

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Quantifying Assignable Microbial Content and Information Loss in Reference- and Assembly-Based Metagenomics

Permalink

<https://escholarship.org/uc/item/8x95607x>

ISBN

9798189226168

Author

Zhao, Zhiying

Publication Date

2026-06-25

Peer reviewed|Thesis/dissertation

Quantifying Assignable Microbial Content and Information Loss in Reference- and Assembly-Based
Metagenomics

By

ZHIYING ZHAO
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

C. Titus Brown, Chair

Tiffany Lowe-Power

Jonathan A. Eisen

Committee in Charge

2026

Quantifying Assignable Microbial Content and Information Loss in Reference- and Assembly-Based Metagenomics

Copyright @ 2026

by

Zhiying Zhao

Acknowledgements

The path to this PhD has been anything but linear. I stepped away early and chose to restart two decades later, with a renewed sense of purpose to finish what I had once begun.

This experience has filled challenges. While raising three children and working full time alongside my graduate studies, I navigated an intense and demanding life with periods of personal and family health difficulties. This dissertation represents not only my research, but also the continuation and fulfillment of a long personal journey.

I am deeply grateful to my advisor, Titus, for his instructive guidance, openness to exploration and trust in unconventional paths, fostered a supportive environment for learning and growth. I thank my committee members, Tiffany and Jonathan, for their timely review and thoughtful feedback, and three other members of the QE committee, Dennis, Sam, and Ben, for sharing their expertise and guidance. I also thank former IGG chair Sean and David for their understanding and support in accommodating my unique situation.

The members of the DIB Lab were invaluable. Tessa guided extensively on Sourmash; Annie contributed to the binning work; Colton developed the saturation analysis; and many others helped.

I am grateful to the LBL/JGI leadership team for supporting my Master's degree, an important step toward this PhD, and my previous academic advisors, Zonghua Wang and Robin Cameron, for their mentorship during my early training.

This journey would not have been possible without my family. I thank my mother and my parents-in-law for their unwavering help with childcare and daily life. To my husband and children, thank you for your love and support, you have been a steady source of strength.

I share this journey with my parents and my brother, who have deeply shaped who I am.

I am grateful to all of you who made it possible.

Abstract

Shotgun metagenomic sequencing enables culture-independent characterization of microbial communities, but most analyses rely on transformed representations of sequencing data rather than raw reads. Reference-based profiling, *de novo* assembly, and genome binning simplify complex datasets while implicitly filtering and reshaping the underlying sequence information. The extent to which these transformations alter observable microbial diversity, and whether biologically meaningful signals are systematically lost, remains poorly quantified across diverse ecosystems.

In this dissertation, I examine how microbial genome content is represented, lost, and biased across common metagenomic analysis workflows using large collections of publicly available metagenomes spanning host-associated, engineered, and environmental biomes. I first quantify the fraction of metagenomic content assignable to known reference genomes using scalable k-mer based methods, revealing strong biome-specific differences and persistent gaps in reference coverage for environmental systems despite continued database expansion. I then show that metagenome assembly removes a substantial fraction of sequence diversity, including strain-level variation that is collapsed during consensus reconstruction. Finally, I demonstrate that genome binning introduces an additional and severe bottleneck, with metagenome-assembled genomes capturing only a small and biased subset of the original sequence content and taxonomic diversity, even in simplified mock communities. Together, these results demonstrate that standard metagenomic representations act as lossy filters rather than comprehensive summaries of microbial communities, motivating the pursuit of integrative approaches that preserve read-level diversity.

Table of Contents

Abstract	iv
Table of Contents	v
List of Figures and Tables	vii
Introduction	1
Chapter 1. A survey of assignable microbial content across biomes	4
Abstract	4
Background	4
Methods	6
a) MGnify metagenome retrieval and dataset assembly	6
b) Sourmash and fastmultigather analysis	7
c) Random selection of 100 samples for genome-grist run	8
d) Comparison of k-mer based and mapping-based reference assignability	8
e) Reproducibility and Implementation	10
Results	11
1.1 Sample selection and host-derived DNA analysis	11
1.2 FMG Filtering and Reference-based Assignability analysis	16
1.3 Impact of GTDB Version Update on Reference Assignability	24
1.4 Validation of k-mer based assignability via read mapping	28
Discussion	32
1) Host-derived sequence as a structural constraint on reference assignability	32
2) Assessing known and unknown microbial genome content across biomes	36
3) Database expansion improves reference-based assignability	39
4) Validation of k-mer based assignability using genome-grist read mapping	40
Chapter 2. Metrics for measuring content loss in metagenome assembly	42
Abstract	42
Background	42
Methods	44
a) Experimental Design and Workflow Implementation	44
b) Conceptual definition of content loss	45
c) Analytical Framework	45
d) Assessing the impact of sequencing depth and breadth on assembly performance	47
Results	48
2.1 Computational workflow for assembly loss and recovery analysis	48
2.2 Pilot analysis of metagenome representations	50
2.3 Assembly-associated content loss of abundance-weighted k-mer signal	52
2.4 Recovery and expansion of sequence representation through assembly	53
2.5 Assembly retention is tightly constrained by sampling completeness	57
Discussion	58
1) Assembly preserves most abundance-weighted genomic signal despite measurable	

