DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
465

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
465

KATERYNA PANTIUKH

From sequences to knowledge: challenges and opportunities of genome-resolved metagenomics

UNIVERSITY OF TARTU
Press
1632

Institute of Genomics, University of Tartu, Estonia

This dissertation is accepted for the commencement of the degree of Doctor of Philosophy in Genomics on June 8, 2026, by the Council of the Institute of Genomics, University of Tartu.

Supervisors: Prof. Elin Org, PhD
Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

Reviewer: Prof. Maido Remm PhD
Institute of Molecular and Cell Biology, University of Tartu, Estonia

Opponent: Prof. Rob Knight, PhD
Professor, Halıcıoğlu Data Science Institute, University of California San Diego

Commencement: Room No. 105, 23B Riia St., Tartu, on August 24, 2026, at 13:15.

The publication of this dissertation is granted by the Institute of Genomics at the University of Tartu, Estonia.

This research was funded by Estonian Research Council grants PUT 1371, PRG1414 and EMBO Installation grant 3573; European Regional Development Fund project no.15-0012 GENTRANSMED; Estonian Center of Genomics / Roadmap II project no.16-0125; the Doctoral School of Biomedicine and Biotechnology scholarship. Data analyses were carried out in part at the High-Performance Computing Center of the University of Tartu, Estonia. The writing of the papers was supported by writing retreats and writing days organized by the Institute of Genomics at the University of Tartu.



ISSN 1024-6479 (print)
ISBN 978-9908-57-255-0 (print)
ISSN 2806-2140 (pdf)
ISBN 978-9908-57-256-7 (pdf)

Copyright: Kateryna Pantiukh, 2026

University of Tartu Press
www.tyk.ee

Сміливі завжди мають щастя

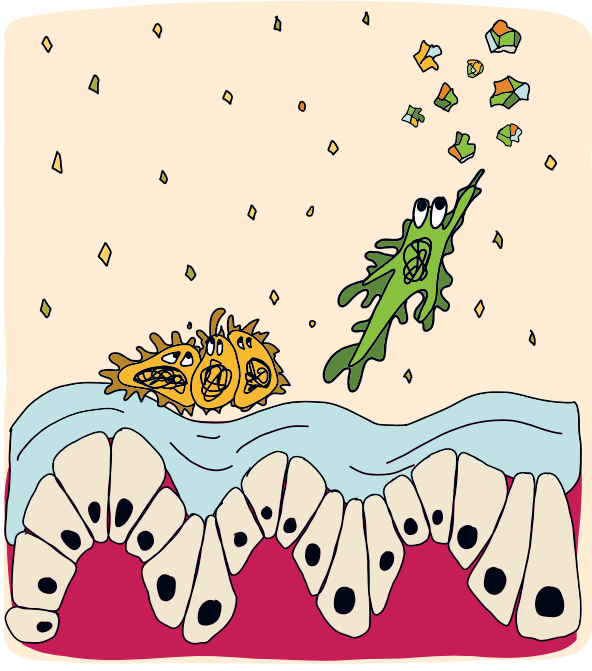


TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS 8
LIST OF ABBREVIATIONS 9
INTRODUCTION 10
1. REVIEW OF THE LITERATURE 11
1.1 Microbiome: core concepts 11
1.2 Early stages of microbiome study 12
1.3 Development of genome-resolved metagenomics 14
1.3.1 Main steps of metagenome assembly 14
1.3.2 History of metagenome assembly 17
1.4 MAG-based databases 19
1.4.1 Databases embedded in taxonomic profiling tools 23
1.5 Bacterial species concept and within-species diversity 23
1.6 Future perspectives 27
1.6.1 Within-species population structure 27
1.6.2 Technical limitations of metagenome sequencing 28
2. AIMS OF THE STUDY 29
3. MATERIAL AND METHODS 30
4. RESULTS AND DISCUSSION 32
4.1 Microbiome community profiling (Ref. I) 32
4.1.1 Taxonomic concordance among sequencing platforms 34
4.1.2 Functional analysis challenges 37
4.2 Genome-resolved metagenomics (GRM) (Ref. II, Ref. III) 38
4.2.1 Metagenome reconstruction 38
4.2.2 Hidden diversity of a new species 40
4.2.3 Assembly-detection gaps 43
4.2.4 Population-specific reference 46
4.2.5 Hidden within-species diversity 48
4.2.6 Hidden within-species within-sample diversity 54
4.3 Genome-resolved microbiome-wide association studies (MWAS) (Ref. II) 55
4.3.1. Species-level MWAS 55
4.3.2. Within-species MWAS 56
4.4 Next challenges to overcome 58
4.4.1 Resolving within-species population structure 58
4.4.2 Reconsidering the species concept 59
CONCLUSIONS 61
SUMMARY IN ESTONIAN 62
REFERENCES 64
ACKNOWLEDGMENTS 71
PUBLICATIONS 75
CURRICULUM VITAE 123
ELULOOKIRJELDUS 126

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications, referred to in the text by Roman numerals (Ref. I to Ref. III):

I Kinga Zielińska*****, **Kateryna Pantiukh*,** Paweł P Łabaj, Tomasz Kosciolek, Elin Org "A large-scale comparative metagenomic analysis of short-read sequencing platforms indicates high taxonomic concordance and functional analysis challenges." mSystems, DOI: 10.1128/msystems.01714-25. https://journals.asm.org/doi/10.1128/msystems.01714-25

II **Pantiukh, Kateryna**; Aasmets, Oliver; Krigul, Kertu Liis; Org, Elin (2026). Metagenome-assembled genomes from a population-based cohort uncover new gut species and within-species diversity, revealing prevalent disease associations, mSystems 2024.07.06.602324. https://journals.asm.org/doi/10.1128/msystems.00114-26

III **Pantiukh, Kateryna**, and Elin Org. "Human gut archaea collection from Estonian population." Scientific data vol. 13,1 366. 6 Feb. 2026, doi:10.1038/s41597-026-06742-1. https://www.nature.com/articles/s41597-026-06742-1

***** These authors contributed equally

The publications listed above have been reprinted with the permission of the copyright owners.

My contributions to the listed publications were as follows:

Ref. I Created the study design, performed the taxonomic profiling, interpreted the results, prepared figures and tables, wrote the manuscript, and participated in the critical review of the paper.

Ref. II Helped with the study design, created the study questions, reconstructed the MAGs, created study hypotheses, interpreted the results, prepared the figures, uploaded MAGs to ENA, participated in discussions, wrote the original manuscript, and participated in the critical review of the paper.

Ref. III Reconstructed the MAGs and created the archaea MAG database, uploaded the data to ENA, wrote the original manuscript, and participated in the critical review of the paper.

LIST OF ABBREVIATIONS

MAG	Metagenome assembled genome
SAG	Single-amplified genome
ncMAG	Near-complete MAG
repMAG	Representative MAG for species
GRM	Genome resolved metagenomics
EstMB	Estonian microbiome cohort
EstMB-deep	Estonian microbiome deep cohort
EstMB-allMAG	Estonian collection of all reconstructed MAGs
EstMB-repMAG	Estonian collection of species representative MAGs
ANI	Average nucleotide identity
GU	Genome unit
GUN	Genome unit number
nGUN	Normalised genome unit number
MWAS	Microbiome wide association study
MIMAG	Minimum Information about a Metagenome-Assembled Genome
MISAG	Minimum Information about a Single Amplified Genome
FISH	Fluorescence in situ hybridization
qPCR	Quantitative polymerase chain reaction

INTRODUCTION

It's hard to grasp both the importance of microorganisms in our lives and their differences from the macro world we've grown accustomed to.

Bacteria and archaea are closely linked to human health. In the gut, they can benefit the host by expanding metabolic capacity, degrading otherwise indigestible dietary compounds, producing metabolites that regulate immunity, and helping protect against invading pathogens. At the same time some microbial activities may erode the colonic mucus barrier or generate harmful molecules associated with inflammation and disease. Despite this recognized importance, we still do not fully understand the mechanisms by which gut microbes influence host health.

Recent advances in genome-resolved metagenomics have transformed microbiome research by enabling the reconstruction of microbial genomes directly from complex communities. With this approach, massive resources have been created, including thousands of metagenome-assembled genomes that have revealed species that were previously unknown along with extensive diversity within species that were already known. Genome-resolved metagenomics therefore provides an opportunity not only to expand microbial reference catalogues, but also to refine how microbiome variation is detected, interpreted and linked to host phenotypes. It provides an opportunity to revisit key concepts in microbiology, including how bacterial species and subspecies-level entities should be defined. In the coming years, genome-resolved metagenomics is likely to bring a new level of understanding of the microbiome and its connection to human health.

However, several important challenges remain. Population-based microbiome studies often combine sequencing data generated on different platforms, making it necessary to evaluate whether platform-specific effects influence taxonomic and functional profiling. Second, global reference databases may not fully capture population-specific microbial diversity, potentially limiting the detection of taxa that are common in particular cohorts but underrepresented in existing catalogues. Third, species-level profiling may obscure biologically relevant within-species variation that may be important for understanding microbiome-host associations.

In this thesis, I use a cohort of 2,504 volunteers from Estonia, including two datasets generated with moderate and deep sequencing, to explore the potential and limitations of current genome-resolved metagenomics in revealing connections between the microbiome and human health. In the first part of the thesis, I show that different sequencing platforms provide comparable microbiome community profiles and can therefore be combined within a single study, although functional profiling is more sensitive to platform-related differences than taxonomic profiling. Second, I demonstrate that population-specific metagenome assembly can reveal new information about both species-level and within-species diversity. Finally, I integrated microbiome profiles with disease diagnosis obtained from electronic health record to identify microbiome-health associations and to show how improved characterization of microbial diversity can provide new insights into microbiome-host interactions.

1. REVIEW OF THE LITERATURE

1.1 Microbiome: core concepts

The **microbiome** can be conceptualized as an ecosystem or a community, composed of numerous interacting microorganisms that collectively respond to internal feedback mechanisms and external perturbations (**Figure 1**).

In this context, the **metagenome** refers to the total genetic material recovered directly from a microbial community within a given sample (Handelsman et al. 1998).

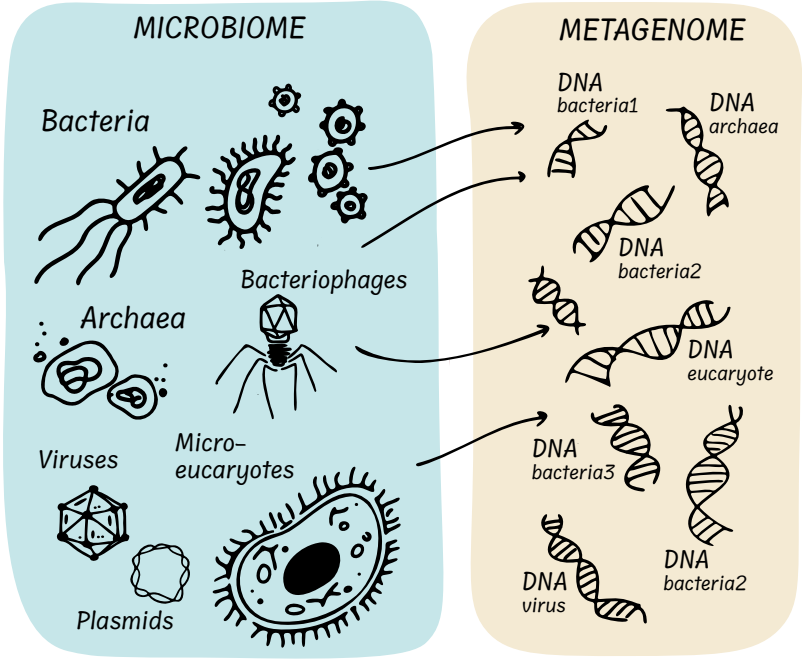


Figure 1. Microbiome and the metagenome as core concepts in microbiome studies. The microbiome represents the microbial community, while the metagenome represents its collective genetic content. Authors' own work.

The **human gut microbiome** refers to the community of microorganisms inhabiting the gastrointestinal tract. This community consists of hundreds of members and is dominated by bacteria, but also includes archaea, viruses, including bacteriophages, and microeukaryotes. Its composition and functional organization are influenced by a wide range of factors, including host genetics (Kurilshikov et al. 2021; Qin et al. 2022), age and geography (Yatsunenko et al. 2012), early-life exposures (Bolte, Moorshead, and Aagaard 2022), diet (Segev et al. 2026), medication use (Aasmets et al. 2022), and environmental factors (Rothschild et al. 2018). In addition to these external drivers, the gut microbiome is shaped by microbe-microbe interactions, host-microbe crosstalk, which together generate community-level properties that cannot be inferred from individual taxa alone.

Microbiome research focuses on communities of microorganisms, their interactions, and the collective properties that arise from these interactions. Microbiology provides the historical and conceptual foundation for microbiome research. A basic understanding of individual microbial components is necessary for studying complex communities, but it is not sufficient, as community-level behaviour is not equivalent to the sum of the behaviours of its individual members. As a result, microbiome research became feasible only when methodological advances enabled the assessment of entire microbial communities as integrated systems.

Genome-resolved metagenomics (GRM) is a metagenomic approach in which genomes of individual microorganisms are directly reconstructed from complex environmental or host-associated microbial DNA sequence data, without requiring cultivation (Kayani et al. 2021). By linking community-level sequencing data to reconstructed microbial genomes, GRM provides a framework for studying both taxonomic diversity and genome-level functional potential. GRM has become a major enabling technology for microbiome research (Pasolli et al. 2019) and now plays an important role in the emerging field of microbiome medicine (N. Kim et al. 2024; Ratiner et al. 2024).

1.2 Early stages of microbiome study

As the human gut microbiome is a complex system, it consists of many interacting components whose interactions are nonlinear, often adaptive. These interactions give rise to emergent, system-level behaviours that cannot be readily inferred from the properties of individual taxa alone. At the same time, a foundational understanding of the individual components and their basic functions remains essential.

Before the establishment of population-scale microbiome research, studies of human-associated microbial communities were largely framed as investigations of "microflora" or "microbiota" and were typically limited to small, specific sample sets. These studies relied mainly on cultivation, microscopy, FISH, targeted PCR, and Sanger sequencing of 16S rRNA genes, providing important early insights but only a partial view of microbial diversity (Tannock et al. 2000; Verhelst et al. 2005; Wilson and Blitchington 1996). A central limitation of these approaches was that many human-associated microorganisms are difficult, or even impossible, to cultivate under standard laboratory conditions. Suau et al. reported that microscopic counts suggested that 60–80% of observable fecal bacteria were not recovered by culture, while 16S rRNA clone libraries revealed numerous previously undescribed gut bacterial lineages (Suau et al. 1999). However, 16S rRNA-based profiling was also not a complete solution, as it relies on a single phylogenetic marker, has limited species- and strain-level resolution, and does not directly capture the functional gene content or metabolic potential of microbial communities.

These limitations motivated a shift toward high-throughput shotgun metagenomics, enabling standardized, population-scale analyses of both microbial community composition and functional potential (MetaHIT Consortium et al. 2010; Tonnelé et al. 2025). This shift was institutionalized through large international initiatives, most notably the NIH Human Microbiome Project (HMP), launched in 2007 (The Integrative HMP (iHMP) Research Network Consortium et al. 2019; The NIH HMP Working Group et al. 2009), and the European MetaHIT initiative (MetaHIT Consortium et al. 2010). These landmark initiatives leveraged advances in high-throughput sequencing to systematically catalogue microbial diversity across body sites and individuals, establishing the first comprehensive reference frameworks for the healthy human microbiome.

The HMP1 and HMP2 projects together generated more than 42 terabytes of publicly available data, including 16S rRNA gene profiles, whole-metagenome shotgun sequencing data and additional molecular measurements (The Integrative HMP (iHMP) Research Network Consortium et al. 2019; The NIH HMP Working Group et al. 2009). Early analyses based on 16S rRNA gene sequencing established a foundational view of microbiome composition across body sites, enabling the identification of broad ecological patterns such as pronounced inter-individual variability alongside site-specific community structures.

In parallel, the European MetaHIT project focused specifically on the human intestinal tract and placed whole-metagenome sequencing at the centre of its strategy. Rather than relying primarily on 16S rRNA gene surveys, its landmark study used Illumina shotgun sequencing of total DNA from faecal samples of 124 European individuals to reconstruct a catalogue of 3.3 million non-redundant microbial genes. This shift from taxonomic profiling towards gene-centred metagenomics was important because it allowed the gut microbiome to be described not only by which organisms were present, but also by the functional potential encoded by the community (MetaHIT Consortium et al. 2010).

Subsequent advances in shotgun metagenomics, especially *de novo* assembly and genome-resolved approaches, expanded reference collections and enabled recovery of metagenome-assembled genomes, thereby revealing previously uncultured and uncharacterized microbial diversity (Almeida et al. 2019).

Early large-scale studies revealed substantial inter-individual variability in taxonomic composition, whereas community functional profiles appeared comparatively more stable across individuals, a pattern also observed in the Estonian microbiome cohort (Aasmets et al. 2022). This apparent discrepancy has often been interpreted as evidence of functional redundancy, whereby different microbial taxa can perform similar ecological roles, thereby buffering taxonomic variation. However, an alternative explanation lies in methodological limitations: current functional annotations are strongly biased toward well-characterized, conserved genes that are widely shared across taxa. As a result, biologically meaningful variation may be obscured when analyses are restricted to broad pathway categories or to genes with established annotations, leaving a substantial fraction of functional diversity unresolved (Bradley and Pollard 2017).

1.3 Development of genome-resolved metagenomics

Metagenome assembly is the central component of genome-resolved metagenomics, as it enables reconstruction of individual microbial genomes directly from mixed-community shotgun sequence data, without the need for cultivation.

1.3.1 Main steps of metagenome assembly

The assembly workflow generally consists of several consecutive steps, from preprocessing of raw reads to genome reconstruction and quality assessment. After read quality control and removal of low-quality or host-derived sequences, short reads from one metagenomic sample are assembled into longer sequences, called **contigs**, based on sequence overlap.

Because these contigs originate from multiple community members, they are then grouped into **bins** using features such as sequence composition, including tetranucleotide frequency and GC content, differential coverage across samples, taxonomic marker genes, and, in some approaches, machine-learning-based classification. This process is called **binning.** To improve the quality of the final genome bins, multiple binning tools can be applied to the same metagenome sample, followed by comparison, refinement, and selection of the highest-quality bins. This strategy is often referred to as multi-binning.

Each bin is assessed for genome completeness and contamination, and, when it meets accepted quality criteria, reported as a **metagenome-assembled genome** (MAG). MAGs are subsequently assigned a taxonomic classification and functionally annotated (Chivian et al. 2023) (**Figure 2**).

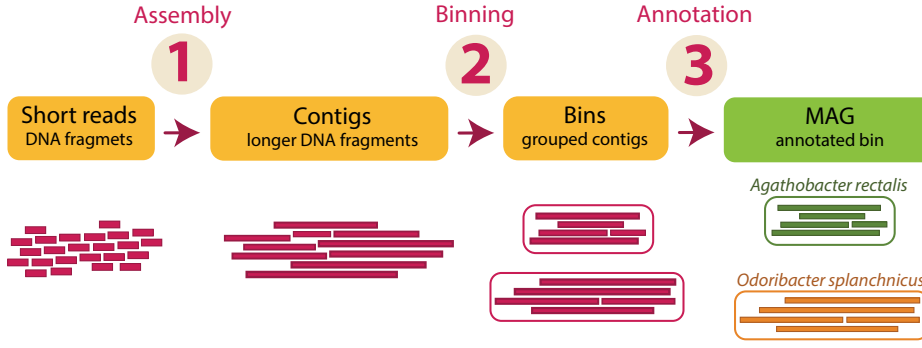


Figure 2. Schematic overview of metagenome assembly. In the first stage, short sequencing reads are assembled into longer DNA fragments called contigs based on sequence overlap. In the second stage, known as binning, contigs are grouped into bins based on their properties. In the final stage, each bin is evaluated for genome completeness, contamination, and taxonomic position. Bins with quality assessment and taxonomic classification are referred to as metagenome-assembled genomes (MAGs). Authors' own work.

MAG quality is commonly evaluated using two complementary metrics: genome **completeness** and genome **contamination**, both reported as percentages. In the original CheckM framework, these metrics were estimated using a lineage-aware approach based on curated sets of single-copy marker genes expected to be ubiquitous within the taxonomic lineage to which a MAG is assigned (Parks et al. 2015). Completeness is defined as the proportion of expected marker genes recovered in a MAG, whereas contamination is inferred from the occurrence of single-copy marker genes in multiple copies, which may indicate the sequences from other genomes (The Genome Standards Consortium et al. 2017) (**Figure 3**).

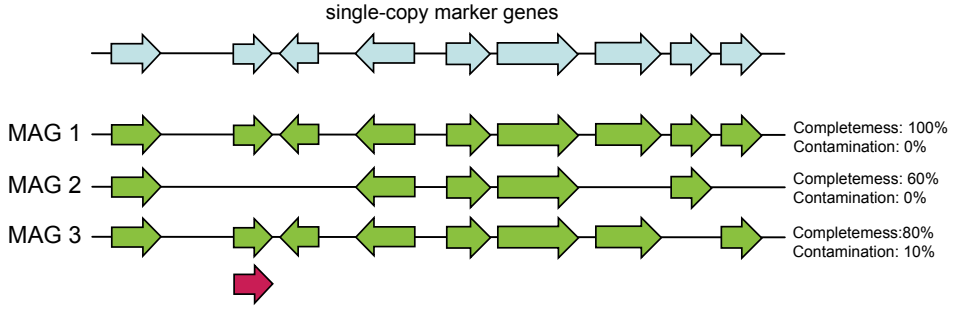


Figure 3. Schematic representation of completeness and contamination estimation. Lineage-aware, curated sets of single-copy marker genes expected to be present in a MAG are shown in blue. Marker genes recovered in the MAG are shown in green, whereas duplicated copies are shown in red. Completeness is defined as the proportion of expected marker genes recovered in the MAG, while contamination is inferred from the presence of single-copy marker genes in multiple copies. Authors' own work.

Although this approach performs well for well-characterized lineages, its accuracy can be reduced for poorly represented, highly divergent or reduced-genome lineages, where the expected marker-gene repertoire may differ from reference-based assumptions. To address these limitations, CheckM2 introduced machine-learning-based models for estimating MAG completeness and contamination, providing improved predictions across bacterial and archaeal lineages, including taxonomically new groups and reduced-genome taxa such as Patescibacteria and DPANN archaea (Chklovski et al. 2023).

According to the Minimum Information about a Metagenome-Assembled Genome (**MIMAG**) standard, MAG quality should be evaluated not only by estimated completeness and contamination, but also by assembly-related features, including the presence of rRNA and tRNA genes. Under this framework, high-quality draft MAGs are expected to be >90% complete and <5% contaminated, and to encode the 5S, 16S and 23S rRNA genes, together with tRNAs for at least 18 of the 20 standard amino acids. The same criteria were applied to single-amplified genomes (SAGs), i.e., genomes reconstructed from individual bacterial cells (**Table 1**) (The Genome Standards Consortium et al. 2017).

Table 1. MIMAG/MISAG genome quality categories and reporting standards for SAGs and MAGs

Genome category	Assembly quality	Complete-ness	Contami-nation	rRNA/tRNA requirements
Finished (SAG/MAG)	Single contiguous sequence without gaps or ambiguities with a consensus error rate equivalent to Q50 or better	~100%	0%	Complete rRNA and tRNA gene set expected.
High-quality draft (SAG/MAG)	Multiple fragments where gaps span repetitive regions.	>90%	<5%	Presence of the 23S, 16S, and at least 18tRNAs.
Medium-quality draft (SAG/MAG)	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.	>50%	<10%	Not required
Low-quality draft (SAG/MAG)	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.	<50%	<10%	Not required

SAG – single-amplified genome; MAG – metagenome-assembled genome; rRNA – ribosomal RNA; tRNA – transfer RNA.

However, rRNA and tRNA genes are often difficult to recover from metagenomic assemblies. These loci can occur in multiple copies and include highly conserved or repetitive regions, which are particularly challenging for short-read assembly. As a result, many MAGs with high estimated completeness and low contamination lack one or more rRNA or tRNA genes, despite otherwise satisfying the genomic completeness criteria. For MAGs with >90% completeness and <5% contamination that do not meet the full MIMAG rRNA/tRNA requirements, the term **near-complete MAGs** is often used (Almeida et al. 2019). Long-read and HiFi-based approaches can improve recovery of these difficult regions, including rRNA loci (C. Y. Kim, Ma, and Lee 2022).

The degree of genome fragmentation, reflected by the number of contigs, assembly gaps and contiguity statistics such as N50/L50, is also an important indicator of MAG quality. Although MIMAG recommends reporting assembly statistics, fragmentation is not defined as an explicit threshold for the main MAG quality categories and therefore should be considered as an additional quality descriptor rather than a primary classification criterion (The Genome Standards Consortium et al. 2017). This is particularly important when MAGs are used for accessory genome analysis or within-species comparisons, where fragmented assemblies may underestimate gene content or distort population structure.

1.3.2 History of metagenome assembly

Metagenome assembly has become a key approach for studying microbial communities in their natural environments and is still undergoing active development (**Figure 4**).The conceptual basis of this approach was formalised with the term *metagenome,* defined as the collective genetic content of a microbial community by Jo Handelsman and colleagues in 1998 (Handelsman et al. 1998). In its modern form, however, genome-resolved metagenomics was established by Gene Tyson et al. in 2004, who reconstructed near-complete genomes of *Leptospirillum* group II and *Ferroplasma* type II, along with partial genomes of three additional populations, from a low-complexity acid mine drainage biofilm (Tyson et al. 2004). This study provided the first clear demonstration that multiple genomes could be recovered directly from environmental samples, although at this stage the approach was mainly applicable to relatively simple communities dominated by a small number of populations with limited within-population diversity.

Subsequent work further clarified both the potential and limitations of metagenome assembly. In particular, Philip Hugenholtz and Gene Tyson highlighted key technical constraints, including the dependence on sequencing depth, community complexity, and accurate binning strategies (Hugenholtz and Tyson 2008). An important methodological advance followed in 2013, when Albertsen et al. applied differential coverage binning across multiple metagenomes to recover 31 bacterial genomes from activated sludge, including taxa with low relative abundance (Albertsen et al. 2013). This was a key step toward scalable metagenome assembly, because it showed that genome recovery was not limited to simple microbial system.

As the number of recovered MAGs increased, consistent quality assessment became essential. Tool such as CheckM a lineage-aware framework for estimating genome completeness and contamination and rapidly became a standard tool for evaluating draft genomes recovered from metagenomes (Parks et al. 2015). This was followed by the **MIMAG** and MISAG standard, which provided shared reporting guidelines for metagenome-assembled genomes (MAG) and single-amplified genomes (SAGs) and introduced a shared terminology for quality categories such as high- and medium-quality draft genomes (The Genome Standards Consortium et al. 2017). Together, these developments transformed metagenome assembly from a proof-of-concept approach into a reproducible framework for large-scale microbial genome recovery.

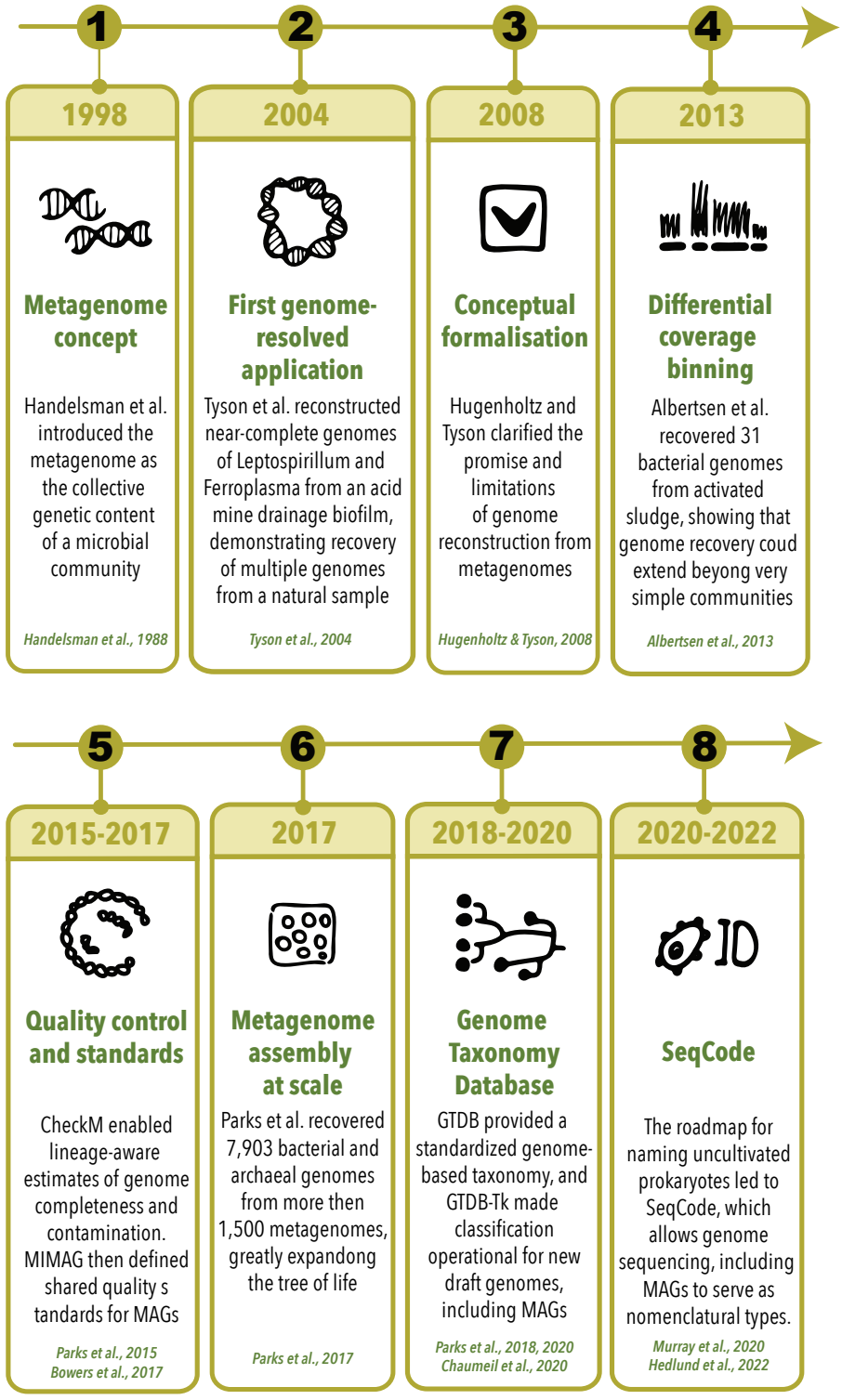


Figure 4. Timeline of major milestones in genome-resolved metagenomics. Authors' own work.

By 2017, the method had reached a new scale. Donovan Parks et al. reconstructed 7,903 bacterial and archaeal genomes from more than 1,500 public metagenomes, substantially expanding the known phylogenetic diversity and providing the first genomic representatives for many previously uncharacterized lineages (Parks et al. 2017). The rapid accumulation of genomes from uncultivated organisms created a strong demand for a standardized, genome-based taxonomic framework. This need was addressed by the Genome Taxonomy Database (GTDB), which applied phylogenomic principles to produce a rank-normalized taxonomy for Bacteria and Archaea (Parks et al. 2018, 2020). GTDB-Tk, a program for assigning GTDB taxonomy to new genomes, subsequently made this framework operational for newly recovered genomes, including MAGs, by enabling consistent taxonomic assignment of draft genomes (Chaumeil et al. 2020).

However, while GTDB provides a classification framework, it does not govern formal nomenclature. Under the traditional International Code of Nomenclature of Prokaryotes (ICNP), uncultivated organisms could not typically receive validly published names, as viable pure cultures were required as nomenclatural types. As the recovery of MAG and single-amplified genomes (SAG) recovery accelerated, this limitation became a major challenge for microbial systematics. A 2020 roadmap for naming uncultivated Archaea and Bacteria outlined the need for an alternative system, and the SeqCode later established such a system by allowing genome sequences themselves, including MAGs and SAGs, to serve as nomenclatural types (Hedlund et al. 2022; Murray et al. 2020; Whitman et al. 2022). In this sense, SeqCode can be viewed as the logical nomenclatural consequence of the metagenome assembly era.

1.4 MAG-based databases

The rapid expansion of genome-resolved metagenomics has driven the accumulation of vast numbers of microbial genomes and, in parallel, the emergence of large reference databases based on metagenome-assembled genomes (MAGs) (**Figure 5**). One of the key resources addressing this challenge is the GTDB, which provides a standardized genome-based taxonomic framework for bacteria and archaea (Parks et al. 2018, 2020). Building on this foundation, large-scale integrative resources have been developed to organize and explore microbial diversity across ecosystems.

GTDB provides a widely used genome-based framework for the taxonomic classification of bacteria and archaea, including both isolate genomes and metagenome-assembled genomes (Parks et al. 2018, 2020). By replacing primarily phenotype- and nomenclature-based assignments with a phylogenetically consistent, rank-normalized taxonomy, GTDB has become a central reference for interpreting prokaryotic diversity in genome-resolved microbiome studies. The most recent release, GTDB R232, issued on 15 April 2026, comprises 901,341 genomes in total, including 878,998 bacterial and 22,343 archaeal genomes. These genomes are organized into 199,923 species clusters, corresponding to 189,801 bacterial and 10,122 archaeal species clusters (Genome Taxonomy

Database. GTDB release R232. Genome Taxonomy Database. Available at: https://gtdb.ecogenomic.org/).

Among the broadest cross-habitat resources, **SPIRE** (Searchable, Planetary-scale mIcrobiome REsource) is currently one of the most comprehensive. SPIRE integrates data from a wide range of habitats, including human-associated, environmental, and engineered ecosystems, thereby substantially expanding known microbial diversity. As of May 2026, SPIRE integrates 99,067 metagenomic samples from 715 studies spanning diverse microbial environments and includes 1.16 million newly reconstructed medium- or high-quality MAGs, together with large-scale protein predictions and species-level genome clustering (Schmidt et al. 2024) (SPIRE. (2026). Searchable Planetary-scale mIcrobiome REsource. European Molecular Biology Laboratory. https://spire.embl.de). While GTDB focuses on standardized taxonomy, (Parks et al. 2022), SPIRE was developed as a searchable, integrative resource for comparative microbiome analysis across ecosystems. A major strength of SPIRE is its interoperability with companion resources. In particular, **VIRE** provides a planetary-scale catalogue of viral genomes reconstructed from a largely overlapping metagenomic universe, while **Metalog** supplies curated sample metadata and precomputed taxonomic profiles. Because these resources are linked through consistent sample identifiers, they enable joint exploration of microbial genomes, viral genomes, and associated contextual metadata from the same or overlapping samples (Kuhn et al. 2026; Nishijima et al. 2026). This interoperability substantially enhances the analytical value of SPIRE for ecological and multi-omic studies.

Such broad resources are valuable because they increase the representation of microbial genome space across environments. However, their broad scope also means that they may not always provide optimal resolution for a specific host population or cohort. While global resources such as the SPIRE database aim to capture microbial diversity across ecosystems, equally important efforts have focused on constructing high-resolution, habitat-specific reference databases.

For the human gut microbiome, the Unified Human Gastrointestinal Genome collection (**UHGG**) remains a foundational reference. Almeida et al. assembled a unified catalogue of nonredundant genomes representing 4,644 gut prokaryotic species and complemented it with the Unified Human Gastrointestinal Protein catalogue (UHGP), thereby consolidating previously fragmented isolate-based and MAG-based resources into a single framework (Almeida et al. 2021). By May 2026, UHGG v2.0.2 included 289,232 genomes representing 4,744 species-level clusters, comprising 4,716 bacterial species and 28 archaeal species (MGnify. (2024). Unified Human Gastrointestinal Genome (UHGG) v2.0.2. EMBL-EBI. Retrieved May 18, 2026, from https://www.ebi.ac.uk/metagenomics/genome-catalogues/human-gut-v2-0-2). This integration enables improved taxonomic resolution and functional annotation within the human gut microbiome. Importantly, more than 70% of the species in UHGG lacked cultured representatives, emphasizing the extent to which our current view of gut microbial diversity still depends on culture-independent genome recovery. Relative to GTDB, UHGG is habitat-specific rather than taxonomy-centred; relative to SPIRE, it exchanges environmental breadth for much deeper representation of a single, highly important host-associated ecosystem.

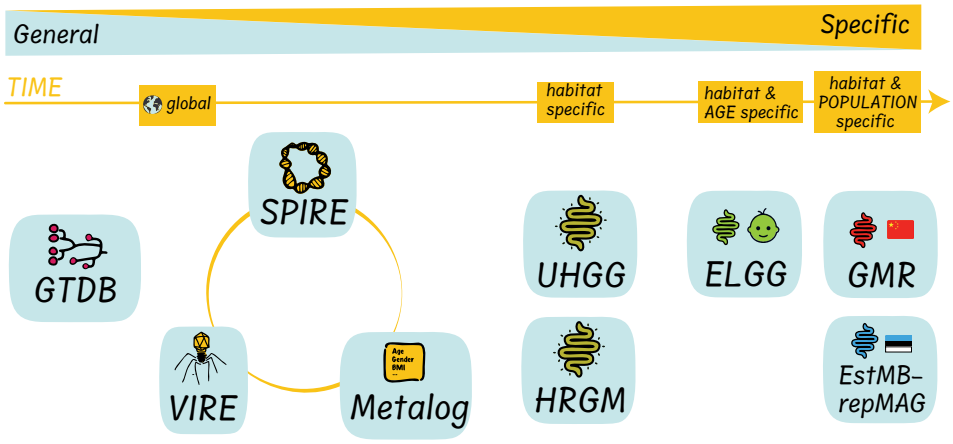


Figure 5. Overview of major MAG databases. Early resources aimed to capture microbial diversity at broad, global or cross-environment scales, as exemplified by the Genome Taxonomy Database (GTDB) and SPIRE, the Searchable Planetary-scale mIcrobiome REsource. More recent databases have increasingly adopted a targeted design, focusing on specific host-associated environments, such as the human gut in the Unified Human Gastrointestinal Genome (UHGG) and the Human Reference Gut Microbiome (HRGM), or combining environmental specificity with additional dimensions, including host age in the Early-Life Gut Genomes (ELGG) catalogue and population background in Gut Microbiome Reference (GMR) and Estonian Microbiome database (EstMB-repMAG). Authors' own work.

A closely related resource is **HRGM**, the Human Reference Gut Microbiome catalog. Kim et al. developed HRGM to address the geographic and population biases in earlier gut genome collections. By incorporating newly assembled genomes from underrepresented Asian metagenomes, HRGM expanded the available gut reference space to 232,098 nonredundant genomes representing 5,414 prokaryotic species, including 780 species do not present in prior catalogues (C. Y. Kim et al. 2021). Thus, the main contribution of HRGM is not simply numerical expansion beyond UHGG, but improved representation of human populations that had been under sampled in earlier resources.

HRGM2 extends this trajectory further by prioritizing genome quality and downstream functional interpretability. In this resource, Ma *et al*. compiled a catalog of near-complete genomes, defined as genomes with at least 90% completeness and no more than 5% contamination, derived from samples collected across 41 countries (Ma et al. 2025). Currently HRGM2 contains 155,211 non-redundant near-complete genomes from 4,824 prokaryotic species and was explicitly designed to improve functional profiling and genome-scale metabolic modelling. Although the total number of genomes is lower than in HRGM, this reflects a stricter quality threshold rather than a narrower biological scope.

Overall, the current trajectory of MAG database development is toward ever-broader genomic coverage, aiming to approximate the full spectrum of microbial species that may occur in a given habitat (Almeida et al. 2021; Schmidt et al.

2024). This approach is well justified, as reference-based profiling performs best when the database contains the genomes that are actually present in the sample. However, database expansion alone is not sufficient. As reference libraries become larger and more redundant, overlap among closely related genomes increases, which can reduce species-level precision by generating ambiguous or high-level taxonomic assignments and by increasing computational burden (Nyström-Persson, Bapatdhar, and Ghosh 2025).

For this reason, an equally important complementary strategy is the construction of population-, region-, or age-specific reference databases. Such resources can better capture lineage diversity underrepresented in broad global catalogues and improve taxonomic and functional profiling in the target cohorts. This is already evident in the human gut field: UHGG produced strong gains in read classification in understudied non-Western populations (Almeida et al. 2021), HRGM improved taxonomic and functional classification by incorporating genomes from underrepresented Asian cohorts (C. Y. Kim et al. 2021).

More recent MAG resources have moved beyond broad environment-specific catalogues by incorporating additional host- or population-level dimensions. For example, the Early-Life Gut Genomes (**ELGG**) catalogue provides an age-specific reference for the human gut microbiome during the first three years of life. ELGG comprises 32,277 MAGs representing 2,172 species reconstructed from 6,122 early-life faecal metagenomes, and increased recruitment of early-life metagenomic reads to 82.8%, supporting the value of life-stage-matched references for improving taxonomic and functional resolution (Zeng et al. 2022). Similarly, Dong et al. constructed the human Gut Microbiome Reference (**GMR**), a population-balanced gut genome resource comprising 478,588 high-quality microbial genomes from Chinese and non-Chinese populations. Their study showed that geographically imbalanced reference databases can underrepresent population-specific diversity: the authors reported that approximately 70% of existing microbial reference data derive from European and North American populations, while integration of new genomes into the taxonomic profiling database improved population-level species profiling by up to 23% (Dong et al. 2025).

In this context, our Estonian Microbiome Cohort-derived MAG resource (**EstMB-repMAG**) provides a further example of a targeted, environment- and population-specific reference, included 84,762 MAGs representing 2,257 species, including 353 previously uncharacterized species [**Ref. II**]. Such cohort-specific resources are not intended to replace global databases, but to complement them by improving detection of microbial diversity that is locally prevalent or underrepresented in broader catalogues.

Thus, the next phase of MAG database development should not be viewed simply as a race toward larger catalogues, but as a balance between breadth and contextual relevance.

1.4.1 Databases embedded in taxonomic profiling tools

It is important to distinguish general MAG-based genome catalogues from the databases embedded in taxonomic profiling tools. Profiling databases are usually transformed into algorithm-specific reference structures and are optimized for the classification strategy of a particular tool.

For example, **MetaPhlAn** uses clade-specific marker genes derived from reference genomes and MAGs from the human gut microbiome and also many other animal and ecological environments (Blanco-Míguez et al. 2023). Instead of operating directly with species as classical taxonomic units, it uses species-level genome bins (SGBs) as profiling units. Each SGB is a group of microbial genomes and MAGs cluster together on species-level. For each SGB, a specific set of marker genes is defined (average of 189 ± 34 unique marker genes per SGB), and species are detected through read mapping to a sufficient fraction of SGB-specific marker genes. The current MetaPhlAn4 database includes 26,970 high-quality species-level genome bins (SGBs).

mOTUs2 similarly relies on universal marker genes, combining reference-derived and metagenome-derived species-level units for microbial profiling, thereby enabling the profiling of more than 7700 microbial species (Milanese et al. 2019). **Kraken2** databases are built from taxonomically labelled genome sequences and represented as k-mers, with the option to use either standard or custom reference collections (Wood, Lu, and Langmead 2019).

These profilers are widely used as primary tools for taxonomic profiling, but their databases differ conceptually from MAG-based genome catalogues and are optimized for the underlying classification strategy, such as marker-gene detection or k-mer matching. As a result, they may provide limited flexibility for incorporating study-specific MAGs, with the exception of tools such as Kraken2 that explicitly support custom databases (Wood, Lu, and Langmead 2019). Moreover, marker-based profiling outputs are not always directly linked to complete genome sequences, which can limit follow-up analyses requiring genome-resolved functional annotation, comparative genomics, or strain-level investigation.

In this context, profiling directly against MAG-based catalogues, for example through read-mapping approaches such as **CoverM**, provides a complementary strategy in which reads are quantified directly against selected genome or MAG catalogues, including study-specific references (Aroney et al. 2025). This approach is particularly relevant when the aim is to quantify cohort-specific MAGs, newly reconstructed species, or within-species genome units.

1.5 Bacterial species concept and within-species diversity

Resources such as SPIRE, UHGG, and HRGM now contain hundreds or thousands of MAGs for many bacterial species. The accumulation of large numbers of genomes per species enables examination of within- species diversity at a scale that was previously impossible when analysis relied on a limited number of

cultured isolates. To understand the importance of within-species diversity, it is necessary to examine how the concept of species is defined in prokaryotes compared to eukaryotes.

In animals and plants, species are often defined using the biological **species concept**, which is based on sexual reproduction and reproductive isolation (Bozdag and Ono 2022; Westram et al. 2022). This framework does not transfer directly to bacteria, where reproduction is primarily clonal and genetic exchange is decoupled from reproduction itself (Bobay, Traverse, and Ochman 2015; Didelot et al. 2010). As a result, **bacterial species** are typically defined in terms of genomic cohesion rather than reproductive isolation (Fraser et al. 2009; Konstantinidis, Ramette, and Tiedje 2006; Rosselló-Mora 2001).

