia sp900544345	0,000006741078006
			***Oscillospiraceae* sp.**	0,00002239091572

Methods

Estonian Microbiome deep cohort

The Estonian Microbiome Cohort (EstMB) was established in 2017, when stool, oral, and blood samples were collected from 2509 Estonian Biobank (EstBB) participants. The EstBB is a volunteer-based population cohort initiated in 1999 that currently includes over 202,282 genotyped adults (≥ 18 years old) across Estonia. All individuals from the EstBB cohort have been genotyped, and multiple additional datasets are available for a subset of the cohort. Extensive information was collected on the EstMB participants, including data from both questionnaires (self-reported) and electronic health records (EHRs) (completed by medical professionals). The EHR data on the diseases, medications, medical procedures, and causes of death were obtained from the Estonian Health Insurance Fund, Estonian National Death Registry, Estonian Cancer Registry, Estonian Causes of Death Registry, and the two largest hospitals in Estonia (University of Tartu Clinic and North Estonia Medical Centre). In addition to the questionnaire and EHR data, the participants' anthropometric measurements (height, weight, blood pressure, and waist and hip circumferences) were taken during a pre-registered visit upon delivering the stool sample.

The Estonian Microbiome Deep cohort (EstMB-deep) includes a subset of samples from the EstMB that have been resequenced with deep coverage (N = 1878). Samples for EstMB-deep were randomly selected from EstMB, with the number of samples based on DNA availability.

Shotgun metagenomic sequencing

Stool sample collection and DNA extraction were performed following a previously published protocol (Aasmets and Krigul et al, 2022). DNA was quantified from all samples using a Qubit 2.0 Fluorometer with a dsDNA Assay Kit (Thermo Fisher Scientific). Metagenomic sequencing was carried out using MGI technology. Library preparation and circularization of equimolarly pooled libraries was carried out with MGIEasy FS DNA Library Prep Set (MGI Tech, Shenzhen, China). The shotgun metagenomic paired-end sequencing was performed by DNBSEQ-G400RS High-throughput Sequencing Set (FCL PE150) according to the manufacturer's instructions (MGI Tech, Shenzhen, China). Sequencing reads' quality control (QC) was performed using Fastqc (v 0.12.1) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and human reads were filtered using Bowtie2 (v 0.6.5)[28] against the GRCH build 38, patch release 14. The remaining reads were assembled into contigs with MEGAHIT v1.2.911[29].

Metagenome assembly and Binning

Binning was performed separately for each individual sample. Initially, contigs were binned using single binners: MetaBAT v2.15[30], MaxBin v2.2.7 (https://sourceforge.net/projects/maxbin2), and VAMB v3.0.7[31]. Given that different binning tools reconstruct genomes differently, a bin aggregation software, i.e., DAS Tool v1.1.4[32] was employed to integrate the bin predictions from VAMB, MetaBAT2, and MaxBin2. This approach

426 optimized the selection of non-redundant, high-quality bin sets using default parameters. MAGs
427 resulting from this process form the EstMB MAG collection.

428 **Representative MAGs and MAG quality estimation**

429 Representative MAGs were selected from the EstMB MAG collection by clustering MAGs from
430 EstMB MAG collection on species level (ANI index = 95) with dRep[33]. Taxonomy of all
431 representative MAGs was assigned with GTDB-Tk v2.3.0[6,17,34]. If the taxonomic classification of a
432 MAG at the species level or higher couldn't be predicted using the GTDB-Tk, the recovered MAG
433 was treated as a MAG belonging to a new species. We referred to this criterion as the GTDB
434 criteria for novelty. MAG quality, including completeness and contamination, were estimated
435 using CheckM v2[27]. MAG statistics, including total size, number of contigs, N50, GC content, etc.,
436 were obtained using seqkit v2.3.1[35].

437 **Comparison with USGG global reference**

438 All 2,257 representative MAGs underwent a second round of clustering with MAGs from the
439 USGG collection[6] to identify species present in both collections as well as species unique to
440 either the USGG or EstMB MAGs collection. Clustering was performed at the species level (ANI
441 index = 95) using dRep[33]. If an EstMB representative MAG did not cluster with any USGG MAG, it
442 was considered absent from the global reference USGG, potentially indicating it as a population-
443 specific gut MAG. We refer to these MAGs as EstMB-specific MAGs.