content loss	58
2) Read mapping reveals that much read-level content remains explainable by assemblies	60
3) Assembly complements reference-based representation of microbial communities	60
4) Implications for genome-centric metagenomic analysis	61
Chapter 3. Quantifying Information Loss During Genome Binning	64
Abstract	64
Background	64
Methods	66
a) Dataset collection	66
b) K-mer content and taxonomic coverage comparison across representations	67
c) Visualization of information flow across metagenomic representations	68
Results	69
3.1 Overview of comparative framework and rationale	69
3.2 Superkingdom-level retention across assembly and genome binning	69
3.3 Progressive loss of taxonomic signal across metagenomes	71
3.4 Phylum-level retention patterns across assembly and genome binning	73
Discussion	77
1) Information loss accumulates across genome-resolved metagenomic workflows	78
2) Taxonomic representation is uneven across microbial lineages	79
3) Genome binning captures dominant taxa but misses community diversity	79
4) Limitations and future directions	81
5) Conclusions	81
Supplementary material	83
Bibliography	87

List of Figures and Tables

Table 1.1: Sourmash manysearch output of eight hosts across all 2867 metagenomes.....	12
Figure 1.1: Scatter plot of eight host-derived signal levels in 2867 metagenomes.....	13
Figure 1.2: Five representative host-derived DNA levels across metagenomes stratified by biome category.....	14
Figure 1.3: Distribution of multi-host signals across metagenomes (>1% threshold).....	15
Table 1.2: Frequency of metagenomes with multi-host signals at increasing thresholds.....	16
Figure 1.4: Workflow for host and non-microbial sample filtering.....	17
Table 1.3: Aggregated Sourmash fastmultigather results before and after the filtering.....	20
Figure 1.5: Biome-specific genome content assignability using GTDB (rs220, k = 21).....	22
Table 1.4: Reference content comparison between GTDB rs220 and rs226.....	25
Figure 1.6: Comparison of total k-mer match counts with GTDB rs220 vs rs226.....	26
Figure 1.7. Comparison of FMG weighted reference fraction using GTDB rs220 and rs226.....	27
Figure 1.8: Workflow of sample selection for Genome-grist analysis.....	29
Figure 1.9: Pairwise comparisons of genome-grist, FMG, and mapping.....	30
Figure 2.1: Computational framework for quantifying sequence retention, loss, and recovery during metagenome assembly.....	49
Figure 2.2: Pilot comparison of sequence representations across four different types of metagenomes.....	51
Figure 2.3: Assembly-associated signal loss across biomes.....	52
Figure 2.4: Read mapping partially recovers abundance-weighted sequence content lost during assembly.....	54
Figure 2.5: Novel abundant sequence recovered by assembly beyond reference-matched content.....	56
Figure 2.6: Relationship between sequencing saturation and assembly retention across metagenomes.....	58
Figure 3.1: Snakemake workflow for k-mer and taxonomic comparison across metagenomic representations.....	68
Figure 3.2: Superkingdom-level retention across 897 metagenome assembly and genome binning.....	70
Figure 3.3: Partitioning of phylum-level taxonomic signals across genome-resolved metagenomic workflows.....	72
Table 3.1: Phylum-level retention across assembly and binning.....	74
Figure 3.4: Taxonomic retention of top 8 phyla across assembly and binning.....	75
Figure 3.5: Median phylum-level retention relative to reads for contigs and bins.....	76
Supplementary Table 1.1: Samples with the highest mouse- and human-derived signals and their study-level grouping.....	83
Supplementary Figure 1.1: Additional abundance-weighted signal recovered by read mapping across biomes.....	84
Supplementary Figure 2.1: Example of Atta Snakefile Upset diagram.....	85
Supplementary Figure 2.2: Assembly Retention vs. Total/Novel Read Count.....	86