Genomic cohesion is commonly operationalized using whole-genome similarity measures, especially average nucleotide identity (ANI), which provides a quantitative estimate of nucleotide-level similarity between genomes (Goris et al. 2007; Konstantinidis and Tiedje 2005). Because ANI is a continuous variable, a key challenge is determining the threshold that separates genomes belonging to the same species from those that do not (Goris et al. 2007; Richter and Rosselló-Móra 2009). Large-scale analyses have shown that genomes assigned to the same species typically share at least about 95% ANI, whereas interspecies comparisons generally fall below this level, producing a clear discontinuity in genome similarity (Goris et al. 2007; Jain et al. 2018). This discontinuity underlies the widely used 95% ANI threshold for species delineation in prokaryotes (**Figure 6**). Thus, bacterial genomes sharing more than 95% ANI are generally considered to belong to the same species.

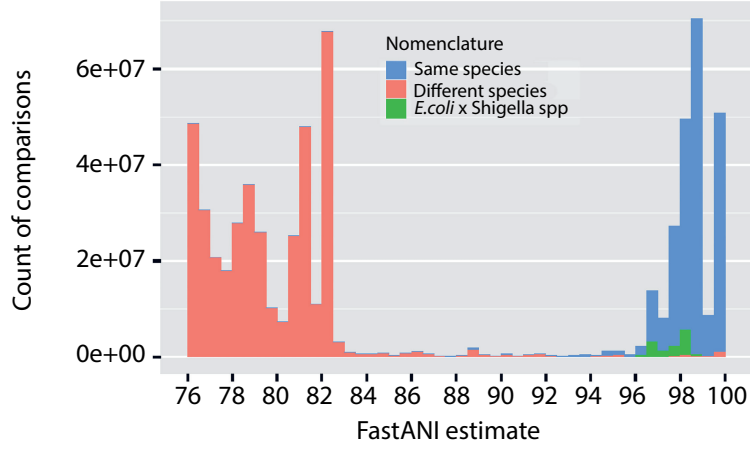


Figure 6. Distribution of average nucleotide identity (ANI) values for within- and between-species genome comparison. The figure shows the frequency distribution of pairwise ANI values across about 90,000 prokaryotic genomes. A clear discontinuity is observed around 95% ANI, separating within-species comparisons (high ANI values) from between-species comparisons (lower ANI values), which underlies the commonly used species boundary threshold (from Jain et.al.; 2018).

Although ANI provides a useful framework for species delineation, it captures nucleotide identity only across the orthologous fraction shared between genomes, rather than across their entire gene repertoire (Jain et al. 2018). This distinction is often described in terms of the core genome, which is shared by all members of a species, and the accessory genome, which varies between strains and may encode functions related to adaptation, metabolism or host interactions. Because homologous recombination, gene gain and loss, and horizontal gene transfer continuously reshape bacterial genomes, strains assigned to the same species on the basis of ANI may still differ extensively in accessory gene content (Diop et al. 2022; Rodriguez-R et al. 2024). As a result, genomes with highly similar core or shared-gene sequence identity can exhibit substantial differences in overall gene composition, highlighting that species assignment based on ANI does not imply functional uniformity (Rodriguez-R et al. 2024) (**Figure 7**).

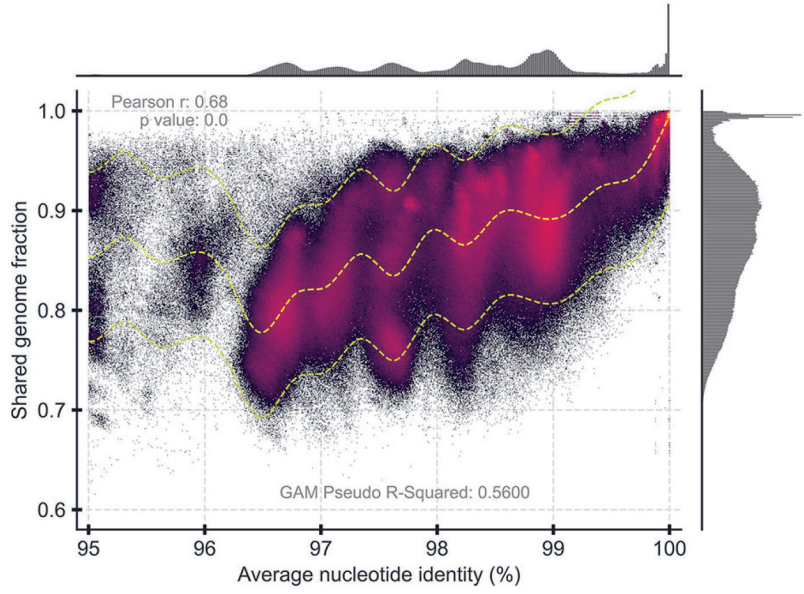


Figure 7. Relationship between averaged nucleotide identity (ANI) and shared genome fraction in prokaryotic genomes. The plot illustrates how genomes with similar ANI values can differ substantially in fraction of shared genes. Even at high ANI (>95%), a considerable variation in shared genome content is observed, reflecting differences in accessory genome composition (from Rodriguez et al.; 2024).

Consistent with these observations, a similar pattern is observed in the Estonian gut microbiome cohort. Among high-quality (HQ) MAGs assigned to the same species-level cluster, reconstructed genome length varies substantially. Across species represented by multiple HQ MAGs, the median normalized standard deviation of genome length is 0.11, reaching up to 0.43 in the most variable cases (**Figure 8**). This observation supports the need for finer-scale analysis beyond species-level classification. However, genome-length variation in MAG should be interpreted cautiously, because it may reflect both biological differences in accessory genome content and technical factors such as incomplete assembly, fragmentation, or binning uncertainty.

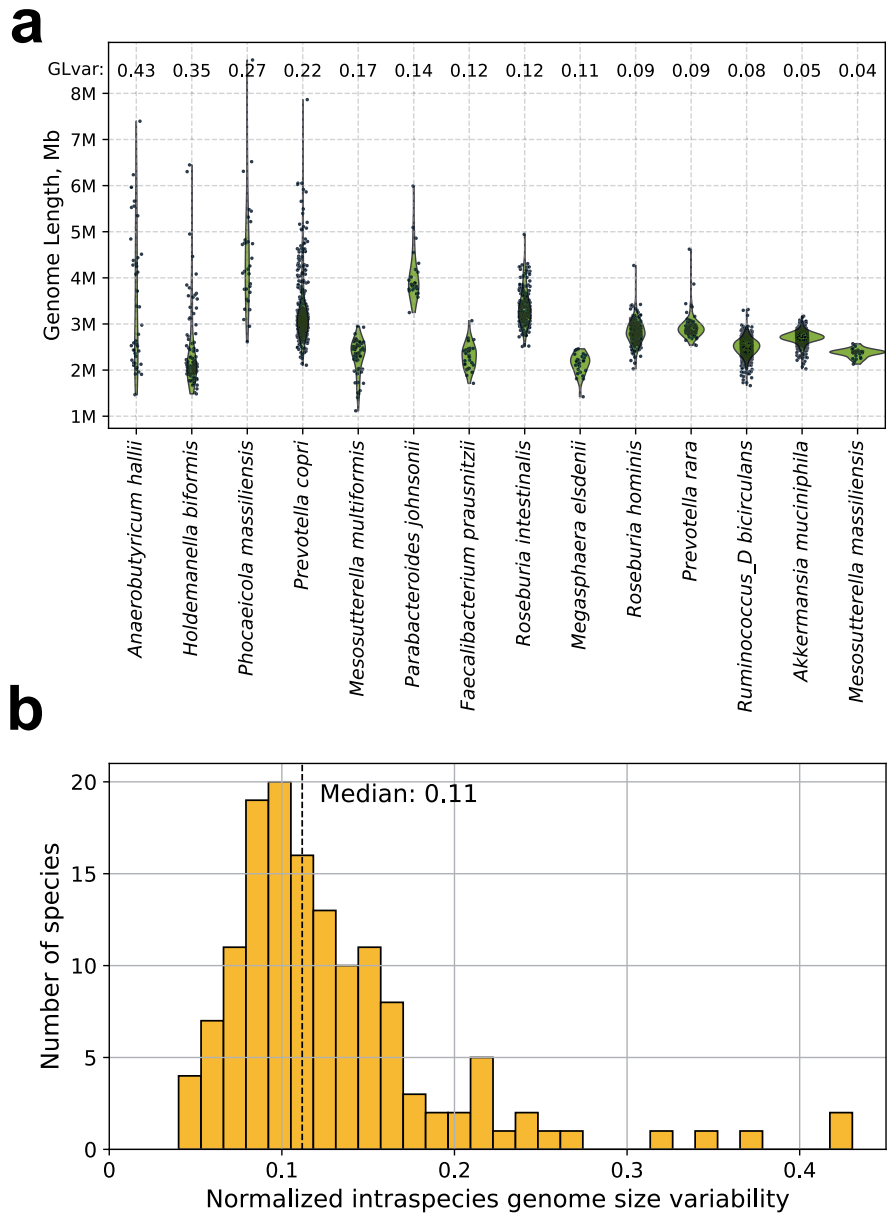


Figure 8. Within-species variation in reconstructed genome length among gut bacterial species in the Estonian microbiome cohort (N=2504 individuals) **a.** Genome length distributions for the most variable and least variable species, illustrating the extent of within-species genomic diversity. These results demonstrate that substantial genome length variation exists within species-level clusters, supporting the need for finer-scale resolution beyond species-level classification. Each dot represents one genome, and the numbers above the plots indicate genome size variability for each species. **b.** Distribution of normalized genome-length variation across species-level clusters represented by more than 100 high-quality (HQ) MAGs. *(not published)*

1.6 Future perspectives

Observations about between-species and within-species differences, we discussed in previous chapter, indicate that bacterial species often function as umbrella categories that encompass multiple genetically and ecologically distinct populations, rather than uniformly cohesive biological entities (Conrad et al. 2024; Viver et al. 2024). For studies aimed at linking the microbiome to host phenotype, species-level of resolution may therefore be insufficient, as substantial within-species genomic variation can translate into differences in microbial function and distinct associations with host traits (Zeevi et al. 2019; Zhernakova et al. 2024).

In this context, it is necessary to move beyond species-level classification and consider finer-scale units. Although these units are often referred to as **strains**, the term strain is traditionally reserved for descendants of a single cultured isolate and is best applied to nearly identical genomes (Van Rossum et al. 2020; Viver et al. 2024). For association analyses, a more suitable framework is to define stable within-species populations with coherent genomic content and, ideally, coherent functional potential, rather than relying solely on species-level classifications.

1.6.1 Within-species population structure

As discussed in the previous chapter, the genome-based concept of prokaryotic species has only recently become established (Jain et al. 2018; Parks et al. 2022). However, defining species boundaries is only the first step. The rapid accumulation of thousands of genomes per species has revealed that the diversity contained within a single named species is often much greater than previously appreciated. Bacterial species, once commonly treated as largely clonal populations (Smith et al. 1993), are now increasingly understood as umbrella categories encompassing multiple genetically and ecologically distinct populations.

This realization opens an important direction for future research: moving beyond species-level comparisons toward a detailed understanding of **within-species structure**. Under current operational definitions, genomes assigned to the same species may differ by up to 5% in shared genome sequence similarity, while differences in accessory gene content can be substantially larger (Conrad et al. 2024; Rodriguez-R et al. 2024). Such variation may reflect differences in ecological niches, host adaptation, recombination dynamics, mobile genetic elements, metabolic capabilities, virulence, or antimicrobial resistance.

A key challenge is to determine which within-species differences represent stable biological populations and which reflect technical artefacts introduced by sequencing depth, fragmentation, binning errors, or database incompleteness. This distinction is especially important when within-species units are used in downstream association analysis.

1.6.2 Technical limitations of metagenome sequencing

To achieve the detailed understanding of within-species structure described above, it will be critical to obtain high-quality genomes representing all major genomic units within each species. In addition, sufficiently large numbers of genomes are required to capture the complete repertoire of accessory genes and to enable statistically robust analyses.

A major technical limitation is that current short-read-based metagenomic assembly approaches remain insufficient for this task. Although metagenome-assembled genomes have greatly expanded the known diversity of uncultivated microorganisms (Nayfach et al. 2021; Parks et al. 2017), their quality is often limited by incomplete assembly, contamination, fragmented genomes, and the formation of composite MAGs when several closely related strains co-occur in the same sample (Chen et al. 2020; Shaiber and Eren 2019).

Genome quality can be improved by selecting only high-quality MAGs based on tools such as CheckM (Parks et al. 2015; The Genome Standards Consortium et al. 2017) and by applying additional quality-control approaches such as GUNC to detect chimerism or taxonomic inconsistency (Orakov et al. 2021). However, these improvements are unlikely to fully resolve the problem. Even high-quality short-read MAGs may fail to capture plasmids, phages, repetitive regions, strain-specific islands, and other accessory elements that are essential for understanding within-species diversity (Bertrand et al. 2019; Moss, Maghini, and Bhatt 2020; Suzuki et al. 2019). Therefore, additional methodological advances will be required.

One promising direction is the increased use of long-read sequencing, which can produce more contiguous assemblies and improve the recovery of complete chromosomes, plasmids, and mobile genetic elements (Bertrand et al. 2019, 2019; Moss, Maghini, and Bhatt 2020). Complementary methods such as Hi-C sequencing can further improve genome reconstruction by linking contigs to the same genome and enhancing binning accuracy (Bickhart et al. 2022; DeMaere and Darling 2019). Hybrid strategies that integrate short reads, long reads, and proximity ligation data are particularly promising for resolving complex microbial communities in which multiple closely related strains coexist (Bertrand et al. 2019; Bickhart et al. 2022).

However, the large-scale application of these approaches in population-based studies remains limited by cost, throughput, and computational demands. A complementary and highly reliable approach is to combine metagenomic analysis with systematic cultivation (Lewis et al. 2021). Where possible, representative isolates could be obtained for major genomic units within a species, followed by high-quality sequencing and complete genome assembly. Such isolate genomes would provide robust references for interpreting MAGs, defining accessory-gene repertoires, resolving within-species population structure, and validating species-specific genomic units. However, many microbial taxa remain difficult to culture, limiting the scalability of this approach. Therefore, cultivation-based strategies will need to be integrated with genome-resolved metagenomics and emerging technologies such as long-read and single-cell sequencing.

2. AIMS OF THE STUDY

This thesis investigates the potential of genome-resolved metagenomics as a framework for studying human microbiome communities and their associations with human health. It first evaluates the comparability of metagenomic profiles generated on two short-read sequencing platforms, which is essential for integrating data across large population-based cohorts. The thesis then focus on the large-scale reconstruction of metagenome-assembled genomes (MAGs) from a population-based cohort, the development of a population- specific MAG reference database, and downstream association analysis. In addition, the thesis introduces a framework for detecting microbiome-health associations at the within-species level.

The specific aims of the thesis are as follows:

- To systematically compare matched human stool metagenomes generated on two short-read sequencing platforms and assess platform-related effects on taxonomic and functional profiling, in order to evaluate the feasibility of integrating data across large population-based cohorts.
- To reconstruct metagenome-assembled genomes (MAGs) from the Estonian Microbiome deep (EstMB-deep) cohort, capturing both bacterial and archaeal diversity, and to establish a population-specific MAG reference database.
- To perform species-level association analyses and develop a framework for within-species-level association studies in large population-based cohorts.

3. MATERIAL AND METHODS

The human gut microbiome samples analysed in this thesis, together with the experimental procedures and analytical workflows, are described in detail in the corresponding original publications and their supporting materials. Here, I provide a brief overview of the study design, datasets and main methodological approaches used in the three studies included in this thesis.

The DNA samples were obtained from unrelated volunteers from Estonian biobank who provided informed consent. Sample collection and data generation were conducted in accordance with the guidelines and approval of the relevant ethics committees.

All three studies included in this thesis were based on the Estonian Microbiome Cohort (EstMB), a population-based cohort of 2,409 adult volunteers from the Estonian Biobank. The cohort includes stool metagenomic sequencing data together with host metadata, and linkage to national electronic health records (EHRs) enables microbiome–health association analyses. In this thesis, EstMB was used in two complementary forms: a broader moderately sequenced dataset for microbiome profiling and association analyses, and a deeply sequenced subset, referred to as EstMB-deep, for metagenome assembly and MAG reconstruction.

An overview of the datasets, sample sizes, methodological approaches, and main outputs of the studies included in this thesis in Table 2.

Table 2. Overview of the datasets and main outputs of the studies included in this thesis

Study	Dataset	Number of samples	Main output
Ref I	matched human stool metagenomes	1,351 samples	Assessment of sequencig platform concordance and platform related effects
Ref II	EstMB-deep and full EstMB cohort	1,878 deeply sequenced samples for MAG reconstrution; 2,504 moderately sequenced samples for downstream profiling and association analysis	EstMB-repMAG population-specific reference database, microbiome-health associations
Ref III	Archaeal MAGs from EstMB-deep	1,878 deeply sequenced samples	Human gut archaeal MAG collection and incorporation into the population-specific reference database

In the first study, paired metagenomic sequencing data generated on two independent sequencing platforms, MGISEQ-2000 and NovaSeq 6000, were compared. The dataset included 1,729 individuals from the Estonian Microbiome Cohort (EstMB), each with a sample sequenced on both platforms. A subset of 53 individuals was additionally sequenced twice on the same platform, allowing the assessment of within-platform technical variability. This study design enabled a systematic comparison of cross-platform and within-platform variability in taxonomic and functional microbiome profiles (**Ref. I**).

In the second study, we used 1,878 deeply sequenced samples from the Estonian Microbiome Cohort, hereafter referred to as EstMB-deep, were used for metagenome reconstruction. Metagenome-assembled genomes (MAGs) reconstructed from these samples were used to generate a cohort-specific reference database for subsequent microbiome profiling. Because these profiles served as the basis for downstream association analyses, maximizing the number of profiled samples was an important consideration. Therefore, the cohort-specific reference database was applied to moderately sequenced samples from the broader EstMB cohort, comprising 2,504 samples, thereby increasing the sample size available for microbiome-health association analysis (**Ref. II**).

In the third study, metagenomic data from the EstMB-deep cohort were re-analysed to specifically reconstruct archaeal metagenome-assembled genomes. For this purpose, the MAG catalogue generated in the second study was revised, with particular attention to genomes that had been excluded during the dereplication step. This strategy was used because excluded genome bins could contain archaeal MAGs that were not retained in the final bacterial MAG database. Targeted curation enabled the recovery of archaeal MAGs, including genomes representing potentially new archeal species. These genomes were subsequently incorporated into the population-specific reference database (**Ref. III**).

4. RESULTS AND DISCUSSION

4.1 Microbiome community profiling (Ref. I)

Microbiome community profiling defines the taxonomic composition of a given microbiome and in sequencing-based studies, typically estimates the relative abundance of the microorganisms present. This information provides the foundation for most downstream microbiome analyses.

The standard output of community profiling is an **abundance table**. In this table, rows usually represent taxa, such as species, genera or strains, and columns represent samples (**Figure 9**). Each value indicates the relative abundance of a given taxon in each sample. Absolute abundance is less commonly available because it requires additional information on the total microbial load in each sample (Morton et al. 2019). This can be estimated using approaches such as flow cytometry (Vandeputte et al. 2017). quantitative PCR (Galazzo et al. 2020) or spike-in standards (Barlow, Bogatyrev, and Ismagilov 2020), but these methods require extra laboratory procedures and increase cost.

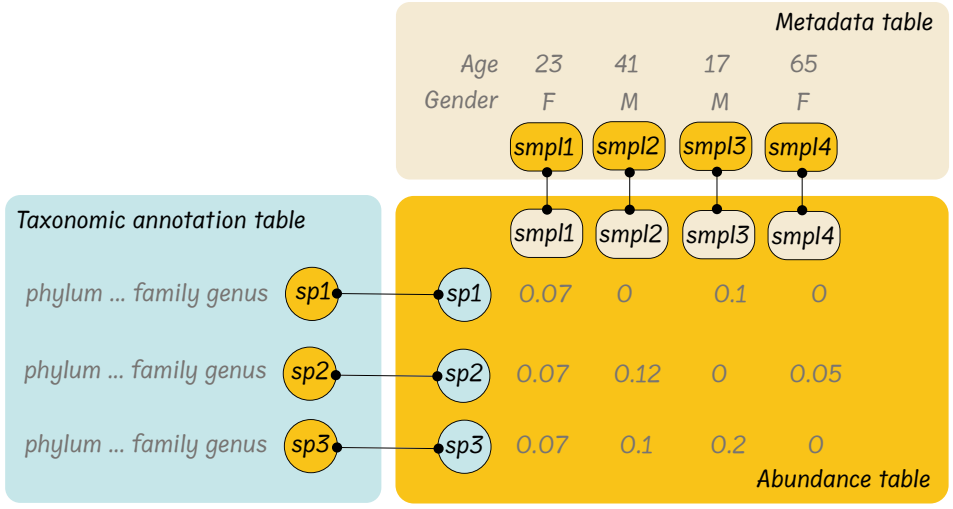


Figure 9. Core input tables for microbiome analysis and their interconnections. The abundance table is accompanied by a taxonomic annotation table, linked through species identifiers, and a metadata table, linked through sample identifiers. Smpl, samples, sp-species. Authors' own work.

Abundance tables are typically accompanied by a **taxonomic annotation table**. This table provides the full taxonomic classification of each feature detected in the abundance table, from higher taxonomic ranks to species or strain level when available. The abundance and taxonomy tables are linked through shared taxon or feature identifiers.

Community profiling data are often further complemented by sample **metadata table**. Metadata describe relevant characteristics of each sample or individual, such as age, sex, body mass index, stool type, disease status, treatment group or other clinical and environmental variables. These metadata are linked to the abundance table through sample identifiers.

Abundance table generation

The abundance tables, accompanied with taxonomic annotation table, serve as a central intermediate for a wide range of downstream analyses, including diversity estimation, microbiome-disease association studies, and functional inference. Despite its critical role, the influence of upstream methodological choices on the resulting abundance estimates remains underappreciated, and analytical attention is frequently directed toward subsequent steps rather than the nuances of the profiling process itself. Differences between DNA extraction protocols, sequencing platforms, profiling tools and profiling databases may introduce systematic biases that affect the reproducibility and comparability of microbiome studies (**Figure 10**).

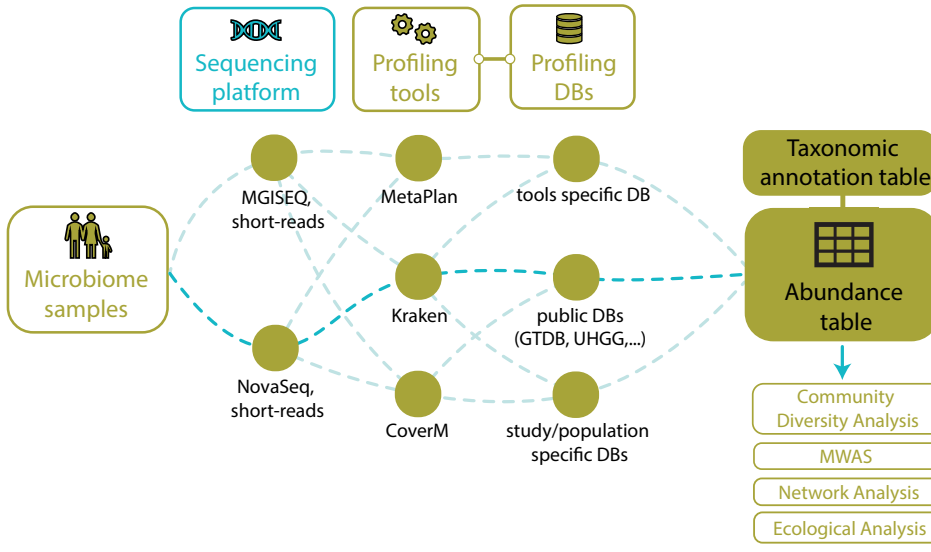


Figure 10. Schematic overview of abundance table generation as a central intermediate step in microbiome analysis workflows. The choice of sequencing platform, taxonomic profiling tool, and profiling database can influence the resulting abundance table and, consequently, all downstream analyses, interpretations, and biological conclusions. Sequencing platforms are highlighted in blue, as that factor is discussed in this chapter. Authors' own work.

In our first study, we aimed to address this gap and performed a large-scale comparative metagenomic analysis of short-read sequencing platforms, including taxonomic concordance and functional analysis.

4.1.1 Taxonomic concordance among sequencing platforms

To evaluate whether abundance tables produced from different sequencing platforms are comparable and suitable for combined downstream analyses, we analysed matched human stool metagenomes generated from the same DNA extracts. Because all samples were profiled using the same bioinformatic pipeline and reference databases, any observed differences in taxonomic abundance profiles could be attributed primarily to the sequencing platform rather than to differences in taxonomic annotation or computational processing.

The analysis included matched human stool metagenome samples from the Estonian Microbiome cohort. In total,1,351 sample paris passed the filtering criteria used for this comparison. Each pair originated from a different individual and was sequenced twice: once on the NovaSeq 6000 platform, hereafter NovaSeq, and once on the MGISEQ-2000 platform from MGI Tech Co., Ltd., hereafter MGI.

Before evaluating **cross-platform** variability, we first estimated the expected background variability introduced by sequencing and taxonomic profiling alone. For this purpose, we analysed an additional set of 53 technical replicate sample pairs generated on the same sequencing platform. This **intra-platform** comparison provided a baseline for the variability observed when the same sample is sequenced and processed under comparable platform conditions. Thus, the central question was not simply whether taxonomic profiles differed between NovaSeq and MGI, but whether the variability between platforms exceeded the background variability observed within a single platform.

The percentage of run-specific species was 3.42% in the intra-platform comparison, corresponding to 37 out of 1,083 detected species. In the cross-platform comparison, this percentage was slightly higher, at 5.89%, corresponding to 174 out of 2,953 detected species. However, the total number of detected species was nearly three times higher in the cross-platform dataset, likely reflecting the substantially larger number of sample pairs included in this comparison: 1,351 cross-platform pairs compared with 53 intra-platform technical replicate pairs. Therefore, the higher number of set-specific species in the cross-platform comparison should be interpreted with caution and may partly reflect the larger sample size rather than a sequencing platform effect alone.

Most platform-specific taxa were rare. Of the 174 species detected only in one platform, 154 were present in only a single sample. Their mean relative abundance was also markedly lower than that of shared species, at 1.30×10^{-6} % compared with 2.85×10^{-4} %, respectively. These results indicate that most platform-specific detections were unlikely to represent major differences in community composition, but rather reflected rare, low-abundance taxa close to the detection limit.

The difference in the number of detected species per sample pair was statistically significant in both the intra-platform comparison (P value = 0.038) and the cross-platform comparison (P value = 1.03×10^{-7}). However, the effect sizes were small in both cases, with Cohen's d = 0.29 for the intra-platform

comparison and Cohen's d = −0.12 for the cross-platform comparison. The negative sign of Cohen's reflects the direction of the platform differences rather than a larger effect. Therefore, despite statistical significance, the magnitude of the differences was limited.

We then performed sample-matched comparisons to assess concordance at the individual sample level. These analyses included the percentage of shared species within sample pairs, the correlation of species prevalence and relative abundance between sequencing runs, and the difference in Shannon diversity index between matched sample pairs, assessed using a paired t-test. The mean percentage of shared species within sample pairs was 96.44% ± 5.96% in the intra-platform comparison and was slightly lower in the cross-platform comparison, at 92.07% ± 5.20% (**Figure 11a, b**).

Species prevalence and relative abundance showed strong agreement between sequencing runs, with R^2 values approaching 1.0 in both the intra-platform and cross-platform comparisons (**Figure 11c–f**). Similarly, no significant difference in Shannon diversity was observed between paired profiles in either comparison (paired t-test, P value > 0.05).

Overall, these results show that taxonomic abundance tables generated from different sequencing platforms exhibit some set-level differences, including differences in the total number of detected species and the presence of platform-specific taxa. However, these differences were only modestly higher than the variability observed within a single platform, and most platform-specific taxa were rare and low in abundance. Cross-platform profiles remained highly concordant, with a high proportion of shared species and strong agreement in species prevalence, relative abundance, and diversity estimates. Together, these findings indicate that taxonomic abundance tables generated from NovaSeq and MGI sequencing data are broadly comparable and can be considered suitable for combined downstream taxonomic analyses when processed using the same taxonomic profiling pipeline and reference database. Nevertheless, comparisons involving rare or very low-abundance taxa should be interpreted with caution, as these taxa are most sensitive to platform-specific detection differences.

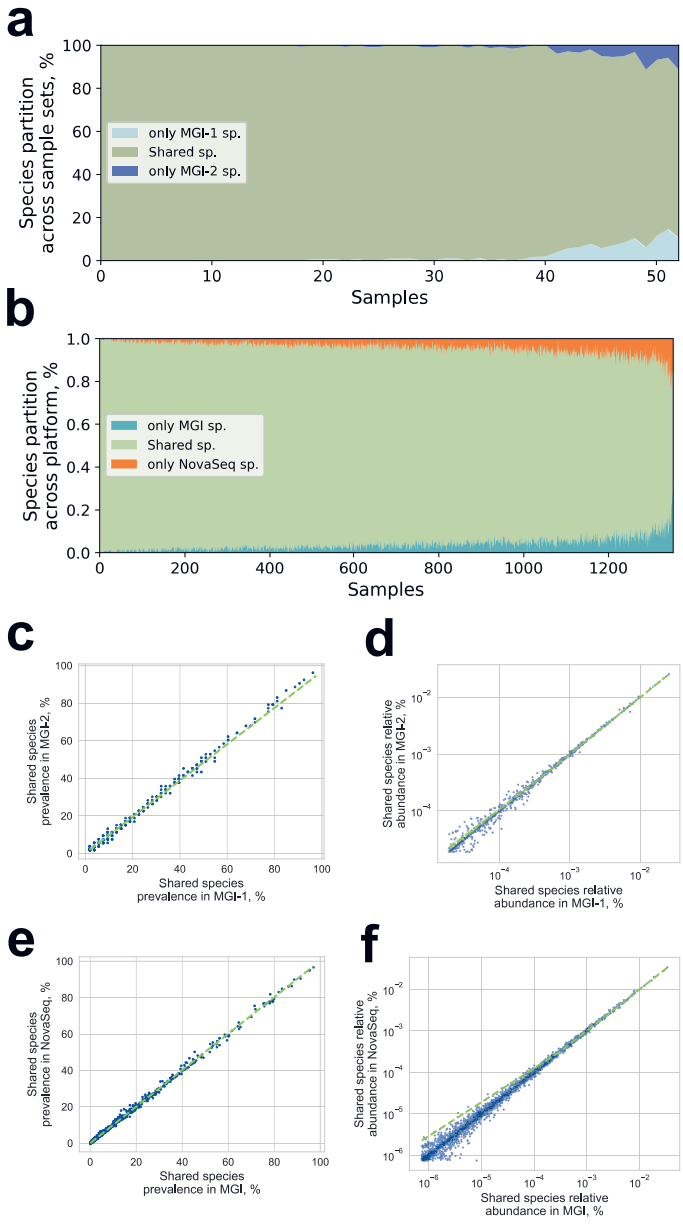


Figure 11. Taxonomic composition comparison within sequencing platforms and between sequencing platforms. **a.** The percentage of shared and unique run-specific species within each sample pair for the intra-platform baseline comparison. **b.** The percentage of shared and unique run-specific species within each sample pair for the cross-platform comparison. **c.** Comparison of the prevalence of shared species between matched samples, intra-platform baseline comparison. **d.** Comparison of the relative abundance of shared species in matched samples, intra-platform baseline comparison. **e.** Comparison of the prevalence of shared species between matched samples, cross-platform comparison. **f.** Comparison of the relative abundance of shared species in matched samples, cross-platform comparison.

4.1.2 Functional analysis challenges

Metagenomic sequencing enables both taxonomic and functional profiling of microbial communities, allowing researchers to characterize not only which microorganisms are present in a sample, but also their potential functional capabilities based on gene and pathway composition.

To assess platform-specific effects on functional profiling, we compared a random subset of 700 matched sample pairs sequenced with MGI and NovaSeq. Functional annotation identified 1,170 unique enzymes, of which only 31 were platform-specific, indicating broad agreement between technologies at the level of enzyme detection.

Nevertheless, MGI consistently recovered a higher number of enzymes and greater pathway recall, particularly in sample pairs with lower functional overlap. These differences were not restricted to rare functions, but also affected common enzymes involved in central metabolic pathways, suggesting a systematic shift in functional representation. K-mer analysis further showed that MGI samples contained substantially greater sequence diversity (**Figure 12a**), while the number of taxonomically assigned reads remained comparable between platforms. In contrast, MGI produced markedly more functionally annotated reads, indicating that higher sequence diversity had a stronger effect on functional annotation than on taxonomic profiling (**Figure 12b**). Functional differences were not explained by taxonomic variation, as taxonomic and functional similarity were not correlated, and coverage analysis revealed platform-specific differences across genomic regions.

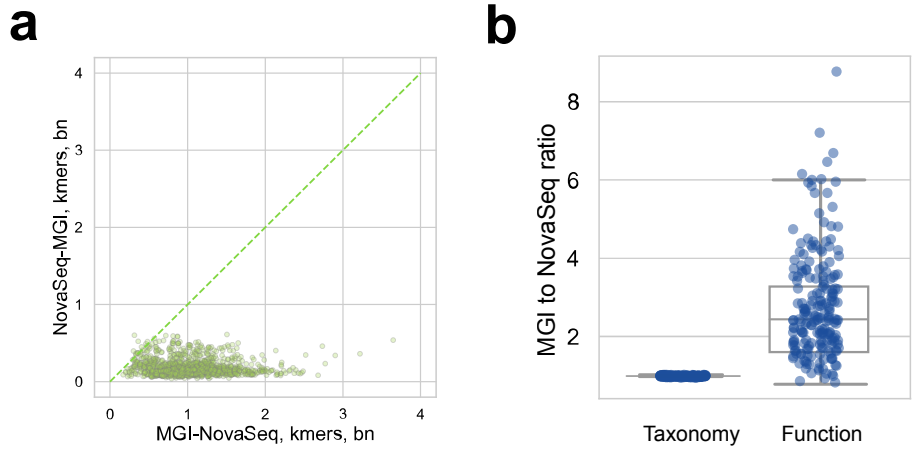


Figure 12. Platform-specific sequence and annotation differences. **a.** Unique k-mers per sample: in NovaSeq versus MGI and vice versa. **b.** Contributions of taxonomic and functional microbiome components found uniquely in MGI or NovaSeq.

Together, these results indicate that MGI data contained more sequence diversity, which was associated with improved functional annotation more than with changes on taxonomic annotation. More broadly, these findings show that functional profiling is more sensitive than taxonomic profiling to sequencing -platform-related differences, including read depth, sequence complexity, and uneven genome coverage. However, these results should not be interpreted as evidence that one sequencing platform is inherently superior for functional profiling. Rather, they support the broader view that a substantial part of the microbiome's functional "dark matter" reflects limitations in library complexity and functional reference catalogues, not simply differences between sequencing platforms (Treichel et al. 2026). Therefore, improving functional metagenomics requires not only platform comparison, but also optimized library preparation, adequate unique coverage, and improved functional gene catalogues. Consequently, although taxonomic profiles generated on MGI and NovaSeq were highly concordant, functional profiles should be interpreted with greater caution when data from different sequencing platforms are combined.

4.2 Genome-resolved metagenomics (GRM) (Ref. II, Ref. III)

Genome-resolved metagenomics has transformed our ability to characterize microbial communities by revealing multiple layers of diversity that remain hidden to classical profiling approaches.

First, it enables the reconstruction of previously **uncharacterized species** directly from metagenomic data, thereby recovering new taxa that would otherwise remain entirely undetected. Second, when applied across large cohorts, genome reconstruction yields multiple genomes per species, enabling investigation of **within-species variation** and its links to functional potential and ecological roles. Third, even within a single sample, microbial populations often consist of closely related co-existing strains. Such **within-sample within-species** variation is typically unresolved by both profiling methods and standard short-read GRM approaches and therefore requires more advanced strategies.

Here, we describe the assembly strategy applied to our cohort and assess its performance in capturing these three layers of previously hidden microbial diversity.

4.2.1 Metagenome reconstruction

MAG reconstruction across the EstMB-deep cohort was based on per-sample assembly with MEGAHIT, followed by consensus binning with MetaBAT v2.15, MaxBin v2.2.7, and VAMB v3.0.7, with bin selection performed by DAS Tool v1.1.4. This strategy was chosen to enable efficient processing of numerous high-coverage samples while maximizing MAG recovery robustness across diverse microbial genomes. Overall, **84,257 MAGs** were recovered, of which

42,224 were classified as near-complete MAGs (completeness >90% and contamination <5%, as assessed by CheckM2).

Species-level clustering with dRep further resolved these genomes into **2,257 species clusters**, each represented by the highest-quality MAG together with additional MAGs assigned to the same species that passed the standard inclusion thresholds applied for dereplication, completeness >50% and contamination <25% (**Figure 13**).

Because no **archaeal** genomes were retained among the final species representative MAGs, we performed an additional screening of MAGs excluded during the dereplication step to assess archaeal representation within the broader MAG collection. Taxonomic annotation of the 27,284 excluded MAGs using GTDB-Tk (release 214) identified 316 candidate archaeal MAGs.

These genomes were subsequently evaluated using CheckM for genome quality assessment, SeqKit for genome size estimation, and contig-level annotation with GUNC to verify taxonomic consistency. MAGs with low quality (completeness <50% or contamination >10%), abnormal genome size (>2 Mb), or inconsistent contig-level annotation (<50% of contigs assigned to the same archaeal taxon) were excluded, yielding a final set of 273 archaeal MAGs. Species-level clustering of these archaeal genomes further resolved them into 21 species clusters.

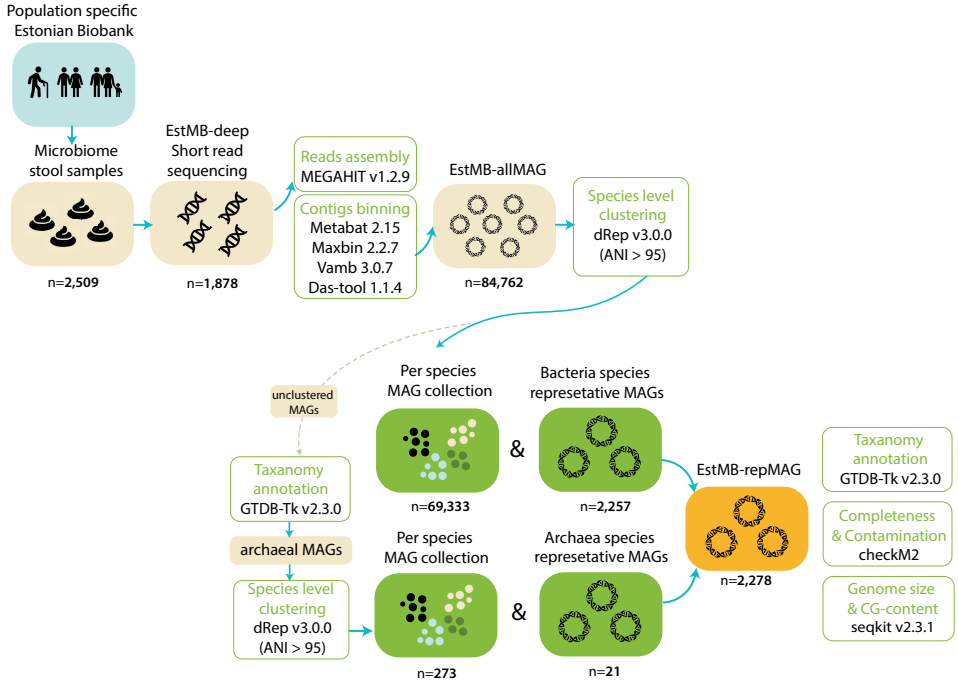


Figure 13. Overview of the MAG reconstruction and species-level clustering workflow applied to the EstMB-deep cohort.

In total, the final representative genome collection from the Estonian population comprised 2,257 bacterial and 21 archaeal species-representative MAGs. This final representative genome catalogue is hereafter referred to as **EstMB-repMAG**.

To ensure consistency across EstMB-repMAG, all MAGs were uniformly annotated using GTDB-Tk, genome statistics were estimated with SeqKit, and genome quality was assessed using CheckM2 (**Figure 13**).

4.2.2 Hidden diversity of a new species

One of the key advantages of genome-resolved metagenomics is the ability to recover previously uncharacterized microbial species directly from sequencing data, thereby expanding microbial diversity beyond the limits of existing reference databases. Unlike reference-based approaches, de novo genome reconstruction can recover both known and previously uncharacterized taxa, making it particularly valuable for identifying population-specific microbial diversity that remains absent from current catalogues.

New species discovery

To evaluate the extent of previously uncharacterized microbial diversity recovered from the EstMB cohort, reconstructed MAG species clusters were compared against the GTDB reference database. We quantified the proportion of new species among all reconstructed species clusters. Among bacteria, 353 species were classified as new, corresponding to 15.6% of all recovered bacterial species clusters. Among archaea, 10 of 21 species clusters (47.6%) were classified as new at the time of assembly (**Figure 14a,b**).

Given that the latest release of the UHGG database contains approximately 4,600 bacterial species and 27 archaeal species, the number of newly identified taxa indicates that substantial unexplored diversity remains even within the human gut microbiome, particularly among archaea, which remain underrepresented in current reference catalogues (**Ref. III**).

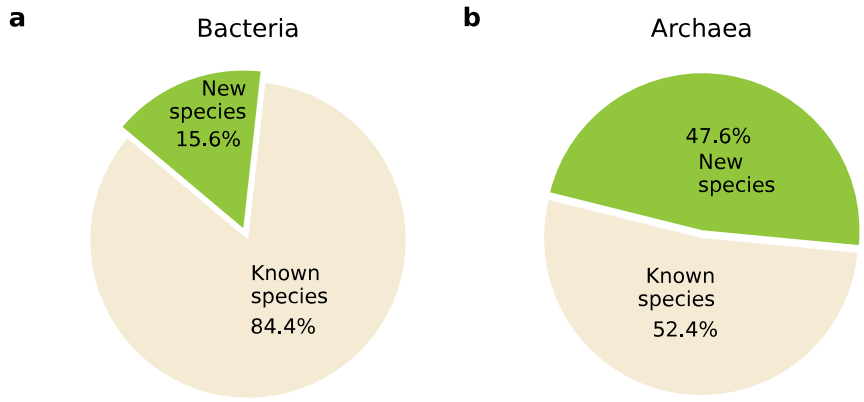


Figure 14. Overview of new species identified from the EstMB MAG collection. **a.** Proportion of new bacterial species among all reconstructed species clusters. **b.** Proportion of new archaeal species among all reconstructed species clusters.

Archaea diversity in Estonian population

Given the complexity of metagenome assembly, we anticipated that not all species present in the population would have corresponding MAGs. To gain a more comprehensive understanding of archaeal diversity in our population, we supplemented our species-level representative MAG collection "Archaea ESTrep-21" (n=21) with publicly available archaeal MAGs and genomes from the UHGG collections v2.0.2 (n=28) (Almeida et al. 2019). We performed species-level dereplication for 21 species from the 'Archaea ESTrep-21' collection and 28 species from the UHGG collection. Following this process, we compiled a refined set of 37 unique species, which we refer to as **'Archaea GUTrep-37'** (**Figure 15a**). Of these, 12 species were shared between the "ESTrep-21" MAG collection and the UHGG collection, while 16 species were exclusive to the UHGG collection, and 9 species were unique to the "ESTrep-21" MAG collection (**Figure 15b**).