444 **Species relative abundance and association analysis**

445 To evaluate species' relative abundance and prevalence, reads from all samples in the EstMB-
446 deep cohort were mapped against representative MAGs using CoverM
447 (https://github.com/wwood/CoverM) and aggregated into a relative abundance table. For the
448 association analysis, we used species' centered log ratio (CLR) transformed relative abundance
449 data and participants' health information from electronic health records (EHRs). We selected 30
450 common diseases based on ICD10 codes, each with over 100 cases within the EstMB-deep
451 cohort. Out of the 2,257 bacterial species observed in our cohort, we selected 1,559 species with
452 a prevalence exceeding 1% for the association analysis. Linear regression models, adjusted for
453 BMI, gender, and age, were constructed to evaluate the association between the selected
454 diseases and species abundance. A stringent Bonferroni correction was applied to the
455 significance level, adjusting for the number of analyzed species, resulting in a corrected alpha of
456 3.13 × 10^-5 (from an original alpha of 0.05).

457 **Data availability**

458 Representative MAGs from EstMB-deep cohort samples have been deposited in the European
459 Nucleotide Archive under study accession ERP161349. The phenotype data contain sensitive
460 information from healthcare registers, and they are available under restricted access through
461 the Estonian biobank upon submission of a research plan and signing a data transfer agreement.
462 All data access to the Estonian Biobank must follow the informed consent regulations of the
463 Estonian Committee on Bioethics and Human Research, which are clearly described in the Data
464 Access section at https://genomics.ut.ee/en/content/estonian-biobank. A preliminary request
465 for raw metagenome and phenotype data must first be submitted via the email address
466 releases@ut.ee.

467 **Code availability**

468 The source code for the analyses is available at GitHub:
469 https://github.com/Chartiza/EstMB_MAGs_db_paper

470

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

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

6. 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.

7. 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.

8. 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.

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

10. 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.

11. 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.

12. 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

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

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

521 15. Brushett, S., Gacesa, R., Vich Vila, A., Brandao Gois, M.F., Andreu-Sánchez, S., Swarte, J.C.,
522 Klaassen, M.A.Y., Collij, V., Sinha, T., Bolte, L.A., et al. (2023). Gut feelings: the relations
523 between depression, anxiety, psychotropic drugs and the gut microbiome. Gut Microbes *15*,
524 2281360. https://doi.org/10.1080/19490976.2023.2281360.

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

530 17. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2022). GTDB-Tk v2: memory
531 friendly classification with the Genome Taxonomy Database (Bioinformatics)
532 https://doi.org/10.1101/2022.07.11.499641.

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

537 19. Parizadeh, M., and Arrieta, M.-C. (2023). The global human gut microbiome: genes, lifestyles,
538 and diet. Trends Mol. Med. *29*, 789–801. https://doi.org/10.1016/j.molmed.2023.07.002.

539 20. Hasan, N., and Yang, H. (2019). Factors affecting the composition of the gut microbiota, and
540 its modulation. PeerJ *7*, e7502. https://doi.org/10.7717/peerj.7502.

541 21. Nishijima, S., Suda, W., Oshima, K., Kim, S.-W., Hirose, Y., Morita, H., and Hattori, M. (2016).
542 The gut microbiome of healthy Japanese and its microbial and functional uniqueness. DNA
543 Res. *23*, 125–133. https://doi.org/10.1093/dnares/dsw002.

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

546 23. Nunes Da Rocha, U., Kasmanas, J.C., Toscan, R., Sanches, D.S., Magnusdottir, S., and Saraiva,
547 J.P. (2023). Simulation of 69 microbial communities indicates sequencing depth and false
548 positives are major drivers of bias in Prokaryotic metagenome-assembled genome recovery.
549 Preprint, https://doi.org/10.1101/2023.05.02.539054
550 https://doi.org/10.1101/2023.05.02.539054.

551 24. Howe, A., and Chain, P.S.G. (2015). Challenges and opportunities in understanding microbial
552 communities with metagenome assembly (accompanied by IPython Notebook tutorial).
553 Front. Microbiol. *6*. https://doi.org/10.3389/fmicb.2015.00678.

25. Lapidus, A.L., and Korobeynikov, A.I. (2021). Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms. Front. Microbiol. *12*, 613791. https://doi.org/10.3389/fmicb.2021.613791.

26. 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.

27. Chklovski, A., Parks, D.H., Woodcroft, B.J., and Tyson, G.W. (2022). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning (Bioinformatics) https://doi.org/10.1101/2022.07.11.499243.

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2021). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. https://doi.org/10.1093/nar/gkab776.

35. 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.