Introduction

Microbial communities drive processes essential to human health, agriculture, industry, and global biogeochemical cycles^{1,2}. Despite their central role, the vast majority of microbial diversity remains poorly understood. Traditional isolation-based approaches, which rely on growing microbes in laboratory culture and studying their physiology, genetics and genomes, have provided critical insights but capture only a small fraction of microbial life³. Most microorganisms cannot be easily isolated and cultured *in vitro* due to their dependence on complex natural habitats, leaving them inaccessible to standard laboratory techniques⁴. As a result, the majority of microbial diversity remains uncharacterized, representing microbial “dark matter” that continues to challenge discovery and classification⁵.

Advances in high-throughput sequencing have made it possible to probe microbial communities directly through metagenomics, bypassing cultivation^{6,7}. This has opened new avenues for the investigation of microbiome taxonomic composition, the study of evolution, the prediction of functional genes, the interpretation of their interactions, and metabolic potentials, from almost everywhere on earth (soil, ocean, permafrost, Antarctica, etc.)^{8–16}, even in the built environment associated with space exploration such as spacecraft and the International Space Station^{17,18}.

Most previous approaches for taxonomic and functional profiling of metagenomes have relied heavily on the availability of reference genomes. These methods typically align sequencing reads directly to curated genome collections, marker gene datasets, or species-specific sequences for taxonomic assignments. Read alignment approaches, in particular, provide high specificity and interpretability when suitable references exist^{19,20}. In addition, k-mer frequency patterns and the read depth of universal single-copy genes are often used to estimate relative abundance²¹.

While reference-based methods are efficient and widely adopted, their performance is fundamentally constrained by the completeness of reference databases²². Human and host-associated microbiomes are generally well represented in reference databases due to their medical importance and extensive sequencing. In contrast, environmental samples often harbor vast microbial diversity that remains uncharacterized. Metagenomes from soil, ocean, and extreme habitats frequently contain novel lineages absent from current references, leaving a large fraction of reads unassigned and key aspects of community structure and function unresolved^{8,23}. Large-scale marker-gene surveys such as SingleM and the Sandpiper resource have profiled over 248,000 metagenomes from diverse environments, greatly advancing our understanding of microbiome diversity²⁴. Yet, they inherently capture only a subset of genomic content and do not quantify the overall assignable fraction across biomes.

Founded on *de novo* assembly, metagenome-assembled genomes (MAGs) has emerged as a promising strategy to overcome the limitation of reference-based profiling by reconstructing genomes directly from sequencing data. In this approach, sequencing reads are first assembled into contigs and scaffolds, which are then grouped into genome bins based on tetranucleotide frequencies (TNFs), abundance patterns, complementary marker genes, and taxonomic alignments²⁵.

Although MAGs have generated valuable reference genomes, including many now used for human microbiome analysis^{26–28}, MAG generation faces challenges that can introduce bias into microbial characterization. The Critical Assessment of Metagenome Interpretation (CAMI) challenge showed that low-abundance or highly similar genomes are often fragmented or missing^{29,30}. Assemblers can also misassemble closely related strains when genomes contain shared conserved regions, mobile elements, or horizontally transferred genes that create ambiguous paths in the assembly graph. Similar ambiguities may also arise from repetitive or low-complexity DNA sequence, even when sequence similarity does not reflect shared ancestry. Such cases can produce chimeric contigs that obscure strain-level diversity and potentially

confound genotype - phenotype associations, particularly for polygenic traits³¹. Complex regions of *de Bruijn* graphs (DBG) caused by sequencing errors, repeats, and uneven coverage further fragment assemblies and reduce recovery of genomes from low-abundance organisms³². Consequently, certain organisms and pathways remain underrepresented²⁷, and even co-assembly or hybrid long-/short-read strategies offer only modest improvements³³. Recent benchmarks of metagenome assemblers reinforce these challenges, showing that while tools like MEGAHIT and metaSPAdes achieve strong scalability and contiguity, their performance varies across datasets^{34,35}. These findings emphasize the importance of systematically evaluating assembly-based recovery together with reference-based mapping, as each approach offers complementary strengths and limitations.