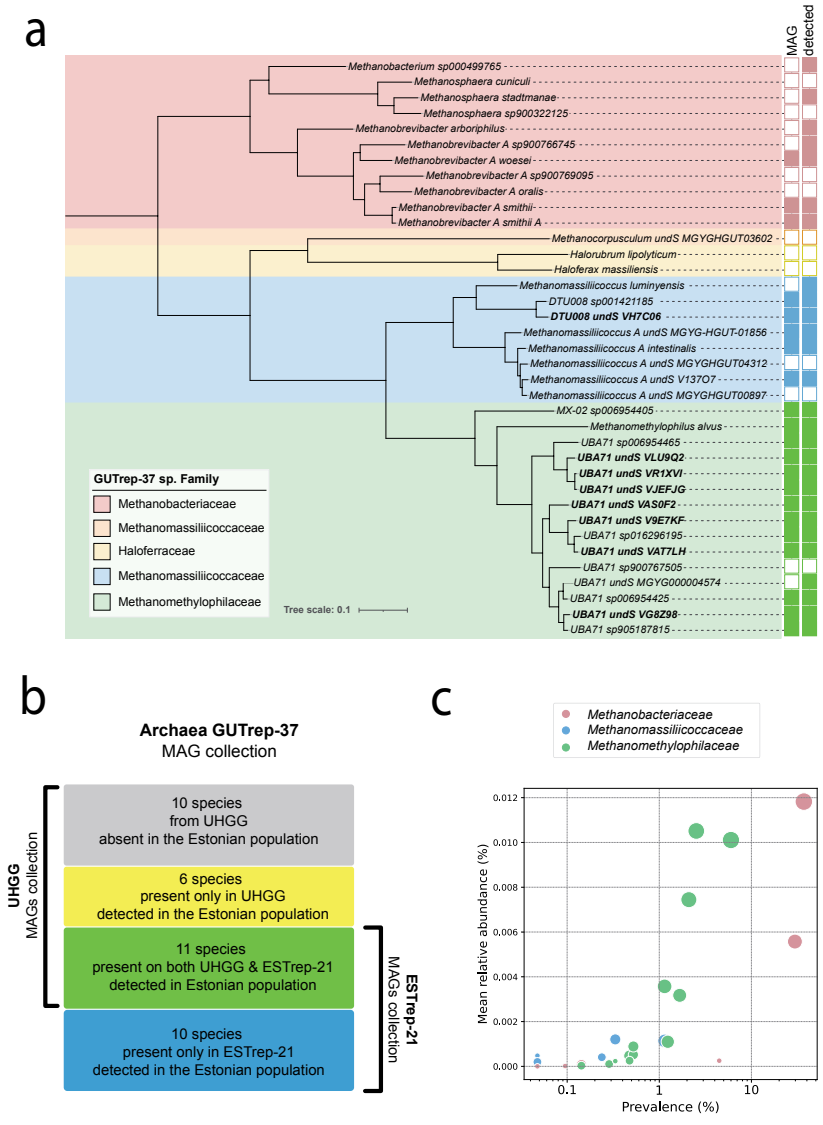


Figure 15. The phylogenetic tree, prevalence and abundance of the human gut archaea in Estonian population. **a.** FastTree approximate maximum-likelihood phylogenetic tree of human gut archaea, including species assembled in this study and those from the UHGG ("Archaea GUTrep-37" collection). The tree was constructed using multiple sequence alignments (MSAs) derived from the concatenation of 53 phylogenetically informative markers (arc53) with GTDB-Tk v2.3.0. Phylogenetic tree was constructed with FastTree and generated with iTOL (https://itol.embl.de), with midpoint rooting. Species are labelled with their respective names, while new species are named by their MAG_ID or UHGG_ID. Adjacent to each species label, two indicator boxes provide additional information: "MAG" – the presence or absence of an assembled MAG for the species in the Estonian population (filled and empty boxes, respectively), and "detected" – the detection of the species through mapping in the Estonian population (filled and empty boxes, respectively). Species are color-coded by their family affiliation: *Methanobacteriaceae*

(pink), *Methanomassiliicoccaceae* (blue), and *Methanomethylophilaceae* (green), *Methanocorpusculaceae* (orange), *Haloferacaceae* (yellow). **b.** Classification of species from "Archaea GUTrep-37" collection based on their origin and their presence in the Estonian population. **c.** Prevalence and abundance of archaeal species in the Estonian population. Species are color-coded by their family affiliation: *Methanobacteriaceae* (pink), *Methanomassiliicoccaceae* (blue), and *Methanomethylophilaceae* (green).

New species discovery rate

To assess whether the observed diversity approached saturation, we plotted the cumulative number of newly detected species against the number of samples included in genome reconstruction (**Figure 16**). Each additional set of 500 samples yielded approximately 70 new species, and no plateau was observed, suggesting that further genome reconstruction will likely continue to uncover previously undescribed diversity. Although this rate will inevitably be influenced by the ongoing expansion of global reference databases, the current trajectory indicates that genome reconstruction remains a productive approach for characterizing previously unrepresented microbial diversity.

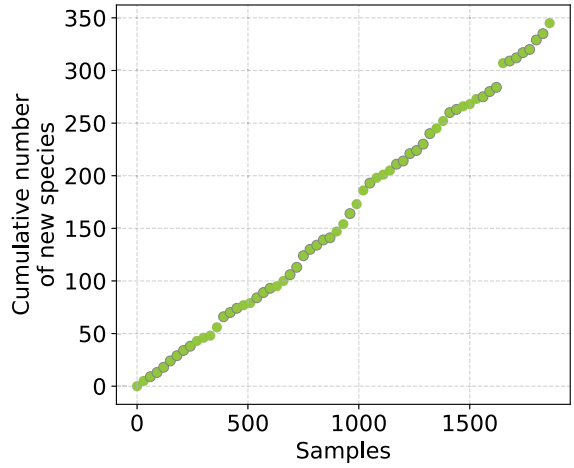


Figure 16. Relationship between the number of analysed samples and the cumulative number of new species identified.

4.2.3 Assembly-detection gaps

Most newly identified MAGs were represented by only one or two assemblies. For many of these species, prevalence estimated by read mapping was similarly low. However, some species exhibited substantially higher prevalence despite being reconstructed in only a few samples. We refer to this discrepancy between the number of samples in which a species was detected by read mapping and the number of samples in which a corresponding MAG was successfully reconstructed as the assembly–detection gap.

To quantify this effect across genera, an assembly-detection gap was calculated as:

$$\text{assembly–detection gap} = (N_{\text{detected}} - N_{\text{assembled}}) / N_{\text{detected}}$$

where N_{detected} represents the number of species detected by profiling and $N_{\text{assembled}}$ the number successfully reconstructed as MAGs. Values close to 1 indicate species that are frequently detected but poorly assembled.

However, some species retained large assembly-detection gaps despite frequent detection in metagenomic reads and moderate relative abundance (>1.5%). These findings suggest that limited assembly representation does not necessarily reflect rarity but may instead arise from biological or technical factors that hinder genome reconstruction, including strain-level complexity, uneven abundance distributions or high genomic similarity among closely related species.

Importantly, some under-assembled yet frequently detected species reached appreciable abundance levels, indicating that they may represent biologically relevant members of the microbiome despite limited representation in MAG collections. Their prevalence and abundance indicate potential functional relevance, underscoring the need to include such taxa in downstream microbiome association analyses despite their limited representation in MAG collections.

Species exhibiting particularly large assembly–detection gaps belonged to genera such as *Phocaeicola, Alistipes, Bacteroides*, CAG-603, CAG-269, *Stercorousia*, RGIG8607, *Mailhella, Nanosynbacter, Butyricimonas* and *Prevotella* (**Figure 17**).

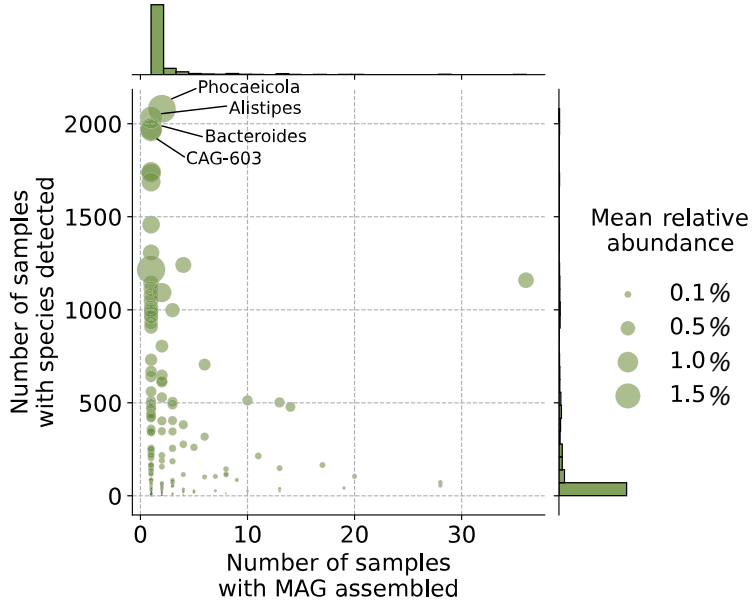


Figure 17. Relationship between the number of MAGs per species, prevalence estimated by read mapping, and mean relative abundance of newly identified species. Each point represents a species, with point size proportional to its mean relative abundance. Several species represented by only one or two MAGs exhibit high prevalence across the EstMB cohort while maintaining substantial mean abundance. (unpublished data)

To place these findings in context, the assembly–detection gap for previously assembled, known species can also vary widely. In contrast to many new species, some known species were represented by relatively large number of MAGs. Mean relative abundance among known species with small assembly–detection gaps was often similar to that of species with substantially larger gaps, suggesting that relative abundance alone does not explain differences in assembly success. Instead, successful genome reconstruction likely depends on taxon-specific genomic or ecological characteristics that influence assembly quality (**Figure 18**).

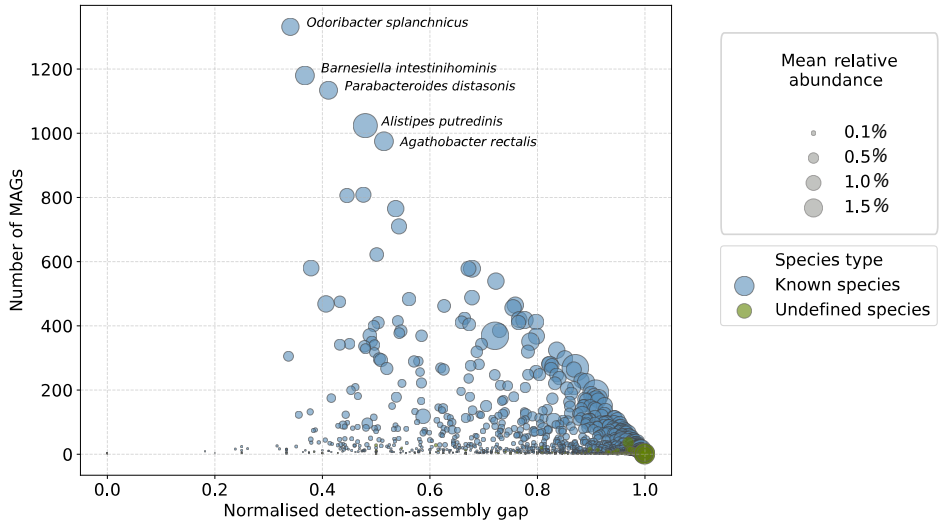


Figure 18. Relationship between the assembly-detection gap and the number of MAGs per species. Each point represents a species, with point size proportional to mean relative abundance. Blue points indicate known species, while green points indicate newly identified species. (unpublished data)

Species with small assembly–detection gaps and large number of reconstructed MAGs included *Odoribacter splanchnicus, Barnesiella intestinihominis, Parabacteroides distasonis, Alistipes putredinis, Agathobacter rectalis*, UBA11524 sp000437595, Ruminococcus_E bromii_B, Ruminococcus_D bicirculans, *Paraprevotella clara, Akkermansia muciniphila,* Limisoma sp000437795, Eubacterium_F sp003491505, *Parabacteroides merdae*, Lachnospira eligens_A, Ruminiclostridium_E siraeum and others.

4.2.4 Population-specific reference

Both newly reconstructed and previously described taxa from the Estonian population cohort were integrated into a population-specific MAG reference collection, EstMB-repMAG. This collection comprises 84,762 MAGs, representing 2,257 bacterial and 21 archaeal species.

Compared to global resources, our population-based EstMB-repMAG collection is approximately twofold smaller than the Unified Human Gastrointestinal Genome (UHGG) catalogue and fourfold smaller than the latest MetaPhlAn database. Because genome assembly does not recover all species present in a community, we hypothesized that a fraction of taxa in our samples remained undetected by assembly-based approaches.

To assess whether incorporating globally derived references improves microbiome profiling, species richness per sample was quantified using three reference databases:

(i) EstMB-repMAG collection (n = 2,278) profiled with CoverM.

(ii) Combined dereplicated UHGG and EstMB-repMAG collections (n = 4,792 species) profiled with CoverM.

(iii) MetaPhlAn4.1 database (8,609 species) profiled with MetaPhlAn.

Interestingly, the smallest, population-specific database yielded the highest number of detected species per sample (**Figure 19**). One possible explanation is that increasing database complexity elevates the frequency of multi-mapping reads, which are subsequently discarded during profiling due to ambiguous assignment, thereby reducing the number of confidently detected taxa.

These findings suggest that an effective reference database should balance comprehensive taxonomic representation with relevance to the target population. Excessively broad reference collections may introduce additional mapping ambiguity, which could reduce the number of confidently detected taxa.

Overall, these results challenge the prevailing strategy of continuously expanding global reference databases for universal application. Instead, population-specific reference collections may provide more accurate and sensitive taxonomic profiling for specific cohorts.

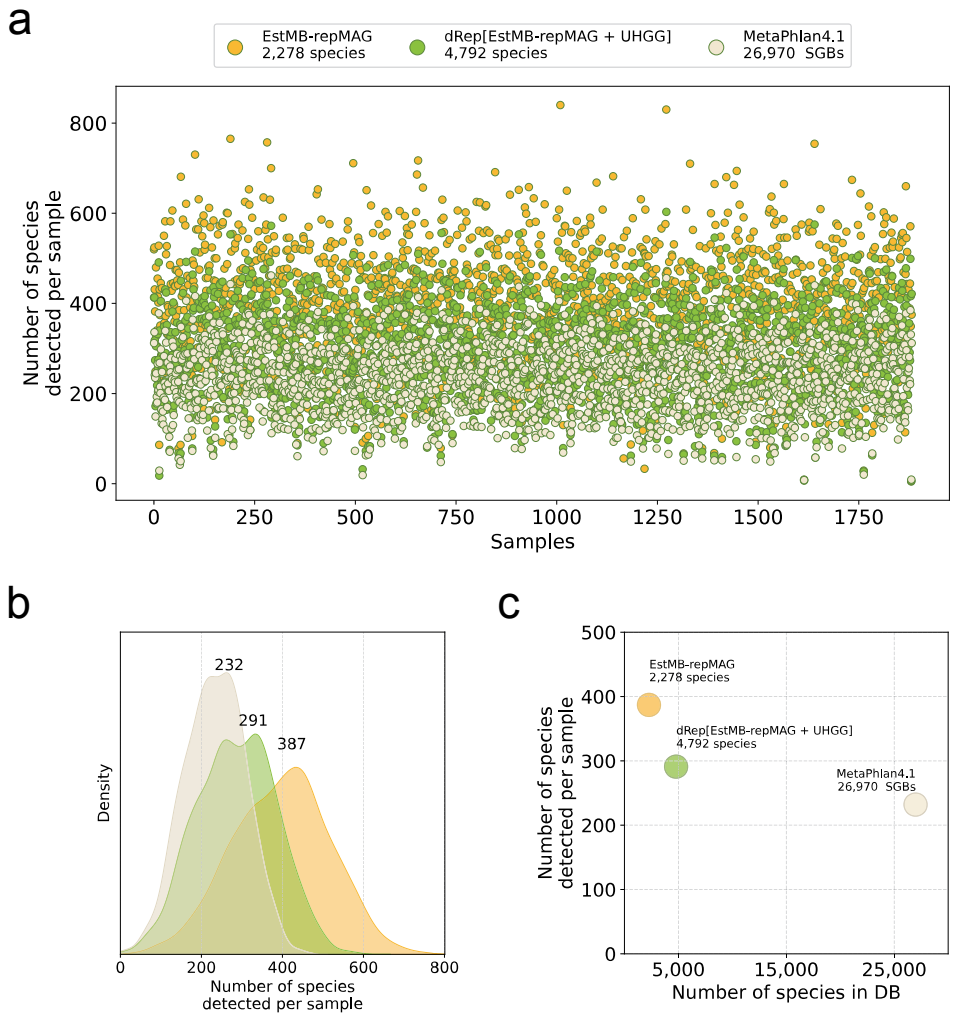


Figure 19. Comparison of species counts or species-level genome bins (SGBs) per sample obtained using different reference databases. **a.** Comparison of the number of species detected per sample with different profiling databases. **b.** Distribution of species counts per sample. Yellow indicates EstMB-repMAG, green indicates the dereplicated UHGG + EstMB-repMAG collection, and beige indicates MetaPhlAn4.1. Numbers above each density curve indicate the mean value of the corresponding distribution. **c.** Relationship between average number of species or SGBs detected per sample and the number of species or SGBs included in the profiling database (unpublished data)

4.2.5 Hidden within-species diversity

Beyond enabling the recovery of previously uncharacterized taxa, large-scale genome reconstruction also generates multiple genomes per species, thereby allowing the investigation of within-species variation. However, this representation is highly uneven across species.

In our dataset of 1,878 deeply sequenced samples, the number of meta-genome-assembled genomes (MAGs) per species varies widely. Most species (1,548; 68.6%) are represented by fewer than 100 MAGs, and a substantial fraction (1,037; 46.0%) by fewer than 10 MAGs. Such limited representation constrains the robust characterization of population structure and reduces statistical power for within-species analyses. As a result, within-species analysis are largely restricted to a relatively small subset of prevalent species for which sufficient numbers of MAGs were reconstructed and the assembly–detection gap was low.

The most extensively represented species, *Odoribacter splanchnicus*, was supported by 1,312 MAGs, followed by a sharp decline to approximately 400 MAGs per species and a more gradual decrease thereafter (**Figure 20**). Four species, *Odoribacter splanchnicus, Barnesiella intestinihominis, Parabacteroides distasonis*, and *Alistipes putredinis*, were each represented by more than 1,000 MAGs. These taxa were not only highly prevalent but also exhibited relatively low assembly–detection gaps. An additional six species were represented by more than 500 MAGs: *Agathobacter rectalis*, UBA11524 sp000437595 (a member of the class *Clostridia*), *Paraprevotella clara, Akkermansia muciniphila*, Limisoma sp000437795, and *Parabacteroides merdae*.

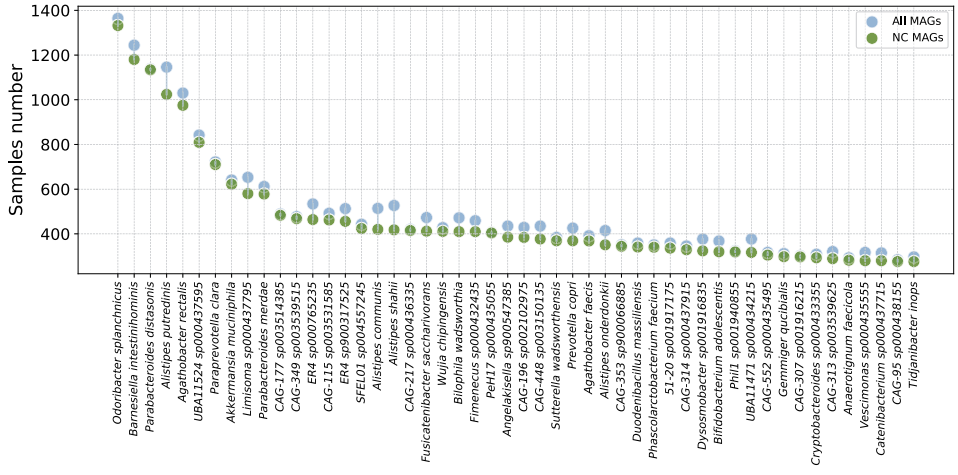


Figure 20. Number of MAGs and near-complete MAGs per species among the 50 species with the highest number of assembled MAGs. Blue dots represent the number of MAGs per species, and green dots represent the number of near-complete MAGs (NC MAGs) per species, defined as MAGs with completeness > 90%, contamination <5% according to CheckM2. (unpublished data)

After establishing that only a subset of species is sufficiently represented by reconstructed MAGs, the next step was to define what should be considered a meaningful within-species unit. To investigate within-species diversity, it is first necessary to define the principle by which genomes are grouped within a species. This is not a trivial issue, as the term “strain”, although widely used in the field, is applied inconsistently and often diverges from the definition established by the International Code of Nomenclature of Prokaryotes (Oren et al. 2023). Under the ICNP framework, a strain is defined as the direct descendant of a single bacterial cell and is therefore expected to have an almost identical genome.

The purpose of exploring within-species diversity is to identify genomic variation associated with specific phenotypic or ecological features. This requires grouping genomes into subpopulations of sufficient size to enable robust association analyses. Therefore, a concept is needed that captures **within-species genomic structure at an intermediate resolution**, broader than near-identical lineages, but finer than species-level classification. For this reason, the formal ICPN strain concept is not suitable for our analytical purpose.

As no existing term adequately describes clusters of genomes defined by a specified level of genomic similarity in the context of population-scale MAG association analysis, we introduce the term **genome unit** to denote such within-species groupings.

> ***Genome unit (GU)*** *– an operational within-species group consisting of genomes assigned to the same species that cluster together based on pairwise genomic similarity above a defined threshold.*

A key remaining question is the choice of an appropriate genome similarity threshold for defining GUs in human gut microbiome species. To address this, we computed a total of 6,942,981 pairwise average nucleotide identity (ANI) values across 376 species represented by more than 100 MAGs (**Figure 21**). The resulting distribution did not exhibit a clear discontinuity at high ANI values above 97% that would allow straightforward delineation of subspecies clusters, in contrast to the well-defined separation around ~95% that is typically used to define species boundaries. However, local minima in the distribution were observed at 95.9%, 96.8%, 98.1%, 99.0%, and 99.7%. The lower values are close to the conventional species boundary and are therefore too permissive, whereas 99.7% approximates a clonal-level definition and would likely yield groups that are too small for association analyses. Based on this, we selected an intermediate threshold of 99.0% ANI to define genome units. The threshold should be interpreted as a pragmatic analytical choice rather than a universal biological boundary.

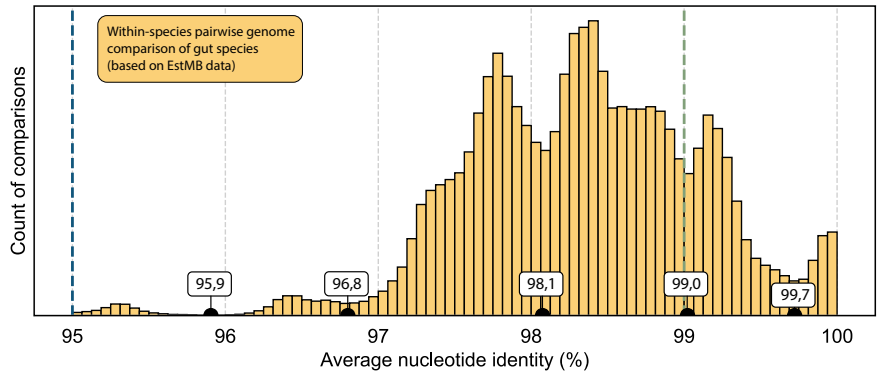


Figure 21. Histogram of within-species pairwise genome comparisons for 376 gut species, each represented by more than 100 MAGs reconstructed in the EstMB dataset. Local minima in the ANI distribution are indicated and were used to guide the selection of pragmatic treshold for genome unit definition (unpublished data)

We define within-species population structure as the set of genome units identified for each species, where each GU consists of genomes that cluster together at a pairwise similarity threshold of ANI ≥ 99.0%.

Within this framework, species diversity can be assessed by the number of GUs per species, referred to here as genome unit number (GUN). To account for differences in the numbers of reconstructed MAGs across species, we introduce both the raw number of genome units per species (GUN) and a normalized metric that scales GUN by the total number of MAGs, referred to as **normalised genome unit number** (nGUN).

GUN = Number of genome units
nGUN = (Number of genome units / Number of MAGs) × 100%

No consistent trend in normalized genome unit number (nGUN) was observed across gut microbial species. Instead, nGUN values varied widely, ranging from 0.4 to 94.0 (**Figure 22**), indicating substantial heterogeneity in within-species population structure.

At the lower end, *Odoribacter splanchnicus* exhibited an nGUN of 0.4, corresponding to approximately one genome unit per 250 MAGs, consistent with a highly cohesive population structure under the selected ANI threshold. In contrast, *Prevotella copri* showed one of the highest nGUN values, 94.0, reflecting its well-documented genetic heterogeneity, where nearly each MAG corresponds to a distinct genome unit. Notably, members of the Alistipes_A genus were observed at both extremes of the distribution, highlighting substantial variability even among closely related taxa.

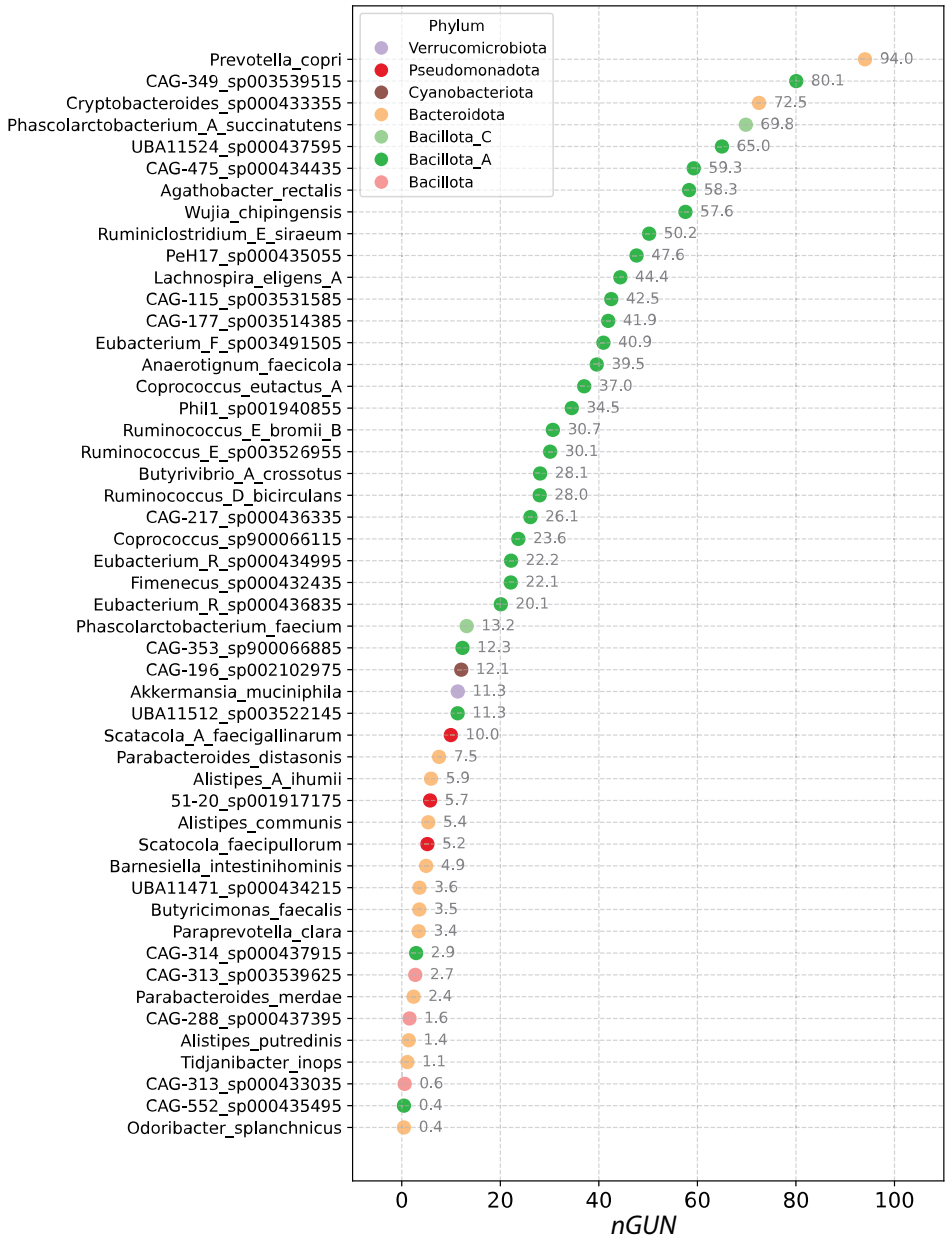


Figure 22. Normalized genome unit number (nGUN) values for the top 50 species with the highest number of metagenome-assembled genomes. nGUN was calculated as the number of genome units divided by the number of MAGs per species and multiplied by 100.

Genome units (GUs) are defined based on genomic similarity, a straightforward metric that can be efficiently computed across large-scale MAG datasets. However, in the context of our primary objective, which is to identify genome groups that differ in their associations with specific phenotypic or ecological features, it is essential to determine whether these GUs also exhibit functional differentiation.

Species with high nGUN values are, in practice, poorly suited for within-species association analyses. Their genomic diversity is distributed across many small units, often lacking sufficiently large subpopulations for meaningful statistical comparisons, regardless of the total number of reconstructed genomes. Such species are better described as exhibiting a continuum of genetic variation rather than discrete, well-defined groups. Conversely, low nGUN values do not necessarily imply low biological diversity, but rather indicate that, under the selected threshold, diversity is concentrated into fewer and larger genome units that are more suitable for association testing.

As mentioned above, within-species analyses are already constrained to a relatively small subset of prevalent species with a low assembly–detection gap, ensuring sufficient numbers of reconstructed MAGs. Our results indicate that an additional constraint must also be considered: low nGUN values are required to ensure the presence of genome units large enough for robust downstream comparisons.

Among the species with the highest numbers of reconstructed MAGs, only a subset meets this criterion. *Odoribacter splanchnicus* exhibited the lowest nGUN (0.4), followed by *Alistipes putredinis* (nGUN = 1.4) and *Parabacteroides merdae* (nGUN = 2.4), making them strong candidates for downstream within-species association analyses. In contrast, other species with similarly large numbers of reconstructed MAGs display high nGUN values and were therefore less suitable. Examples include *Agathobacter rectalis* (nGUN = 58.3), *UBA11524 sp000437595* (nGUN = 65.0), and *Limisoma sp000437795* (nGUN = 74.7).

To address this, we selected *Odoribacter splanchnicus* as a model species, as it represents the most suitable candidate for within-species association analyses based on our criteria: a high number of reconstructed MAGs, a low assembly–detection gap, and a low nGUN value. Within this species, we performed functional annotation of the two largest genome units GU-N1 and GU-N2. Although four GUs were identified in total, the remaining two contained too few genomes to support meaningful analysis and were therefore excluded (**Figure 23a**).

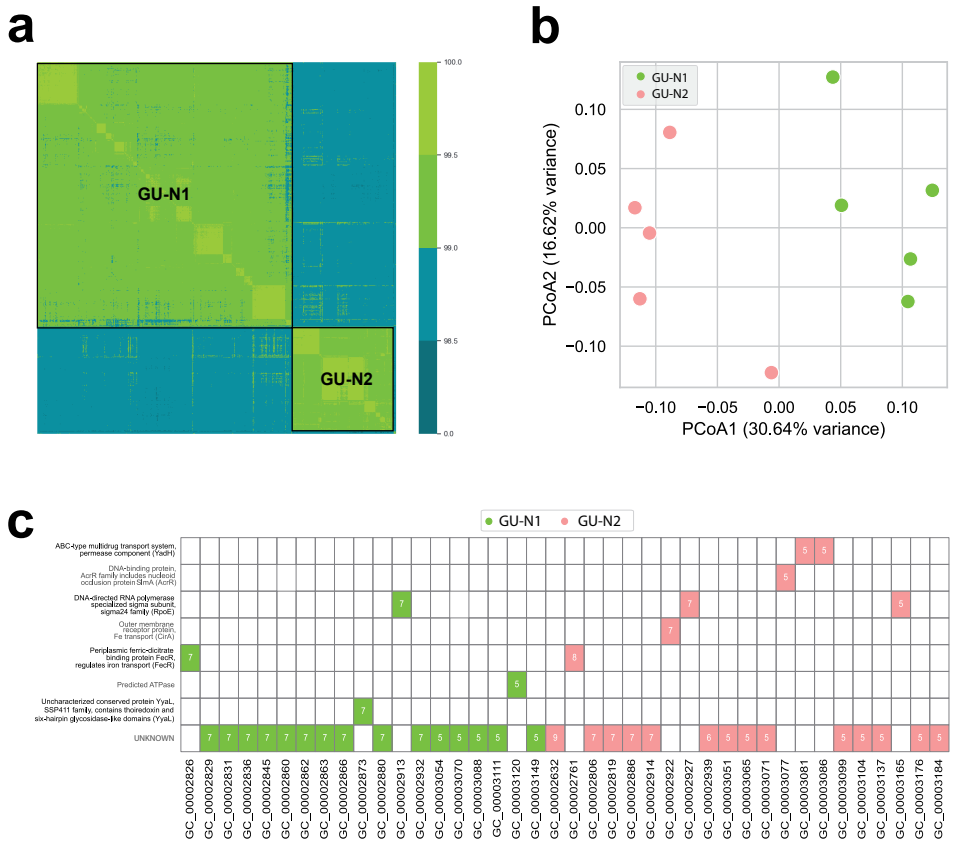


Figure 23. Genomic and functional variation across two major *O. splanchnicus* genome units. **a.** Heatmap of average nucleotide identity (ANI) values among *O. splanchnicus* MAGs, revealing two distinct genome units. **b.** Principal coordinates analysis (PCoA) of *O. splanchnicus* representative MAGs based on predicted gene cluster presence/absence profiles. **c.** Gene clusters uniquely present in only one of the two major *O. splanchnicus* genome units.

To explore functional differences, we performed a principal coordinates analysis (PCoA) of GU-N1 and GU-N2. As shown in **Figure 23b**, the two genome units were separated in the ordination space, indicating pronounced genomic differentiation at the level of gene content.

In total, 40 gene clusters were identified as unique to one of the two genome units (**Figure 23c**). While the majority of these clusters encoded hypothetical or uncharacterized proteins, several were assigned putative functions based on the Clusters of Orthologous Groups annotation framework. GU-N2 exhibited a broader repertoire of genes associated with stress response, iron acquisition, and antimicrobial resistance, features consistent with enhanced fitness under inflammatory conditions in the gastrointestinal tract. Notably, this included increased representation of the extracytoplasmic stress sigma factor RpoE (σ^E), iron uptake components such as FecR and CirA, and multidrug resistance determinants

including AcrR and an ABC-type efflux pump (YadH). In contrast, GU-N1 was enriched in proteins involved in redox homeostasis, such as YyaL/DsbD, suggesting a distinct adaptive strategy centered on oxidative stress mitigation.

Together, these results show that genome units defined by ANI-based clustering can also correspond to differences in gene content and predicted functional potential. This supports the use of GUs as biologically informative within-species units for downstream microbiome-host association analysis, while also emphasizing that their interpretation should remain species-specific and dependent on sufficient MAG representation.

4.2.6 Hidden within-species within-sample diversity

Apart from concerns related to completeness and contamination, a fundamental limitation of MAG-based analyses arises when multiple strains of the same species coexist within a single sample. Short-read assembly, which remains the most widely used approach in genome-resolved metagenomics, generally lacks the resolution to separate closely related strains. Instead of reconstructing distinct genomes, it may produce a consensus or **"composite" assembly** that predominantly reflects the most abundant strain genotype, while still incorporating signals from less abundant strains. As a result, these composite MAGs can obscure true within-sample strain diversity and may display intermediate genomic similarity patterns, depending on the relative abundance of the underlying strains and the genome regions incorporated into the assembly.

This limitation is particularly relevant for the interpretation of genome units. Composite MAGs may bias estimates of the genome unit number (GUN) per species, either by creating artificial intermediate genomes or by blurring the separation between otherwise distinct genome units. Consequently, the observed within-species population structure may partly reflect technical limitations of short-read assembly rather than true biological structure. Several approaches can mitigate this issue, including long-read sequencing–based assemblies, Hi-C–assisted metagenomics, single-cell genomics, or the integration of genome-resolved metagenomics with isolate sequencing. Nevertheless, short-read assemblies remain valuable for providing an initial, coarse-grained view of species structure, which can subsequently be refined using more advanced methodologies.

In some cases, the presence of multiple strains within a single sample may lead to fragmented assemblies that fall below MAG quality thresholds or **fail to be binned** into a coherent genome, resulting in the complete absence of a reconstructed genome for this particular species. If this pattern is consistent across samples, it can produce a pronounced assembly–detection gap. Such a gap may indicate species that tend to occur as diverse strain mixtures rather than as clonal populations, whereas species with minimal gaps are more likely to exist as relatively homogeneous populations or as populations dominated by a single strain per sample. However, the assembly-detection gap should not be interpreted as direct evidence of within-sample strain diversity on its own, as it may also

result from low abundance, uneven coverage, repetitive genomic regions, mobile genetic elements, high similarity to related species, or binning limitations.

Overall, within-sample within-species diversity represents an important hidden layer of microbial variation that is only partly captured by standard short-read MAG reconstruction. Recognizing this limitation is essential for interpreting both assembly-detection gaps and genome unit-based analysis, and highlighting the need for complementary methods to resolve strain mixtures within individual metagenomic samples.

4.3 Genome-resolved microbiome-wide association studies (MWAS) (Ref. II)

Our goal is to use genome-resolved metagenomes (GRM) to better understand how the microbiome relates to human health. A practical and accessible starting point toward this objective is association analysis, which enables the identification of links between microbiome composition and host phenotypes. While not definitive, this step provides a useful way to prioritize species that may be involved in underlying mechanisms, host-microbiome interactions or the development of predictive models.

GRM enables, first, the inclusion of newly reconstructed and previously uncharacterized species in MWAS and, second, within-species association analyses for a subset of species with high prevalence, low assembly–detection gaps, and low nGUN. Thus, GRM can extend microbiome-wide association studies beyond known reference species and beyond species-level resolution.

We leveraged the rich medical data available for the EstMB cohort and conducted MWAS across all prevalent diseases recorded in electronic health records (EHR) at both the species and within-species levels.

4.3.1 Species-level MWAS

For species-level MWAS, we selected diseases based on their prevalence in the cohort, including all conditions with more than 100 cases among the 2,504 individuals. This resulted in a set of 33 diseases, each defined using ICD-10 codes derived from EHRs linked to the Estonian Biobank cohort (**Ref. II, Supplementary Table S4**).

To limit the multiple testing burden, we restricted the analysis to species present in at least 1% of samples, yielding 1,595 species. In total, we identified 105 significant associations (Bonferroni-adjusted P value $< 2.71 \times 10^{-5}$) involving 96 bacterial species across 25 diseases. Notably, newly assembled species contributed to associations with 8 of the 33 diseases. For example, one of the strongest associations with chronic ischemic heart disease was observed for a previously uncharacterized species within the genus *Nanosynbacter* species (ID: H2144_Nanosynbacter_undS; adjusted P value = 3.13×10^{-6}) (**Figure 24**). This

illustrates that population-specific genome reconstruction can reveal disease-associated microbial signals that would be missed or poorly resolved using only existing reference databases.

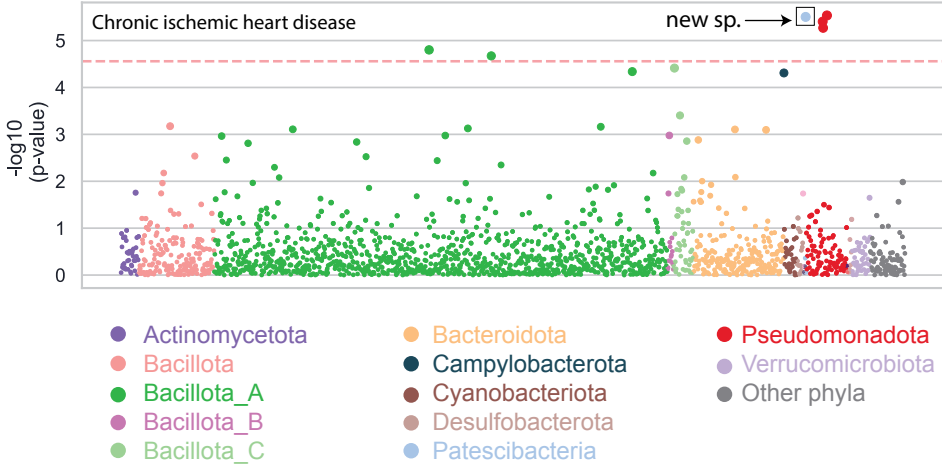


Figure 24. Metagenome-wide association results between EstMB-repMAG species abundance and chronic ischemic heart disease. Each data point corresponds to a single species, with vertical position reflecting the log-transformed P-value from the association model. Significant associations involving newly reconstructed species are highlighted with a box.

4.3.2 Within-species MWAS

For within-species MWAS, we selected *Odoribacter splanchnicus,* which stood out as the best candidate for within-species analysis based on our selection criteria (**Table 3**).

At the species level, *O. splanchnicus* did not show a significant association with any of the analysed diseases. To determine whether this species showed disease association at the within-species level, we performed logistic regression models testing the presence or absence of the two major *O. splanchnicus* units, GU-N1 and GU-N2, against the same 33 disease phenotypes previously analyzed in the species-level MWAS. Models were adjusted for BMI, age, and sex. This analysis identified a significant association between the presence of GU-N1 and two different diseases- gastritis and duodenitis, and hypertensive heart disease (**Figure 25**). The odds ratio for GU-N1 was below 1 for both diseases (gastritis and duodenitis OR = 0.56, hypertensive heart disease OR = 0.63). This indicates that the presence of GU-N1 was associated with lower odds of these diagnoses in the cohort. Importantly, this should be interpreted as an association rather than evidence of a causal protective effect.

Table 3. Characteristics of Odoribacter splanchnicus supporting its selection for within-species MWAS

Parameter	Value	Position among other species
Species prevalence	96%	Among the 3% most prevalent species
Number of MAGs	1332	Ranked first
Assembly-detection gap	23%	Among species with low assembly-detection gaps
GUN	4	Ranked first
nGUN	0.4	Lowest among analysed species

These results demonstrate that within-species MWAS can reveal associations that are not detectable at the species level. In the case of *O. splanchnicus,* the species as a whole showed no significant disease associations, whereas one of its genome units was significantly associated with two disease phenotypes. This finding supports the central premise that species-level microbiome profiling can mask biologically relevant within-species heterogeneity, and that genome unit-based analysis may improve the resolution of microbiome-health association studies.

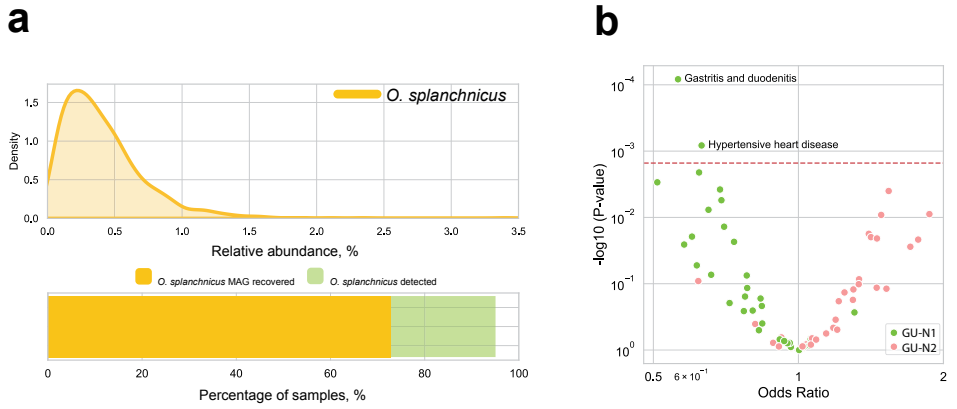


Figure 25. Within-species MWAS of *Odoribacter splanchnicus.* **a.** *O.splanchnicus* relative abundance, number of recovered MAGs and prevalence across samples. **b.** Volcano plot of associations between the two major *O. splanchnicus* genome units and 33 disease phenotypes. The red line indicates the Bonferroni-corrected significance threshold.

4.4 Next challenges to overcome

The study of within-species diversity is still in its early stages. A deeper understanding of how bacterial species are internally structured, together with the resolution of current technical limitations in metagenomic assembly, will be essential for advancing the field. Advancing this area will be essential for moving microbiome research beyond descriptive analysis toward a mechanistic understanding of microbial community function and host-microbiome interactions.

4.4.1 Resolving within-species population structure

A major challenge for future studies will be to determine how within-species diversity is structured. Even among species inhabiting relatively stable environments, such as the human gut, species differ markedly in their population structure. Some species harbour large, well-separated genomic clusters, whereas others exhibit more continuous variation, with little or no obvious clustering (**Figure 26**). Using the data from the Estonian Microbiome cohort (EstMB), we observed that, *Odoribacter splanchnicus* shows several large, clearly separated within-species clusters that can be distinguished using an ANI threshold of approximately 99.0% (**Figure 26a**). In contrast, *Agathobacter rectalis* displays even more pronounced cluster separation, but the relevant boundary occurs at a much lower ANI threshold, around 97.6%. Applying a standard ANI threshold of 99.0% to this species fails to recover the biologically meaningful larger clusters, instead fragmenting the species into many small genomic units (**Figure 26b**). A similar pattern is observed for *Akkermansia muciniphila,* where the major within-species clusters are separated at approximately 97.7% ANI (**Figure 26c**). For some other species, such as UBA11524 sp000437595, the situation is less clear, and it remains uncertain whether discrete within-species genomic units can be identified at all (**Figure 26d**).

These examples are preliminary and should be interpreted with caution. They nevertheless illustrate an important point: within-species structure varies substantially between bacterial species and cannot always be captured by a single universal ANI threshold. In the future, a more systematic, careful, and detailed analysis of within-species structure across a broad range of bacterial species will be needed.

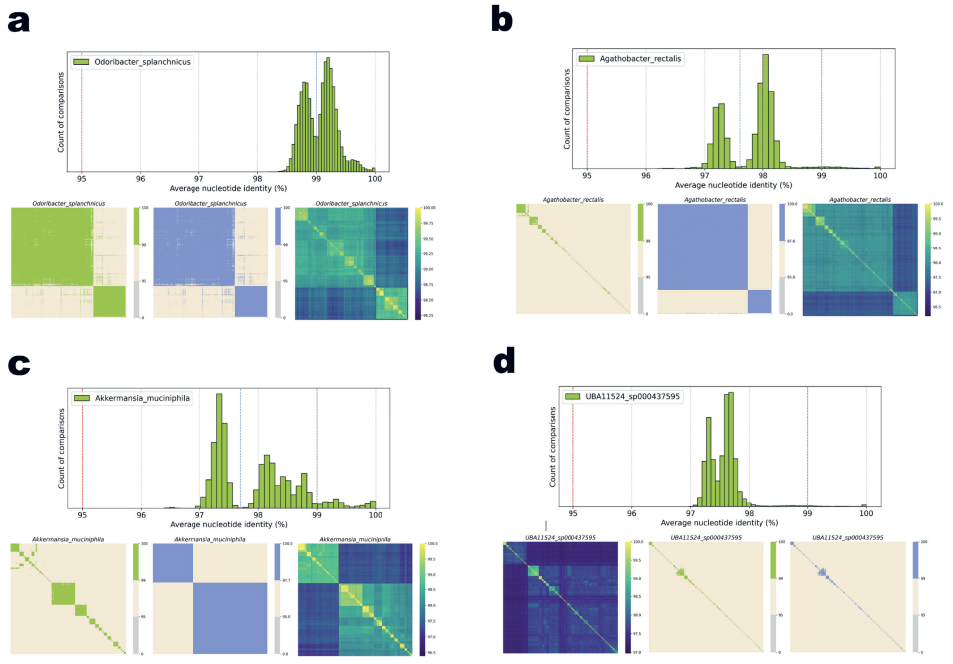


Figure 26. Examples of within-species population structure in selected gut bacterial species reconstructed from the Estonian Microbiome (EstMB) cohort MAG database. (unpublished data)

4.4.2 Reconsidering the species concept

These observations suggest that future work should move toward species-specific frameworks for defining within-species populations. Such frameworks should integrate multiple dimensions of variation, including nucleotide similarity, accessory genome composition, and functional potential. A key open question is whether genomic clusters identified in this way correspond to biologically meaningful units, such as populations with distinct ecological roles or host associations.

This question is particularly important for microbiome-wide association studies (MWAS), where biologically relevant signals may reside below the species level and remain undetected in species-based analyses. Resolving within-species structure therefore has the potential to substantially improve the resolution and interpretability of microbiome–host association studies.

Such frameworks would provide a more realistic view of bacterial species as structured and variable genomic populations, rather than as homogeneous taxonomic entities (**Figure 27**).

Developing robust frameworks for identifying these units represents a key challenge for future microbiome research.

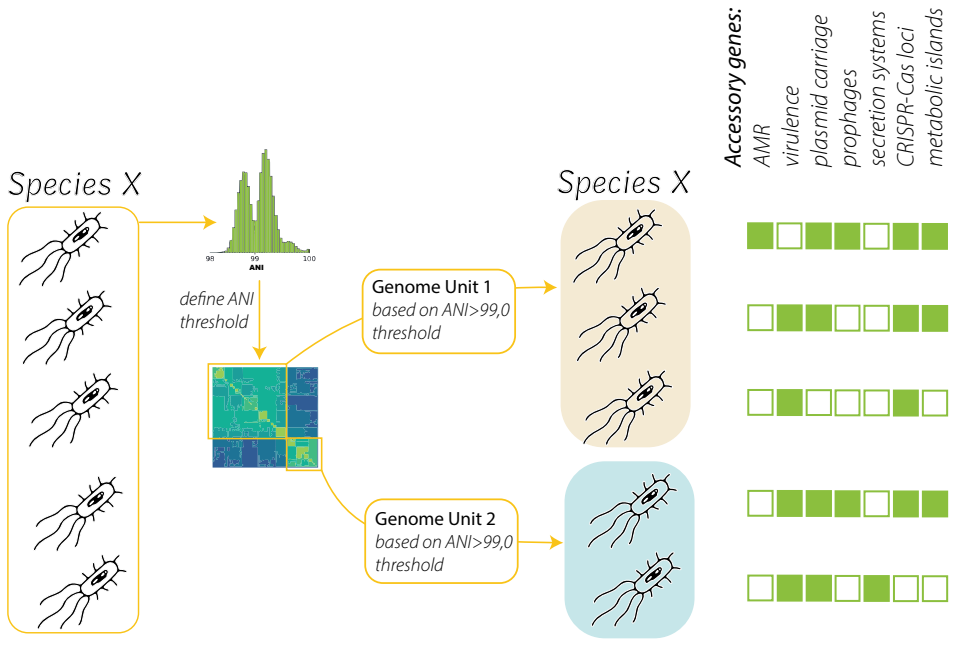


Figure 27. Schematic representation of bacterial species as a structured and genomically variable population. Authors' own work.

CONCLUSIONS

Therefore, the main conclusions drawn from this thesis are as follows:

- The two major short-read sequencing platforms used for metagenomic sequencing are broadly compatible for taxonomic profiling and can be applied to complementary samples from the same cohort. However, functional profiling was more sensitive to platform-related differences, indicating that taxonomic concordance does not necessarily imply equivalent functional profiles.
- Population-specific MAG reference databases can improve microbiome profiling by balancing taxonomic coverage with relevance to the target cohort. In this work, the EstMB-repMAG collection provided a cohort-specific reference resource that incorporated both newly reconstructed and previously described taxa from the Estonian population.
- Databases of reconstructed bacterial and archaeal MAGs represent a valuable resource for microbiome research. They provide an essential foundation not only for association studies but also for investigating fundamental biological questions, including the nature of within-species diversity. The recovery or archaeal MAGs further highlights that archaeal diversity in the human gut remains underrepresented in current reference collections.
- Genome-resolved metagenomics revealed multiple hidden layers of microbial diversity, including previously uncharacterized species, within-species genomic structure, and within-sample strain-level complexity that is only partly resolved by standard short-read MAG reconstruction. These findings emphasise both the value and the current limitations of short-read-based genome reconstruction.
- In this work, we identified significant disease associations at the species level, developed a framework for selecting species suitable for within-species analysis, and demonstrated that functionally distinct genomic units within a species can be associated with disease in ways that are not detectable at the species level alone. This supports the use of genome-unit-level analysis as a complementary strategy to species-level MWAS for improving the resolution of microbiome-health association studies.

SUMMARY IN ESTONIAN

Mikrobioomi genoomipõhine analüüsi: võimalused ja väljakutsed

Soolestiku mikroorganismid, sealhulgas bakterid ja arhed, on inimese tervisega tihedalt seotud. Nad võivad mõjutada peremeesorganismi ainevahetust, lagundada muidu seedimatuid toidukomponente, toota immuunsust mõjutavaid metaboliite ning aidata kaitsta organismi patogeenide eest. Samas võivad mõned mikroobsed protsessid kahjustada soole limabarjääri või viia selliste ühendite tekkeni, mis on seotud põletiku ja haigustega. Seetõttu on mikrobioomi koosseisu ja funktsiooni täpsem kirjeldamine oluline inimese tervise mõistmiseks.

Viimastel aastatel on genoomipõhine metagenoomika oluliselt muutnud mikrobioomiuuringuid. See lähenemine võimaldab rekonstrueerida mikroobide genoome otse keerukatest kooslustest ilma mikroobe kultiveerimata. Sellisel viisil saadud metagenoomi abil kokku pandud genoomid ehk MAG-id on avanud suure hulga varem kirjeldamata mikroobset mitmekesisust ning näidanud, et ka juba tuntud bakteriliikide sees võib esineda ulatuslik genoomne varieeruvus. Seetõttu pakub genoomipõhine metagenoomika võimalust uurida nii uusi liike kui ka liikide sisemist struktuuri ning hinnata, kas traditsiooniline liigitaseme käsitlus on piisav mikrobioomi ja tervise vaheliste seoste uurimiseks.

Selles väitekirjas kasutasin Eesti Mikrobioomi kohordi andmeid, mis hõlmavad 2504 Eesti Geenivaramu osaleja soolestiku metagenoomseid andmeid. Töö eesmärk oli hinnata genoomipõhise metagenoomika võimalusi ja piiranguid inimese soolestiku mikrobioomi kirjeldamisel ning mikrobioomi ja tervise vaheliste seoste leidmisel.

Töö esimeses osas võrdlesin kahe lühikeste lugemite sekveneerimisplatvormi, NovaSeq 6000 ja MGISEQ-2000, tulemusi. Taksonoomilised profiilid olid platvormide vahel üldiselt hästi võrreldavad, mis näitab, et eri platvormidel genereeritud andmeid saab sama töövoo ja referentsandmebaaside kasutamisel taksonoomiliseks profiilimiseks koos analüüsida. Funktsionaalne profiilimine oli aga platvormiga seotud erinevuste suhtes tundlikum. MGI andmetes tuvastati rohkem funktsionaalselt annoteeritud lugemeid, viidates sellele, et funktsionaalsete geenide ja radade tuvastamine sõltub rohkem sekveneerimissügavusest, järjestusraamatukogu keerukusest ja genoomi katvuse ühtlusest.

Töö teises osas rekonstrueerisin sügavalt sekveneeritud EstMB-deep andmestikust mikroobseid genoome ning lõin Eesti populatsioonile kohandatud MAG-referentsandmebaasi. See andmebaas, EstMB-repMAG, sisaldab nii varem tuntud kui ka Eesti kohordis rekonstrueeritud uusi mikroobseid liike. Populatsioonispetsiifiline referentsandmebaas parandas mikrobioomi profileerimist. Lisaks bakteritele võimaldas täiendav arhedele suunatud analüüs rekonstrueerida inimese soolestiku arhede MAG-e, mis näitab, et arhede mitmekesisus on praegustes referentskogudes endiselt alakaetud.

Genoomipõhine metagenoomika paljastas mitu varjatud mikroobse mitmekesisuse tasandit. Esiteks tuvastati varem kirjeldamata liike, mida olemasolevad

referentsandmebaasid ei esindanud. Teiseks võimaldas suur hulk sama liigi MAG-e uurida liikide sisemist genoomset varieeruvust. Tulemused näitasid, et bakteriliigid ei ole alati homogeensed üksused, vaid võivad koosneda mitmest selgelt eristatavast genoomsest alarühmast. Kolmandaks ilmnes, et ühe proovi sees võib sama liigi piires esineda mitu lähedast tüve või genoomset varianti, mida standardne lühikeste lugemite põhine MAG-rekonstruktsioon lahendab ainult osaliselt. Need tulemused rõhutavad nii genoomipõhise metagenoomika väärtust kui ka selle praeguseid tehnilisi piiranguid.

Töö kolmandas osas kasutasin EstMB kohordi elektrooniliste terviseandmetega seotud infot, et uurida mikrobioomi ja haiguste vahelisi seoseid. Liigitaseme mikrobioomiülene assotsiatsiooniuuring tuvastas mitmeid olulisi seoseid bakteriliikide ja haiguste vahel, sealhulgas seoseid ka varem kirjeldamata liikidega. See näitab, et kohordispetsiifiline MAG-andmebaas võib avada haigustega seotud mikroobseid signaale, mis jääksid tavapäraste referentsandmebaasidega märkamatuks või halvasti kirjeldatuks.

Lisaks arendasin välja raamistiku liikide valimiseks liigisisese assotsiatsioonianalüüsi jaoks. Selle lähenemise abil näitasin, et ühe bakteriliigi sees olevad genoomsed üksused võivad olla haigustega seotud viisil, mida kogu liigi tasemel analüüs ei tuvasta. Näiteks *Odoribacter splanchnicus* ei näidanud liigitasemel olulisi haigusseoseid, kuid selle ühe genoomse üksuse esinemine oli seotud kahe haigusfenotüübiga. See toetab ideed, et liigitaseme profiilimine võib varjata bioloogiliselt olulist liigisisest heterogeensust ning et genoomsete üksuste tasemel analüüs võib parandada mikrobioomi ja tervise vaheliste seoste lahutusvõimet.

Kokkuvõttes näitab see väitekiri, et genoomipõhine metagenoomika võimaldab liikuda mikrobioomi kirjeldamiselt suurema bioloogilise lahutusvõimega analüüside poole. Populatsioonispetsiifilised MAG-referentsandmebaasid parandavad mikrobioomi kirjeldamist, aitavad tuvastada varem kirjeldamata mikroobset mitmekesisust ja loovad aluse liigisisese struktuuri uurimiseks. Tulevikus on vaja paremaid eksperimentaalseid ja bioinformaatilisi meetodeid, sealhulgas sügavamat sekveneerimist, kvaliteetsemaid referentsandmebaase, pikemate lugemite kasutamist ja süstemaatilisemat lähenemist liigisiseste genoomsete üksuste määratlemisele. Sellised arengud aitavad muuta mikrobioomiuuringud täpsemaks ja bioloogiliselt tõlgendatavamaks ning võivad parandada meie arusaamist mikrobioomi rollist inimese tervisele.

REFERENCES

Aasmets, Oliver, Kertu Liis Krigul, Kreete Lüll, Andres Metspalu, and Elin Org. 2022. "Gut Metagenome Associations with Extensive Digital Health Data in a Volunteer-Based Estonian Microbiome Cohort." *Nature Communications* 13(1): 869. doi:10.1038/s41467-022-28464-9.

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson, and Per H Nielsen. 2013. "Genome Sequences of Rare, Uncultured Bacteria Obtained by Differential Coverage Binning of Multiple Metagenomes." *Nature Biotechnology* 31(6): 533–38. doi:10.1038/nbt.2579.

Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. 2019. "A New Genomic Blueprint of the Human Gut Microbiota." *Nature* 568(7753): 499–504. doi:10.1038/s41586-019-0965-1.

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2021. "A Unified Catalog of 204,938 Reference Genomes from the Human Gut Microbiome." *Nature Biotechnology* 39(1): 105–14. doi:10.1038/s41587-020-0603-3.

Aroney, Samuel T N, Rhys J P Newell, Jakob N Nissen, Antonio Pedro Camargo, Gene W Tyson, and Ben J Woodcroft. 2025. "CoverM: Read Alignment Statistics for Metagenomics" ed. Can Alkan. *Bioinformatics* 41(4): btaf147. doi:10.1093/bioinformatics/btaf147.

Barlow, Jacob T., Said R. Bogatyrev, and Rustem F. Ismagilov. 2020. "A Quantitative Sequencing Framework for Absolute Abundance Measurements of Mucosal and Lumenal Microbial Communities." *Nature Communications* 11(1): 2590. doi:10.1038/s41467-020-16224-6.

Bertrand, Denis, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M. Senthil Kumar, Chenhao Li, Mirta Dvornicic, et al. 2019. "Hybrid Metagenomic Assembly Enables High-Resolution Analysis of Resistance Determinants and Mobile Elements in Human Microbiomes." *Nature Biotechnology* 37(8): 937–44. doi:10.1038/s41587-019-0191-2.

Bickhart, Derek M., Mikhail Kolmogorov, Elizabeth Tseng, Daniel M. Portik, Anton Korobeynikov, Ivan Tolstoganov, Gherman Uritskiy, et al. 2022. "Generating Lineage-Resolved, Complete Metagenome-Assembled Genomes from Complex Microbial Communities." *Nature Biotechnology* 40(5): 711–19. doi:10.1038/s41587-021-01130-z.

Blanco-Míguez, Aitor, Francesco Beghini, Fabio Cumbo, Lauren J. McIver, Kelsey N. Thompson, Moreno Zolfo, Paolo Manghi, et al. 2023. "Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species Using MetaPhlAn 4." *Nature Biotechnology* 41(11): 1633–44. doi:10.1038/s41587-023-01688-w.

Bobay, Louis-Marie, Charles C. Traverse, and Howard Ochman. 2015. "Impermanence of Bacterial Clones." *Proceedings of the National Academy of Sciences* 112(29): 8893–8900. doi:10.1073/pnas.1501724112.

Bolte, Erin E., David Moorshead, and Kjersti M. Aagaard. 2022. "Maternal and Early Life Exposures and Their Potential to Influence Development of the Microbiome." *Genome Medicine* 14(1): 4. doi:10.1186/s13073-021-01005-7.

Bozdag, G Ozan, and Jasmine Ono. 2022. "Evolution and Molecular Bases of Reproductive Isolation." *Current Opinion in Genetics & Development* 76: 101952. doi:10.1016/j.gde.2022.101952.

Bradley, Patrick H., and Katherine S. Pollard. 2017. “Proteobacteria Explain Significant Functional Variability in the Human Gut Microbiome.” *Microbiome* 5(1): 36. doi:10.1186/s40168-017-0244-z.

Chaumeil, Pierre-Alain, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. 2020. “GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database” ed. John Hancock. *Bioinformatics* 36(6): 1925–27. doi:10.1093/bioinformatics/btz848.

Chen, Lin-Xing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. 2020. “Accurate and Complete Genomes from Metagenomes.” *Genome Research* 30(3): 315–33. doi:10.1101/gr.258640.119.

Chivian, Dylan, Sean P. Jungbluth, Paramvir S. Dehal, Elisha M. Wood-Charlson, Richard S. Canon, Benjamin H. Allen, Mikayla M. Clark, et al. 2023. “Metagenome-Assembled Genome Extraction and Analysis from Microbiomes Using KBase.” *Nature Protocols* 18(1): 208–38. doi:10.1038/s41596-022-00747-x.

Chklovski, Alex, Donovan H. Parks, Ben J. Woodcroft, and Gene W. Tyson. 2023. “CheckM2: A Rapid, Scalable and Accurate Tool for Assessing Microbial Genome Quality Using Machine Learning.” *Nature Methods* 20(8): 1203–12. doi:10.1038/s41592-023-01940-w.

Conrad, Roth E., Catherine E. Brink, Tomeu Viver, Luis M. Rodriguez-R, Borja Aldeguer-Riquelme, Janet K. Hatt, Stephanus N. Venter, et al. 2024. “Microbial Species and Intraspecies Units Exist and Are Maintained by Ecological Cohesiveness Coupled to High Homologous Recombination.” *Nature Communications* 15(1): 9906. doi:10.1038/s41467-024-53787-0.

DeMaere, Matthew Z., and Aaron E. Darling. 2019. “bin3C: Exploiting Hi-C Sequencing Data to Accurately Resolve Metagenome-Assembled Genomes.” *Genome Biology* 20(1): 46. doi:10.1186/s13059-019-1643-1.

Didelot, Xavier, Daniel Lawson, Aaron Darling, and Daniel Falush. 2010. “Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences.” *Genetics* 186(4): 1435–49. doi:10.1534/genetics.110.120121.

Diop, Awa, Ellis L. Torrance, Caroline M. Stott, and Louis-Marie Bobay. 2022. “Gene Flow and Introgression Are Pervasive Forces Shaping the Evolution of Bacterial Species.” *Genome Biology* 23(1): 239. doi:10.1186/s13059-022-02809-5.

Dong, Quanbin, Beining Ma, Xiaofeng Zhou, Pan Huang, Mengke Gao, Shuai Yang, Yuwen Jiao, et al. 2025. “Expanded Gut Microbial Genomes from Chinese Populations Reveal Population-Specific Genomic Features Related to Human Physiological Traits.” *Genome Medicine* 17(1): 137. doi:10.1186/s13073-025-01566-x.

Fraser, Christophe, Eric J. Alm, Martin F. Polz, Brian G. Spratt, and William P. Hanage. 2009. “The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity.” *Science* 323(5915): 741–46. doi:10.1126/science.1159388.

Galazzo, Gianluca, Niels Van Best, Birke J. Benedikter, Kevin Janssen, Liene Bervoets, Christel Driessen, Melissa Oomen, et al. 2020. “How to Count Our Microbes? The Effect of Different Quantitative Microbiome Profiling Approaches.” *Frontiers in Cellular and Infection Microbiology* 10: 403. doi:10.3389/fcimb.2020.00403.

Goris, Johan, Konstantinos T. Konstantinidis, Joel A. Klappenbach, Tom Coenye, Peter Vandamme, and James M. Tiedje. 2007. “DNA–DNA Hybridization Values and Their Relationship to Whole-Genome Sequence Similarities.” *International Journal of Systematic and Evolutionary Microbiology* 57(1): 81–91. doi:10.1099/ijs.0.64483-0.

Handelsman, Jo, Michelle R. Rondon, Sean F. Brady, Jon Clardy, and Robert M. Goodman. 1998. “Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products.” *Chemistry & Biology* 5(10): R245–49. doi:10.1016/S1074-5521(98)90108-9.

Hedlund, Brian P., Maria Chuvochina, Philip Hugenholtz, Konstantinos T. Konstantinidis, Alison E. Murray, Marike Palmer, Donovan H. Parks, et al. 2022. “SeqCode: A Nomenclatural Code for Prokaryotes Described from Sequence Data.” *Nature Microbiology*. doi:10.1038/s41564-022-01214-9.

Hugenholtz, Philip, and Gene W Tyson. 2008. “Ten Years after the Term Metagenomics Was Coined, the Approach Continues to Gather Momentum. This Culture-Independent, Molecular Way of Analysing Environmental Samples of Cohabiting Microbial Populations Has Opened up Fresh Perspectives on Microbiology.” *Nature* 455.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. “High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries.” *Nature Communications* 9(1): 5114. doi:10.1038/s41467-018-07641-9.

Kayani, Masood Ur Rehman, Wanqiu Huang, Ru Feng, and Lei Chen. 2021. “Genome-Resolved Metagenomics Using Environmental and Clinical Samples.” *Briefings in Bioinformatics* 22(5): bbab030. doi:10.1093/bib/bbab030.

Kim, Chan Yeong, Muyoung Lee, Sunmo Yang, Kyungnam Kim, Dongeun Yong, Hye Ryun Kim, and Insuk Lee. 2021. “Human Reference Gut Microbiome Catalog Including Newly Assembled Genomes from Under-Represented Asian Metagenomes.” *Genome Medicine* 13(1): 134. doi:10.1186/s13073-021-00950-7.

Kim, Chan Yeong, Junyeong Ma, and Insuk Lee. 2022. “HiFi Metagenomic Sequencing Enables Assembly of Accurate and Complete Genomes from Human Gut Microbiota.” *Nature Communications* 13(1): 6367. doi:10.1038/s41467-022-34149-0.

Kim, Nayeon, Junyeong Ma, Wonjong Kim, Jungyeon Kim, Peter Belenky, and Insuk Lee. 2024. “Genome-Resolved Metagenomics: A Game Changer for Microbiome Medicine.” *Experimental & Molecular Medicine* 56(7): 1501–12. doi:10.1038/s12276-024-01262-7.

Konstantinidis, Konstantinos T, Alban Ramette, and James M Tiedje. 2006. “The Bacterial Species Definition in the Genomic Era.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 361(1475): 1929–40. doi:10.1098/rstb.2006.1920.

Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. “Genomic Insights That Advance the Species Definition for Prokaryotes.” *Proceedings of the National Academy of Sciences* 102(7): 2567–72. doi:10.1073/pnas.0409727102.

Kuhn, Michael, Thomas Sebastian B Schmidt, Pamela Ferretti, Anna Głazek, Shahriyar Mahdi Robbani, Wasiu Akanni, Anthony Fullam, et al. 2026. “Metalog: Curated and Harmonised Contextual Data for Global Metagenomics Samples.” *Nucleic Acids Research* 54(D1): D826–34. doi:10.1093/nar/gkaf1118.

Kurilshikov, Alexander, Carolina Medina-Gomez, Rodrigo Bacigalupe, Djawad Radjabzadeh, Jun Wang, Ayse Demirkan, Caroline I. Le Roy, et al. 2021. “Large-Scale Association Analyses Identify Host Factors Influencing Human Gut Microbiome Composition.” *Nature Genetics* 53(2): 156–65. doi:10.1038/s41588-020-00763-1.

Lewis, William H., Guillaume Tahon, Patricia Geesink, Diana Z. Sousa, and Thijs J. G. Ettema. 2021. “Innovations to Culturing the Uncultured Microbial Majority.” *Nature Reviews Microbiology* 19(4): 225–40. doi:10.1038/s41579-020-00458-8.

Ma, Junyeong, Nayeon Kim, Jun Hyung Cha, Wonjong Kim, Chan Yeong Kim, Yongho Lee, Han Sang Kim, et al. 2025. “A Human Gut Metagenome-Assembled Genome Catalogue Spanning 41 Countries Supports Genome-Scale Metabolic Models.” *Nature Microbiology* 11(1): 317–34. doi:10.1038/s41564-025-02206-1.

MetaHIT Consortium, Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” *Nature* 464(7285): 59–65. doi:10.1038/nature08821.

Milanese, Alessio, Daniel R Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, et al. 2019. “Microbial Abundance, Activity and Population Genomic Profiling with mOTUs2.” *Nature Communications* 10(1): 1014. doi:10.1038/s41467-019-08844-4.

Morton, James T., Clarisse Marotz, Alex Washburne, Justin Silverman, Livia S. Zaramela, Anna Edlund, Karsten Zengler, and Rob Knight. 2019. “Establishing Microbial Composition Measurement Standards with Reference Frames.” *Nature Communications* 10(1): 2719. doi:10.1038/s41467-019-10656-5.

Moss, Eli L., Dylan G. Maghini, and Ami S. Bhatt. 2020. “Complete, Closed Bacterial Genomes from Microbiomes Using Nanopore Sequencing.” *Nature Biotechnology* 38(6): 701–7. doi:10.1038/s41587-020-0422-6.

Murray, Alison E., John Freudenstein, Simonetta Gribaldo, Roland Hatzenpichler, Philip Hugenholtz, Peter Kämpfer, Konstantinos T. Konstantinidis, et al. 2020. “Roadmap for Naming Uncultivated Archaea and Bacteria.” *Nature Microbiology* 5(8): 987–94. doi:10.1038/s41564-020-0733-x.

Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, et al. 2021. “A Genomic Catalog of Earth’s Microbiomes.” *Nature Biotechnology* 39(4): 499–509. doi:10.1038/s41587-020-0718-6.

Nishijima, Suguru, Anthony Fullam, Thomas S B Schmidt, Michael Kuhn, and Peer Bork. 2026. “VIRE: A Metagenome-Derived, Planetary-Scale Virome Resource with Environmental Context.” *Nucleic Acids Research* 54(D1): D902–11. doi:10.1093/nar/gkaf1225.

Nyström-Persson, Johan, Nishad Bapatdhar, and Samik Ghosh. 2025. “Precise and Scalable Metagenomic Profiling with Sample-Tailored Minimizer Libraries.” *NAR Genomics and Bioinformatics* 7(2): lqaf076. doi:10.1093/nargab/lqaf076.

Orakov, Askarbek, Anthony Fullam, Luis Pedro Coelho, Supriya Khedkar, Damian Szklarczyk, Daniel R. Mende, Thomas S. B. Schmidt, and Peer Bork. 2021. “GUNC: Detection of Chimerism and Contamination in Prokaryotic Genomes.” *Genome Biology* 22(1): 178. doi:10.1186/s13059-021-02393-0.

Oren, Aharon, David R. Arahal, Markus Göker, Edward R. B. Moore, Ramon Rossello-Mora, and Iain C. Sutcliffe. 2023. “International Code of Nomenclature of Prokaryotes. Prokaryotic Code (2022 Revision).” *International Journal of Systematic and Evolutionary Microbiology* 73(5a). doi:10.1099/ijsem.0.005585.

Parks, Donovan H., Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, and Philip Hugenholtz. 2020. “A Complete Domain-to-Species Taxonomy for Bacteria and Archaea.” *Nature Biotechnology* 38(9): 1079–86. doi:10.1038/s41587-020-0501-8.

Parks, Donovan H, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2022. “GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy.” *Nucleic Acids Research* 50(D1): D785–94. doi:10.1093/nar/gkab776.

Parks, Donovan H, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. “A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life.” *Nature Biotechnology* 36(10): 996–1004. doi:10.1038/nbt.4229.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research* 25(7): 1043–55. doi:10.1101/gr.186072.114.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. “Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life.” *Nature Microbiology* 2(11): 1533–42. doi:10.1038/s41564-017-0012-7.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.” *Cell* 176(3): 649-662.e20. doi:10.1016/j.cell.2019.01.001.

Qin, Youwen, Aki S. Havulinna, Yang Liu, Pekka Jousilahti, Scott C. Ritchie, Alex Tokolyi, Jon G. Sanders, et al. 2022. “Combined Effects of Host Genetics and Diet on Human Gut Microbiota and Incident Disease in a Single Population Cohort.” *Nature Genetics* 54(2): 134–42. doi:10.1038/s41588-021-00991-z.

Ratiner, Karina, Dragos Ciocan, Suhaib K. Abdeen, and Eran Elinav. 2024. “Utilization of the Microbiome in Personalized Medicine.” *Nature Reviews Microbiology* 22(5): 291–308. doi:10.1038/s41579-023-00998-9.

Richter, Michael, and Ramon Rosselló-Móra. 2009. “Shifting the Genomic Gold Standard for the Prokaryotic Species Definition.” *Proceedings of the National Academy of Sciences* 106(45): 19126–31. doi:10.1073/pnas.0906412106.

Rodriguez-R, Luis M., Roth E. Conrad, Tomeu Viver, Dorian J. Feistel, Blake G. Lindner, Stephanus N. Venter, Luis H. Orellana, et al. 2024. “An ANI Gap within Bacterial Species That Advances the Definitions of Intra-Species Units” ed. Igor B. Jouline. *mBio* 15(1): e02696-23. doi:10.1128/mbio.02696-23.

Rosselló-Mora, R. 2001. “The Species Concept for Prokaryotes.” *FEMS Microbiology Reviews* 25(1): 39–67. doi:10.1016/S0168-6445(00)00040-1.

Rothschild, Daphna, Omer Weissbrod, Elad Barkan, Alexander Kurilshikov, Tal Korem, David Zeevi, Paul I. Costea, et al. 2018. “Environment Dominates over Host Genetics in Shaping Human Gut Microbiota.” *Nature* 555(7695): 210–15. doi:10.1038/nature25973.

Schmidt, Thomas S B, Anthony Fullam, Pamela Ferretti, Askarbek Orakov, Oleksandr M Maistrenko, Hans-Joachim Ruscheweyh, Ivica Letunic, et al. 2024. “SPIRE: A Searchable, Planetary-Scale mIcrobiome REsource.” *Nucleic Acids Research* 52(D1): D777–83. doi:10.1093/nar/gkad943.

Segev, Tomer, Daniel Barak, Liron Zahavi, Anastasia Godneva, Michal Rein, David Krongauz, Dorit Samocha-Bonet, et al. 2026. “Diet–Microbiome Associations in 10,068 Individuals from the Human Phenotype Project to Guide Personalized Nutrition.” *Nature Medicine* 32(5): 1884–94. doi:10.1038/s41591-026-04312-x.

Shaiber, Alon, and A. Murat Eren. 2019. “Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories” ed. David A. Relman. *mBio* 10(3): e00725-19. doi:10.1128/mBio.00725-19.

Smith, J M, N H Smith, M O'Rourke, and B G Spratt. 1993. "How Clonal Are Bacteria?" *Proceedings of the National Academy of Sciences* 90(10): 4384–88. doi:10.1073/pnas.90.10.4384.

Suau, Antonia, Régis Bonnet, Malène Sutren, Jean-Jacques Godon, Glenn R. Gibson, Matthew D. Collins, and Joel Doré. 1999. "Direct Analysis of Genes Encoding 16S rRNA from Complex Communities Reveals Many Novel Molecular Species within the Human Gut." *Applied and Environmental Microbiology* 65(11): 4799–4807. doi:10.1128/AEM.65.11.4799-4807.1999.

Suzuki, Yoshihiko, Suguru Nishijima, Yoshikazu Furuta, Jun Yoshimura, Wataru Suda, Kenshiro Oshima, Masahira Hattori, and Shinichi Morishita. 2019. "Long-Read Metagenomic Exploration of Extrachromosomal Mobile Genetic Elements in the Human Gut." *Microbiome* 7(1): 119. doi:10.1186/s40168-019-0737-z.

Tannock, G. W., K. Munro, H. J. M. Harmsen, G. W. Welling, J. Smart, and P. K. Gopal. 2000. "Analysis of the Fecal Microflora of Human Subjects Consuming a Probiotic Product Containing *Lactobacillus Rhamnosus* DR20." *Applied and Environmental Microbiology* 66(6): 2578–88. doi:10.1128/AEM.66.6.2578-2588.2000.

The Genome Standards Consortium, Robert M Bowers, Nikos C Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T B K Reddy, et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35(8): 725–31. doi:10.1038/nbt.3893.

The Integrative HMP (iHMP) Research Network Consortium, Lita M. Proctor, Heather H. Creasy, Jennifer M. Fettweis, Jason Lloyd-Price, Anup Mahurkar, Wenyu Zhou, et al. 2019. "The Integrative Human Microbiome Project." *Nature* 569(7758): 641–48. doi:10.1038/s41586-019-1238-8.

The NIH HMP Working Group, Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A. Schloss, et al. 2009. "The NIH Human Microbiome Project." *Genome Research* 19(12): 2317–23. doi:10.1101/gr.096651.109.

Tonnelé, Hélène, Denghui Chen, Felipe Morillo, Jorge Garcia-Calleja, Apurva S. Chitre, Benjamin B. Johnson, Thiago Missfeldt Sanches, et al. 2025. "Genetic Architecture and Mechanisms of Host-Microbiome Interactions from a Multi-Cohort Analysis of Outbred Laboratory Rats." *Nature Communications* 16(1): 10126. doi:10.1038/s41467-025-66105-z.

Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, et al. 2004. "Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment." *Nature* 428(6978): 37–43. doi:10.1038/nature02340.

Van Rossum, Thea, Pamela Ferretti, Oleksandr M. Maistrenko, and Peer Bork. 2020. "Diversity within Species: Interpreting Strains in Microbiomes." *Nature Reviews Microbiology* 18(9): 491–506. doi:10.1038/s41579-020-0368-1.

Vandeputte, Doris, Gunter Kathagen, Kevin D'hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, et al. 2017. "Quantitative Microbiome Profiling Links Gut Community Variation to Microbial Load." *Nature* 551(7681): 507–11. doi:10.1038/nature24460.

Verhelst, Rita, Hans Verstraelen, Geert Claeys, Gerda Verschraegen, Leen Van Simaey, Catharine De Ganck, Ellen De Backer, Marleen Temmerman, and Mario Vaneechoutte. 2005. "Comparison between Gram Stain and Culture for the Characterization of Vaginal Microflora: Definition of a Distinct Grade That Resembles Grade I Microflora and Revised Categorization of Grade I Microflora." *BMC Microbiology* 5(1): 61. doi:10.1186/1471-2180-5-61.

Viver, Tomeu, Roth E. Conrad, Luis M. Rodriguez-R, Ana S. Ramírez, Stephanus N. Venter, Jairo Rocha-Cárdenas, Mercè Llabrés, et al. 2024. “Towards Estimating the Number of Strains That Make up a Natural Bacterial Population.” *Nature Communications* 15(1): 544. doi:10.1038/s41467-023-44622-z.

Westram, Anja M., Sean Stankowski, Parvathy Surendranadh, and Nick Barton. 2022. “What Is Reproductive Isolation?” *Journal of Evolutionary Biology* 35(9): 1143–64. doi:10.1111/jeb.14005.

Whitman, William B., Maria Chuvochina, Brian P. Hedlund, Philip Hugenholtz, Konstantinos T. Konstantinidis, Alison E. Murray, Marike Palmer, et al. 2022. “Development of the SeqCode: A Proposed Nomenclatural Code for Uncultivated Prokaryotes with DNA Sequences as Type.” *Systematic and Applied Microbiology* 45(5): 126305. doi:10.1016/j.syapm.2022.126305.

Wilson, K H, and R B Blitchington. 1996. “Human Colonic Biota Studied by Ribosomal DNA Sequence Analysis.” *Applied and Environmental Microbiology* 62(7): 2273–78. doi:10.1128/aem.62.7.2273-2278.1996.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. “Improved Metagenomic Analysis with Kraken 2.” *Genome Biology* 20(1): 257. doi:10.1186/s13059-019-1891-0.

Yatsunenko, Tanya, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, et al. 2012. “Human Gut Microbiome Viewed across Age and Geography.” *Nature* 486(7402): 222–27. doi:10.1038/nature11053.

Zeevi, David, Tal Korem, Anastasia Godneva, Noam Bar, Alexander Kurilshikov, Maya Lotan-Pompan, Adina Weinberger, et al. 2019. “Structural Variation in the Gut Microbiome Associates with Host Health.” *Nature* 568(7750): 43–48. doi:10.1038/s41586-019-1065-y.

Zeng, Shuqin, Dhrati Patangia, Alexandre Almeida, Zhemin Zhou, Dezhi Mu, R. Paul Ross, Catherine Stanton, and Shaopu Wang. 2022. “A Compendium of 32,277 Metagenome-Assembled Genomes and over 80 Million Genes from the Early-Life Human Gut Microbiome.” *Nature Communications* 13(1): 5139. doi:10.1038/s41467-022-32805-z.

Zhernakova, Daria V., Daoming Wang, Lei Liu, Sergio Andreu-Sánchez, Yue Zhang, Angel J. Ruiz-Moreno, Haoran Peng, et al. 2024. “Host Genetic Regulation of Human Gut Microbial Structural Variation.” *Nature* 625(7996): 813–21. doi:10.1038/s41586-023-06893-w.

ACKNOWLEDGMENTS



I did not plan to do a PhD. In 2022, I found myself in Estonia, and at that time I honestly believed that the world would probably end soon. I want to thank the Ukrainian Armed Forces for holding the line and giving all of us a chance to continue living our lives. I also want to thank Estonia and the Estonian people for restoring my faith in humanity and for bringing me into science. The enormous amount of kindness and help I received here surprised me, and Estonia will always be my second favourite country after my own beautiful and strong Ukraine.

In March 2022, I had no idea what I was going to do next. I knew only one person in Estonia, my landlady, Merike. Merike showed me one of the most impressive sides of Estonia: the power of community. I needed a job, so Merike wrote about me on Facebook, and somehow her post reached the Tartu Biobank. I was invited for an interview. Thank you, Merike. It is probably because of you that I am a researcher now.

I came to Tartu by train and saw the Institute of Genomics for the first time. There, I met Lili, who introduced me to my current PI, Elin, showed me the Aparaat, and bought me soup for lunch. I remember that this was the first time I thought that everything was going to be okay. Thank you, Lili!

Soon, I joined the Tartu microbiome group and started working on bacterial genome reconstruction under Elin's supervision. I was very lucky because Elin turned out to be the best PI anyone could have. Elin, I want to thank you for so many things. Thank you for the care I always felt, not only at work but also outside the office. I always knew that if I faced a problem, I would not be completely alone and could ask you for help. That meant a lot to me. Despite being extremely busy, you always managed to find time for me and listen to both my project ideas and my concerns about everything else. Thank you for never pressuring me to follow the original plan and instead giving me the freedom to develop the project in the directions I found interesting. Thank you also for supporting my side projects. Whenever I wanted to attend a conference, summer school, or research visit, you supported me. You not only helped make these trips possible but also introduced me to interesting people and made conference travel not only useful, but also fun.

When I first met Elin, she introduced me to Kertu and Oliver so that I could privately ask them anything I wanted about the group. I remember that when we entered the room, they were sitting at their computers and looking very serious.

At the same time, they were wearing brightly colored wigs and huge heart-shaped glasses. That was the moment I understood that this was my place.

Thank you, Oliver and Kertu, for being so supportive and for sharing your knowledge and skills with me whenever I needed help. A special thank you for providing me with my favorite support dog, Nuru.



During my PhD, I was able to participate in many conferences and schools. They gave me not only new knowledge but also the opportunity to meet interesting people from all over the world and see places I had never seen before. I travelled from Oulu in Finland, where the sun simply does not go down in June, to Brisbane in Australia, where bin chickens walk around the streets instead of pigeons and you can travel to work by city boat. I am grateful to all the grants and scholarships that made them possible: the EMBO|EMBL Symposium Travel Grant, the SymbNET Travel Grant, the Erasmus+ Scholarship Grant, the Estonian Doctoral School Mobility Grant, the CWT Estonia Kaleva Travel Scholarship, the EMBO Special Travel Grant, and the Mobility Grant for Doctoral Students of the Graduate School of Biomedicine and Biotechnology.

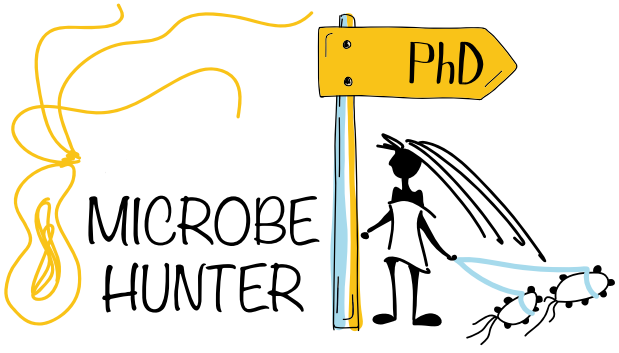
I have always loved data analysis. I remember that when I was a child, my father showed me Excel and taught me how to calculate salaries for several people in it. I was fascinated and played with it for days. Writing, however, was a skill I had to develop during my PhD. The writing retreats organized by our institute were an enormous help. The simple idea of leaving the office and dedicating almost an entire week only to writing, without allowing yourself to be distracted by other tasks, is a game changer for someone like me, who started by struggling to write even a single sentence. Thank you, Monika, for organizing these retreats. I wrote all my papers during them.

I met many friends during my PhD. They were all very different, but at the same time so kind and intelligent that I believe it is worth doing a PhD just to meet people like them. Thank you, Valentina, for teaching me how to play bingo. Galadriel and Jelisaveta, thank you for all the great times. A special thank you to Kai and Madeleine. You supported me when I had a bad moment, and I value that very much. Our Christmas tree was the most beautiful one I have ever had!

I am especially grateful to Anastasiia for making it possible for me to feel at home and for giving me someone to talk to who shares the same values and has similar experiences. Something so simple becomes priceless when you are far from home. Thank you for helping me organize our charity art workshop. We collected much more money than I expected to help protect the Ukrainian sky, and it would not have been possible without you. Thank you also for the wonderful time in Dublin. I truly believe that good results are impossible without a great vacation!

I would like to thank my opponent, Rob Knight, for agreeing to read my thesis and provide feedback. When we spoke in Lisbon, I was impressed by how deeply you understood the technical side of metagenomics and, at the same time, how big and strategic is your vision for the future of our field. This combination is rare, and I greatly value the opportunity to discuss my work with you.

Finally, I want to thank my parents. They are very different, and I feel that I am a mixture of their different characters. Thank you for supporting me, believing in me, and never forcing me to follow the path you wanted instead of the one I wanted. I am very lucky to have parents like you.



PUBLICATIONS

I

Kinga Zielińska*, **Kateryna Pantiukh*,** et. al.
"A large-scale comparative metagenomic analysis of
short-read sequencing platforms indicates high
taxonomic concordance and functional analysis challenges."
mSystems, DOI: 10.1128/msystems.01714-25.

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

| Open Peer Review | Computational Biology | Research Article

A large-scale comparative metagenomic analysis of short-read sequencing platforms indicates high taxonomic concordance and functional analysis challenge

Kinga Zielińska,[1] Kateryna Pantiukh,[2] Pawel P. Łabaj,[1] Tomasz Kosciolek,[3] Elin Org[2]

AUTHOR AFFILIATIONS See affiliation list on p. 14.