Both mapping-based and assembly-based analyses are essential for comprehensive metagenomic interpretation. Clear metrics to evaluate how well these approaches perform across different types of datasets would establish empirical guidelines for selecting methods in microbiome profiling. Ultimately, achieving maximum sensitivity will likely require integrating both strategies while continuing to refine and expand reference catalogs, enabling more accurate and complete characterization of microbial diversity.

Chapter 1. A survey of assignable microbial content across biomes

Abstract

In this Chapter, I analyze 2,867 publicly available metagenomes spanning host-associated, engineered, and environmental biomes. After contamination filtering, 1,901 microbial-dominated samples are retained for analysis. Using k-mer based containment against successive releases of the Genome Taxonomy Database (GTDB), I quantify the fraction of metagenomic sequence content that can be linked to known reference genomes. Assignability varies strongly by biome, with host-associated microbiomes, particularly vertebrate digestive systems, showing high reference coverage, while environmental biomes such as soil, sediment, freshwater, and marine systems remain substantially underrepresented. Comparison across GTDB versions demonstrates that reference expansion systematically improves assignability, but with diminishing returns and persistent biome-specific gaps. Additional validation using read-mapping (genome-grist) confirms that k-mer based estimates capture large-scale trends in reference recoverability while highlighting methodological differences between containment and alignment-based approaches.

Background

Shotgun metagenomic sequencing enables direct characterization of microbial communities across a wide range of environments without the need for cultivation^{6,36}. By sampling total DNA from host-associated, engineered, and environmental systems, metagenomics has revealed extensive microbial diversity and functional potential^{37,38}. A

common approach to interpreting these data relies on reference-based analysis, in which sequencing reads are compared to existing genome databases^{36,39}. However, a persistent limitation of reference-based metagenomic analysis is that a substantial fraction of sequenced content cannot be confidently assigned to known genomes, despite recent large-scale efforts to recover genomes from metagenomes⁴⁰. The proportion of unassigned content varies widely across biomes and reflects both biological novelty and uneven representation of microbial diversity in reference genome databases^{22,41,42}.

Reference genome collections underpin taxonomic and functional interpretation of metagenomes²², and their completeness directly constrains the fraction of genome content that can be recovered from sequencing data²⁶. The GTDB provides a standardized, phylogenomic consistent framework for microbial taxonomy and has expanded rapidly in recent years through the inclusion of isolated genomes and MAGs²⁷. Despite this growth, it remains unclear how effectively current GTDB releases capture microbial genome content across different biome types, and whether database expansion has reduced long-standing disparities between host-associated, engineered, and environmental systems.

Most existing assessments of reference coverage are limited in scale or confounded by assembly and binning biases^{30,43}. This chapter focuses on quantifying genome content assignability directly from trimmed sequencing reads. K-mer based containment methods provide a scalable approach for estimating the fraction of metagenomic content shared with reference genomes by comparing k-mer composition rather than relying on assembly or read alignment^{44,45}. These approaches are commonly implemented using MinHash⁴⁴, a probabilistic sketching technique that compresses large k-mer sets into compact representations while preserving similarity relationships, enabling efficient comparison of large sequencing datasets. Sourmash (<https://sourmash.readthedocs.io/en/latest/index.html>) builds on this framework to support large-scale sequence comparison and taxonomic profiling using MinHash sketches⁴⁶. Sourmash fastmultigather (FMG) was used to summarize genome containment and

abundance-weighted reference overlap, enabling large-scale analysis across thousands of samples^{47,48}.

Because k-mer containment does not require read-level alignment or genomic positional coherence, it represents an indirect measure of reference recoverability. To evaluate the robustness of containment-based estimates, this chapter incorporates an additional validation using genome-grist^{49,50}, a read-mapping based workflow that acts downstream of sourmash to quantify genome coverage, effective abundance, and the fraction of reads that can be directly aligned to reference genomes. Comparing these approaches using the same reference database allows methodological differences to be assessed independently of database composition.

Accurate estimation of microbial genome content also requires control for host and non-microbial contamination. Host-associated and engineered metagenomes frequently contain substantial animal- or plant-derived DNA, which can inflate apparent reference coverage. This chapter therefore begins with systematic filtering of samples dominated by animal- or plant-associated sequences, ensuring that downstream analyses reflect microbial genome content.