ABSTRACT Driven by the increasing scale of microbiome studies and the rise of large, continuously expanding population cohorts, the volume of sequencing data is growing rapidly. As such, ensuring the comparability of data generated across different sequencing platforms has become a pressing concern in efforts to uncover robust links between the microbiome and human health. In this study, we conducted a comprehensive comparison of taxonomic and functional profiles from 1,351 matched human gut microbiome sample pairs, sequenced using both the MGISEQ-2000 (MGI) and NovaSeq 6000 (Illumina NovaSeq) platforms. Taxonomic profiles showed high concordance within and between platforms: 96.44% ± 5.96% of species were shared between MGI–MGI pairs, and 92.07 %± 5.20% were shared between MGI and NovaSeq pairs. The proportion of platform-specific species was low, at 3.42% for MGI–MGI comparisons and 5.89% for MGI–NovaSeq comparisons. No significant differences in Shannon diversity were observed for either within-platform or between-platform comparisons. However, functional profiles revealed notable discrepancies between platforms, which were attributed to differences in pre-sequencing protocols.

IMPORTANCE Our findings demonstrate robust taxonomic comparability between MGI and NovaSeq platforms, while revealing systematic functional differences that should be carefully considered in cross-platform metagenomic studies.

KEYWORDS metagenomics, short-read sequencing, microbiome

Metagenomics has emerged as a powerful tool for uncovering key roles that microorganisms play in human health and environmental ecosystems (1, 2). Recent advancements in Next generation sequencing (NGS) technologies have enabled increasingly fine-scale taxonomic resolution while also significantly expanding our understanding of microbial functions and their dynamic interactions with hosts and other microbes. However, the expansion of NGS has been followed by new protocols and sequencing platforms. This introduced biases, complicating interpretation and reducing the reproducibility of signals identified through metagenomic analyses.

A recent study by Forry et al. compared the Firmicutes:Bacteroides ratios defined from the same biological samples independently by 44 participating laboratories (3). The results revealed a substantial impact of a protocol choice, introducing both biases in measurements tied to specific methodological decisions, as well as influencing measurement robustness, as evidenced by the variability in results across labs employing similar methods. Metagenomic results—specifically microbial load and composition—are highly susceptible to biases stemming from sample storage. This is particularly true for low-biomass samples, underscoring the necessity of an efficient, universal storage method to ensure robust cross-study comparisons (4).

Editor Xiaoquan Su, Qingdao University, Qingdao, Shandong, China

Peer Reviewers Wei-Hua Chen, Huazhong University of Science and Technology, Wuhan, China; Zheng Sun, Brigham and Women's Hospital, Boston, Massachusetts, USA

Address correspondence to Elin Org, elin.org@ut.ee, or Tomasz Kosciolek, t.kosciolek@sanoscience.org.

Kinga Zielińska and Kateryna Pantiukh contributed equally to this article.

T.K. and P.Ł. are co-founders and shareholders of Onebiome sp. z o.o. microbiome characterization-based diet and lifestyle recommendations.

See the funding table on p. 14.

Received 19 December 2025
Accepted 19 January 2026

Copyright © 2026 Zielińska et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

The major providers of massive parallel sequencing technologies in metagenomics are Illumina (sequencing by synthesis) and MGI/BGI (DNBSEQ) technology, both claiming accuracy but differing substantially in underlying chemistry and amplification design. These methodological distinctions can introduce systematic variation in metagenomic data. A key difference lies in the DNA amplification strategy. Illumina's sequencing-by-synthesis technology employs bridge amplification, a solid-phase exponential PCR on a glass flow cell that is known to be susceptible to GC-content bias and the propagation of early-cycle PCR errors (5). In contrast, MGI's DNBSEQ technology utilizes rolling-circle amplification (RCA) to generate DNA nanoballs (DNBs) from single DNA molecules in solution (6). This linear amplification process minimizes amplification bias and reduces PCR duplication compared with exponential bridge amplification. Additional differences exist in the sequencing chip architecture, where MGI loads DNBs onto high-density patterned arrays, which may reduce optical noise compared with the random cluster distribution on Illumina flow cells (7). Beyond sequencing chemistry, upstream steps, such as DNA extraction and library preparation represent a well-documented source of experimental bias that can influence metagenomic profiles (8). Moreover, differences in multiplexing levels between sequencing runs can influence the relative representation of libraries, leading to systematic variation in sequencing depth and read distribution across samples. Collectively, these technical distinctions are potential sources of systematic variation that can influence the results of metagenomic studies.

Comparative analyses of sequencing by synthesis and DNBSEQ technologies conclude that the two could be used interchangeably. They cover a range of fields, starting with whole-genome DNA sequencing (9), through paleogenomics (10) and bacterial genome assembly. Multiple sequencing-specific biases have been discovered, however, including a higher number of insertions detected in (11) and more deletions in BGISEQ-500 genome assemblies, as well as a bias toward genes with higher GC content on the HiSeq platforms (11). These could have profound impacts on the downstream analyses, but they have not been further explored.

Here, we take advantage of a large population-based metagenomic data set (1,351 individuals) simultaneously sequenced using the Illumina NovaSeq 6000 (NovaSeq) and MGISEQ-2000 (MGI) platforms, allowing us to directly compare the sequencing technologies for metagenomic applications (12). We begin by evaluating the robustness of MGI sequencing through an analysis of samples sequenced twice (N = 53), assessing the degree of intra-sequencing variability. We then conduct a comprehensive comparative analysis of the full data set, examining matched samples from MGI and NovaSeq platforms to identify sequencing-specific biases (N = 1,351).

RESULTS

Study design and cohort overview

Our aim was to evaluate the consistency and comparability of microbiome sequencing outputs across the platforms, with a specific focus on the taxonomic and functional profiles. We therefore utilized a subset of the Estonian microbiome (EstMB) cohort, sequenced on NovaSeq and MGI platforms.

The original EstMB cohort is a population-based cohort consisting of 2,509 gut metagenomic samples, sequenced on the NovaSeq 6000 platform (hereafter NovaSeq) to 30.63 ± 3.12 million paired reads per sample (Fig. 1a) (12). A subset of the EstMB cohort was sequenced on the MGISEQ-2000 platform by MGI Tech Co, Ltd (hereafter MGI) with deeper coverage (108.71 ± 42.08 million paired reads per sample). This subset, referred to as the Estonian Microbiome deep (EstMB-deep) cohort , includes 1,729 gut metagenomic samples (53 samples sequenced twice), all of which are also represented in the EstMB cohort (Fig. 1a and b).

To evaluate within-platform and cross-platform variability, we first analyzed the MGI–MGI pairs (53 samples randomly selected and sequenced on the MGI platform twice). We then expanded the analysis to include 1,729 samples sequenced on both platforms: MGI–NovaSeq pairs. To ensure a fair comparison between platforms, MGI reads were

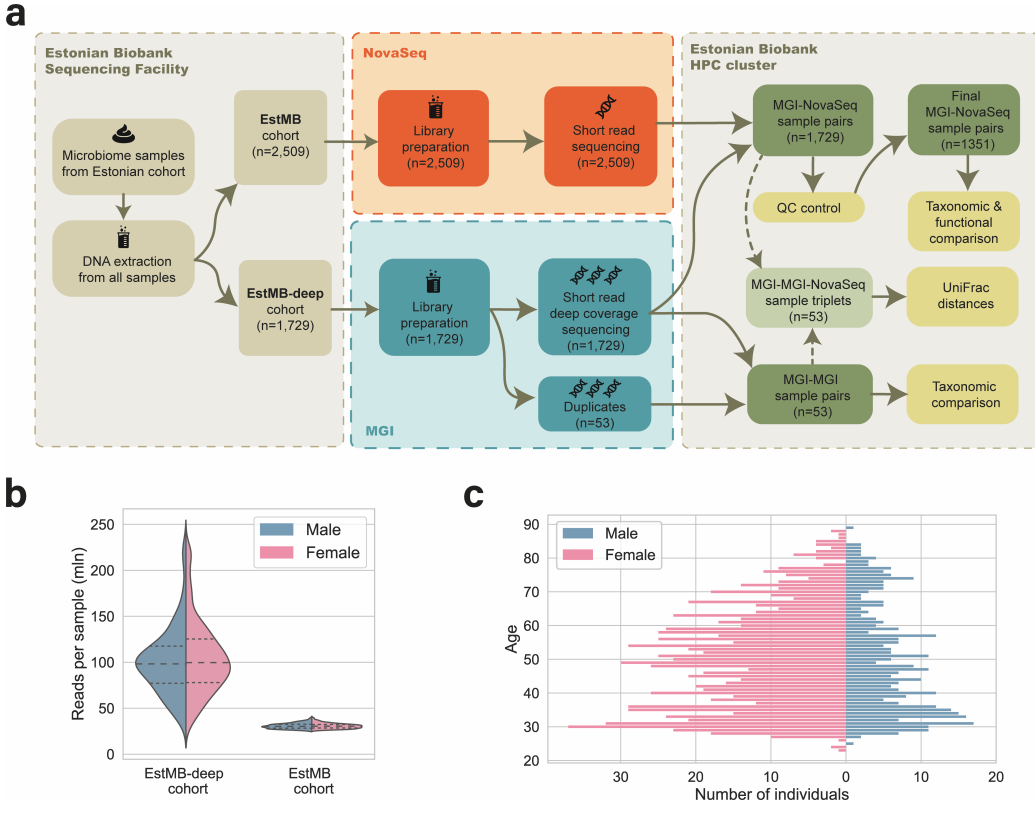


FIG 1 Study overview. (a). Study design and cohort characteristics (b). Distribution of the number of reads per sample for MGI (EstMB-deep cohort) and NovaSeq (EstMB cohort) sequenced samples from the final sample set (*N* = 1,351). (c) Gender and age distributions of individuals in the final sample set (*N* = 1,351).

subsampled first to match the read depth of their NovaSeq counterparts. As a result of the initial quality assessment, the high-quality data set comprised 1,351 MGI–NovaSeq pairs and consisted of 71.87% females, ranging in age from 23 to 89 years with a mean age of 49.7 years (Fig. 1c). The samples then underwent downstream profiling and subsequent cross-platform comparisons (Fig. 1a).

Technical reproducibility: intra-platform comparison of MGI replicates

To establish a baseline for platform consistency, we first evaluated intra-platform variability by examining whether replicate sequencing runs introduced systematic variation. Specifically, we analyzed 53 biological samples sequenced in duplicate on the MGISEQ-2000 platform within the same library preparation batch. This approach ensured the observed reproducibility reflected technical consistency of sequencing within a single batch, isolated from library preparation variation. To assess intra-platform taxonomic consistency, we analyzed MGI–MGI sample pairs (replicates from MGI-1 and MGI-2), which were rarefied to equal read counts (Fig. 1a). Consistency was evaluated using metrics comparing the two sets, including the percentage of shared species, the difference in the number of detected species, and the differences in species prevalence, relative abundance, and Shannon diversity index.

Across the two MGI sequencing runs, a total of 1,083 species (380 genera) were identified. Only 3.42% (37 species) were detected exclusively in one run: 17 species unique to MGI-1 and 20 to MGI-2 (Fig. 2a; Table S1). These set-specific species exhibited low prevalence and low relative abundance. Specifically, 36 of the 37 were found in only one sample (one species, *Lactococcus cremoris*, was found in two). The mean relative abundance of set-specific species was 2.25×10^{-5} %, which contrasts with the 7.67×10^{-4} % mean relative abundance for the shared species (Fig. 2b; Fig. S1).

To assess the consistency of species detection, we calculated the difference in species counts for each MGI–MGI sample pair and visualized the distribution (Fig. 2c). The distribution was centered around a mean difference of 0.85 species (SD = 2.88). The

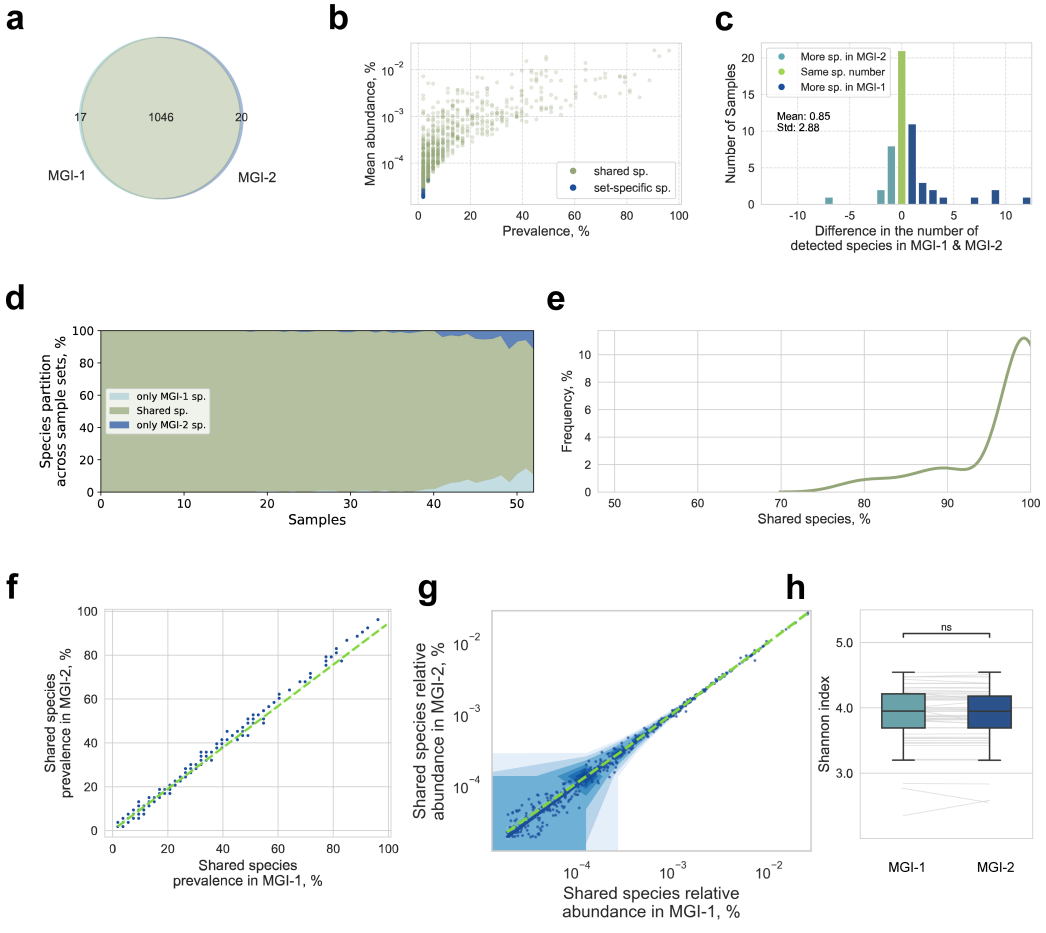


FIG 2 Intra-platform baseline comparison ($N = 53$). (a). Venn diagram of the number of species detected in MGI-1 and MGI-2. (b) Prevalence versus mean relative abundance of detected species, with species found only in one run highlighted in blue. (c) Distribution of differences in the number of detected species within sample pairs. Pairs with identical species counts are represented by green bars, while sample pairs with more species in MGI-1 or MGI-2 are represented by light or dark blue bars, respectively. (d) The percentage of shared and unique run-specific species within each sample pair. (e) Distribution of the percentage of shared species across sample pairs. (f) Comparison of the prevalence of shared species between matched samples. (g) Comparison of the relative abundance of shared species in matched samples. (h) Comparison of the Shannon diversity index within sample pairs.

Wilcoxon signed-rank test showed a significant difference between runs (P = 0.032), but with a small effect size (Cohen's d = 0.29). This suggests the observed difference was minor and likely due to random variation (e.g., stochasticity in library preparation, lane/batch effects, or minor read assignment fluctuations) rather than true biological differences.

The overlap of species detected within MGI–MGI sample pairs demonstrated a high level of concordance. Most pairs shared more than 90% of species (mean = 96.44% ± 5.96%, Fig. 2d and e). Furthermore, species prevalence and relative abundance showed strong agreement across sequencing runs, with high correlation indicated by R^2 values approaching 1.0 (Fig. 2f and g). Finally, Shannon diversity index values were consistent across replicates, with a paired *t*-test revealing no significant differences (T-statistic = 0.086, $P > 0.05$, Cohen's d: 0.017, Fig. 2h). This indicated consistent diversity estimates across the technical replicates.

Cross-platform reproducibility of taxonomic composition

Following the assessment of intra-platform variability, we addressed the central question: whether MGI and NovaSeq sequencing platforms produce biologically comparable results. We conducted a comprehensive analysis of paired samples generated on both platforms. This evaluation included a comparison of taxonomic composition, raw sequencing read quality, and functional profiling.

Assessment of intra- and inter-platform variability using MGI-MGI-NovaSeq triplets

Prior to the large-scale analysis, we assessed intra- and inter-platform variability using Shannon entropy and UniFrac distances on a subset of 53 samples sequenced three times: twice on MGI and once on NovaSeq (MGI–MGI–NovaSeq triads). We used the UniFrac metric to quantify differences based on phylogenetic relationships. We found no statistically significant difference between intra-platform (MGI vs MGI) and inter-platform (MGI vs NovaSeq) UniFrac distances (PERMANOVA *P*-value=1; Fig. S2a), and matched samples showed substantial overlap in the three-dimensional PCoA space (Fig. S2b). However, in 79% of the triads, the two MGI replicates were more similar to each other than to the corresponding NovaSeq replicate. Furthermore, the mean distance between MGI and NovaSeq samples was approximately 2.5 times greater than the distance between the MGI replicates, demonstrating that intra-platform variability was consistently lower than cross-platform variability.

Raw reads sequencing quality assessment

To assess raw sequencing quality, we ran FastQC on all samples from the MGI–NovaSeq sample pairs (1,729 samples per platform). All samples passed key quality metrics, including per-base sequence quality and adapter content. One sample from each platform was flagged for "overrepresented sequences" (those samples were from different pairs). GC content distributions were overall similar between the platforms, with a mean of 45.27% ± 1.82% for MGI and 45.60% ± 1.61% for NovaSeq. Despite the similarity, a paired Wilcoxon test indicated a statistically significant difference in the GC content ($P = 4.25 \times 10^{-25}$), although the effect size was small (Cohen's d ≈ 0.24) (Fig. 3a). The slightly higher GC content observed in NovaSeq samples aligns with the platform's known bias against low-GC regions, likely attributable to PCR amplification and cluster generation artifacts inherent to Illumina's bridge amplification (13). In contrast, DNBSEQ technology, utilizing rolling circle amplification to generate DNA nanoballs before loading onto a patterned array, aims to minimize such amplification biases.

Based on sequence duplication levels, 12.2% of all samples did not meet quality thresholds and were flagged as "warn" or "fail," indicating potential issues such as PCR amplification bias, low library complexity, or contamination. The failure rate was substantially higher for MGI (15.9%) compared with NovaSeq (8.4%) (Fig. 3b). Notably,

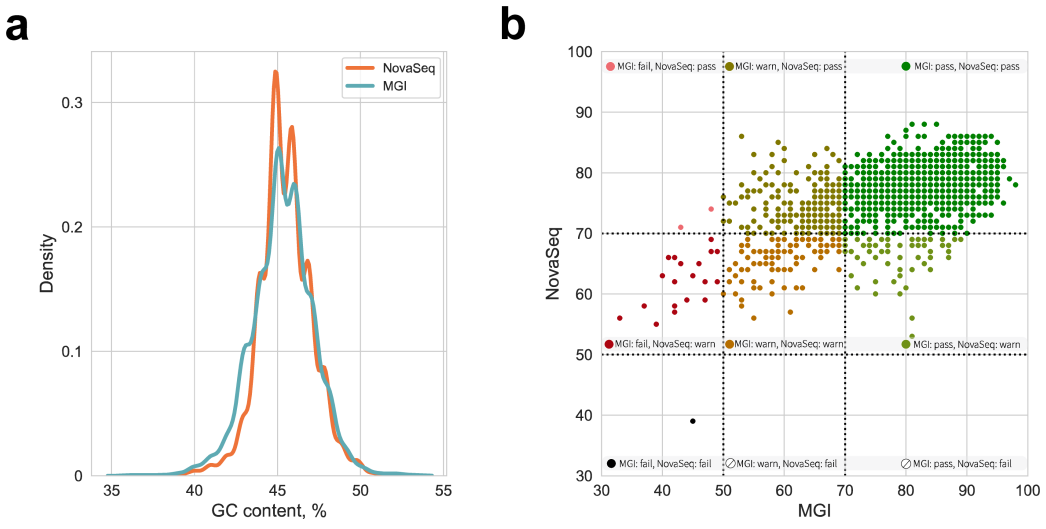


FIG 3 Inter-platform raw sequencing quality comparison (n = 1,729) (a) Distribution of the GC content across MGI and NovaSeq samples. (b) Sample pairs with their corresponding quality tags, assigned based on sequence duplication levels.

the failed samples did not consistently overlap across platforms, suggesting that the observed quality issues likely reflected platform-specific biases, differences in library preparation, or other technical artifacts. If this had been due to early-stage sample processing errors, both platforms would have been affected in a similar manner.

To ensure the robustness of the comparison, sample pairs containing at least one quality warning ("warn" or "fail," N = 378) were excluded from downstream analysis. This yielded a final data set of 1,351 high-quality sample pairs, which were subsequently used for taxonomic and functional assessments.

Cross-platform concordance in taxonomic profiles

We performed taxonomic comparisons for all sample pairs from the final data set (N = 1,351). Across both sequencing platforms, we identified a total of 2,953 species spanning 808 genera. Among these, 5.89% (174 species) were detected exclusively by one platform—88 species were unique to MGI and 86 to NovaSeq (Fig. 4a; Table S2). Most of the platform-specific taxa were rare: 154 out of 174 appeared in only one sample, and their mean relative abundance (1.30×10^{-6} %) was markedly lower than that of the shared species (2.85×10^{-4} %) (Fig. 4b).

To assess the consistency in species detection, we compared the difference in species counts for each paired sample and visualized the distribution of these differences (Fig. 4c). The distribution was centered around a mean difference of −0.50, with a standard deviation of 4.13. The distribution failed the Shapiro–Wilk test ($P \ll 0.05$), prompting rejection of its normality and the use of the Wilcoxon signed-rank test, which showed a significant difference in species counts between platforms ($P = 1.03 \times 10^{-7}$). However, the effect size was small (Cohen's d = −0.12), and visual inspection suggested the difference was minimal and likely due to random variation.

To extend our comparison, we assessed the overlap of species detected within the matched samples. As shown in Fig. 4d and e, most pairs exhibited a high level of concordance, with more than 90% of species shared between the replicates (mean = 92.07% ± 5.20%). In addition, species prevalence and relative abundance exhibited

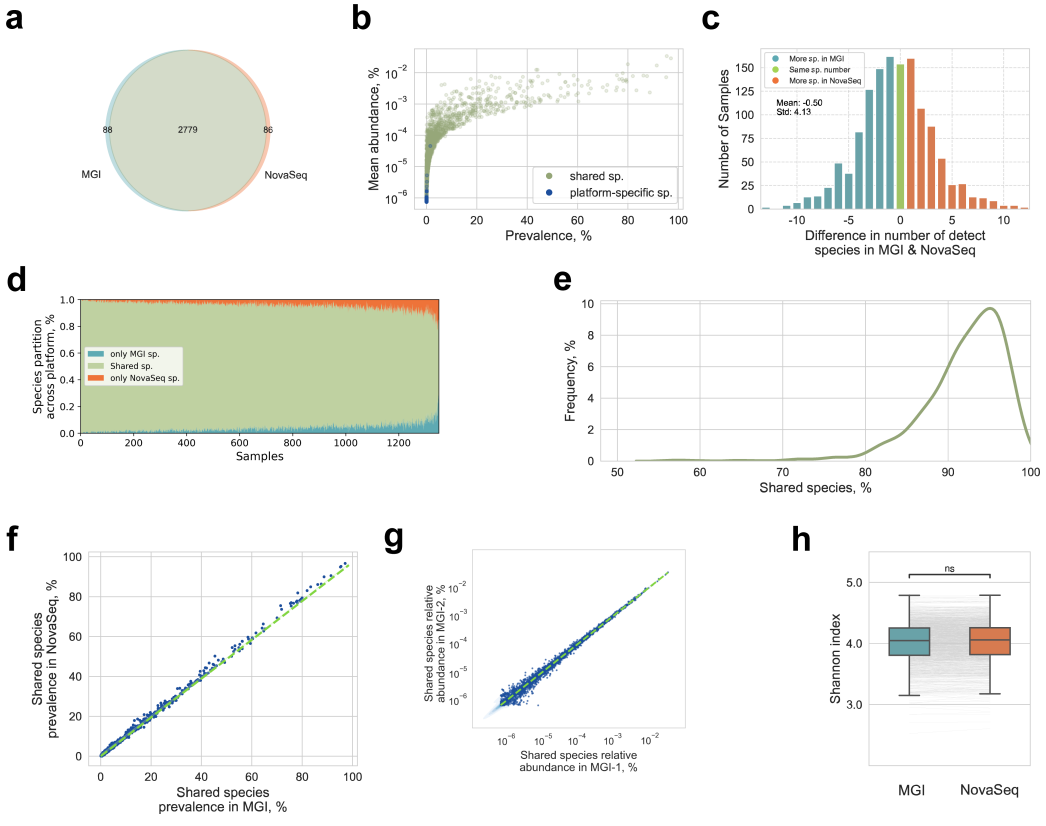


FIG 4 Taxonomic composition comparison between sequencing platforms (n = 1,351) (a) Proportion of species shared between platforms and those detected exclusively in MGI or NovaSeq. (b) Prevalence and relative abundance of all detected species, with platform-specific species (found only in one platform) highlighted in blue. (c) Distribution of differences in the number of species detected per sample pairs. Pairs with identical species counts are shown in green, those with more species in MGI are shown in light blue and with more species in NovaSeq are shown in orange. (d) Percentage of shared and platform-specific species across sample pairs. (e) Distribution of the percentage of shared species within each sample pair. (f) Species prevalence for taxa shared between platforms within sample pairs. (g) Relative abundance of shared species within sample pairs. (h) Comparison of Shannon diversity index within each sample pair.

strong agreement across sequence runs, demonstrating high correlation with R^2 values approaching 1.0 (Fig. 4f and g; Fig. S3).

No significant difference in Shannon diversity was observed between the platforms (paired t-test, $P > 0.05$) (Fig. 4h). Additionally, Procrustes analysis demonstrated excellent concordance between sample pairs across the two platforms ($M^2 = 0.02$, $P = 0.0001$; Fig. S4). We also identified no statistically significant beta diversity differences (Bray-Curtis and Euclidean distances, PERMANOVA P-values > 0.05). Differential abundance analysis with MaAsLin2 initially revealed three species enriched in either platform (two from the *ER4* genus family *Oscillospiraceae* and one *Paraprevotella*), but they all failed to pass the correction for multiple testing (Table S3).

To test the biological interpretability of the results, we examined associations between microbial taxa and health-related variables using linear regression, separately for each sequencing platform. Taxonomic associations were overall consistent between

TABLE 1 Comparison summary: intra-platform (MGI-MGI) and inter-platform (NovaSeq-MGI) taxonomy metrics

Metric	Intra-platform comparison (MGI-MGI sample pairs)	Inter-platform comparison (NovaSeq-MGI sample pairs)
Number of sample pairs	53	1,351
Percentage of set-specific species	3.42% (37 out of 1,083 total species)	5.89% (174 out of 2953 total species)
Difference in the number of species per sample pair (Wilcoxon signed-rank test)	significant difference ($P = 0.038$) Effect size: small (Cohen's $d = 0.29$)	significant difference ($P = 1.03n \times 10^{-7}$) Effect size: very small (Cohen's $d = -0.12$)
Mean percentage of shared species within sample pairs	96.44% ± 5.96%	92.07% ± 5.20%
Correlation of species' prevalence and relative abundance between platforms within sample pairs	$R^2 = 0.999$	$R^2 = 0.999$
Difference in Shannon diversity index between platforms within sample pairs (paired *t*-test)	$P > 0.05$	$P > 0.05$

the MGI and NovaSeq data sets, with only two taxa showing platform-specific significance but with negligible effect sizes (Table S4).

The results of the intra- and inter-platform taxonomic comparisons are summarized in Table 1.

Platform-specific functional differences reflect sequencing depth and complexity

To systematically assess differences in functional profiling between the sequencing platforms, we randomly selected 700 matched sample pairs from our final data set of 1,351 metagenomic samples. Functional annotation was performed with *mi*-faser, which resulted in the identification of 1,170 unique enzymes (EC numbers) across all samples. Of these, 2.65% (31 enzymes) were detected exclusively by one platform—17 enzymes uniquely to MGI and 14 uniquely to NovaSeq (Fig. 5a and b).

The functional similarity (based on enzyme presence/absence) within matched pairs mirrored the taxonomic pattern, ranging from 80% to 97% (Fig. 5c). Despite the high similarity, there were statistically significant differences in beta diversity between platforms (Euclidean distance PERMANOVA $P < 0.0001$; Bray-Curtis distance PERMANOVA $P < 0.001$). In pairs with lower functional overlap, MGI consistently yielded a higher number of detected enzymes per sample. Consequently, MGI systematically exhibited higher pathway recall compared with Illumina when assessed against a pseudo-reference (Fig. 5i). A differential abundance analysis (MaAsLin2) revealed 241 enzymes significantly enriched in one platform (FDR-adjusted $P < 0.05$). The bias showed a clear directional trend: 78% of the 100 most differentially abundant enzymes showed higher relative abundance in Illumina samples. The magnitude varied substantially. For example, trimethylamine-N-oxide reductase (X3.5.4.33) was significantly enriched in Illumina (coefficient = +0.38, q < 0.001), while quinate/shikimate 5-dehydrogenase (X1.1.1.308) was severely underrepresented (coefficient = −1.22, q < 0.001). Notably, these functional biases were most pronounced in central metabolic pathways, including carbohydrate metabolism, amino acid biosynthesis, and energy/cofactor metabolism.

To assess if the functional differences were caused by detection limits of rare functions, we examined the prevalence of the differentially abundant enzymes. We found the bias affected both ubiquitous and low-prevalence functions alike. For instance, many of the strongest discrepancies involved core functions present in samples (e.g., X3.5.4.33 and X1.2.1.12), confirming the bias wasn't due to sensitivity toward low-abundance features. Biases were also observed for sparser functions (e.g., serine endopeptidase, X3.4.21.19, present in 161 samples). This indicates a pervasive systematic shift affecting the entire functional landscape, extending beyond simple enzyme counts.

We hypothesized that sequence diversity underlies the observed functional discrepancies. To test this, a k-mer analysis was performed using the Jellyfish tool to evaluate sequence complexity. We found that MGI samples consistently contained 5–10 times more unique k-mers than their NovaSeq counterparts (Fig. 5d). Despite this pronounced difference in k-mer diversity, the number of reads assigned taxonomic labels

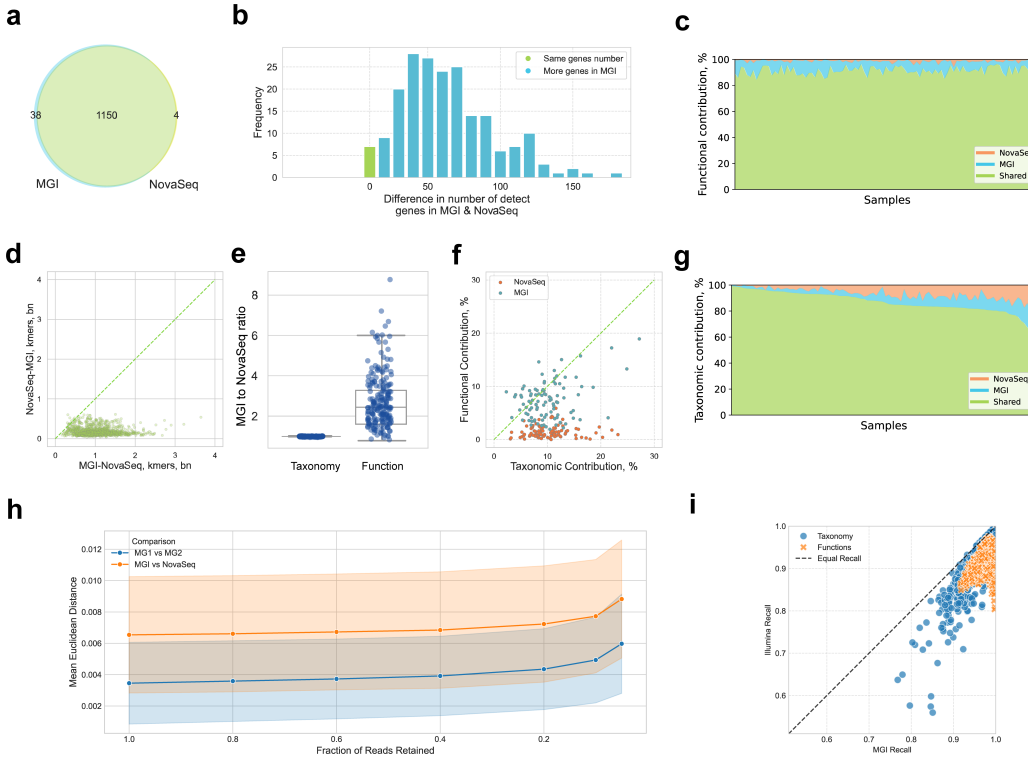


FIG 5 Function-based comparison of the MGI and NovaSeq batches (n = 700). (a) Proportion of genes shared between platforms and platform-specific genes detected exclusively in MGI or NovaSeq platforms. (b) Overlap of functions in matched MGI and NovaSeq samples. (c) A functional overlap in 200 exemplary samples with different levels of taxonomic overlap, sorted by descending functional overlap. (d) Unique k-mers per sample: in NovaSeq versus MGI and vice versa. (e) Contributions of taxonomic and functional microbiome components found uniquely in MGI or NovaSeq. (f) Percentage of reads with taxonomic and functional assignments per sample, presented as MGI to NovaSeq ratio. (g) A taxonomic overlap of matched samples, sorted by descending functional overlap in panel c. (h) Mean Euclidean distance between functional profiles under simulated rarefaction, showing greater systematic distance between platforms than between technical replicates. (i) Recall the comparison of the sequencing technologies, computed pairwise within matched samples using the union as a pseudo-reference.

was similar across platforms. However, MGI samples yielded several-fold more functionally annotated reads (Fig. 5e). This suggests that the increased sequence diversity in MGI data has limited impact on taxonomic profiling, but substantially improves functional annotation.

To investigate if functional differences were linked to taxonomic variation, we ordered sample pairs by decreasing taxonomic overlap and compared their functional similarity. No correlation was observed between the two (Spearman's $\rho = 0.005$, $P > 0.05$, Fig. 5c and g), indicating that functional differences are not simply a consequence of taxonomic variation. Further analysis of genome coverage across a representative genomic fragment revealed platform-specific biases: MGI and Illumina data sets exhibited uneven and non-overlapping coverage patterns (Fig. S5). Specifically, regions highly covered by one platform were entirely missing in the other, which likely explains the functional discrepancies. While we observed no strong clustering by sequencing platform in a plot representing shared, MGI-unique, and Illumina-unique enzyme fractions, the values broadly overlapped, with MGI points consistently slightly higher (Fig. 5f). This suggests

the differences are not solely due to the sequencing technologies themselves, but are more likely linked to pre-sequencing protocols and uneven coverage.

To quantify differences between sequencing replicates and technologies, we simulated the effect of decreasing sequencing depth on the similarity of functional profiles. By iteratively subsampling reads and recalculating the Euclidean distance between profiles, we assessed the stability of intra- and inter-platform comparisons (Fig. 5h). The results clearly demonstrate a systematic difference between the MGI and NovaSeq workflows, with the "MGI vs NovaSeq" comparison showing a consistently higher (approximately double) baseline distance than the technical replicate comparison ("MG1 vs MG2"). Importantly, this significant inter-platform distance persists even after rarefaction was applied to match the sequencing depth of the NovaSeq samples. This highlights that rarefaction alone is insufficient to fully mitigate systematic technological biases. Therefore, while technical replicates on the MGI platform are highly concordant, the choice of sample processing protocol introduces a more substantial source of variation.

DISCUSSION

Previous comparative analyses of MGI and Illumina sequencing technologies largely suggested their interchangeability for applications in metagenomics and clinical diagnostics (2, 14). Existing investigations, however, often suffer from limited sample sizes and fail to rigorously isolate the impact of the sequencing technology from other upstream methodological variables (e.g., library preparation or bioinformatics). Our study addresses these gaps by providing one of the first large-scale comparisons of MGI and Illumina NovaSeq sequencing technologies. We offer a comprehensive evaluation of MGI's inherent robustness and consistency through extensive intra-platform comparisons and a direct, detailed assessment against the NovaSeq platform, aiming to clarify the reliability and comparability of these systems for advancing metagenomics.

Deeper analysis of sequencing quality revealed differences in GC content distribution and sequence duplication between the MGI and NovaSeq platforms, indicating potential sequencing biases. Published research has shown that GC content is a critical challenge, with workflows like MiSeq and NextSeq exhibiting major biases leading to falsely low coverage in GC-rich and especially GC-poor sequences (13). Despite these differences, our observed discrepancies, primarily within the 40%–50% GC coverage range, were not systematic enough to influence major biological interpretations. We found strong agreement in taxonomic profiles, with over 94% overlap in species, genera, and families. Furthermore, key diversity metrics (Shannon entropy, UniFrac distances) and disease-microbiome associations did not differ significantly between the sequencing technologies. However, MGI samples tended to cluster more closely together than with their matched NovaSeq counterparts, suggesting that while intra-platform variation is minimal, inter-platform differences do exist.

Our observations revealed differences in the number and amount of enzymes detected in matched samples (the degree of functional similarity within some matched pairs was as low as 80% and differential enrichment analysis revealed 241 enzymes significantly more abundant in either platform). Overall, the functional overlap was considerably lower and did not align with the taxonomic deviations (Spearman correlation $\rho > 0.05$). This suggests that factors beyond mere taxonomic presence are influencing the functional profiles we observed. A recent study by the ABRF consortium sheds light on this issue, demonstrating how the initial amount of the starting material and subsequent PCR duplication rates can introduce significant biases (15). Their research, though focused on RNA-seq, found that low input material leads to a reduced number of detected genes and increased noise in expression counts. These conclusions are highly relevant to metagenomics, especially with the rise of higher multiplexing strategies.

Higher multiplexing, while cost-effective for large studies, reduces the input material per sample, often resulting in higher duplication rates and uneven coverage of the

studied genomes. Taxonomic annotation tools are generally less affected by this uneven coverage because they use multiple genomic positions and rely on copy number for abundance determination. In contrast, functional profiling is much more sensitive; it depends on smaller genomic fragments and requires reasonably equal coverage across the genome, a condition difficult to maintain when input is limited and duplication is high. This supports our findings, where MGI samples, sequenced in smaller sets at higher depth, maintained higher and more consistent original coverage. Consistent with published work (16), which showed that low-prevalence functions are disproportionately lost at shallow sequencing depths, we found that although rarefaction was applied to normalize read depth, this correction was insufficient to eliminate disparity in functional detection.

Conclusions

Our study confirms that MGI sequencing is a reliable and robust alternative to NovaSeq for taxonomic microbiome profiling; however, we observed differences at the functional level. Our findings underscore the critical need for methodological consistency when comparing results, particularly in studies utilizing a gene catalog.

For accurate and comparable outcomes, standardized sample preparation is a critical first step. Furthermore, experimental design should prioritize sequencing depth. We strongly recommend limiting the degree of sample multiplexing; this increases the amount of input material per sample in the prepared sequencing library, which is essential for minimizing coverage gaps and achieving a more uniform representation of the bacterial genomes. This comprehensive breadth of coverage is required for reliable functional annotations.

MATERIALS AND METHODS

Cohort overview

The EstMB cohort (n = 2,509) was established in 2017 (12) as a part of the Estonian Biobank (EstBB), currently exceeding 202,282 genotyped adults across Estonia (17). The EstMB-deep cohort includes a subset of samples from the EstMB that have been deeply resequenced with MGISEQ-2000 platform (N = 1,729). Samples for EstMB-deep were randomly selected from EstMB.

Microbiome sample collection and DNA extraction

Participants collected a fresh stool sample immediately after defecation using a sterile Pasteur pipette and placed it in a 15 mL conical polypropylene tube. They were instructed to collect the sample as close as possible to their study center visit. Samples were transported and stored at −80°C until DNA extraction. The median time between sampling and freezing was 3 h 25 min (mean 4 h 34 min), and transport time was not significantly associated with alpha (Spearman correlation, P-value of 0.949 for observed richness and 0.464 for Shannon index) or beta diversity (P-value of 0.061, R-squared of 0.0005). Microbial DNA was extracted from 200 mg of stool using a QIAamp DNA Stool Mini Kit (Qiagen), following manufacturer's instructions, after all samples were collected. DNA was quantified using a Qubit 2.0 Fluorometer with a dsDNA Assay Kit (Thermo Fisher Scientific).

Metagenomic data processing

Illumina NovaSeq 6000 platform

The metagenomic paired-end sequencing was performed for whole EstMB cohort (N = 2,509) by Novogene Bioinformatics Technology Co., Ltd., using the Illumina NovaSeq6000 platform, resulting in 4.62 ± 0.44 Gb of data per sample (insert size, 350 bp; read length,

2 × 250 bp). A NEBNext Ultra DNA Library Prep Kit (NEB, USA) was used for generating sequencing libraries, according to the manufacturer's recommendations. Sequencing reads' quality control was performed using FastQC (v 0.12.1) (18). Read trimming and adapter removal was performed using fastp (v0.23.2) (19) with default parameters before the rarefaction/subsampling step for both cohorts. The host reads were removed using Bowtie2 (v 0.6.5) (20) against the GRCH build 38, patch release (20, 21).

MGISEQ-2000 platform

Metagenomic paired-end sequencing was performed using DNBSEQI technology, resulting in 22.47 ± 4.2 Gb of data per sample (insert size, 350 bp; read length, 2 × 250 bp). Library preparation and circularization of equimolarly pooled libraries were done with MGIEasy FS DNA Library Prep Set (MGI Tech Co, Ltd., Shenzhen, China) according to the standard protocol. The sequencing was performed with MGISEQ-2000 High-throughput Sequencing Set (FCL PE150) according to the manufacturer's instructions (MGI Tech, Shenzhen, China). Read quality control was performed using FastQC (v 0.12.1) (18). Read trimming and adapter removal were performed using fastp (v0.23.2) (19) with default parameters before the rarefaction/subsampling step for both cohorts. The hosts were removed using Bowtie2 (v 0.6.5) (20) against the GRCH build 38, patch release (21).

Bioinformatics and statistical analysis

Sample selection

For downstream analysis, we selected 1,729 samples sequenced with both platforms from the EstMB and EstMB-deep cohorts. To minimize bias, we first excluded sample pairs with fewer than 13 million or more than 20 million NovaSeq reads, as the number of NovaSeq reads was consistently lower than MGI (post QC). We then adjusted the number of MGI reads to match the NovaSeq read count within each sample pair using seqtk v1.3 (22). The final data set consisted of 1,351 samples with approximately 20 million reads per sample. For the separate intra-sequencing comparison, we used the 53 samples sequenced twice with the MGI platform.

Read quality comparison

The FastQC files were aggregated and analyzed using Python. The quality control was conducted to assess base sequence quality, adapter content, sequence length distribution, GC content, sequence duplication levels, and overrepresented sequences. Differences in GC content between sequencing platforms were evaluated using the Wilcoxon signed-rank test, and correlation was assessed using Pearson's correlation coefficient. Samples exceeding sequence duplication thresholds—indicative of PCR amplification bias, low library complexity, or potential contamination—were flagged. Overrepresented sequences were identified based on FastQC's criteria for sequence enrichment beyond expected distributions. For all downstream analyses, we included only sample pairs where both samples passed all quality criteria (N = 1,351).

Taxonomic profiling

Taxonomic profiling was done using Kraken2 v2.1.1 and Bracken v2.8 with a custom database (GTDB release 214) (23, 24). Kraken report files were processed using custom scripts (parse_kraken_NovaSeq.py, parse_kraken_mgi.py), retaining only taxa with relative abundance exceeding 0.01%. Relative abundance was defined based on the Kraken's "fraction_total_reads" values and prevalence was calculated as the fraction of samples in which the species was detected.

Statistical analysis

Differences in species counts between sample pairs were tested for normality using the Shapiro–Wilk test. As distributions deviated significantly from normality ($P \ll 0.05$), comparisons between runs and platforms were performed using the Wilcoxon signed-rank test. Effect sizes were calculated using Cohen's d.

Further analyses

Alpha and beta diversity measures (UniFrac and Procrustes analyses) were performed using QIIME 2 with default parameters (25). Functional predictions were made on a random subset of 700 samples with mi-faser with standard parameters (26). K-mers were calculated with jellyfish v.2.3.0 (27). Differential enrichment analysis was done with MaAsLin2 in its default mode (28). The association between centered log-ratio transformed species abundances and 33 common diseases (defined by ICD10 codes with cases in the EstMB cohort) was investigated using linear regression models. For each disease, the remaining samples served as controls. The models were adjusted for BMI, gender, and age, and the significance level was strictly controlled using a Bonferroni correction for the number of analyzed species.

Genome coverage

To evaluate sequencing coverage of key microbial species across samples, we selected four representative gut bacteria: *Akkermansia muciniphila*, *Bacteroides uniformis*, *Faecalibacterium prausnitzii*, and *Prevotella copri*. For each species, reference genomes were downloaded from the NCBI genome database (accessions: GCA_000020225.1, GCA_025147485.1, GCA_003324185.1, and GCA_025151535.1, respectively). Genome sequences in FASTA format were concatenated into a single reference file and indexed using Bowtie2 to create a searchable index (20).

To determine coverage, post-QC paired-end sequencing reads from each sample were individually aligned to the combined reference using Bowtie2. The alignment was converted to BAM format and sorted using SAMtools (29). Per-base genome coverage was then calculated with BEDTools' *genomecov* function, using the reference genome file to define genome sizes (30). This process was repeated for each sample to quantify species-level coverage separately across all conditions.

Genomic coverage comparisons between the two sequencing methodologies were performed in Integrative Genomics Viewer (31). For illustrative purposes, a representative genomic fragment was randomly selected, clearly demonstrating the observed differences in coverage uniformity.

ACKNOWLEDGMENTS

The authors express our gratitude to all individuals who made valuable contributions to the Estonian Microbiome cohort—Oliver Aasmets, Kertu Liis Krigul, Kreete Lüll, and Steven Smith. The authors thank the Estonian Biobank research team: Andres Metspalu, Lili Milani, Tõnu Esko, and Mait Metspalu for providing health data from Estonian Biobank participants. 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. Last but not least, we thank Łukasz Szydłowski for useful discussions during the early stages of this study and Rafał Mostowy for inviting Elin Org as a part of his EMBO Installation Grant, which established this project.

This work was funded by Estonian Research Council grant (PUT 1371, PRG1414 to E.O) and an EMBO Installation grant (No. 3573 to E.O.). EstMB sample collection was supported by Estonian Center of Genomics/Roadmap II project No 16-0125. This research was also conducted as a part of the NCN Sonata BIS grant number 2020/38/E/NZ2/00598. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grants no. PLG/2023/016234 and no. PLG/2024/017180. T.K. is funded by the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796.

AUTHOR AFFILIATIONS

[1]Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[2]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
[3]Sano Centre for Computational Medicine, Kraków, Poland

AUTHOR ORCIDs

Tomasz Kosciolek http://orcid.org/0000-0002-9915-7387
Elin Org http://orcid.org/0000-0001-8688-9717

FUNDING

Funder	Grant(s)	Author(s)
Estonian Research Competency Council	PUT 1371,PRG1414	Elin Org
Narodowe Centrum Nauki	2020/38/E/NZ2/00598	Pawel P. Łabaj
Ministerstwo Edukacji i Nauki	MEiN/2023/DIR/3796	Tomasz Kosciolek
Infrastruktura PL-Grid	PLG/2023/016234,PLG/2024/017180	Pawel P. Łabaj
Estonian Center of Genomics/Roadmap II	16-0125	Elin Org

AUTHOR CONTRIBUTIONS

Kinga Zielińska, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review and editing | Kateryna Pantiukh, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review and editing | Pawel P. Łabaj, Investigation, Supervision, Writing – review and editing | Tomasz Kosciolek, Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review and editing | Elin Org, Conceptualization, Funding acquisition, Project administration, Resources, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

The source code for the analyses is available at GitHub: https://github.com/Chartiza/2024_Illumina_vs_BGI

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental Information (mSystems01714-25-s0001.docx). Supplemental figures and text

Supplemental Tables (mSystems01714-25-s0002.xlsx). Tables S1–S5.

REFERENCES

1. Pinto Y, Bhatt AS. 2024. Sequencing-based analysis of microbiomes. Nat Rev Genet 25:829–845. https://doi.org/10.1038/s41576-024-00746-6
2. Gilbert JA, Azad MB, Bäckhed F, Blaser MJ, Byndloss M, Chiu CY, Chu H, Dugas LR, Elinav E, Gibbons SM, Gilbert KE, Henn MR, Ishaq SL, Ley RE, Lynch SV, Segal E, Spector TD, Strandwitz P, Suez J, Tropini C, Whiteson K, Knight R. 2025. Clinical translation of microbiome research. Nat Med 31:1099–1113. https://doi.org/10.1038/s41591-025-03615-9
3. Forry SP, Servetas SL, Kralj JG, Soh K, Hadjithomas M, Cano R, Carlin M, Amorim MG de, Auch B, Bakker MG, et al. 2024. Variability and bias in microbiome metagenomic sequencing: an interlaboratory study comparing experimental protocols. Sci Rep 14:9785. https://doi.org/10.1038/s41598-024-57981-4
4. Marotz C, Cavagnero KJ, Song SJ, McDonald D, Wandro S, Humphrey G, Bryant M, Ackermann G, Diaz E, Knight R. 2021. Evaluation of the effect of storage methods on fecal, saliva, and skin microbiome composition. mSystems 6:e01329-20. https://doi.org/10.1128/mSystems.01329-20
5. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. https://doi.org/10.1038/nature07517
6. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H, Qu S, Mei X, Chen H, Yu T, Sun N, Rao J, Wang J, Zhang W, Chen Y, Liao S, Jiang H, Liu X, Yang Z, Mu F, Gao S. 2017. A reference human genome dataset of the BGISEQ-500 sequencer. Gigascience 6:1–9. https://doi.org/10.1093/gigascience/gix024
7. Jeon SA, Park JL, Park S-J, Kim JH, Goh S-H, Han J-Y, Kim S-Y. 2021. Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. Genes Genomics 43:713–724. https://doi.org/10.1007/s13258-021-01096-x
8. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351. https://doi.org/10.1038/nrg.2016.49
9. Korostin D, Kulemin N, Naumov V, Belova V, Kwon D, Gorbachev A. 2020. Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. PLoS One 15:e0230301. https://doi.org/10.1371/journal.pone.0230301
10. Zhu K, Du P, Xiong J, Ren X, Sun C, Tao Y, Ding Y, Xu Y, Meng H, Wang C-C, Wen S-Q. 2021. Comparative performance of the MGISEQ-2000 and illumina X-ten sequencing platforms for paleogenomics. Front Genet 12:745508. https://doi.org/10.3389/fgene.2021.745508
11. Hu T, Chen J, Lin X, He W, Liang H, Wang M, Li W, Wu Z, Han M, Jin X, Kristiansen K, Xiao L, Zou Y. 2024. Comparison of the DNBSEQ platform and Illumina HiSeq 2000 for bacterial genome assembly. Sci Rep 14:1292. https://doi.org/10.1038/s41598-024-51725-0
12. Aasmets O, Krigul KL, Lüll K, Metspalu A, 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
13. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, Rasmussen M, Zervas A, Hansen LH. 2020. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. Gigascience 9:giaa008. https://doi.org/10.1093/gigascience/giaa008
14. Kumar A, Robertsen EM, Willassen NP, Fu J, Hjerde E. 2023. Comparative analysis of HiSeq3000 and BGISEQ-500 sequencing platform with shotgun metagenomic sequencing data. Genomics Inform 21:e49. https://doi.org/10.5808/gi.23072
15. Foox J, Tighe SW, Nicolet CM, Zook JM, Byrska-Bishop M, Clarke WE, Khayat MM, Mahmoud M, Laaguiby PK, Herbert ZT, et al. 2021. Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. Nat Biotechnol 39:1129–1140. https://doi.org/10.1038/s41587-021-01049-5
16. Liu J, Coker MO, Osazuwa-Peters N, Peter O, Idemudia NL, Schlecht NF, Obuekwe O, Eki-Udoko FE, Bromberg Y. 2025. Whole metagenome sequencing: not deep enough for complete microbial function recovery. bioRxiv. https://doi.org/10.1101/2025.11.04.685665
17. 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
18. FastQC. 2015
19. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560
20. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
21. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR, 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393
22. Shen W, Le S, Li Y, 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
23. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46
24. Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci 3:e104. https://doi.org/10.7717/peerj-cs.104
25. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9
26. Zhu C, Miller M, Marpaka S, Vaysberg P, Rühlemann MC, Wu G, Heinsen F-A, Tempel M, Zhao L, Lieb W, Franke A, Bromberg Y. 2018. Functional sequencing read annotation for high precision microbiome analysis. Nucleic Acids Res 46:e23. https://doi.org/10.1093/nar/gkx1209
27. Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27:764–770. https://doi.org/10.1093/bioinformatics/btr011
28. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, Tickle TL, Weingart G, Ren B, Schwager EH, Chatterjee S, Thompson KN, Wilkinson JE, Subramanian A, Lu Y, Waldron L, Paulson JN, Franzosa EA, Bravo HC, Huttenhower C. 2021. Multivariable association discovery in population-scale meta-omics studies. PLoS Comput Biol 17:e1009442. https://doi.org/10.1371/journal.pcbi.1009442
29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
30. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033
31. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26. https://doi.org/10.1038/nbt.1754

II

Pantiukh, Kateryna et al.
“Metagenome-assembled genomes from
a population-based cohort uncover new g
ut species and within-species diversity,
revealing prevalent disease associations.”
mSystems, e0011426. 16 Mar. 2026,
doi:10.1128/msystems.00114-26

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

| Open Peer Review | Host-Microbial Interactions | Research Article

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]

AUTHOR AFFILIATION See affiliation list on p. 14.

ABSTRACT Metagenomic profiling has advanced the 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, within-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 collection to create an expanded reference (GUTrep), enabling profiling of 2,509 EstMB individuals and testing associations with 33 prevalent diseases. Of the 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 within-species analyses. Based on normalized GUN, we prioritized *Odoribacter splanchnicus,* a prevalent species with the lowest within-species heterogeneity, yielding sufficient power for a within-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, 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 at the within-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 within-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 within-species clusters as recovered genomes, making association studies at the within-species level essentially intractable. To address this, we introduce the genome unit number (GUN), a scalable metric for quantifying within-species structure. Using GUN, we demonstrate that only species with limited within-species diversity, such as *Odoribacter splanchnicus*, currently allow for robust within-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 within-species analyses in highly diverse species.

Editor Alejandro Reyes Munoz, Universidad de Los Andes, Bogota, Colombia

Peer Reviewer Luis Miguel Rodriguez-Rojas, Universität Innsbruck, Innsbruck, Austria

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

The authors declare no conflict of interest.

See the funding table on p. 15.

Received 31 January 2026
Accepted 10 February 2026
Published 16 March 2026

Copyright © 2026 Pantiukh et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

KEYWORDS gut microbiome, metagenome-assembled genomes, genome unit number, population microbiome, metagenomics, metagenome-wide association study, within-species diversity

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 toward 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 MAG 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 data sets 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 Estonian Microbiome (EstMB) Cohort from the Estonian Biobank (EstBB) (10). Analyses of these data sets have demonstrated that gut microbiome composition is associated with a range of environmental and lifestyle factors, particularly diet and medication use (10, 11), and that variation in the microbiome is associated with several diseases, such as cardiovascular diseases (12, 13), mental health disorders (14), and cancers (15, 16). Furthermore, emerging evidence suggests that the gut microbiome has predictive power, as demonstrated in the context of incident heart failure (17). However, many of these studies still rely solely on reference databases. These databases may lack representatives for many uncultured or underrepresented population-specific microbial species, leading to incomplete or biased interpretations.

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

RESULTS

Study design and cohort overview

The study aimed to first recover MAGs from a deeply sequenced Estonian gut microbiome cohort (EstMB-deep, N = 1,878) and expand the reference database by combining these newly assembled genomes with the existing public Unified Human

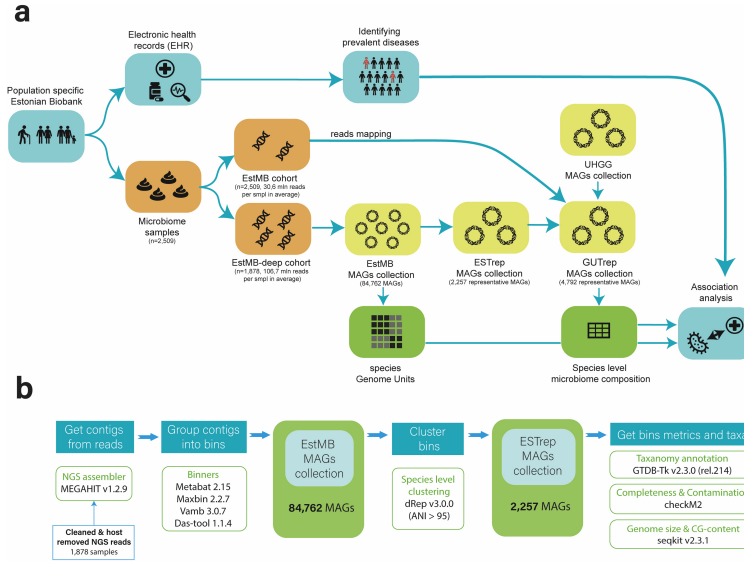


FIG 1 Study overview and cohort description. (a) Study workflow overview. (b) Overview of the MAG recovery pipeline.

Gastrointestinal Genome collection and demonstrate the added value of genome-resolved, within-species analysis in identifying disease associations (3) (Fig. 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 (Fig. 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 23–89 (Fig. 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; Fig. S2b). A detailed description of the EstMB, including omics and phenotypic data, is provided in Aasmets et al. (10).

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

To characterize population-specific microbes and expand publicly available human gut microbiome databases with microbial genomes from the Estonian population, we performed *de novo* MAG reconstruction from all 1,878 samples in the EstMB-deep cohort. The MAG reconstruction pipeline is summarized in Fig. 1b. We successfully reconstructed 84,762 MAGs from EstMB-deep, with an average of 45.13 MAGs per sample. Among these, 42,049 (49.61%) were almost complete MAGs, i.e., MAGs with completeness > 90% and contamination < 5% according to CheckM. To describe the Estonian population species pool, we clustered all MAGs with dRep using a 95.0% average nucleotide identity (ANI) threshold, ensuring that the final clusters represent distinct species.

The species-level clustering procedure yielded 2,257 clusters (Table S1). For each cluster, the representative MAG was selected based on genome completeness, minimal contamination, strain heterogeneity, and N_{50} (a parameter reflecting assembly fragmentation level). We refer to these 2,257 species representative MAGs as the "ESTrep" collection. The majority of ESTrep MAGs from the ESTrep collection (72.97%, *n*

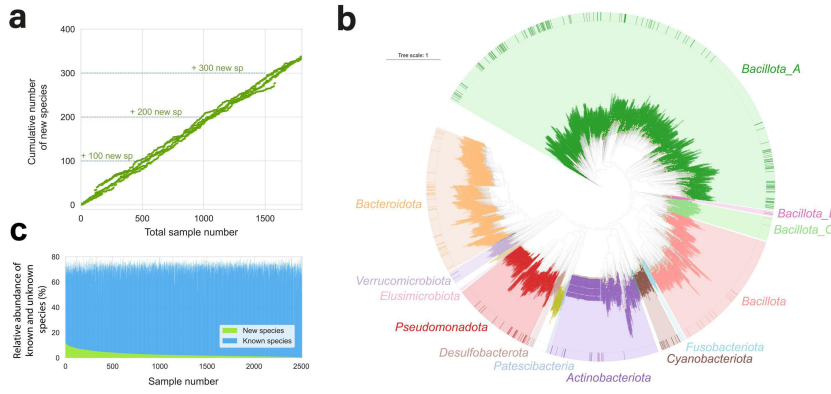


FIG 2 Overview of novel species from the EstMB MAG collection. (a) The relationship between the number of samples analyzed and the cumulative number of novel species identified. (b) Phylogenetic tree of the ESTrep species. The inner circle displays a phylogenetic tree of species, with branches colored by phylum (according to the Genome Taxonomy Database [GTDB-Tk version 2.3.0]), the outer ring highlights novel species assembled in this study. (c) Relative abundances of known and novel species.

= 1,647 MAGs) were almost complete MAGs (MAGs >90% complete and <5% contaminated). Under the MIMAG standard (18), 475 representative MAGs (21.05%) qualified as high-quality, exhibiting >90% completeness, <5% contamination, and the presence of the 5S, 16S, and 23S rRNA genes together with at least 18 tRNAs. The remaining genomes included 1,714 MAGs (75.94%) of medium-quality (MQ) and 68 MAGs (3.01%) of low-quality (LQ) (Fig. S3).

MAG assembly remains essential for detecting novel population-specific species

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

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

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

As the success of metagenome assembly and genome reconstruction depends on multiple technical and analytical factors, we did not expect to recover all microbial

genomes present in the gut. Therefore, we constructed an integrated species-level reference by combining newly reconstructed MAGs from the Estonian population with publicly available human gut-associated species. This integrated reference, called the GUTrep collection (Fig. 1a), was generated by deduplicating the ESTrep MAG collection and UHGG MAG 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 data set. 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 data set (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 (Fig. 3a). The most prevalent (>95%) species detected by read mapping were all well-known gut microbes: *Phocaeicola dorei*, *Bacteroides* species (*Bacteroides uniformis*, *Bacteroides xylanisolvens*, and *Bacteroides ovatus*), *Faecalibacterium prausnitzii*, and *Odoribacter splanchnicus*. *P. dorei* and *B. uniformis* were also among the most abundant species, in addition to *Prevotella copri* (>2% on average) (Fig. S4). We observe that samples with more species detected by mapping also tended to have more MAGs recovered (Fig. S5). However, the number of assembled genomes per species did not clearly correlate with species prevalence or mean abundance (Fig. 3b; Table S2). Moreover, the difference between these values can range from minimal to substantial. For example, despite being one of the most prevalent species, *Bacteroides xylanisolvens* was detected in 97.13% of samples, with a mean relative abundance of 0.39%. However, 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 (Fig. 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 (Fig. S6), illustrating the challenges of comprehensive genome reconstruction.

Newly assembled species provide valuable input for association studies

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

We identified 105 significant associations (Bonferroni-adjusted $P < 2.71 \times 10^{-5}$) between 96 bacterial species and 25 diseases (Tables S4 and S5). Notably, newly assembled species were associated with 8 out of the 33 diseases, including asthma, chronic ischemic heart disease, chronic rhinitis, nasopharyngitis and pharyngitis, female infertility, heart failure, hemorrhoids, iron deficiency anemia, and vitamin D deficiency. For example, one of the strongest associations was observed for chronic ischemic heart disease, involving a newly assembled species from the *Nanosynbacter* genus (species ID: H2144_Nanosynbacter_undS, adjusted $P = 3.13 \times 10^{-6}$) (Fig. 3d; Table S4). These findings emphasize the importance of population-specific reference databases for detecting disease-associated microbiome changes. However, further studies are needed to confirm whether these associations generalize beyond the Estonian cohort.

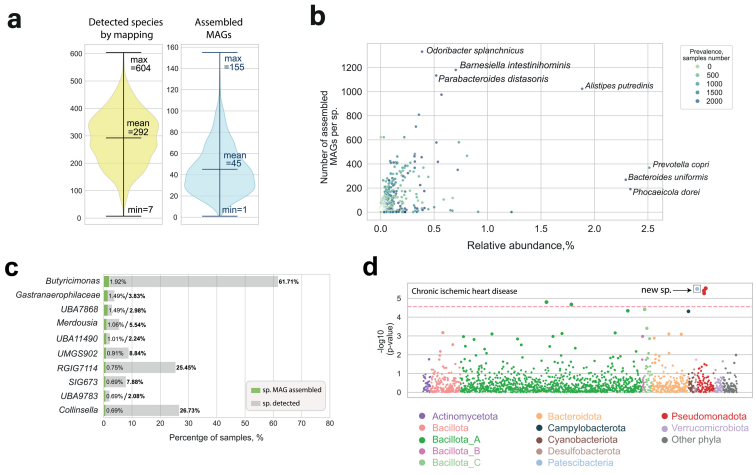


FIG 3 Assembly–mapping comparison and microbiome-wide association study associations. (a) Average number of species detected by read mapping (yellow) versus number of recovered MAGs per sample (blue). (b) Relationship between species prevalence, mean relative abundance, and number of assembled MAGs per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

MAG data enable within-species genetic diversity analysis across species

Most large-scale microbiome association studies are conducted at the species level, although strain-level analysis is often recommended for understanding the functional insights (4, 23, 24). However, strain definitions vary substantially across studies, and many species do not exhibit genomic clustering patterns that provide sufficient sample sizes for robust within-species-level association testing. Metagenome assembly enables high-resolution reconstruction of species-level genomic diversity and allows the identification of taxa for which such analyses are feasible, thereby broadening the range of detectable biological and clinical associations. To capture this variation systematically and select species suitable for within-species-level MWAS, we introduce the genome unit number metric. GUN quantifies genomic diversity within a species by counting the number of highly similar genome units per species (ANI > 99%). We further define the normalized GUN (nGUN), which scales the GUN by the total number of MAGs recovered for the analyzed species, allowing comparison across species with differing MAG counts.

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

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

The nGUN values varied widely, ranging from 0.4 to 100 (Fig. 4a; Table S6), indicating substantial differences in species population genetic diversity. *Odoribacter splanchnicus* exhibited the lowest nGUN (0.4), with one genome unit per ~250 MAGs, reflecting low diversity despite high prevalence. In contrast, *Prevotella copri* had one of the highest nGUN (94.0), consistent with its well-documented heterogeneity, where nearly every MAG represents a unique genome unit, making it difficult to conduct within-species

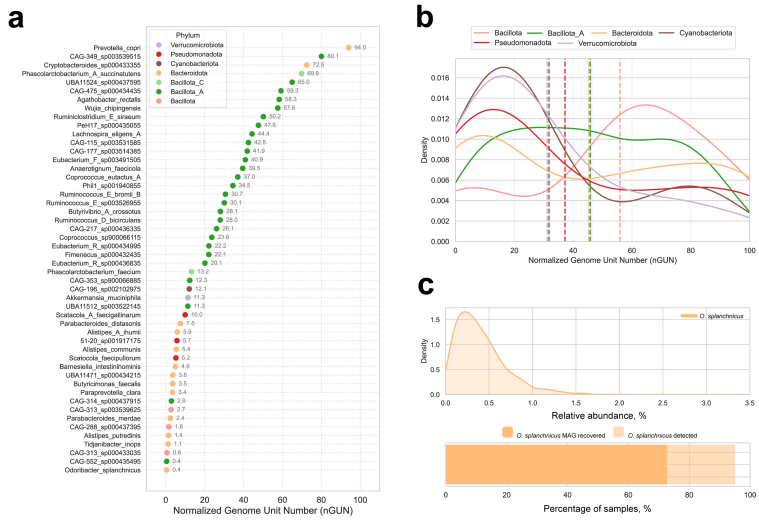


FIG 4 Within-species diversity and within-species analysis of *Odoribacter splanchnicus*. (a) Normalized genome unit number values for the top 50 species with the highest number of MAGs. (b) Distribution of nGUN values across major gut bacteria phyla. (c) *Odoribacter splanchnicus* relative abundance, number of recovered MAGs, and prevalence across samples.

association analysis in the population. Interestingly, species from Alistipes_A genus appeared in both the lowest and highest nGUN groups.

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

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

In order to demonstrate the value of nGUN-guided within-species metagenome-wide association analisys 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%; Fig. 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 = 974 samples, original cluster ID: 1_2.3.4.6.9) and genome unit N2 (GU-N2; n = 335 samples, original cluster ID: 1_1) (Fig. 5a).

Logistic regression models adjusted for BMI, gender, and age were used to assess the association between the presence or absence of *O. splanchnicus* GU-N1 and GU-N2 and the same 33 diseases previously analyzed at the species-level MWAS. Our analysis identified a significant association between the presence of GU-N1 and two different diseases—gastritis and duodenitis, and hypertensive heart disease (Fig. 5b). The odds ratio for GU-N1 was less than 1 in both diseases (gastritis and duodenitis OR = 0.56, hypertensive heart disease OR = 0.63), indicating that its presence is associated with a

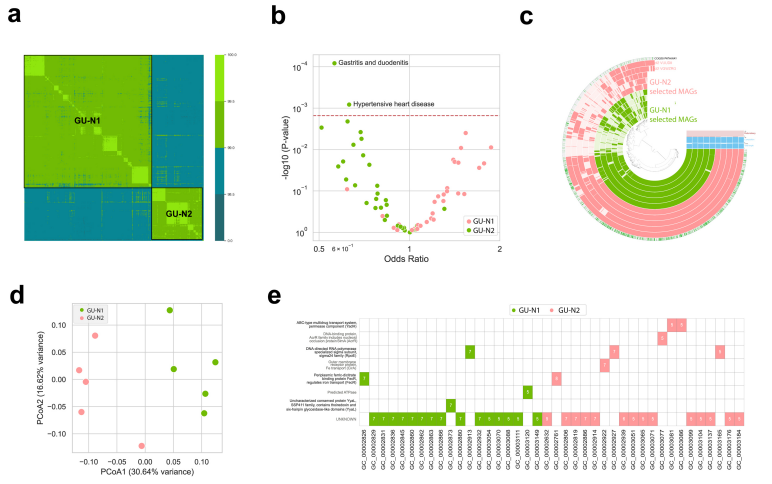


FIG 5 Genomic structure and disease-associated variation across two major *O. splanchnicus* genome units. (a) Heatmap of average nucleotide identity values among *O. splanchnicus* MAGs, revealing two distinct genome units. (b) Volcano plot of associations between the two major *O. splanchnicus* genome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

reduced likelihood of having the disease. Notably, these associations were not detected at the species level, highlighting the added resolution of within-species-level analysis.

To explore functional differences, we performed a pangenome analysis of GU-N1 and GU-N2. We carried out principal coordinate analysis of predicted gene cluster presence/absence, which showed clear separation between the genome units (Fig. 5c). We identified that the two genome units formed distinct clusters, indicating clear genomic differentiation based on gene content (Fig. 5d), with 40 gene clusters unique to one of the two (Fig. 5e; Table S7). While most encoded hypothetical or uncharacterized proteins, some were annotated with putative functions based on the Clusters of Orthologous Genes (COG20) (25). GU-N2 harbored a broader repertoire of genes associated with stress response, iron acquisition, and antimicrobial resistance— traits consistent with enhanced survival in inflammatory gastrointestinal environments. These included elevated copy numbers of the extracytoplasmic stress sigma factor RpoE (σ^E), iron transport components FecR and CirA, and multidrug resistance elements such as AcrR and an ABC-type efflux pump (YadH). In contrast, GU-N1 was enriched for redox maintenance proteins such as YyaL/DsbD, suggesting a distinct strategy centered on oxidative stress mitigation.

DISCUSSION

Our study presents a scalable, genome-resolution framework for population-scale microbiome analysis, enabling improved species and within-species characterization and discovery of disease associations. By expanding the gut microbial reference database with thousands of metagenome-assembled genomes, including novel bacterial species, as well as within-species genome units of known taxa, we address a major limitation in current reference data sets, which often underrepresent global microbiome diversity. We demonstrate that genome-resolution microbiome analysis, coupled with population-specific MAG catalogs, enables more comprehensive species- and within-species-level

association studies. Furthermore, we introduce the genome unit number, a novel quantitative metric of within-species genetic diversity, which we applied across 378 gut species to guide candidate selection for within-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 (26, 27). Alternatively, MAGs offer a culture-independent, reference-free approach to recover community structure (28). 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. This highlights that even abundant taxa are not always fully recoverable by MAG 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 (29–31). 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 (32) and underscore the need for a hybrid strategy combining MAG assembly and high-resolution read mapping against population-specific reference. Our GUTrep database exemplifies such a strategy, integrating local MAGs with the Unified Human Gastrointestinal Genome collection to improve reference coverage. Notably, 31% of dereplicated GUTrep species originated from our Estonian-specific MAGs, illustrating the added value of local assembly efforts.

A common argument against investing in resource-intensive *de novo* assembly is that the rapidly growing and regularly updated public gut genome catalogs increasingly capture known microbial diversity, suggesting that most gut species will soon be represented, and further assembly may become redundant. However, our findings challenge this assumption and support continued *de novo* assembly in new population studies. Early efforts reported high proportions of novel taxa, with 77% of MAGs classified as novel in Pasolli et al. (33) and 66% in Almeida et al. (22). More recent studies, such as Leviatan et al. (34), still report 310 novel species out of 3,594 assembled (8.6%). In our cohort, we recovered 353 novel species from 2,257 MAGs (15.6%), with ~102 additional novel species per 500 samples, and no indication of a discovery plateau. Moreover, we confirm a previously reported finding that many novel species assembled from a few samples were nonetheless widespread by mapping, suggesting that assembly remains essential even in well-characterized industrialized populations (32). These results reinforce the idea that local assembly efforts complement global references and remain critical for uncovering the full spectrum of microbial diversity.

Another advantage of *de novo* assembly is its ability to uncover within-species-level variation (4, 5). Strains of the same species can differ significantly in function and disease associations (35). As a classic example, well-known gut microbe *Escherichia coli* species includes strains that can be pathogenic (e.g., enterohemorrhagic O157:H7), probiotic (Nissle 1917), or commensal (K-12), and this demonstrates how it can be insufficient to study the microbe at the species level (36). While strain-level taxonomic profilers such as MetaPhlAn 4.0 (37) offer efficient resolution, they lack the genomic content necessary for detailed functional analyses, as they are usually based on marker gene analysis only. Other powerful tools, such as inStrain, operate at the whole-genome level and can report presence of different strains even among strains coexisting within the same sample (38).

In contrast, reconstructing MAGs directly from samples linked to host metadata allows for in-depth investigation of within-species genomic variation in relation to

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

Our normalized genome unit number helps to assess whether within-species analysis is feasible by quantifying within-species diversity. High nGUN value of the species might indicate that each individual harbors a unique genome unit, complicating population-level associations. In this study, we analyzed within-species diversity across 378 gut species, expanding upon previous work that showed a strain richness of 92 gut species based on pure isolates from at least three different individuals (39). Our results significantly expand on this by including a broader range of species, each represented by more than 10 MAGs. Consistent with earlier findings, we observe substantial variability in within-species diversity across species (39). However, our data provide a more detailed and comprehensive picture due to the larger number of species. Among the 14 overlapping species between the two studies, some show similar patterns of genomic diversity—for example, *Odoribacter splanchnicus* and *Barnesiella intestinihominis* exhibit consistently low diversity, while others like *Bifidobacterium longum* remain highly diverse. Other overlapping examples, such as *Fusicatenibacter saccharivorans* and *Coprococcus eutactus*, display divergent diversity estimates, likely reflecting methodological differences (culture isolate versus metagenome-based) and highlighting the need for further comparative research. Highly diverse species such as *Prevotella copri,* well known from other studies (40, 41), are absent from the Chen-Liaw data set, but prominent in ours. Our larger data set also allowed the investigation of phylum-level patterns, suggesting that the common phyla, such as *Bacteroidota and Bacillota_A* species, tend to have higher nGUN values, while *Verrucomicrobiota, Cyanobacteriota,* and *Pseudomonadota* exhibit lower diversity. These phylum-level differences in within-species diversity suggest possible evolutionary or ecological constraints. However, further studies are needed, particularly for the less common phyla, to validate and understand the underlying mechanisms.

Understanding microbiome diversity at both species and within-species levels enhances resolution in metagenome-wide association studies. At the species level, we identified 96 bacterial species significantly associated with 25 common diseases, of which 8 diseases involved previously uncharacterized species, highlighting the limits of relying solely on global references. These differences may reflect population-specific variation driven by local differences in diet, genetics, and lifestyle (11). At the within-species level, we identified associations that were undetectable at the species level, such as the negative association of the *Odoribacter splanchnicus* genome unit GU-N1 with gastritis, duodenitis, and hypertensive heart disease. Previous research has shown that the abundance of the genus *Odoribacter* is negatively correlated with systolic blood pressure in overweight and obese pregnant women, suggesting that short-chain fatty acid-producing taxa may influence host blood pressure (42). Comparative genomic analysis of the MAGs from two *O. splanchnicus* genome units helped us to identify a set of gene clusters that differed between the two groups. These genomic features suggest that GU-N1 is functionally better adapted to conditions characteristic of gastritis and duodenitis, such as oxidative stress, nutrient limitation, and host antimicrobial pressure. Together, our findings highlight the value of genome-resolved metagenomic approaches in revealing disease-relevant microbial functions that would remain hidden in broader taxonomic analysis. While our study focused on a single, ethnically homogeneous Northern European population, several key findings, such as the detection of widespread yet previously uncharacterized species and species structures within common gut taxa, are likely to extend beyond the Estonian population. However, microbiome composition

is influenced by genetic, dietary, and environmental factors that vary between populations. Future studies in diverse cohorts will be essential to evaluate the generalizability of our results and validate the species- and within-species disease associations uncovered here. The genome-resolved analytical framework we present is scalable and readily applicable to other population-scale microbiome data sets, enabling cross-cohort comparison and discovery.

Our study also has some limitations that should be considered. First, the reliance on short-read sequencing may reduce assembly contiguity and within-species resolution compared to long-read approaches (43). Although long-read technologies offer higher genomic completeness, their current cost limits their use in large-scale population studies. A practical compromise could involve using short-read sequencing for most samples and applying long-read sequencing to key or novel taxa. Second, the observed associations are correlative, and further validation in longitudinal and experimental studies is needed to assess causality. Third, our association study on within-species level is focused on one species due to sample representation and analytical feasibility, and broader application across species remains a key next step. Finally, the genome unit number is influenced by multiple biological and technical factors that may bias its ability to accurately reflect within-species microdiversity. These include pangenome size, genome length, the propensity for horizontal gene transfer, and the level of homologous recombination. In addition, when a species is present within a sample as several highly similar genomic variants that may belong to different genome units, metagenomic assembly becomes more challenging. In such cases, MAGs may fail to be recovered or may represent composite genomes spanning multiple genome variants, which can lead to underestimation or conflation of true within-species diversity. Moreover, while functional differences between genome units of one species were identified, interpretation was limited by the prevalence of unannotated genes. Future improvements in genome annotation and cross-cohort replication will be essential to build on these findings. Despite these challenges, our findings demonstrate the value of population-scale metagenomics in uncovering novel microbial diversity and functional signatures relevant to human health.

Conclusion

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

MATERIALS AND METHODS

Estonian Microbiome Cohort description

The Estonian Microbiome Cohort was established in 2017, when stool, oral, and blood samples were collected from 2,509 EstBB participants (10). The EstBB is a volunteer-based population cohort initiated in 1999 that currently includes over 212,000 adults of European ancestry (≥18 years old) across Estonia (44). Extensive information is available for the EstMB participants, including data from self-reported questionnaires and EHRs (completed by medical professionals) covering diseases, medication use, and medical procedures both before and after sample collection. In addition to the questionnaire and EHR data, the participants' anthropometric measurements (e.g., height, weight, blood pressure, and waist and hip circumferences) were taken during a pre-registered visit upon delivering the stool sample. The Estonian Microbiome Deep cohort includes a subset of stool samples from the EstMB cohort that have been resequenced with over three times deeper coverage (N = 1,878).

Microbiome sample collection and DNA extraction

The participants collected a fresh stool sample immediately after defecation with a sterile Pasteur pipette and placed it inside a polypropylene conical 15 mL tube. The participants were instructed to time their sample collection as close as possible to the visiting time in the study center. The samples were stored at −80°C until DNA extraction. The median time between sampling and arrival at the freezer in the core facility was 3 h 25 min (mean 4 h 34 min), and the transport time was not significantly associated with alpha diversity (Spearman correlation, *P*-value 0.949 for observed richness and 0.464 for Shannon index) or beta diversity (Bray–Curtis, *P*-value 0.061, R^2 = 0.0005). Microbial DNA extraction was performed after all samples were collected using a QIAamp DNA Stool Mini Kit (Qiagen, Germany). For the extraction, approximately 200 mg of stool was used as a starting material for the DNA extraction kit, according to the manufacturer's instructions. DNA was quantified from all samples using a Qubit 2.0 Fluorometer with a dsDNA Assay Kit (Thermo Fisher Scientific).

Shotgun metagenomic sequencing

Sequencing for the main EstMB cohort was done using shotgun metagenomic paired-end sequencing on the Illumina NovaSeq 6000 platform and described in detail in reference 10. The EstMB-deep cohort samples were selected based on DNA quality and resequenced at higher depth using paired-end shotgun metagenomic sequencing on the MGISEQ-2000 platform. Sequencing reads' quality control (QC) was performed using FastQC (version 0.12.1) (45), and human reads were filtered using Bowtie2 (version 0.6.5) (46) against the GRCh38.p14 human genome reference. While following the QC, the EstMB cohort had an average of 30.63 ± 3.12 million reads per sample, whereas the EstMB-deep cohort resulted in 106.70 ± 42.1 million reads per sample, indicating over three times deeper sequencing coverage. Sequencing depth was evaluated with Nonpareil version 3.5.5 (47).

EstMB MAG metagenome assembly and binning

The EstMB MAG collection refers to all MAGs recovered from the EstMB-deep cohort, which comprises 1,878 samples sequenced at deep coverage. Reads were assembled into contigs with MEGAHIT (version 1.2.9) (48). Binning was performed separately for each sample from the EstMB-deep cohort. Contigs were binned using binners: MetaBAT (version 2.15) (49), MaxBin (version 2.2.7) (50), and VAMB (version 3.0.7) (51), with further refining with DAS Tool (version 1.1.4) (52). MAGs resulting from this process form the EstMB MAGs collection. MAG quality, including completeness and contamination, was estimated using CheckM (version 2.3.1) (53).

ESTrep MAG collection

ESTrep MAG collection refers to representative MAGs from the EstMB MAG collection described earlier. Representative MAGs were selected from the EstMB MAG collection by clustering MAGs from the EstMB MAG collection on the species level (ANI index = 95) with dRep (54). Taxonomy of all representative MAGs was assigned using GTDB-Tk (version 2.3.0) (19), a software toolkit for assigning objective taxonomic classifications to bacterial and archaeal genomes based on the Genome Taxonomy Database (GTDB) (19, 55). If a MAG could not be taxonomically classified at the species level or higher using GTDB-Tk, this indicates that the genome does not closely match any existing entries in the GTDB reference database. Therefore, it was treated as a novel species. This criterion is widely used in studies involving MAG assembly (20, 21). MAG completeness and contamination were estimated using CheckM (version 2.3.1) (53). Ribosomal RNA genes were identified with Barrnap version 0.8 (56), and tRNA genes were predicted using tRNAscan-SE version 2.0.0 (57).

MAGs were classified into three quality tiers. High-quality MAGs were defined as those with >90% completeness and <5% contamination and by the presence of ≥18 tRNA genes and a full complement of rRNA genes (5S, 16S, and 23S) (18). Medium-quality MAGs were defined as those with >50% completeness and <10% contamination. MAGs not meeting HQ or MQ thresholds were classified as LQ. Assembly statistics, including total assembly size, number of contigs, N_{50}, and GC content, were calculated using SeqKit (version 2.3.1) (58).

Population-based reference GUTrep MAG collection

The GUTrep MAG collection is a non-redundant set of representative MAGs, created by combining MAGs from the current study (ESTrep MAGs) with those from the Unified Human Gastrointestinal Genome collection (22). This integrated reference includes both population-specific taxa identified in our cohort and globally distributed species that, while detected in our samples, could not be completely assembled but are present in public databases. To remove redundancy, MAGs from both collections were clustered at the species level using an average nucleotide identity threshold of 95% with dRep (54). For each species cluster containing MAGs from both sources, the higher-quality MAG—based on completeness, contamination, and assembly statistics—was retained as the representative.

Species relative abundance and prevalence estimation

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

Species-level association study

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

Genome unit number estimation

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

The corresponding formula is shown below:

$$\mathrm{nGUN} = \frac{\text{species GUN}}{\text{total number of MAGs}} \times 100\%$$

Within-species association study

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

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

ACKNOWLEDGMENTS

We express our gratitude to all individuals who made valuable contributions to the Estonian Microbiome cohort, as well as to those who developed the software and databases utilized in this study. We thank Mait Metspalu, Andres Metspalu, Lili Milani, and Tõnu Esko from the Estonian Biobank research team for the Estonian Biobank health data collection. Data analysis was carried out in part in the High-Performance Computing Centre of the University of Tartu, and we thank the HPC Support Team of the Institute of Computer Science at the University of Tartu for delivering exceptional service and assistance in installing the necessary programs on the cluster. This work was written at writing retreats and writing days organized by the Institute of Genomics, University of Tartu. We also thank Million Metagenomes of Human Project (MMHP) for providing the sequencing for the EstMB-deep cohort.

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.

K.P., K.L.K., and E.O. conceptualized the study; K.P. and O.A. designed the methodology; K.P. analyzed the data, visualized the study, and prepared the first draft of the manuscript, which all authors reviewed and edited. 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.

AUTHOR AFFILIATION

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

AUTHOR ORCIDs

Kateryna Pantiukh http://orcid.org/0000-0002-2595-0673
Kertu Liis Krigul http://orcid.org/0000-0002-4195-7357
Oliver Aasmets https://orcid.org/0009-0001-9872-6031

Elin Org http://orcid.org/0000-0001-8688-9717

FUNDING

Funder	Grant(s)	Author(s)
Estonian Research Council Grant	PRG1414	Elin Org
EMBO Installation grant	3573	Elin Org
Estonian Center of Genomics/Roadmap II	16-0125	Elin Org

AUTHOR CONTRIBUTIONS

Kateryna Pantiukh, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft | Kertu Liis Krigul, Conceptualization, Validation, Writing – review and editing | Oliver Aasmets, Methodology, Writing – review and editing | Elin Org, Conceptualization, Supervision, Writing – review and editing

DATA 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.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental Figures (mSystems00114-26-s0001.docx). Fig. S1 to S6.
Supplemental Tables (mSystems00114-26-s0002.xlsx). Tables S1 to S7.

Open Peer Review

PEER REVIEW HISTORY (review-history.pdf). An accounting of the reviewer comments and feedback.