Methods

a) MGnify metagenome retrieval and dataset assembly

To comprehensively evaluate extractable genomic content from publicly available metagenomes, we enumerated all eligible samples (as of date October 3, 2024) from annotated metagenome database of EMBL-EBI MGnify⁵¹ (<https://www.ebi.ac.uk/metagenomics/>), ensuring broad coverage across diverse environmental and host-associated biomes. To streamline downstream genome-grist and assembly workflows, we restricted our analysis to metagenomic

datasets generated using Illumina paired-end sequencing technologies. For each sample, three hierarchical levels of biome classification curated by MGnify were extracted and retained.

Metagenome metadata were programmatically retrieved from MGnify using the REST API. Biome identifiers were first extracted from the MGnify biome hierarchy and used to query associated samples implemented in the script (<https://github.com/ctb/2024-ebi-mgnify-rest>) `get-runs-by-biome.py`. The resulting run-level metadata were post-processed using `digest-runs-by-biome.py` to produce a flattened accession table `zyzhao-list-oct3.csv` containing one record per metagenome sequencing run. For each metagenome sequence, the table records the Illumina platform and instrument model, experiment type, accession identifier, and up to three levels of MGnify biome annotation. This workflow yielded a total of 2,867 unique Illumina-sequenced metagenomic run accessions.

Raw sequencing data was not downloaded as part of this accession enumeration workflow. Instead, the consolidated run list served as input for downstream data retrieval and analysis pipelines.

b) Sourmash and fastmultigather analysis

Sourmash `manysearch` (<https://sourmash.readthedocs.io/en/latest/command-line.html>) at standard scale 10,000 was used for the initial screening survey of eight representative host reference genomes (human, mouse, cattle, pig, chicken, dog, cat, and sheep) on all 2,867 metagenomes. For each sample, we quantified the proportion of metagenomic k-mers matching each host genome and summarized contamination levels.

We created the `oct4-sigpath-list.txt` file containing the path of signature files generated from Sourmash `Wort` (<https://github.com/sourmash-bio/wort>), a software that automatically sketches newly released SRA and stores signature files in the database. We then applied Sourmash `fastmultigather` (<https://github.com/sourmash-bio/sourmash/issues/3095>) with

an indexed eukaryotic or GTDB rocksdb reference

(<https://sourmash.readthedocs.io/en/latest/databases.html>). The analysis was configured with a scale of 10,000 and k-mer size of 51 (to minimize false positives) or 21 (to maximize sensitivity across the diverse sample types). For each metagenome, the information of the fraction of weighted (ref_f_weighted) multiplicity k-mers matched to reference genomes were compiled for the analysis.

To lay out an overall picture of genome content recovery from the reference-based approach, we plot the average assignable k-mer percentage on metagenomes grouped by biome type in a color-coded bar chart.

c) Random selection of 100 samples for genome-grist run

Random subsampling of metagenomes was performed using a custom script [subsample2.py](#) (with a fixed random seed to promote reproducibility) on the table of metagenome FMG summary 1901_summary_rs226_k21.csv and the master list of eligible samples zyzhao-list-oct3.csv. For each biome category, up to three samples were selected. Biomes containing three or more samples were subsampled uniformly at random to select three samples, whereas biomes containing one or two samples contributed all available samples without subsampling. Because the number of biomes and their sample counts limited the maximum achievable total under this constraint, the final subsample comprised 92 metagenomes rather than a fixed target of 100.

d) Comparison of k-mer based and mapping-based reference assignability

To validate k-mer based estimates of reference assignability, we compared results from FMG with those from genome-grist, the read-mapping based workflow, using the same GTDB database. This comparison evaluates whether k-mer containment metrics capture patterns

consistent with direct read alignment.

FMG (k-mer based) reference assignability

FMG estimates reference representation by measuring k-mer containment between metagenomes and reference genomes. For each metagenome, FMG reports per-genome and summary statistics describing the fraction and abundance of matched k-mers. The primary FMG metrics used in this comparison were:

- **f_match**, representing the fraction of reference k-mers detected (genome containment);
- **median_abund**, reflecting the relative abundance of matched k-mers;
- **n_unique_weighted_found**, a weighted measure of total k-mer matches per reference genome.

At the sample level, **ref_f_weighted** is calculated as the sum of abundance-weighted unique k-mers that match to reference genomes, normalized by the total number of k-mers in the metagenome sketch. It therefore serves as an estimate of the fraction of the metagenome content explained by known references.

Genome-grist (mapping-based) reference assignability

Genome-grist quantifies reference representation through direct read mapping to the same GTDB reference genomes. For each metagenome, genome-level summaries contains the report of:

- **f_covered_bp**, the fraction of reference genome bases covered by at least one mapped read, representing genome containment by read alignment;
- **effective_coverage**, a weighted coverage metric that accounts for read depth distribution;
- **n_mapped_reads**, the number of reads mapped to each reference genome.