REFERENCES

1. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, Dunn M, Mkandawire TT, Zhu A, Shao Y, Pike LJ, Louie T, Browne HP, Mitchell AL, Neville BA, Finn RD, Lawley TD. 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 AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. 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 6:101128msystems0088121. https://doi.org/10.1128/msystems.00881-21
5. Chen L, Wang D, Garmaeva S, Kurilshikov A, Vich Vila A, Gacesa R, Sinha T, Lifelines Cohort Study, Segal E, Weersma RK, Wijmenga C, Zhernakova A, Fu J. 2021. The long-term genetic stability and individual specificity of the human gut microbiome. Cell 184:2302–2315. https://doi.org/10.1016/j.cell.2021.03.024
6. Gacesa R, Kurilshikov A, Vich Vila A, Sinha T, Klaassen MAY, Bolte LA, 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, Wolf BC, Lotan-Pompan M, Levi K, Cohen

O, Saranga H, Weinberger A, Segal E. 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 31:3191–3203. https://doi.org/10.1038/s41591-025-03790-9
9. Salosensaari A, Laitinen V, Havulinna AS, Meric G, Cheng S, Perola M, Valsta L, Alfthan G, Inouye M, Watrous JD, Long T, Salido RA, Sanders K, Brennan C, Humphrey GC, Sanders JG, Jain M, Jousilahti P, Salomaa V, Knight R, Lahti L, Niiranen T. 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 KL, Lüll K, Metspalu A, 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 PI, Godneva A, Kalka IN, 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 AS, Cheng S, Inouye M, Jain M, Salido RA, Sanders K, Brennan C, Humphrey GC, Sanders JG, Vartiainen E, Laatikainen T, Jousilahti P, Salomaa V, Knight R, Lahti L, Niiranen TJ. 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 KF, Hammar U, Nguyen D, Nielsen N, Eklund AC, Varotsis G, Holm JB, Nielsen HB, Lind L, Bergström G, Smith JG, Engström G, Ärnlöv J, Sundström J, Orho-Melander M, Fall T. 2025. The association between the gut microbiome and 24-h blood pressure measurements in the SCAPIS study. Commun Med (London) 5:276. https://doi.org/10.1038/s43856-025-00980-x
14. Brushett S, Gacesa R, Vich Vila A, Brandao Gois MF, Andreu-Sánchez S, Swarte JC, Klaassen MAY, Collij V, Sinha T, Bolte LA, Wu J, Swertz M, de Kroon MLA, Reijneveld SA, Wijmenga C, Weersma RK, Fu J, van Loo HM, Kurilshikov A, Zhernakova A. 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 TSB, Molina-Montes E, Rodríguez-Perales S, Wirbel J, Maistrenko OM, Akanni WA, Alashkar Alhamwe B, Alves RJ, 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 PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, 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 PP, Kartal E, iñares-Blanco J, The FINRISK Microbiome DREAM Challenge and ML4Microbiome. 2023. Microbiome-based risk prediction in incident heart failure a community challenge. Preprint. https://doi.org/10.1101/2023.10.12.23296829
18. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, 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 AJ, Hugenholtz P, Parks DH. 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 IC, Lecomte A, Shishido TK, Laine P, Suppula J, Paulin L, Scheperjans F, Pereira PAB, 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 UE, Castelán-Sánchez HG, Madrigal-Trejo D, Eguiarte LE, 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 ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 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 OM, 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. Carrow HC, Batachari LE, Chu H. 2020. Strain diversity in the microbiome: lessons from Bacteroides fragilis. PLoS Pathog 16:e1009056. https://doi.org/10.1371/journal.ppat.1009056
25. Galperin MY, Vera Alvarez R, Karamycheva S, Makarova KS, Wolf YI, Landsman D, Koonin EV. 2025. COG database update 2024. Nucleic Acids Res 53:D356–D363. https://doi.org/10.1093/nar/gkae983
26. Jin D-M, Morton JT, 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
27. Maghini DG, Oduaran OH, Olubayo LAI, Cook JA, Smyth N, Mathema T, Belger CW, Agongo G, Boua PR, Choma SSR, et al. 2025. Expanding the human gut microbiome atlas of Africa. Nature 638:718–728. https://doi.org/10.1038/s41586-024-08485-8
28. Yang C, Chowdhury D, Zhang Z, Cheung WK, Lu A, Bian Z, 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
29. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 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
30. Vicedomini R, Quince C, Darling AE, Chikhi R. 2021. Strainberry: automated strain separation in low-complexity metagenomes using long reads. Nat Commun 12:4485. https://doi.org/10.1038/s41467-021-24515-9
31. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. 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
32. Feng X, Li H. 2024. Evaluating and improving the representation of bacterial contents in long-read metagenome assemblies. Genome Biol 25:92. https://doi.org/10.1186/s13059-024-03234-6
33. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176:649–662. https://doi.org/10.1016/j.cell.2019.01.001
34. Leviatan S, Shoer S, Rothschild D, Gorodetski M, 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
35. Ravichandar JD, Rutherford E, Chow C-ET, Han A, Yamamoto ML, Narayan N, Kaplan GG, Beck PL, Claesson MJ, Dabbagh K, Iwai S, DeSantis TZ. 2022. Strain level and comprehensive microbiome analysis in inflammatory bowel disease *via* multi-technology meta-analysis identifies key bacterial influencers of disease. Front Microbiol 13:961020. https://doi.org/10.3389/fmicb.2022.961020
36. Leimbach A, Hacker J, Dobrindt U. 2013. *E. coli* as an all-rounder: the thin line between commensalism and pathogenicity. Edited by U. Dobrindt, J. H. Hacker, and C. Svanborg. Curr Top Microbiol Immunol 358:3–32. https://doi.org/10.1007/82_2012_303
37. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, Manghi P, Dubois L, Huang KD, Thomas AM, et al. 2023. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol 41:1633–1644. https://doi.org/10.1038/s41587-023-01688-w
38. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. 2021. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol 39:727–736. https://doi.org/10.1038/s41587-020-00797-0
39. Chen-Liaw A, Aggarwala V, Mogno I, Haifer C, Li Z, Eggers J, Helmus D, Hart A, Wehkamp J, Lamousé-Smith ESN, et al. 2025. Gut microbiota strain richness is species specific and affects engraftment. Nature 637:422–429. https://doi.org/10.1038/s41586-024-08242-x
40. Metwaly A, 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

41. Fehlner-Peach H, Magnabosco C, Raghavan V, Scher JU, Tett A, Cox LM, Gottsegen C, Watters A, Wiltshire-Gordon JD, Segata N, Bonneau R, Littman DR. 2019. Distinct polysaccharide utilization profiles of human intestinal prevotella copri isolates. Cell Host Microbe 26:680–690. https://doi.org/10.1016/j.chom.2019.10.013
42. Gomez-Arango LF, Barrett HL, McIntyre HD, Callaway LK, Morrison M, Dekker Nitert M, SPRING Trial Group. 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
43. Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, 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
44. 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
45. Andrews S. 2023 FastQC: a quality control tool for high throughput sequence data. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
46. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
47. Rodriguez-R LM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT. 2018. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. mSystems 3:e00039-18. https://doi.org/10.1128/mSystems.00039-18
48. Li D, Liu C-M, Luo R, Sadakane K, 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 DD, Li F, Kirton E, Thomas A, Egan R, An H, 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 BA, Singer SW. 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 JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O, Rasmussen S. 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 CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 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 DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 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 MR, Brown CT, Brooks B, Banfield JF. 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 DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, 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 from: https://github.com/tseemann/barrnap
57. Chan PP, Lin BY, Mak AJ, Lowe TM. 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, 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 STN, Newell RJP, Nissen JN, Camargo AP, Tyson GW, Woodcroft BJ. 2025. CoverM: read alignment statistics for metagenomics. Bioinformatics 41:btaf147. https://doi.org/10.1093/bioinformatics/btaf147
60. Musiał K, Petruńko L, Gmiter D. 2024. Simple approach to bacterial genomes comparison based on Average Nucleotide Identity (ANI) using fastANI and ANIclustermap. FBOe 18:66–71. https://doi.org/10.18778/1730-2366.18.10
61. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319. https://doi.org/10.7717/peerj.1319

III

Pantiukh, Kateryna, and Elin Org.
“Human gut archaea collection from
Estonian population.”
Scientific data vol. 13,1 366. 6 Feb. 2026,
doi:10.1038/s41597-026-06742-1

scientific data

OPEN

DATA DESCRIPTOR

Human gut archaea collection from Estonian population

Kateryna Pantiukh & Elin Org

While microbiota plays a crucial role in maintaining overall health, archaea, a component of microbiota, remain relatively unexplored. Here, we present a newly assembled set of archaeal metagenome-assembled genomes (MAGs) from 1,878 fecal microbiome samples. These MAGs were reconstructed from metagenomic reads of the Estonian Microbiome Deep (EstMB-deep) cohort, which were reused here specifically for archaeal MAG reconstruction. We identified 273 archaeal MAGs, representing 21 species and 144 strains which we curated into the "EstMB MAGdb Archaea-273" MAGs collection.

Background & Summary

The human gut microbiota plays a crucial role in maintaining overall health, yet the diversity of archaea within it remains relatively unexplored. The latest release of Unified Human Gastrointestinal Genome (UHGG) collection (v2.0.2, released on January 23, 2024) contains 28 archaeal species[1]. The most prevalent archaeal species in the human gut belong to the orders *Methanobacteriales* (87.15%) and *Methanomassiliicoccales* (12.43%), with much lower representation from *Methanomicrobiales* (0.26%) and *Halobacteriales* (0.17%). The most prevalent archaeal species detected in the human gut is methanogenic archaea *Methanobrevibacter smithii* (proposed reclassification as *Candidatus* Methanocatella smithii), from the order *Methanobacteriales*, which is found in some studies in over 95% of adults[2]. Methanogens act at the terminal step of fermentation, converting H_2 and CO_2, into methane[3]. Methanogenic archaea prevalence has been linked to slow intestinal transit[4] and several diseases[5,6]. In irritable bowel syndrome, higher prevalence is associated with constipation, whereas lower prevalence is observed in diarrheal cases[7]. In colorectal cancer, methanogens exhibit a complex pattern, with species of the phylum *Methanobacteriota* enriched and members of the order *Methanomassiliicoccales* depleted, some showing a progressive decline from healthy controls to adenoma to CRC[8]. In obesity, *M. smithii* is elevated when coupled with increased hydrogen production[9]. Consequently, much of the research to date has focused on methanogenic archaea, particularly *M. smithii* species[10,11]. However, further research, especially in underrepresented populations are needed to uncover the broader diversity of archaeal species in the human gut, as these may have important implications for human health.

Our goal was to assemble MAGs representing potentially new species as well as new strains of known species in order to expand the existing collection of human gut archaeal genomes present in Estonian and potentially in other closely related populations (Fig. 1a). To achieve this, we analysed a total of 1,878 individuals from Estonian Microbiome cohort (EstMB)[12], covering a range of ages and both sexes: 1,308 samples (69.65%) were from females and 570 samples (30.35%) from males, with mean age of 50.05 years (range 23–89 years). The samples were sequenced at high depth, yielding on average 16.49 ± 6.2 Gb of total base pairs per sample, corresponding to 56.07 ± 19.44 million paired reads per sample.

Metagenome assembly was performed individually for each sample, and a multi-binning strategy (MaxBin, MetaBAT, and VAMB) followed by refinement with DAS-Tool produced a total of 84,762 metagenome-assembled genomes (MAGs). Among these, 316 archaeal MAGs were identified based on taxonomic classification using GTDB-Tk v2 (release 226). (Fig. 1b). We evaluated completeness and contamination of all MAGs using CheckM2 and determined genome size with SeqKit. Low quality MAGs (completeness $< 50\%$, contamination $> 10\%$) and those labelled as lacking "consensus contig annotation" based on GUNC annotation were removed from the final set, resulting in 273 archaeal MAGs. This final curated set is hereafter referred to as "**EstMB MAGdb Archaea-273**".

Functional annotation of all MAGs in the 'EstMB MAGdb Archaea-273' collection was performed using Prokka.

Archaeal MAGs from the "EstMB MAGdb Archaea-273" collection were then clustered at the species level (ANI > 95), yielding 21 species. A single genome was selected as representative for each species cluster, forming

Institute of Genomics, Estonian Genome Centre, University of Tartu, Tartu, Estonia. e-mail: pantiukh@ut.ee; elin.org@ut.ee



Fig. 1 Overview of the creation of the Estonian archaeal MAGs collection, MAG quality assessment, and description of the origin samples. (**a**) Study overview. (**b**) Taxonomic placement of representative MAGs from the "Archaea ESTrep-21" MAGs collection. (**c**) Quality estimation for the archaea MAGs from the "Archaea ESTrep-21" MAGs collection. High-quality (HQ) MAGs (completeness > 90, contamination < 5)– 16; Medium-quality (MQ) MAGs (completeness > 50, contamination < 10) – 5. (**d**) Completeness of the MAGs from the "Archaea ESTrep-21" MAGs collection (%). (**e**) Contamination of the MAGs from the "Archaea ESTrep-21" MAGs collection (%) f. Distribution of the number of MAGs assembled per sample.

the final "**Archaea ESTrep-21**" MAG collection. Taxonomic annotation of these representative MAGs was performed with GTDB-Tk v2.3.0 using GTDB Release 226 (Fig. 1b).

Out of the 21 species representative MAGs, 16 (76.19%) met high-quality (HQ) criteria (completeness > 90%, contamination < 5%), while five were medium quality (MQ) (completeness > 50%, contamination < 10% (Fig. 1c). Overall, the **Archaea ESTrep-21** MAG collection had a median completeness of 96.85% (Interquartile Range [IQR], 91.49–98.39) and a median contamination of 0.81% (IQR, 0.00–1.61) (Fig. 1d,e).

Typically, only one archaeal MAG was recovered per sample. The distribution of the number of archaeal species assembled per individual sample is presented in Fig. 1f.

Methods

MAG assembly and binning. Short-read sequencing data, already filtered to remove low-quality sequences and host-derived reads, were obtained from the European Nucleotide Archive database (https://www.ebi.ac.uk/ega/), accession codes: EGAS00001008448, EGAS50000001538. Detailed descriptions of the samples and sequencing data are provided in our previous publication[12]. The reads were assembled into contigs with MEGAHIT v1.2.911[13]. Binning was performed separately for each individual sample. Initially, contigs were binned using single binners: MetaBAT v2.15[14], MaxBin v2.2.7 (https://sourceforge.net/projects/maxbin2), and VAMB v3.0.7[15]. Given that different binning tools reconstruct genomes differently, a bin aggregation software, i.e., DAS Tool v1.1.4[16] was employed to integrate the bin predictions from VAMB, MetaBAT2, and MaxBin2. This approach optimized the selection of non-redundant, high-quality bin sets using default parameters. MAGs resulting from this process form the EstMB MAG collection (84,762 MAGs). All MAGs were clustered on species level (ANI index $\geq$ 95) with dRep[17] where 69,302 were clustered into 2,257 species clusters and 15,460 MAGs failed to cluster due to not meeting the built-in dRep quality check threshold.

Taxonomic annotation. Taxonomy of all MAGs was assigned with GTDB-Tk v2.3.0[18,19]. Any MAGs that could not be assigned to any known species based on GTDB-Tk v2.3.0[20] taxonomic annotation was identified as potentially new species.

Evaluation of MAG quality. The quality of all MAGs was evaluated based on CheckM2[21] results. Here, high quality (HQ) MAGs exhibited completeness exceeding 90% and contamination below 5%, medium quality (MQ) MAGs failed to classify as a HQ MAGs but displayed completeness greater than 50% and contamination below 10%, while all other MAGs were categorized as low quality (LQ) MAGs. Additionally, to assess the quality of the recovered MAGs, we examined genome size distributions and taxonomic consistency of contigs. Three MAGs annotated as unknown species within the order *Candidatus* Korarchaeia exhibited exceptionally large genome sizes (4.9 Mb, 10.2 Mb, and 13.2 Mb), and two MAGs from the genus Methanocatella showed similarly atypical sizes (6–8 Mb). Given that the only sequenced *Candidatus* Korarchaeia genome, *Candidatus* Korarchaeum cryptofilum, is approximately 1.6 Mb[22], and Methanocella genomes are typically 1.7–2.0 Mb, these inflated genomes sizes suggested possible misassembly or contamination. To further evaluate this, we annotated the contigs of each MAG using GUNC[23]. MAGs in which more than 50% of contigs were assigned taxonomic annotations inconsistent with the overall MAG classification were labelled as lacking "consensus contig annotation". All MAGs with unusually large genome sizes also received this label, supporting concerns about their assembly quality.

"EstMB MAGdb Archaea-273" MAGs collection. For all MAGs classified as archaea (n = 316), we assessed completeness and contamination using CheckM2[21], determined genome size and number of contigs with seqkit/2.3.1[24] and annotated each MAG contigs separately with GUNC[23]. During subsequent quality control steps, we excluded low quality MAGs (completeness $<$ 50%, contamination $>$ 10%), MAG with abnormal genome size ($>$2 Mb) and MAGs with non-consensus contig annotation (consensus contig annotation refers to the situation where more than half of the contig lengths are annotated with the same genus). This final filtering step allowed us to eliminate questionable MAGs with unusually large genome sizes and heterogeneous contig origins, even when completeness was high c and contamination were low, resulting in a final set of 273 archaeal MAGs.

"Archaea ESTrep-21" MAGs collection. All MAGs from the "EstMB MAGdb Archaea-273" MAGs collection were clustered at the species level (ANI index $\geq$ 95) using dRep (parameters: -sa 0.95 --checkM_method taxonomy_wf -comp 25 -con 50 --S_algorithm fastANI --multiround_primary_clustering --greedy_secondary_clustering --run_tertiary_clustering)[25], resulting in 21 species level clusters with a representative genome from each cluster. These representative genomes form the "Archaea ESTrep-21" MAGs collection, which is a subset of "EstMB MAGdb Archaea-273" MAGs collection. The MAGs from this collection are not stored separately, as they are included in the larger "EstMB MAGdb Archaea-273" collection. If needed, representative MAGs can be retrieved from the "EstMB MAGdb Archaea-273" collection using the MAG_ID. To evaluate strain diversity within the "EstMB MAGdb Archaea-273" MAGs collection, MAGs from this collection were also clustered on strain level (ANI index $\geq$ 99) with dRep[17].

Estimation of prevalence and abundance of archaeal species in Estonian population. For mapping-based community profiling, we used the "Archaea **ESTrep-21**" MAGs collection as a reference. All reads were mapped against the reference with CoverM v0.6.1 (https://github.com/wwood/CoverM) and archaeal species abundance was estimated for all samples. To calculate species prevalence, we aggregated single sample results into a summary table (script "parse_abud_results.py") and performed further analysis with standard Python packages.

Functional annotation. Protein-coding sequences (CDSs), tRNA genes and CRISPR were predicted and annotated with Prokka v.1.14.6[26] using the parameters '–kingdom Archaea'. Annotation depth was defined based on Prokka results as the proportion of coding sequences (CDSs) with assigned putative functions relative to the total number of CDSs.

Tools used for data visualization. All visualisations were conducted using the Python packages *pandas*[27], *matplotlib*[28] and *seaborn*[29].

Data Record

This Data Record describes two related datasets of archaeal metagenome assembled genomes (MAGs) derived from the Estonian Microbiome (EstMB) cohort.

EstMB MAGdb Archaea-273. The "EstMB MAGdb Archaea-273" collection comprises a final curated set of 273 archaeal MAGs reconstructed from shotgun metagenomic sequencing data generated within the EstMB cohort.

All MAG genome assemblies are publicly available from the European Nucleotide Archive (ENA) under the study accession PRJEB81541. Genome sequences are provided in standard FASTA format.

A comprehensive description of the collection is publicly available on Figshare. This includes MAG quality metrics, taxonomic annotations, and genome characteristics, as well as functional annotation results generated using Prokka. Metadata are provided as structured tabular files in XLSX format, and Prokka outputs are provided in their standard file formats.

Archaea ESTrep-21. The "Archaea ESTrep-21" collection is a curated species level representative subset of the "EstMB MAGdb Archaea-273" dataset. It consists of 21 archaeal MAGs selected as species representatives based on taxonomic classification.

Representative MAGs are not stored separately, as they are included within the larger "EstMB MAGdb Archaea-273" collection. Individual representative genomes can be retrieved using their corresponding MAG_ID values.

The associated description file for the "Archaea ESTrep-21" collection is publicly available on Figshare and includes MAG quality parameters, taxonomic annotation, genome characteristics, the number of MAGs per species, species prevalence, and species mean relative abundance within the Estonian population. These data are provided as structured tabular files in XLSX format.

Technical Validation

Within the initial set of 316 archaeal MAGs, three genomes were annotated as unknown species belonging to the class *Candidatus* Korarchaeia (IDs: A316K001, A316K002, A316K003). As the presence of this lineage in the human gut was unexpected, we performed additional verification of these assemblies. Two MAGs showed acceptable quality estimates according to CheckM2 (A316K001: 97.2% completeness, 5.2% contamination; A316K002: 91.2% completeness, 5.7% contamination), whereas A316K003 displayed extremely low completeness (7.4%) despite low contamination (2.7%).

Genome sizes for the higher-quality MAGs (A316K001: 13.2 Mb; A316K002: 10.2 Mb) substantially exceeded the expected size for members of Korarchaeia lineage, as the only sequenced reference genome, *Candidatus* Korarchaeum cryptofilum, is approximately 1.6 Mb[13]. The low-quality MAG (A316K003) had an intermediate genome size of 4.9 Mb. Furthermore, contig-level annotations using GUNC revealed that all three MAGs were highly fragmented, with the majority of contigs assigned to diverse, often bacterial, taxa.

Taken together, these results indicate that the three MAGs most likely represent technical artifacts rather than genuine members of the Korarchaeia lineage. This example highlights that even MAGs passing conventional quality thresholds may still reflect misassembled or chimeric sequences. We therefore recommend complementing completeness and contamination estimates with additional validation steps, such as genome size assessment, evaluation of assembly fragmentation, and contig-level taxonomic annotation, when assessing MAG reliability.

Usage Notes

The "EstMB MAGdb Archaea-273" collection represents an expanded set of archaeal species and strains derived from the human gut microbiome. We recommend its use in studies of human gut microbial diversity, particularly in association analyses, to enhance reference database coverage and improve the detection and characterization of archaeal- and more broadly, microbiome- diversity across samples. In addition, the inclusion of multiple genomes for known species enables detailed investigations of strain-level structure, evolutionary dynamics, and ecological patterns.

All MAGs are publicly available for download from the European Nucleotide Archive (ENA) without any access restrictions under study accession PRJEB81541. Functional annotation is also provided at Figshare.

Notably, we also share our experience regarding the occurrence of false-positive MAGs that can arise during metagenome assembly and binning. To improve the reliability of MAG datasets, we provide the following recommendations for additional post-processing quality checks:

1. **Manual inspection** of not only completeness and contamination metrics (e.g., via CheckM2), but also genome size and GUNC-based contig-level annotations.
2. **Filtering based on genome size**, excluding MAGs with abnormal genome sizes defined as more than ± 3 standard deviations from the mean genome size within the corresponding phylum.
3. **Taxonomic consistency check** using GUNC: MAGs in which more than 50% of contigs are annotated with a lineage inconsistent with the GTDB-Tk-based taxonomic classification should be considered unreliable and excluded.

These additional quality control steps help reduce the inclusion of chimeric or misclassified assemblies, ultimately improving the robustness and interpretability of downstream analyses.

Data availability

The MAG sequences from "EstMB MAGdb Archaea-273" MAGs collection have been deposited in the European Nucleotide Archive (ENA). Study accession PRJEB81541. Available at: https://www.ebi.ac.uk/ena/browser/view/PRJEB81541 (accessed 22 October 2024)[30].

"EstMB MAGdb Archaea-273" MAGs collection description, including quality parameters, taxonomic annotation, and genome characteristics are publicly available on Figshare at https://doi.org/10.6084/m9.figshare.30920087[31].

"Archaea ESTrep-21" MAGs collection description, including quality parameters, taxonomic annotation, and genome characteristics are publicly available on Figshare at https://doi.org/10.6084/m9.figshare.30920126[32].

Prokka annotation results for the MAGs from the 'EstMB MAGdb Archaea-273' collection, containing all assembled archaeal genomes, are publicly available on Figshare at https://doi.org/10.6084/m9.figshare.29329166[33].

Code availability

All commands and pipelines used in data processing were executed according to the manual and protocols of the corresponding bioinformatics software. The options and parameters of all tools used for the analysis are described in the main text.

The code for creating figures and tables is accessible on GitHub (https:// github.com/Chartiza/ArchaeaDraft Genomes).

Received: 9 December 2024; Accepted: 27 January 2026;
Published online: 06 February 2026

References

1. Almeida, A. *et al*. 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 (2021).
2. Chibani, C. M. *et al*. A catalogue of 1,167 genomes from the human gut archaeome. *Nat Microbiol* **7**, 48–61 (2021).
3. Kumpitsch, C. *et al*. Reduced B12 uptake and increased gastrointestinal formate are associated with archaeome-mediated breath methane emission in humans. *Microbiome* **9**, 193 (2021).
4. Attaluri, A., Jackson, M., Valestin, J. & Rao, S. S. Methanogenic Flora Is Associated With Altered Colonic Transit but Not Stool Characteristics in Constipation Without IBS. *American Journal of Gastroenterology* **105**, 1407–1411 (2010).
5. Kuehnast, T. *et al*. Exploring the human archaeome: its relevance for health and disease, and its complex interplay with the human immune system. *The FEBS Journal febs*. 17123, https://doi.org/10.1111/febs.17123 (2024).
6. Basseri, R. J. *et al*. Intestinal Methane Production in Obese Individuals Is Associated with a Higher Body Mass Index.
7. Ghoshal, U., Shukla, R., Srivastava, D. & Ghoshal, U. C. Irritable Bowel Syndrome, Particularly the Constipation-Predominant Form, Involves an Increase in *Methanobrevibacter smithii*, Which Is Associated with Higher Methane Production. *Gut and Liver* **10**, 932–938 (2016).
8. Li, T. *et al*. Multi-Cohort Analysis Reveals Altered Archaea in Colorectal Cancer Fecal Samples Across Populations. *Gastroenterology* **168**(3), 525–538.e2 (2025).
9. Mathur, R. *et al*. Methane and Hydrogen Positivity on Breath Test Is Associated With Greater Body Mass Index and Body Fat. *The Journal of Clinical Endocrinology & Metabolism* **98**, E698–E702 (2013).
10. Ghavami, S. B. *et al*. Alterations of the human gut Methanobrevibacter smithii as a biomarker for inflammatory bowel diseases. *Microbial Pathogenesis* **117**, 285–289 (2018).
11. Mbakwa, C. A. *et al*. Gut colonization with *methanobrevibacter smithii* is associated with childhood weight development. *Obesity* **23**, 2508–2516 (2015).
12. Aasmets, O., Krigul, K. L., Lüll, K., Metspalu, A. & Org, E. 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 (2022).
13. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676, https://doi.org/10.1093/bioinformatics/btv033 (2015).
14. Kang, D. D. *et al*. 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 (2019).
15. Nissen, J. N. *et al*. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560, https://doi.org/10.1038/s41587-020-00777-4 (2021).
16. Sieber, C. M. K. *et al*. 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 (2018).
17. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868, https://doi.org/10.1038/ismej.2017.126 (2017).
18. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. *GTDB-Tk v2: memory friendly classification with the Genome Taxonomy Database*. http://biorxiv.org/lookup/doi/10.1101/2022.07.11.499641, https://doi.org/10.1101/2022.07.11.499641 (2022).
19. Parks, D.H., *et al* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gkab776 (2021).
20. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510, https://doi.org/10.1038/s41586-019-1058-x (2019).
21. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. *CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning*. http://biorxiv.org/lookup/doi/10.1101/2022.07.11.499243, https://doi.org/10.1101/2022.07.11.499243 (2022).
22. Elkins, J. G. *et al*. A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc. Natl. Acad. Sci. USA*. **105**, 8102–8107 (2008).
23. Orakov, A. N. *et al*. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biology* **22**, 178–178 (2021).
24. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **11**, e0163962 (2016).
25. Lü, Z. & Lu, Y. Methanocella conradii sp. nov., a Thermophilic, Obligate Hydrogenotrophic Methanogen, Isolated from Chinese Rice Field Soil. *PLoS ONE* **7**, e35279 (2012).

26. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
27. McKinney, W. Data Structures for Statistical Computing Python. 56–61, https://doi.org/10.25080/Majora-92bf1922-00a (Austin, Texas, 2010).
28. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
29. Waskom, M. seaborn: statistical data visualization. *JOSS* **6**, 3021 (2021).
30. *ENA European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB81541 (2024).
31. Pantiukh, P. & Org, E. EstMB MAGdb Archaea-273. *Figshare* https://doi.org/10.6084/m9.figshare.30920087 (2025).
32. Pantiukh, P. & Org, E. Archaea ESTrep-21. *Figshare* https://doi.org/10.6084/m9.figshare.30920126 (2025).
33. Pantiukh, P. & Org, E. Archaea. Prokka annotation. *Figshare* https://doi.org/10.6084/m9.figshare.29329166 (2025).

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

Author contributions

K.P. designed the study and pipeline, performed the data analysis, interpreted the data, prepared the figures and tables and wrote the manuscript. E.O. conceptualised and supervised the study, reviewed and edited a draft.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.P. or E.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

CURRICULUM VITAE

Name: Kateryna Pantiukh
Date of birth: 11.05.1988
Contact: Estonian Genome Center, Institute of Genomics, University of Tartu
Riia 23b, 51010, Tartu
E-mail: pantiukh@gmail.com

Education:

2022–	PhD in Genomics, Institute of Genomics, University of Tartu
2009–2011	Master's Degree in Cytology, Histology and Embryology, National Taras Shevchenko University of Kyiv
2005–2009	Bachelor's Degree in Biology, National Taras Shevchenko University of Kyiv

Professional employment:

2026–2026	Kyiv School of Economics (Lecturer in Metagenomics)
2022–…	University of Tartu, Institute of Genomics (Junior Research Fellow)
2022-2022	University of Tartu, Institute of Genomics (Specialist)

Administrative work:

2022–... Estonian Society for Microbiology

Publications:

Pantiukh, K.; Aasmets, O.; Krigul, K. L.; Org, E. (2026). Metagenome-assembled genomes from a population-based cohort uncover new gut species and within-species diversity, revealing prevalent disease associations. mSystems, e00114-26. DOI: 10.1128/msystems.00114-26.

Pantiukh, K.; Org, E. (2026). Human gut archaea collection from Estonian population. Scientific Data, 13, 366. DOI: 10.1038/s41597-026-06742-1.

Zielińska K*, **Pantiukh K*,** Łabaj, P. P.; Kosciolek, T.; Org, E. (2026). A large-scale comparative metagenomic analysis of short-read sequencing platforms indicates high taxonomic concordance and functional analysis challenges. mSystems. DOI: 10.1128/msystems.01714-25.

Zielińska K*, **Pantiukh K*,** Org, E., Łabaj, P. P., Kosciolek, T. (2026) Moving from a taxonomic to a functional perspective in global microbiome analysis requires optimizing multiplexing ratios. Su X, editor. mSystems. 29;e00144-26. doi:10.1128/msystems.00144-26

Milani, Lili; Alver, Maris; Laur, Sven; Reisberg, Sulev; Haller, Toomas; Aasmets, Oliver; Abner, Erik; Alavere, Helene; Allik, Annely; Annilo, Tarmo; Fischer, Krista; Hudjashov, Georgi; Jõeloo, Maarja; Kals, Mart; Karo-Astover, Liis; Kasela, Silva; Kolde, Anastassia; Krebs, Kristi; Krigul, Kertu Liis; Kronberg, Jaanika; Kruusmaa, Karoliina; Kukuškina, Viktorija; Kõiv, Kadri; Lehto, Kelli; Leitsalu, Liis; Lind, Sirje; Luitva, Laura Birgit; Läll, Kristi; Lüll, Kreete; Metsalu, Kristjan; Metspalu, Mait; Mõttus, René; Nelis, Mari; Nikopensius, Tiit; Nurm, Miriam; Nõukas, Margit; Oja, Marek; Org, Elin; Palover, Marili; Palta, Priit; Pankratov, Vasili; **Pantiukh, Kateryna**; Pervjakova, Natalia; Pujol-Gualdo, Natàlia; Reigo, Anu; Reimann, Ene; Smit, Steven; Sokurova, Diana; Taba, Nele; Talvik, Harry-Anton; Teder-Laving, Maris; Tõnisson, Neeme; Vaht, Mariliis; Vainik, Uku; Võsa, Urmo; Esko, Tõnu; Kolde, Raivo; Mägi, Reedik; Vilo, Jaak; Laisk, Triin; Metspalu, Andres (2025). From Biobanking to The Estonian Biobank's journey from biobanking to personalized medicine, Nature communications, vol. 16,1 3270. https://doi.org/10.1038/s41467-025-58465-3

Krigul, Kertu Liis & Feeney, Rachel H.; Wongkuna, Supapit; Aasmets, Oliver; Holmberg, Sandra M.; Andreson, Reidar; Puértolas Balint, Fabiola; **Pantiukh, Kateryna**; Sootak, Linda; Org, Tõnis; Tenson, Tanel; Org, Elin & Schroeder, Björn O. (2024). A history of repeated antibiotic usage leads to microbiota-dependent mucus defects. Gut Microbes, 16(1). https://doi.org/10.1080/19490976.2024.2377570

Popular science articles:

Kateryna Pantiukh (2024). Hidden bacteria may shape your health. Research in Estonia https://researchinestonia.eu/2024/10/24/bacteria-and-genome/

Kateryna Pantiukh (2023). Tartu genomic researcher finds 300 new bacterial species. Novaator. https://news.err.ee/1609141072/3-minute-lecture-tartu-genomic-researcher-finds-300-new-bacterial-species

Supervised dissertations:

Oleksandr Kholonivskyi	Bachelor's Degree, Institute of Genomics, University of Tartu, Estonia, 2026, "Characterizing the genomic diversity of prevalent gut microbiome bacterial species using metagenome-assembled genomes"
Vanessa Viiding	Bachelor's Degree, Institute of Genomics, University of Tartu, Estonia, 2026, "Arhede tuvastamist mõjutavad tegurid inimese soolestiku mikrobioomi uuringutes" and in english "Factors influencing the detection of archaea in human gut microbiome studies"

Awards and scholarships:

2025 **The 1st Prize for the best poster presentation**
SymbNET PhD Summer School on Host-Microbe Symbioses

2025 **Outstanding Poster presentation**
Institute of Molecular and Cell Biology & Institute of Genomics Annual Conference 2024

2025 **Acknowledgement for the best poster presentation in microbiology**
Institute of Molecular and Cell Biology & Institute of Genomics Annual Conference 2024. Estonian Society for Microbiology

2025 **EMBO|EMBL Symposium travel grant**
EMBO | EMBL Symposium: The human microbiome taking place at EMBL Heidelberg (€400)

2025 **SymbNET Travel grant**
SymbNET PhD Summer School on Host-Microbe Symbioses (€350)

2024 **Artur Lind Scholarship**
Institute of Genomics of University of Tartu, in collaboration with the Estonian Genome Foundation

2024 **Outstanding Talk award**
Institute of Molecular and Cell Biology & Institute of Genomics Annual Conference 2023

2024 **Audience award for the Best Talk**
Institute of Molecular and Cell Biology & Institute of Genomics Annual Conference 2023

2024 **Winner of the PhD student competition**
Training Camp 2024: The Art of Giving a Popular Science Talk

2024 **Erasmus+ Scholarship Grant**
Erasmus+EU programme for education, training, youth and sport (€2,100)

2024 **Estonian Doctoral School mobility grant**
Research visit to Centre for Microbiome Research, Queensland University of Technology, Brisbane, Australia (€1,100)

2024 **CWT Estonia (Kaleva Travel) travel scholarship**
Support attendance at the 10th International Human Microbiome Consortium (IHMC) Congress 2024, Rome, Italy (€1,200)

2024 **EMBO Special Travel Grant**
Support attendance at the EMBO Practical Course: Integrative analysis of multi-omics data taking place at EMBL Heidelberg, Germany (€750)

2023 **Mobility grant for doctoral students of graduate school of the bio-medicine and biotechnology**
EMBO | EMBL Symposium "The human microbiome", Heidelberg, Germany (€700)

ELULOOKIRJELDUS

Nimi: Kateryna Pantiukh
Sünniaeg: 11.08.1988
Aadress: Eesti Geenivaramu, Genoomika instituut, Tartu Ülikool
Riia 23b, 51010, Tartu
E-post: pantiukh@gmail.com

Haridus:

2021–	Doktorikraad bioinformaatikas, Tartu Ülikooli Genoomika Instituut
2009-2011	Magistrikraad tsütoloogias, histoloogias ja embrüoloogias, Kiievi Riiklik Taras Ševtšenko Ülikool
2005–2009	Bakalaureusekraad bioloogias, Kiievi Riiklik Taras Ševtšenko Ülikool

Teenistuskäik:

2026–2026	Kiievi Majanduskool (metagenoomika lektor)
2022–…	Tartu Ülikool, Genoomika Instituut (noorem teadur)
2022-2022	Tartu Ülikool, Genoomika Instituut (spetsialist)

Teadusorganisatsiooniline ja- administratiivne tegevus:

2022–...	Eesti Mikrobioloogide Ühendus

Teaduspublikatsioonid:

Pantiukh, K.; Aasmets, O.; Krigul, K. L.; Org, E. (2026). Metagenome-assembled genomes from a population-based cohort uncover new gut species and within-species diversity, revealing prevalent disease associations. mSystems, e00114-26.
DOI: 10.1128/msystems.00114-26.

Pantiukh, K.; Org, E. (2026). Human gut archaea collection from Estonian population. Scientific Data, 13, 366. DOI: 10.1038/s41597-026-06742-1.

Zielińska K*, **Pantiukh K*,** Łabaj, P. P.; Kosciolek, T.; Org, E. (2026). A large-scale comparative metagenomic analysis of short-read sequencing platforms indicates high taxonomic concordance and functional analysis challenges. mSystems.
DOI: 10.1128/msystems.01714-25.

Zielińska K*, **Pantiukh K*,** Org, E., Łabaj, P. P., Kosciolek, T. (2026) Moving from a taxonomic to a functional perspective in global microbiome analysis requires optimizing multiplexing ratios. Su X, editor. mSystems. 29;e00144-26. doi:10.1128/msystems.00144-26

Milani, Lili; Alver, Maris; Laur, Sven; Reisberg, Sulev; Haller, Toomas; Aasmets, Oliver; Abner, Erik; Alavere, Helene; Allik, Annely; Annilo, Tarmo; Fischer, Krista; Hudjashov, Georgi; Jõeloo, Maarja; Kals, Mart; Karo-Astover, Liis; Kasela, Silva; Kolde, Anastassia; Krebs, Kristi; Krigul, Kertu Liis; Kronberg, Jaanika; Kruusmaa, Karoliina; Kukuškina, Viktorija; Kõiv, Kadri; Lehto, Kelli; Leitsalu, Liis; Lind, Sirje; Luitva, Laura Birgit; Läll, Kristi; Lüll, Kreete; Metsalu, Kristjan; Metspalu, Mait; Mõttus, René; Nelis, Mari; Nikopensius, Tiit; Nurm, Miriam; Nõukas, Margit; Oja, Marek; Org, Elin; Palover, Marili; Palta, Priit; Pankratov, Vasili; **Pantiukh, Kateryna**; Pervjakova, Natalia; Pujol-Gualdo, Natàlia; Reigo, Anu; Reimann, Ene; Smit, Steven; Sokurova, Diana; Taba, Nele; Talvik, Harry-Anton; Teder-Laving, Maris; Tõnisson, Neeme; Vaht, Mariliis; Vainik, Uku; Võsa, Urmo; Esko, Tõnu; Kolde, Raivo; Mägi, Reedik; Vilo, Jaak; Laisk, Triin; Metspalu, Andres (2025). From Biobanking to The Estonian Biobank's journey from biobanking to personalized medicine, Nature communications, vol. 16,1 3270. https://doi.org/10.1038/s41467-025-58465-3

Krigul, Kertu Liis & Feeney, Rachel H.; Wongkuna, Supapit; Aasmets, Oliver; Holmberg, Sandra M.; Andreson, Reidar; Puértolas Balint, Fabiola; **Pantiukh, Kateryna**; Sootak, Linda; Org, Tõnis; Tenson, Tanel; Org, Elin & Schroeder, Björn O. (2024). A history of repeated antibiotic usage leads to microbiota-dependent mucus defects. Gut Microbes, 16(1). https://doi.org/10.1080/19490976.2024.2377570

Populaarteaduslikud artiklid:

Kateryna Pantiukh (2024). Varjatud bakterid võivad teie tervist mõjutada. Uuringud Eestis https://researchinestonia.eu/2024/10/24/bacteria-and-genome/

Kateryna Pantiukh (2023). Tartu genoomikauurija leidis 300 uut bakteriliiki. Novaator. https://news.err.ee/1609141072/3-minute-lecture-tartu-genomic-researcher-finds-300-new-bacterial-species

Juhendatud väitekirjad:

Oleksandr Kholonivskyi	Bakalaureusekraad, Tartu Ülikooli Genoomika Instituut, 2026, "Levinud soolemikrobioomi bakteriliikide genoomse mitmekesisuse iseloomustamine metagenoomi abil kokkupandud genoomide abil"
Vanessa Viiding	Bakalaureusekraad, Tartu Ülikooli Genoomika Instituut, 2026, "Arhede tuvastamist mõjutavad tegurid inimese soolestiku mikrobioomi uuringutes" and in english "Factors influencing the detection of archaea in human gut microbiome studies"

Auhinnad ja stipendiumid:

2025 **The 1st Prize for the best poster presentation**
SymbNET PhD Summer School on Host-Microbe Symbioses

2025 **Outstanding Poster presentation**
Institute of Molecular and Cell Biology & Institute of Genomics Annual Conference 2024

2025 **Acknowledgement for the best poster presentation in microbiology**
Institute of Molecular and Cell Biology & Institute of Genomics Annual Conference 2024. Estonian Society for Microbiology

2025 **EMBO|EMBL Symposium travel grant**
EMBO | EMBL Symposium: The human microbiome taking place at EMBL Heidelberg (€400)

2025 **SymbNET Travel grant**
SymbNET PhD Summer School on Host-Microbe Symbioses (€350)

2024 **Artur Lind Scholarship**
Institute of Genomics of University of Tartu, in collaboration with the Estonian Genome Foundation

2024 **Outstanding Talk award**
Institute of Molecular and Cell Biology & Institute of Genomics Annual Conference 2023

2024 **Audience award for the Best Talk**
Institute of Molecular and Cell Biology & Institute of Genomics Annual Conference 2023

2024 **Winner of the PhD student competition**
Training Camp 2024: The Art of Giving a Popular Science Talk

2024 **Erasmus+ Scholarship Grant**
Erasmus+EU programme for education, training, youth and sport (€2,100)

2024 **Estonian Doctoral School mobility grant**
Research visit to Centre for Microbiome Research, Queensland University of Technology, Brisbane, Australia (€1,100)

2024 **CWT Estonia (Kaleva Travel) travel scholarship**
Support attendance at the 10th International Human Microbiome Consortium (IHMC) Congress 2024, Rome, Italy (€1,200)

2024 **EMBO Special Travel Grant**
Support attendance at the EMBO Practical Course: Integrative analysis of multi-omics data taking place at EMBL Heidelberg, Germany (€750)

2023 **Mobility grant for doctoral students of graduate school of the biomedicine and biotechnology**
EMBO | EMBL Symposium "The human microbiome", Heidelberg, Germany (€700)

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida.* Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand**. The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro.* Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous grassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan**. Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.
41. **Sulev Ingerpuu**. Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.

42. **Veljo Kisand**. Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa**. Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa**. Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik**. Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo**. Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo**. Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots**. Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero**. Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees**. Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks**. Cholecystokinin (CCK) – induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and serotonin. Tartu, 1999, 123 p.
52. **Ebe Sild**. Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva**. Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna**. Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro**. Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane**. Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm**. Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg**. Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild**. The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu**. Studies of the TOL plasmid transcription factor XylS. Tartu, 2000, 88 p.
61. **Dina Lepik**. Modulation of viral DNA replication by tumor suppressor protein p53. Tartu, 2000, 106 p.
62. **Kai Vellak**. Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu, 2000, 122 p.

63. **Jonne Kotta**. Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu, 2000, 160 p.
64. **Georg Martin**. Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000, 139 p.
65. **Silvia Sepp**. Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira**. On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000, 96 p.
67. **Priit Zingel**. The role of planktonic ciliates in lake ecosystems. Tartu, 2001, 111 p.
68. **Tiit Teder**. Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu, 2001, 122 p.
69. **Hannes Kollist**. Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu, 2001, 80 p.
70. **Reet Marits**. Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu, 2001, 112 p.
71. **Vallo Tilgar**. Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major,* breeding in Nothern temperate forests. Tartu, 2002, 126 p.
72. **Rita Hõrak**. Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002, 108 p.
73. **Liina Eek-Piirsoo**. The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002, 74 p.
74. **Krõõt Aasamaa**. Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002, 110 p.
75. **Nele Ingerpuu**. Bryophyte diversity and vascular plants. Tartu, 2002, 112 p.
76. **Neeme Tõnisson**. Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002, 124 p.
77. **Margus Pensa**. Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003, 110 p.
78. **Asko Lõhmus**. Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003, 168 p.
79. **Viljar Jaks**. p53 – a switch in cellular circuit. Tartu, 2003, 160 p.
80. **Jaana Männik**. Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003, 140 p.
81. **Marek Sammul**. Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003, 159 p
82. **Ivar Ilves**. Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003, 89 p.
83. **Andres Männik**. Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003, 109 p.

84. **Ivika Ostonen**. Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003, 158 p.
85. **Gudrun Veldre**. Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003, 199 p.
86. **Ülo Väli.** The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004, 159 p.
87. **Aare Abroi**. The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004, 135 p.
88. **Tiina Kahre**. Cystic fibrosis in Estonia. Tartu, 2004, 116 p.
89. **Helen Orav-Kotta**. Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004, 117 p.
90. **Maarja Öpik**. Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004, 175 p.
91. **Kadri Tali**. Species structure of *Neotinea ustulata*. Tartu, 2004, 109 p.
92. **Kristiina Tambets**. Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004, 163 p.
93. **Arvi Jõers**. Regulation of p53-dependent transcription. Tartu, 2004, 103 p.
94. **Lilian Kadaja**. Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004, 103 p.
95. **Jaak Truu**. Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004, 128 p.
96. **Maire Peters**. Natural horizontal transfer of the *pheBA* operon. Tartu, 2004, 105 p.
97. **Ülo Maiväli**. Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004, 130 p.
98. **Merit Otsus**. Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004, 103 p.
99. **Mikk Heidemaa**. Systematic studies on sawflies of the genera *Dolerus, Empria,* and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004, 167 p.
100. **Ilmar Tõnno**. The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004, 111 p.
101. **Lauri Saks**. Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004, 144 p.
102. **Siiri Rootsi**. Human Y-chromosomal variation in European populations. Tartu, 2004, 142 p.
103. **Eve Vedler**. Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.
104. **Andres Tover**. Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 126 p.
105. **Helen Udras**. Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005, 100 p.