For each metagenome, the total number of mapped reads was calculated by summing **n_mapped_reads** across all reference genomes, and the fraction of mapped reads was computed relative to the total number of sequenced reads for that sample.

Summary-level comparison

The following correspondences were used to compare the two approaches, providing mapping-based analogs to FMG's k-mer based containment and abundance estimates:

- FMG **ref_f_weighted** ↔ genome-grist **fraction of mapped reads** (fraction of metagenome explained by references)
- FMG **f_match** ↔ genome-grist **f_covered_bp** (genome containment)
- FMG **median_abund** ↔ genome-grist **effective_coverage** (relative abundance)
- FMG **n_unique_weighted_found** ↔ genome-grist **n_mapped_reads** (reference representation)

All comparisons were performed on per-sample summaries derived independently from FMG and genome-grist outputs, without reliance on assembly-based reconstruction.

e) Reproducibility and Implementation

All analyses were conducted using reproducible workflows implemented in Snakemake. Parameters, databases, and software versions were fixed across all samples. Intermediate results were retained to enable per-sample and cohort-wide comparisons.

All commands and analysis can be found through the github repository (<https://github.com/jeanzzhao/thesis>). [Sourmash](#) version 4.8.11 and [sourmash-plugin-branchwater](#) version 0.9.9 are used in this study.

Results

1.1 Sample selection and host-derived DNA analysis

Metagenomic samples were drawn from MGnify, an open-access resource that provides standardized processing, metadata annotation, and biome classification for publicly available microbiome datasets. MGnify was selected to enable large-scale, reproducible comparisons across biomes while minimizing technical heterogeneity arising from inconsistent preprocessing or metadata curation. We restricted our analysis to shotgun metagenomic datasets with publicly available raw sequencing reads and included all 2,867 eligible samples available as of October 3, 2024. These samples span host-associated, engineered, and environmental biomes as defined by MGnify's curated biome ontology.

Because this study aims to quantify the fraction of metagenomic sequence assignable to known microbial reference genomes using k-mer based methods, it is critical to distinguish microbial signals from host-derived DNA. In host-associated metagenomes, substantial portions of sequencing data may originate from the host organism rather than microbial community members. If unaccounted for, host DNA can artificially inflate the proportion of unassigned k-mers, bias sequence saturation estimates, and confound interpretations of microbial reference coverage. Therefore, prior to sample selection and assignability analysis, we systematically screened all metagenomes for host-derived sequences.

1.1.1 Pilot survey of eight hosts across 2,867 metagenomes

All 2,867 metagenomes from MGnify were initially screened against eight representative (human and common livestock species) host reference genomes using Sourmash manysearch as a pilot assessment of potential host-derived signals. The proportion of k-mers matching each host genome was quantified for every sample and summarized in Table 1.1.

query	host	p_metag %		No. of Metagenomes with p_metag		
		mean	max	>1%	>5%	>50%
bosTau9.fa.gz	Cattle	0.2	89.2	21	6	4
canFam6.fa.gz	Dog	0.1	8.7	71	2	0
felCat9.fa.gz	Cat	0.2	8.7	81	28	0
galGal6.fa.gz	Chicken	0.3	57.8	76	20	3
hg38.all	Human	1.3	88.6	127	66	29
mm39.fa.gz	Mouse	1.9	88.8	388	195	24
oviAri4.fa.gz	Sheep	0.2	38.5	39	7	0
susScr11.fa.gz	Pig	0.2	39.3	62	6	0
Avg/Total		0.5	52.5	865	330	60

Table 1.1: Sourmash manysearch output of eight hosts across all 2867 metagenomes.

Although the average level of host-derived DNA (p_metag %) for all eight hosts surveyed is less than 5%, some samples exhibit host signals exceeding 80% (e.g., cattle: 89.2%, human: 88.6%, mouse: 88.8%). Among the eight host genomes, human (hg38) and mouse (mm39) genomes showed the highest average levels of host-derived sequence and were the most frequently detected hosts across samples. Specifically, mouse-derived DNA exceeded 1% in 388 metagenomes and 5% in 195 metagenomes, with 24 samples exceeding 50%. Human-derived DNA exceeded 1% in 127 metagenomes and 5% in 66 metagenomes, with 29 samples exceeding 50%. These results indicate that although host-derived DNA is generally modest, host-dominated datasets are not uncommon.

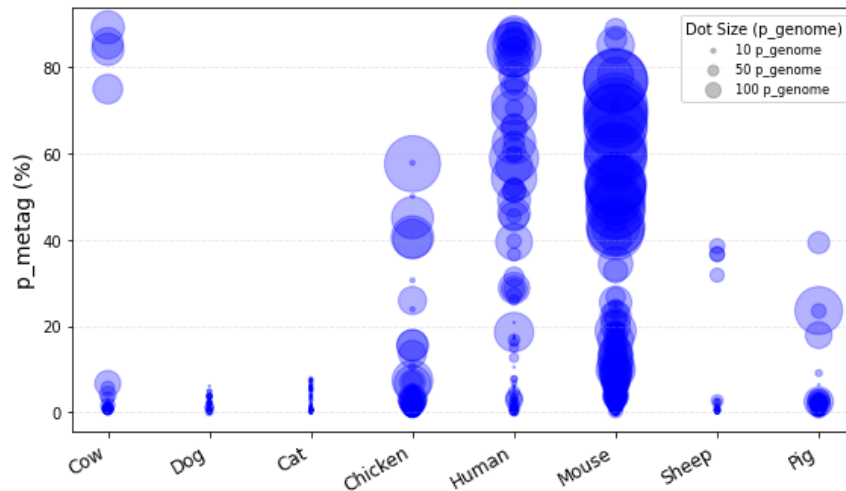


Figure 1.1: 1

X-axis: Eight reference host genomes. Y-axis: Percentage of genomes detected in the metagenome (p_metag %). Dot size: Represents the fraction of the genome detected within the corresponding metagenome.

This visualization highlights the extent of host-derived sequence detected for each reference genome, with larger dots indicating higher representation of the host genome within the sample. Notably, human and mouse genomes show significantly higher signal levels than other hosts, likely reflecting the prevalence of human- and mouse- associated human- and mouse-associated microbiome samples in MGnify, as well as possible contributions from incomplete host DNA removal during sample preparation.

1.1.2 Study-level clustering and biome dependence

Samples with high levels of host-derived DNA were not randomly distributed but frequently clustered within specific studies. For example, all 24 samples exhibiting >50% mouse-derived sequence originated from a single mouse gut metagenome study (BioProject: PRJNA815731, Supplementary Table 1.1), while elevated human-derived signal was enriched in particular human-associated datasets (Supplementary Table 1.1). These clustering patterns suggest that host DNA largely reflects biological context (e.g., host-associated sampling) and experimental design rather than stochastic noise.

To evaluate whether the detected host-derived signal was consistent with sample annotations, we examined metagenomes with p_metag > 5% and compared the identified host genome to corresponding host metadata and Biome2 classifications (Figure 1.2). In most cases, host signals aligned with expected biological origins. For example, mouse-derived DNA was predominantly observed in Host-associated:Mammals samples, and human signals were enriched in Host-associated:Human datasets. Similarly, cattle-derived DNA was detected mainly

in Mammal- or Animal-labeled samples, and chicken matches were largely found in Animal or Birds categories.