106. **Ave Suija**. Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005, 162 p.
107. **Piret Lõhmus**. Forest lichens and their substrata in Estonia. Tartu, 2005, 162 p.
108. **Inga Lips**. Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005, 156 p.
109. **Krista Kaasik**. Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005, 121 p.
110. **Juhan Javoiš**. The effects of experience on host acceptance in ovipositing moths. Tartu, 2005, 112 p.
111. **Tiina Sedman**. Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005, 103 p.
112. **Ruth Aguraiuja**. Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005, 112 p.
113. **Riho Teras**. Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 106 p.
114. **Mait Metspalu**. Through the course of prehistory in India: tracing the mtDNA trail. Tartu, 2005, 138 p.
115. **Elin Lõhmussaar**. The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006, 124 p.
116. **Priit Kupper**. Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006, 126 p.
117. **Heili Ilves**. Stress-induced transposition of Tn*4652* in *Pseudomonas Putida.* Tartu, 2006, 120 p.
118. **Silja Kuusk**. Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006, 126 p.
119. **Kersti Püssa**. Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006, 90 p.
120. **Lea Tummeleht**. Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006, 94 p.
121. **Toomas Esperk**. Larval instar as a key element of insect growth schedules. Tartu, 2006, 186 p.
122. **Harri Valdmann**. Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.
123. **Priit Jõers**. Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli**. Gata3 and Gata2 in inner ear development. Tartu, 2007, 123 p.
125. **Kai Rünk**. Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007, 143 p.

126. **Aveliina Helm**. Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007, 89 p.
127. **Leho Tedersoo**. Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007, 233 p.
128. **Marko Mägi**. The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007, 135 p.
129. **Valeria Lulla**. Replication strategies and applications of Semliki Forest virus. Tartu, 2007, 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007, 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007, 171 p.
132. **Tatjana Krama**. Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007, 112 p.
133. **Signe Saumaa**. The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida.* Tartu, 2007, 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007, 96 p.
135. **Priit Kilgas**. Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007, 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007, 95 p.
137. **Kärt Padari**. Protein transduction mechanisms of transportans. Tartu, 2008, 128 p.
138. **Siiri-Lii Sandre**. Selective forces on larval colouration in a moth. Tartu, 2008, 125 p.
139. **Ülle Jõgar**. Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008, 99 p.
140. **Lauri Laanisto**. Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008, 133 p.
141. **Reidar Andreson**. Methods and software for predicting PCR failure rate in large genomes. Tartu, 2008, 105 p.
142. **Birgot Paavel**. Bio-optical properties of turbid lakes. Tartu, 2008, 175 p.
143. **Kaire Torn**. Distribution and ecology of charophytes in the Baltic Sea. Tartu, 2008, 98 p.
144. **Vladimir Vimberg**. Peptide mediated macrolide resistance. Tartu, 2008, 190 p.
145. **Daima Örd**. Studies on the stress-inducible pseudokinase TRB3, a novel inhibitor of transcription factor ATF4. Tartu, 2008, 108 p.
146. **Lauri Saag**. Taxonomic and ecologic problems in the genus *Lepraria* (*Stereocaulaceae*, lichenised *Ascomycota*). Tartu, 2008, 175 p.

147. **Ulvi Karu**. Antioxidant protection, carotenoids and coccidians in greenfinches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm**. Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks**. Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu**. Acclimation of stomatal structure and function in tree canopy: effect of light and CO_2 concentration. Tartu, 2008, 108 p.
151. **Janne Pullat**. Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš**. Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtšenko**. Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast**. Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats**. Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova**. The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida.* Tartu, 2009, 124 p.
157. **Tsipe Aavik**. Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2009, 112 p.
158. **Kaja Kiiver**. Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja**. Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast**. Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.
161. **Ain Vellak**. Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.
162. **Triinu Remmel**. Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.
163. **Jaana Salujõe**. Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.
164. **Ele Vahtmäe**. Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.
165. **Liisa Metsamaa**. Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.

166. **Pille Säälik**. The role of endocytosis in the protein transduction by cell-penetrating peptides. Tartu, 2009, 155 p.
167. **Lauri Peil**. Ribosome assembly factors in *Escherichia coli.* Tartu, 2009, 147 p.
168. **Lea Hallik**. Generality and specificity in light harvesting, carbon gain capacity and shade tolerance among plant functional groups. Tartu, 2009, 99 p.
169. **Mariliis Tark**. Mutagenic potential of DNA damage repair and tolerance mechanisms under starvation stress. Tartu, 2009, 191 p.
170. **Riinu Rannap**. Impacts of habitat loss and restoration on amphibian populations. Tartu, 2009, 117 p.
171. **Maarja Adojaan**. Molecular variation of HIV-1 and the use of this knowledge in vaccine development. Tartu, 2009, 95 p.
172. **Signe Altmäe**. Genomics and transcriptomics of human induced ovarian folliculogenesis. Tartu, 2010, 179 p.
173. **Triin Suvi**. Mycorrhizal fungi of native and introduced trees in the Seychelles Islands. Tartu, 2010, 107 p.
174. **Velda Lauringson**. Role of suspension feeding in a brackish-water coastal sea. Tartu, 2010, 123 p.
175. **Eero Talts**. Photosynthetic cyclic electron transport – measurement and variably proton-coupled mechanism. Tartu, 2010, 121 p.
176. **Mari Nelis**. Genetic structure of the Estonian population and genetic distance from other populations of European descent. Tartu, 2010, 97 p.
177. **Kaarel Krjutškov**. Arrayed Primer Extension-2 as a multiplex PCR-based method for nucleic acid variation analysis: method and applications. Tartu, 2010, 129 p.
178. **Egle Köster**. Morphological and genetical variation within species complexes: *Anthyllis vulneraria* s. l. and *Alchemilla vulgaris* (coll.). Tartu, 2010, 101 p.
179. **Erki Õunap**. Systematic studies on the subfamily Sterrhinae (Lepidoptera: Geometridae). Tartu, 2010, 111 p.
180. **Merike Jõesaar**. Diversity of key catabolic genes at degradation of phenol and *p*-cresol in pseudomonads. Tartu, 2010, 125 p.
181. **Kristjan Herkül**. Effects of physical disturbance and habitat-modifying species on sediment properties and benthic communities in the northern Baltic Sea. Tartu, 2010, 123 p.
182. **Arto Pulk**. Studies on bacterial ribosomes by chemical modification approaches. Tartu, 2010, 161 p.
183. **Maria Põllupüü**. Ecological relations of cladocerans in a brackish-water ecosystem. Tartu, 2010, 126 p.
184. **Toomas Silla**. Study of the segregation mechanism of the Bovine Papillomavirus Type 1. Tartu, 2010, 188 p.
185. **Gyaneshwer Chaubey**. The demographic history of India: A perspective based on genetic evidence. Tartu, 2010, 184 p.

186. **Katrin Kepp**. Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.
187. **Virve Sõber**. The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.
188. **Kersti Kangro**. The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.
189. **Joachim M. Gerhold**. Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.
190. **Helen Tammert**. Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.
191. **Elle Rajandu**. Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.
192. **Paula Ann Kivistik**. ColR-ColS signalling system and transposition of Tn*4652* in the adaptation of *Pseudomonas putida.* Tartu, 2010, 118 p.
193. **Siim Sõber**. Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.
194. **Kalle Kipper**. Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.
195. **Triinu Siibak**. Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.
196. **Tambet Tõnissoo**. Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.
197. **Helin Räägel**. Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus**. Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius**. Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värv**. Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Välk**. Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.
202. **Arno Põllumäe**. Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.
203. **Egle Tammeleht**. Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.
205. **Teele Jairus**. Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.

206. **Kessy Abarenkov**. PlutoF – cloud database and computing services supporting biological research. Tartu, 2011, 125 p.
207. **Marina Grigorova**. Fine-scale genetic variation of follicle-stimulating hormone beta-subunit coding gene (*FSHB*) and its association with reproductive health. Tartu, 2011, 184 p.
208. **Anu Tiitsaar**. The effects of predation risk and habitat history on butterfly communities. Tartu, 2011, 97 p.
209. **Elin Sild**. Oxidative defences in immunoecological context: validation and application of assays for nitric oxide production and oxidative burst in a wild passerine. Tartu, 2011, 105 p.
210. **Irja Saar**. The taxonomy and phylogeny of the genera *Cystoderma* and *Cystodermella* (Agaricales, Fungi). Tartu, 2012, 167 p.
211. **Pauli Saag**. Natural variation in plumage bacterial assemblages in two wild breeding passerines. Tartu, 2012, 113 p.
212. **Aleksei Lulla**. Alphaviral nonstructural protease and its polyprotein substrate: arrangements for the perfect marriage. Tartu, 2012, 143 p.
213. **Mari Järve**. Different genetic perspectives on human history in Europe and the Caucasus: the stories told by uniparental and autosomal markers. Tartu, 2012, 119 p.
214. **Ott Scheler**. The application of tmRNA as a marker molecule in bacterial diagnostics using microarray and biosensor technology. Tartu, 2012, 93 p.
215. **Anna Balikova**. Studies on the functions of tumor-associated mucin-like leukosialin (CD43) in human cancer cells. Tartu, 2012, 129 p.
216. **Triinu Kõressaar**. Improvement of PCR primer design for detection of prokaryotic species. Tartu, 2012, 83 p.
217. **Tuul Sepp**. Hematological health state indices of greenfinches: sources of individual variation and responses to immune system manipulation. Tartu, 2012, 117 p.
218. **Rya Ero**. Modifier view of the bacterial ribosome. Tartu, 2012, 146 p.
219. **Mohammad Bahram**. Biogeography of ectomycorrhizal fungi across different spatial scales. Tartu, 2012, 165 p.
220. **Annely Lorents**. Overcoming the plasma membrane barrier: uptake of amphipathic cell-penetrating peptides induces influx of calcium ions and downstream responses. Tartu, 2012, 113 p.
221. **Katrin Männik**. Exploring the genomics of cognitive impairment: whole-genome SNP genotyping experience in Estonian patients and general population. Tartu, 2012, 171 p.
222. **Marko Prous**. Taxonomy and phylogeny of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae). Tartu, 2012, 192 p.
223. **Triinu Visnapuu**. Levansucrases encoded in the genome of *Pseudomonas syringae* pv. tomato DC3000: heterologous expression, biochemical characterization, mutational analysis and spectrum of polymerization products. Tartu, 2012, 160 p.
224. **Nele Tamberg**. Studies on Semliki Forest virus replication and pathogenesis. Tartu, 2012, 109 p.

225. **Tõnu Esko**. Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies. Tartu, 2012, 149 p.
226. **Timo Arula**. Ecology of early life-history stages of herring *Clupea harengus membras* in the northeastern Baltic Sea. Tartu, 2012, 143 p.
227. **Inga Hiiesalu**. Belowground plant diversity and coexistence patterns in grassland ecosystems. Tartu, 2012, 130 p.
228. **Kadri Koorem**. The influence of abiotic and biotic factors on small-scale plant community patterns and regeneration in boreonemoral forest. Tartu, 2012, 114 p.
229. **Liis Andresen**. Regulation of virulence in plant-pathogenic pectobacteria. Tartu, 2012, 122 p.
230. **Kaupo Kohv**. The direct and indirect effects of management on boreal forest structure and field layer vegetation. Tartu, 2012, 124 p.
231. **Mart Jüssi**. Living on an edge: landlocked seals in changing climate. Tartu, 2012, 114 p.
232. **Riina Klais**. Phytoplankton trends in the Baltic Sea. Tartu, 2012, 136 p.
233. **Rauno Veeroja**. Effects of winter weather, population density and timing of reproduction on life-history traits and population dynamics of moose (*Alces alces*) in Estonia. Tartu, 2012, 92 p.
234. **Marju Keis**. Brown bear (*Ursus arctos*) phylogeography in northern Eurasia. Tartu, 2013, 142 p.
235. **Sergei Põlme**. Biogeography and ecology of *alnus*- associated ectomycorrhizal fungi – from regional to global scale. Tartu, 2013, 90 p.
236. **Liis Uusküla**. Placental gene expression in normal and complicated pregnancy. Tartu, 2013, 173 p.
237. **Marko Lõoke**. Studies on DNA replication initiation in *Saccharomyces cerevisiae.* Tartu, 2013, 112 p.
238. **Anne Aan**. Light- and nitrogen-use and biomass allocation along productivity gradients in multilayer plant communities. Tartu, 2013, 127 p.
239. **Heidi Tamm**. Comprehending phylogenetic diversity – case studies in three groups of ascomycetes. Tartu, 2013, 136 p.
240. **Liina Kangur**. High-Pressure Spectroscopy Study of Chromophore-Binding Hydrogen Bonds in Light-Harvesting Complexes of Photosynthetic Bacteria. Tartu, 2013, 150 p.
241. **Margus Leppik**. Substrate specificity of the multisite specific pseudouridine synthase RluD. Tartu, 2013, 111 p.
242. **Lauris Kaplinski**. The application of oligonucleotide hybridization model for PCR and microarray optimization. Tartu, 2013, 103 p.
243. **Merli Pärnoja**. Patterns of macrophyte distribution and productivity in coastal ecosystems: effect of abiotic and biotic forcing. Tartu, 2013, 155 p.
244. **Tõnu Margus**. Distribution and phylogeny of the bacterial translational GTPases and the Mqsr/YgiT regulatory system. Tartu, 2013, 126 p.
245. **Pille Mänd**. Light use capacity and carbon and nitrogen budget of plants: remote assessment and physiological determinants. Tartu, 2013, 128 p.

246. **Mario Plaas**. Animal model of Wolfram Syndrome in mice: behavioural, biochemical and psychopharmacological characterization. Tartu, 2013, 144 p.
247. **Georgi Hudjašov**. Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations. Tartu, 2013, 115 p.
248. **Mari Lepik**. Plasticity to light in herbaceous plants and its importance for community structure and diversity. Tartu, 2013, 102 p.
249. **Ede Leppik**. Diversity of lichens in semi-natural habitats of Estonia. Tartu, 2013, 151 p.
250. **Ülle Saks**. Arbuscular mycorrhizal fungal diversity patterns in boreonemoral forest ecosystems. Tartu, 2013, 151 p.
251. **Eneli Oitmaa**. Development of arrayed primer extension microarray assays for molecular diagnostic applications. Tartu, 2013, 147 p.
252. **Jekaterina Jutkina**. The horizontal gene pool for aromatics degradation: bacterial catabolic plasmids of the Baltic Sea aquatic system. Tartu, 2013, 121 p.
253. **Helen Vellau**. Reaction norms for size and age at maturity in insects: rules and exceptions. Tartu, 2014, 132 p.
254. **Randel Kreitsberg**. Using biomarkers in assessment of environmental contamination in fish – new perspectives. Tartu, 2014, 107 p.
255. **Krista Takkis**. Changes in plant species richness and population performance in response to habitat loss and fragmentation.Tartu, 2014, 141 p.
256. **Liina Nagirnaja**. Global and fine-scale genetic determinants of recurrent pregnancy loss. Tartu, 2014, 211 p.
257. **Triin Triisberg**. Factors influencing the re-vegetation of abandoned extracted peatlands in Estonia. Tartu, 2014, 133 p.
258. **Villu Soon**. A phylogenetic revision of the *Chrysis ignita* species group (Hymenoptera: Chrysididae) with emphasis on the northern European fauna. Tartu, 2014, 211 p.
259. **Andrei Nikonov**. RNA-Dependent RNA Polymerase Activity as a Basis for the Detection of Positive-Strand RNA Viruses by Vertebrate Host Cells. Tartu, 2014, 207 p.
260. **Eele Õunapuu-Pikas**. Spatio-temporal variability of leaf hydraulic conductance in woody plants: ecophysiological consequences. Tartu, 2014, 135 p.
261. **Marju Männiste**. Physiological ecology of greenfinches: information content of feathers in relation to immune function and behavior. Tartu, 2014, 121 p.
262. **Katre Kets**. Effects of elevated concentrations of CO_2 and O_3 on leaf photosynthetic parameters in *Populus tremuloides*: diurnal, seasonal and interannual patterns. Tartu, 2014, 115 p.
263. **Külli Lokko**. Seasonal and spatial variability of zoopsammon communities in relation to environmental parameters. Tartu, 2014, 129 p.
264. **Olga Žilina**. Chromosomal microarray analysis as diagnostic tool: Estonian experience. Tartu, 2014, 152 p.

265. **Kertu Lõhmus**. Colonisation ecology of forest-dwelling vascular plants and the conservation value of rural manor parks. Tartu, 2014, 111 p.
266. **Anu Aun**. Mitochondria as integral modulators of cellular signaling. Tartu, 2014, 167 p.
267. **Chandana Basu Mallick**. Genetics of adaptive traits and gender-specific demographic processes in South Asian populations. Tartu, 2014, 160 p.
268. **Riin Tamme**. The relationship between small-scale environmental heterogeneity and plant species diversity. Tartu, 2014, 130 p.
269. **Liina Remm**. Impacts of forest drainage on biodiversity and habitat quality: implications for sustainable management and conservation. Tartu, 2015, 126 p.
270. **Tiina Talve**. Genetic diversity and taxonomy within the genus *Rhinanthus.* Tartu, 2015, 106 p.
271. **Mehis Rohtla**. Otolith sclerochronological studies on migrations, spawning habitat preferences and age of freshwater fishes inhabiting the Baltic Sea. Tartu, 2015, 137 p.
272. **Alexey Reshchikov**. The world fauna of the genus *Lathrolestes* (Hymenoptera, Ichneumonidae). Tartu, 2015, 247 p.
273. **Martin Pook**. Studies on artificial and extracellular matrix protein-rich surfaces as regulators of cell growth and differentiation. Tartu, 2015, 142 p.
274. **Mai Kukumägi**. Factors affecting soil respiration and its components in silver birch and Norway spruce stands. Tartu, 2015, 155 p.
275. **Helen Karu**. Development of ecosystems under human activity in the North-East Estonian industrial region: forests on post-mining sites and bogs. Tartu, 2015, 152 p.
276. **Hedi Peterson**. Exploiting high-throughput data for establishing relationships between genes. Tartu, 2015, 186 p.
277. **Priit Adler**. Analysis and visualisation of large scale microarray data, Tartu, 2015, 126 p.
278. **Aigar Niglas**. Effects of environmental factors on gas exchange in deciduous trees: focus on photosynthetic water-use efficiency. Tartu, 2015, 152 p.
279. **Silja Laht**. Classification and identification of conopeptides using profile hidden Markov models and position-specific scoring matrices. Tartu, 2015, 100 p.
280. **Martin Kesler**. Biological characteristics and restoration of Atlantic salmon *Salmo salar* populations in the Rivers of Northern Estonia. Tartu, 2015, 97 p.
281. **Pratyush Kumar Das**. Biochemical perspective on alphaviral nonstructural protein 2: a tale from multiple domains to enzymatic profiling. Tartu, 2015, 205 p
282. **Priit Palta**. Computational methods for DNA copy number detection. Tartu, 2015, 130 p.
283. **Julia Sidorenko**. Combating DNA damage and maintenance of genome integrity in pseudomonads. Tartu, 2015, 174 p.

284. **Anastasiia Kovtun-Kante**. Charophytes of Estonian inland and coastal waters: distribution and environmental preferences. Tartu, 2015, 97 p.
285. **Ly Lindman**. The ecology of protected butterfly species in Estonia. Tartu, 2015, 171 p.
286. **Jaanis Lodjak**. Association of Insulin-like Growth Factor I and Corticosterone with Nestling Growth and Fledging Success in Wild Passerines. Tartu, 2016, 113 p.
287. **Ann Kraut**. Conservation of Wood-Inhabiting Biodiversity – Semi-Natural Forests as an Opportunity. Tartu, 2016, 141 p.
288. **Tiit Örd**. Functions and regulation of the mammalian pseudokinase TRIB3. Tartu, 2016, 182. p.
289. **Kairi Käiro**. Biological Quality According to Macroinvertebrates in Streams of Estonia (Baltic Ecoregion of Europe): Effects of Human-induced Hydromorphological Changes. Tartu, 2016, 126 p.
290. **Leidi Laurimaa**. *Echinococcus multilocularis* and other zoonotic parasites in Estonian canids. Tartu, 2016, 144 p.
291. **Helerin Margus**. Characterization of cell-penetrating peptide/nucleic acid nanocomplexes and their cell-entry mechanisms. Tartu, 2016, 173 p.
292. **Kadri Runnel**. Fungal targets and tools for forest conservation. Tartu, 2016, 157 p.
293. **Urmo Võsa**. MicroRNAs in disease and health: aberrant regulation in lung cancer and association with genomic variation. Tartu, 2016, 163 p.
294. **Kristina Mäemets-Allas**. Studies on cell growth promoting AKT signaling pathway – a promising anti-cancer drug target. Tartu, 2016, 146 p.
295. **Janeli Viil**. Studies on cellular and molecular mechanisms that drive normal and regenerative processes in the liver and pathological processes in Dupuytren's contracture. Tartu, 2016, 175 p.
296. **Ene Kook**. Genetic diversity and evolution of *Pulmonaria angustifolia* L. and *Myosotis laxa sensu lato* (Boraginaceae). Tartu, 2016, 106 p.
297. **Kadri Peil**. RNA polymerase II-dependent transcription elongation in *Saccharomyces cerevisiae*. Tartu, 2016, 113 p.
298. **Katrin Ruisu**. The role of RIC8A in mouse development and its function in cell-matrix adhesion and actin cytoskeletal organisation. Tartu, 2016, 129 p.
299. **Janely Pae**. Translocation of cell-penetrating peptides across biological membranes and interactions with plasma membrane constituents. Tartu, 2016, 126 p.
300. **Argo Ronk**. Plant diversity patterns across Europe: observed and dark diversity. Tartu, 2016, 153 p.
301. **Kristiina Mark**. Diversification and species delimitation of lichenized fungi in selected groups of the family Parmeliaceae (Ascomycota). Tartu, 2016, 181 p.
302. **Jaak-Albert Metsoja**. Vegetation dynamics in floodplain meadows: influence of mowing and sediment application. Tartu, 2016, 140 p.

303. **Hedvig Tamman**. The GraTA toxin-antitoxin system of *Pseudomonas putida*: regulation and role in stress tolerance. Tartu, 2016, 154 p.
304. **Kadri Pärtel**. Application of ultrastructural and molecular data in the taxonomy of helotialean fungi. Tartu, 2016, 183 p.
305. **Maris Hindrikson**. Grey wolf (*Canis lupus*) populations in Estonia and Europe: genetic diversity, population structure and -processes, and hybridization between wolves and dogs. Tartu, 2016, 121 p.
306. **Polina Degtjarenko**. Impacts of alkaline dust pollution on biodiversity of plants and lichens: from communities to genetic diversity. Tartu, 2016, 126 p.
307. **Liina Pajusalu**. The effect of CO_2 enrichment on net photosynthesis of macrophytes in a brackish water environment. Tartu, 2016, 126 p.
308. **Stoyan Tankov**. Random walks in the stringent response. Tartu, 2016, 94 p.
309. **Liis Leitsalu**. Communicating genomic research results to population-based biobank participants. Tartu, 2016, 158 p.
310. **Richard Meitern**. Redox physiology of wild birds: validation and application of techniques for detecting oxidative stress. Tartu, 2016, 134 p.
311. **Kaie Lokk**. Comparative genome-wide DNA methylation studies of healthy human tissues and non-small cell lung cancer tissue. Tartu, 2016, 127 p.
312. **Mihhail Kurašin**. Processivity of cellulases and chitinases. Tartu, 2017, 132 p.
313. **Carmen Tali**. Scavenger receptors as a target for nucleic acid delivery with peptide vectors. Tartu, 2017, 155 p.
314. **Katarina Oganjan**. Distribution, feeding and habitat of benthic suspension feeders in a shallow coastal sea. Tartu, 2017, 132 p.
315. **Taavi Paal**. Immigration limitation of forest plants into wooded landscape corridors. Tartu, 2017, 145 p.
316. **Kadri Õunap**. The Williams-Beuren syndrome chromosome region protein WBSCR22 is a ribosome biogenesis factor. Tartu, 2017, 135 p.
317. **Riin Tamm**. In-depth analysis of factors affecting variability in thiopurine methyltransferase activity. Tartu, 2017, 170 p.
318. **Keiu Kask**. The role of RIC8A in the development and regulation of mouse nervous system. Tartu, 2017, 184 p.
319. **Tiia Möller**. Mapping and modelling of the spatial distribution of benthic macrovegetation in the NE Baltic Sea with a special focus on the eelgrass *Zostera marina* Linnaeus, 1753. Tartu, 2017, 162 p.
320. **Silva Kasela**. Genetic regulation of gene expression: detection of tissue- and cell type-specific effects. Tartu, 2017, 150 p.
321. **Karmen Süld**. Food habits, parasites and space use of the raccoon dog *Nyctereutes procyonoides*: the role of an alien species as a predator and vector of zoonotic diseases in Estonia. Tartu, 2017, p.
322. **Ragne Oja**. Consequences of supplementary feeding of wild boar – concern for ground-nesting birds and endoparasite infection. Tartu, 2017, 141 p.
323. **Riin Kont**. The acquisition of cellulose chain by a processive cellobiohydrolase. Tartu, 2017, 117 p.

324. **Liis Kasari**. Plant diversity of semi-natural grasslands: drivers, current status and conservation challenges. Tartu, 2017, 141 p.
325. **Sirgi Saar**. Belowground interactions: the roles of plant genetic relatedness, root exudation and soil legacies. Tartu, 2017, 113 p.
326. **Sten Anslan**. Molecular identification of Collembola and their fungal associates. Tartu, 2017, 125 p.
327. **Imre Taal**. Causes of variation in littoral fish communities of the Eastern Baltic Sea: from community structure to individual life histories. Tartu, 2017, 118 p.
328. **Jürgen Jalak**. Dissecting the Mechanism of Enzymatic Degradation of Cellulose Using Low Molecular Weight Model Substrates. Tartu, 2017, 137 p.
329. **Kairi Kiik**. Reproduction and behaviour of the endangered European mink (*Mustela lutreola*) in captivity. Tartu, 2018, 112 p.
330. **Ivan Kuprijanov**. Habitat use and trophic interactions of native and invasive predatory macroinvertebrates in the northern Baltic Sea. Tartu, 2018, 117 p.
331. **Hendrik Meister**. Evolutionary ecology of insect growth: from geographic patterns to biochemical trade-offs. Tartu, 2018, 147 p.
332. **Ilja Gaidutšik**. Irc3 is a mitochondrial branch migration enzyme in *Saccharomyces cerevisiae.* Tartu, 2018, 161 p.
333. **Lena Neuenkamp**. The dynamics of plant and arbuscular mycorrhizal fungal communities in grasslands under changing land use. Tartu, 2018, 241 p.
334. **Laura Kasak**. Genome structural variation modulating the placenta and pregnancy maintenance. Tartu, 2018, 181 p.
335. **Kersti Riibak**. Importance of dispersal limitation in determining dark diversity of plants across spatial scales. Tartu, 2018, 133 p.
336. **Liina Saar**. Dynamics of grassland plant diversity in changing landscapes. Tartu, 2018, 206 p.
337. **Hanna Ainelo**. Fis regulates *Pseudomonas putida* biofilm formation by controlling the expression of *lapA*. Tartu, 2018, 143 p.
338. **Natalia Pervjakova**. Genomic imprinting in complex traits. Tartu, 2018, 176 p.
339. **Andrio Lahesaare**. The role of global regulator Fis in regulating the expression of *lapF* and the hydrophobicity of soil bacterium *Pseudomonas putida*. Tartu, 2018, 124 p.
340. **Märt Roosaare**. *K*-mer based methods for the identification of bacteria and plasmids. Tartu, 2018, 117 p.
341. **Maria Abakumova**. The relationship between competitive behaviour and the frequency and identity of neighbours in temperate grassland plants. Tartu, 2018, 104 p.
342. **Margus Vilbas**. Biotic interactions affecting habitat use of myrmecophilous butterflies in Northern Europe. Tartu, 2018, 142 p.

343. **Liina Kinkar**. Global patterns of genetic diversity and phylogeography of *Echinococcus granulosus* sensu stricto – a tapeworm species of significant public health concern. Tartu, 2018, 147 p.
344. **Teivi Laurimäe**. Taxonomy and genetic diversity of zoonotic tapeworms in the species complex of *Echinococcus granulosus* sensu lato. Tartu, 2018, 143 p.
345. **Tatjana Jatsenko**. Role of translesion DNA polymerases in mutagenesis and DNA damage tolerance in Pseudomonads. Tartu, 2018, 216 p.
346. **Katrin Viigand**. Utilization of α-glucosidic sugars by *Ogataea* (*Hansenula*) *polymorpha.* Tartu, 2018, 148 p.
347. **Andres Ainelo**. Physiological effects of the *Pseudomonas putida* toxin grat. Tartu, 2018, 146 p.
348. **Killu Timm**. Effects of two genes (DRD4 and SERT) on great tit (*Parus major*) behaviour and reproductive traits. Tartu, 2018, 117 p.
349. **Petr Kohout**. Ecology of ericoid mycorrhizal fungi. Tartu, 2018, 184 p.
350. **Gristin Rohula-Okunev**. Effects of endogenous and environmental factors on night-time water flux in deciduous woody tree species. Tartu, 2018, 184 p.
351. **Jane Oja**. Temporal and spatial patterns of orchid mycorrhizal fungi in forest and grassland ecosystems. Tartu, 2018, 102 p.
352. **Janek Urvik**. Multidimensionality of aging in a long-lived seabird. Tartu, 2018, 135 p.
353. **Lisanna Schmidt**. Phenotypic and genetic differentiation in the hybridizing species pair *Carex flava* and *C. viridula* in geographically different regions. Tartu, 2018, 133 p.
354. **Monika Karmin**. Perspectives from human Y chromosome – phylogeny, population dynamics and founder events. Tartu, 2018, 168 p.
355. **Maris Alver**. Value of genomics for atherosclerotic cardiovascular disease risk prediction. Tartu, 2019, 148 p.
356. **Lehti Saag**. The prehistory of Estonia from a genetic perspective: new insights from ancient DNA. Tartu, 2019, 171 p.
357. **Mari-Liis Viljur**. Local and landscape effects on butterfly assemblages in managed forests. Tartu, 2019, 115 p.
358. **Ivan Kisly**. The pleiotropic functions of ribosomal proteins eL19 and eL24 in the budding yeast ribosome. Tartu, 2019, 170 p.
359. **Mikk Puustusmaa**. On the origin of papillomavirus proteins. Tartu, 2019, 152 p.
360. **Anneliis Peterson**. Benthic biodiversity in the north-eastern Baltic Sea: mapping methods, spatial patterns, and relations to environmental gradients. Tartu, 2019, 159 p.
361. **Erwan Pennarun**. Meandering along the mtDNA phylogeny; causerie and digression about what it can tell us about human migrations. Tartu, 2019, 162 p.

362. **Karin Ernits**. Levansucrase Lsc3 and endo-levanase BT1760: characterization and application for the synthesis of novel prebiotics. Tartu, 2019, 217 p.
363. **Sille Holm**. Comparative ecology of geometrid moths: in search of contrasts between a temperate and a tropical forest. Tartu, 2019, 135 p.
364. **Anne-Mai Ilumäe**. Genetic history of the Uralic-speaking peoples as seen through the paternal haplogroup N and autosomal variation of northern Eurasians. Tartu, 2019, 172 p.
365. **Anu Lepik**. Plant competitive behaviour: relationships with functional traits and soil processes. Tartu, 2019, 152 p.
366. **Kunter Tätte**. Towards an integrated view of escape decisions in birds under variable levels of predation risk. Tartu, 2020, 172 p.
367. **Kaarin Parts**. The impact of climate change on fine roots and root-associated microbial communities in birch and spruce forests. Tartu, 2020, 143 p.
368. **Viktorija Kukuškina**. Understanding the mechanisms of endometrial receptivity through integration of 'omics' data layers. Tartu, 2020, 169 p.
369. **Martti Vasar**. Developing a bioinformatics pipeline gDAT to analyse arbuscular mycorrhizal fungal communities using sequence data from different marker regions. Tartu, 2020, 193 p.
370. **Ott Kangur**. Nocturnal water relations and predawn water potential disequilibrium in temperate deciduous tree species. Tartu, 2020, 126 p.
371. **Helen Post**. Overview of the phylogeny and phylogeography of the Y-chromosomal haplogroup N in northern Eurasia and case studies of two linguistically exceptional populations of Europe – Hungarians and Kalmyks. Tartu, 2020, 143 p.
372. **Kristi Krebs**. Exploring the genetics of adverse events in pharmacotherapy using Biobanks and Electronic Health Records. Tartu, 2020, 151 p.
373. **Kärt Ukkivi**. Mutagenic effect of transcription and transcription-coupled repair factors in *Pseudomonas putida.* Tartu, 2020, 154 p.
374. **Elin Soomets**. Focal species in wetland restoration. Tartu, 2020, 137 p.
375. **Kadi Tilk**. Signals and responses of ColRS two-component system in *Pseudomonas putida.* Tartu, 2020, 133 p.
376. **Indrek Teino**. Studies on aryl hydrocarbon receptor in the mouse granulosa cell model. Tartu, 2020, 139 p.
377. **Maarja Vaikre**. The impact of forest drainage on macroinvertebrates and amphibians in small waterbodies and opportunities for cost-effective mitigation. Tartu, 2020, 132 p.
378. **Siim-Kaarel Sepp**. Soil eukaryotic community responses to land use and host identity. Tartu, 2020, 222 p.
379. **Eveli Otsing**. Tree species effects on fungal richness and community structure. Tartu, 2020, 152 p.
380. **Mari Pent**. Bacterial communities associated with fungal fruitbodies. Tartu, 2020, 144 p.

381. **Einar Kärgenberg**. Movement patterns of lithophilous migratory fish in free-flowing and fragmented rivers. Tartu, 2020, 167 p.
382. **Antti Matvere**. The studies on aryl hydrocarbon receptor in murine granulosa cells and human embryonic stem cells. Tartu, 2021, 163 p.
383. **Jhonny Capichoni Massante**. Phylogenetic structure of plant communities along environmental gradients: a macroecological and evolutionary approach. Tartu, 2021, 144 p.
384. **Ajai Kumar Pathak**. Delineating genetic ancestries of people of the Indus Valley, Parsis, Indian Jews and Tharu tribe. Tartu, 2021, 197 p.
385. **Tanel Vahter**. Arbuscular mycorrhizal fungal biodiversity for sustainable agroecosystems. Tartu, 2021, 191 p.
386. **Burak Yelmen**. Characterization of ancient Eurasian influences within modern human genomes. Tartu, 2021, 134 p.
387. **Linda Ongaro**. A genomic portrait of American populations. Tartu, 2021, 182 p.
388. **Kairi Raime**. The identification of plant DNA in metagenomic samples. Tartu, 2021, 108 p.
389. **Heli Einberg**. Non-linear and non-stationary relationships in the pelagic ecosystem of the Gulf of Riga (Baltic Sea). Tartu, 2021, 119 p.
390. **Mickaël Mathieu Pihain**. The evolutionary effect of phylogenetic neighbourhoods of trees on their resistance to herbivores and climatic stress. Tartu, 2022, 145 p.
391. **Annika Joy Meitern**. Impact of potassium ion content of xylem sap and of light conditions on the hydraulic properties of trees. Tartu, 2022, 132 p.
392. **Elise Joonas**. Evaluation of metal contaminant hazard on microalgae with environmentally relevant testing strategies. Tartu, 2022, 118 p.
393. **Kreete Lüll**. Investigating the relationships between human microbiome, host factors and female health. Tartu, 2022, 141 p.
394. **Triin Kaasiku**. A wader perspective to Boreal Baltic coastal grasslands: from habitat availability to breeding site selection and nest survival. Tartu, 2022, 141 p.
395. **Meeli Alber**. Impact of elevated atmospheric humidity on the structure of the water transport pathway in deciduous trees. Tartu, 2022, 170 p.
396. **Ludovica Molinaro**. Ancestry deconvolution of Estonian, European and Worldwide genomic layers: a human population genomics excavation. Tartu, 2022, 138 p.
397. **Tina Saupe**. The genetic history of the Mediterranean before the common era: a focus on the Italian Peninsula. Tartu, 2022, 165 p.
398. **Mari-Ann Lind**. Internal constraints on energy processing and their consequences: an integrative study of behaviour, ornaments and digestive health in greenfinches. Tartu, 2022, 137 p.
399. **Markus Valge**. Testing the predictions of life history theory on anthropometric data. Tartu, 2022, 171 p.
400. **Ants Tull**. Domesticated and wild mammals as reservoirs for zoonotic helminth parasites in Estonia. Tartu, 2022, 152 p.

401. **Saleh Rahimlouye Barabi**. Investigation of diazotrophic bacteria association with plants. Tartu, 2022, 137 p.
402. **Farzad Aslani**. Towards revealing the biogeography of belowground diversity. Tartu, 2022, 124 p.
403. **Nele Taba**. Diet, blood metabolites, and health. Tartu, 2022, 163 p.
404. **Katri Pärna**. Improving the personalized prediction of complex traits and diseases: application to type 2 diabetes. Tartu, 2022, 190 p.
405. **Silva Lilleorg**. Bacterial ribosome heterogeneity on the example of bL31 paralogs in *Escherichia coli.* Tartu, 2022, 189 p.
406. **Oliver Aasmets**. The importance of microbiome in human health. Tartu, 2022, 123 p.
407. **Henel Jürgens**. Exploring post-translational modifications of histones in RNA polymerase II-dependent transcription. Tartu, 2022, 147 p.
408. **Mari Tagel**. Finding novel factors affecting the mutation frequency: a case study of tRNA modification enzymes TruA and RluA. Tartu, 2022, 176 p.
409. **Marili Sell**. The impact of environmental change on ecophysiology of hemiboreal tree species – acclimation mechanisms in belowground. Tartu, 2022, 163 p.
410. **Kaarin Hein**. The hissing behaviour of Great Tit (*Parus major*) females reflects behavioural phenotype and breeding success in a wild population. Tartu, 2022, 96 p.
411. **Maret Gerz**. The distribution and role of mycorrhizal symbiosis in plant communities. Tartu, 2022, 206 p.
412. **Kristiina Nõomaa**. Role of invasive species in brackish benthic community structure and biomass changes. Tartu, 2023, 151 p.
413. **Anton Savchenko**. Taxonomic studies in Dacrymycetes: *Cerinomyces* and allied taxa. Tartu, 2023, 181 p.
414. **Ahto Agan**. Interactions between invasive pathogens and resident mycobiome in the foliage of trees. Tartu, 2023, 155 p.
415. **Diego Pires Ferraz Trindade**. Dark diversity dynamics linked to global change: taxonomic and functional perspective. Tartu, 2023, 134 p.
416. **Madli Jõks**. Biodiversity drivers in oceanic archipelagos and habitat fragments, explored by agent-based simulation models. Tartu, 2023, 116 p.
417. **Ciara Baines**. Adaptation to oncogenic pollution and natural cancer defences in the aquatic environment. Tartu, 2023, 164 p.
418. **Rain Inno**. Placental transcriptome and miRNome in normal and complicated pregnancies. Tartu, 2023, 145 p.
419. **Daniyal Gohar**. Diversity, genomics, and potential functions of fungus-inhabiting bacteria. Tartu, 2023, 138 p.
420. **Sirli Rosendahl**. Fitness effects of chromosomal toxin-antitoxin systems in *Pseudomonas putida.* Tartu, 2023, 154 p.
421. **Mathilde Frédérique E. André**. New Guinea, a hotspot for Human evolution: settlement history and adaptation in northern Sahul. Tartu, 2023, 202 p.

422. **Vlad-Julian Piljukov**. Biochemical characterization of Irc3 helicase. Tartu, 2023, 137 p.
423. **Gerli Albert**. Carbon use strategies of macrophyte communities in the northeastern Baltic Sea: implications for a high CO_2 environment. Tartu, 2023, 128 p.
424. **Mariann Koel**. The molecular interactions between trophoblast and endometrial cells in embryo implantation. Tartu, 2023, 171 p.
425. **Robin Gielen**. Diversity and ecological role of pathogenic fungi in insect populations. Tartu, 2023, 139 p.
426. **Kaspar Reier**. Quantity, stability and disparity of ribosomal components in *Escherichia coli* stationary phase. Tartu, 2023, 151 p.
427. **Linda Rusalepp**. The impact of environmental drivers and competition on phenolic metabolite profiles in hybrid aspen and silver birch. Tartu, 2023, 153 p.
428. **Eliisa Pass**. The effect of managed forest-wetland landscapes on forest grouse and nest predation. Tartu, 2023, 115 p.
429. **Sanni Färkkilä**. Methods for studying plant-fungal interactions – reflecting on the old, the new and the upcoming. Tartu, 2024, 147 p.
430. **Maarja Jõeloo**. Advances in microarray-based copy number variation discovery and phenotypic associations. Tartu, 2024, 209 p.
431. **Natàlia Pujol Gualdo**. Decoding genetic associations of female reproductive health traits. Tartu, 2024, 205 p.
432. **Sirelin Sillamaa**. The role of helicases Hmi1 and Irc3 in yeast mitochondrial DNA maintenance. Tartu, 2024, 189 p.
433. **Iris Reinula**. Genetic variation of grassland plants in changing landscapes. Tartu, 2024, 201 p.
434. **Vi Ngan Tran**. The cellular dynamics and epithelial morphogenesis in *Drosophila* wing development. Tartu, 2024, 158 p.
435. **Slendy Julieth Rodríguez Alarcón**. Intraspecific trait diversity in plants: characterizing effects of trait variation on community assembly and ecosystem functioning. Tartu, 2024, 129 p.
436. **Arun Kumar Devarajan**. Microbes and climate change: insights from plant-microbe interactions in rice phyllosphere and soil microbiomes in subarctic grasslands. Tartu, 2024, 224 p.
437. **Leonard Owuraku Opare**. Rearing density effects on a commercially important insect species. Tartu, 2024, 145 p.
438. **Siqiao Liu**. The effect of anthropogenic disturbance on soil fungal communities. Tartu, 2024, 172 p.
439. **Kertu Liis Krigul**. The gut microbiome at the interface of human health and disease. Tartu, 2024, 158 p.
440. **Danat Yermakovich**. The evolutionary history of complex traits: implications of archaic admixture. Tartu, 2024, 153 p.
441. **Yiming Meng**. Plant mycorrhizal type and status in the global flora. Tartu, 2024, 200 p.
442. **Iryna Yatsiuk**. Evolution, species delimitation and diversity in myxomycetes: *Arcyria* and allied genera. Tartu, 2024, 193 p.

443. **Daniela León Velandia**. Mycorrhizal trait distribution and composition in plant communities under natural gradients. Tartu, 2024, 121 p.
444. **Bruno Paganeli**. Dark diversity methods for prioritization of areas and species in nature conservation. Tartu, 2024, 155 p.
445. **Mario Reiman**. Placental transcriptome in normal and complicated pregnancies. Tartu, 2025, 167 p.
446. **Maarja Kõrkjas**. Dynamics of tree-related microhabitats in live forest trees and its links with biodiversity. Tartu, 2025, 134 p.
447. **Eleonora Beccari**. Mapping and exploring trait spaces across the tree of life. Tartu, 2025, 190 p.
448. **Jack R. Hall**. Dissolved organic carbon dynamics of Baltic Sea macroalgae: production, bioavailability and ecosystem effects. Tartu, 2025, 135 p.
449. **Artjom Stepanjuk**. Function of adhesion molecules and signalling pathways in human endometrial and embryonic models. Tartu, 2025, 247 p.
450. **Marianne Kivastik**. Heterostylous plants in an era of global change: the role of local, landscape and climatic actors. Tartu, 2025, 167 p.
451. **Yehor Yatsiuk**. Large tree-cavities as key structures for forest biodiversity. Tartu, 2025, 215 p.
452. **Ovidiu Copoț**. Relevance of eDNA, citizen science, and species distribution modelling for fungal conservation. Tartu, 2025, 198 p.
453. **Tarmo Puurand**. Human genome studies with k-mer frequencies. Tartu, 2025, 184 p.
454. **Stênio Ítalo Araújo Foerster**. Phylogenetic comparative studies of body size in insects and arachnids: from predictions to applications. Tartu, 2025, 171 p.
455. **Hanna Maria Kariis**. Improving pharmacotherapy outcomes in psychiatric and cardiovascular conditions. Tartu, 2025, 193 p.
456. **Elisabeth Prangel**. The impact of land-use change and ecological restoration on biodiversity and ecosystem service supply in semi-natural grasslands. Tartu, 2025, 233 p.
457. **Nidal Fetnassi**. Determinants of moth assemblages across human-modified landscapes of Estonia and Morocco. Tartu, 2025, 162 p.
458. **Vineesh Nedumpally**. Assembling the phylogenetic tree of northern European macroheteroceran moths. Tartu, 2025, 171 p.
459. **Ali Hakimzadeh**. Long-read metabarcoding: from available tools to reference databases. Tartu, 2026, 124 p.
460. **John Yangyuoru Kupagme**. Biodiversity of African soil fungi. Tartu, 2026, 162 p.
461. **Bariş Yaşar**. Advanced chromosomal testing tools for embryo quality and fetal health. Tartu, 2026, 248 p.
462. **Biancamaria Bonucci**. Reading the archaeological record through ancient biomolecules: preservation, disease landscapes, and human-microbe interactions in the past. Tartu, 2026, 286 p.
463. **Stefania Sasso**. Population dynamics and health in medieval Europe: an archaeogenomic perspective. Tartu, 2026, 212 p.
464. **Harleen Kaur.** Performance of antimicrobial surfaces under application-relevant conditions. Tartu, 2026, 202 p.