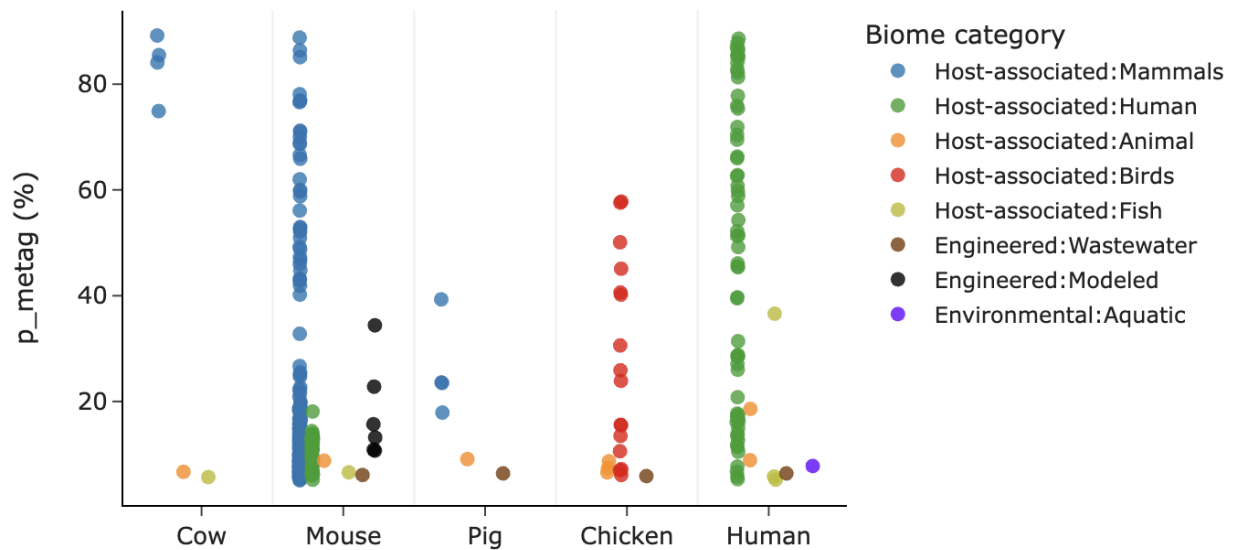


Figure 1.2: Five representative host-derived DNA levels across metagenomes stratified by biome category.

Scatter plot showing the proportion of metagenomic k-mers matching five representative host reference genomes. The x-axis indicates the queried host genome, and the y-axis represents the fraction of the host genome detected within each metagenome. Points are colored according to the Biome2 classification of each sample.

However, some discrepancies were observed. Cattle, mouse, and human DNA was detected in one fish-associated metagenome; mouse signals appeared in a small number of human-associated and engineered samples; and low levels of human DNA were identified in fish-associated datasets. Most cross-biome detections occurred at low abundance and may reflect shared conserved k-mers, minor sample handling effects, or limitations in biome annotation specificity. In contrast, the moderate mouse-derived signal observed in two engineered samples (~22% and ~35%) and human DNA detected in fish-associated (~38%) are unlikely to be explained solely by cross-mapping artifacts. These cases more plausibly

represent genuine host-derived sequences, potentially arising from mouse-associated experimental systems or sample handling during collection or processing rather than from reference genome impurity.

1.1.3 Multi-host signals

We further examined the presence of multiple host genomes within individual samples (Figure 1.3 and Table 1.2). At a 1% threshold, approximately 500 metagenomes exhibited detectable signals from a single host genome, while 108 showed matches from multiple hosts. At more stringent thresholds ($>5\%$), multi-host signals were less common (47 samples). In rare cases, low-level matches to all eight hosts were observed, likely reflecting shared conserved k-mers among host genomes or technical cross-sample signal during sequencing and processing. Although multi-host signals were uncommon at high abundance levels, their presence highlights the need for careful interpretation of k-mer matches in host-associated environments.

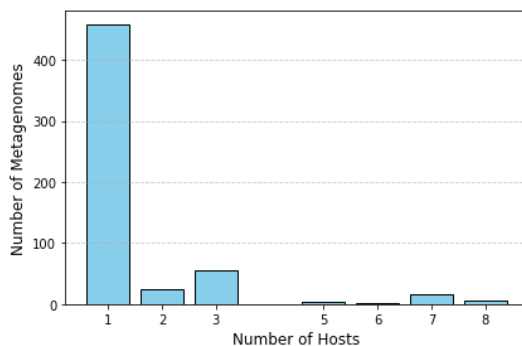


Figure 1.3: Distribution of multi-host signals across metagenomes ($>1\%$ threshold).

Bar plot showing the number of host genomes detected per metagenome at a host-derived DNA threshold of $p_{\text{metag}} > 1\%$. The x-axis indicates the number of host genomes detected within a single metagenome, and the y-axis shows the number of metagenomes meeting this criterion.

Most samples exhibit signals from a single host genome, while progressively fewer samples show matches to multiple hosts. Multi-host signals are uncommon at higher host counts, indicating that high-level host-derived DNA levels are typically dominated by a single primary host source.

No. of host detected	# of Metagenomes w/ p_metag		
	>1%	>2%	>5%
2	24	11	19
3	56	46	26
4	0	0	0
5	4	1	0
6	2	0	0
7	16	10	2
8	6	1	0
Total	108	69	47

Table 1.2: Frequency of metagenomes with multi-host signals at increasing thresholds.

Table summarizing the number of metagenomes containing multiple host genomes exceeding thresholds of >1%, >2%, and >5% of total metagenomic k-mers (p_metag). Rows indicate the number of distinct host genomes detected per metagenome, and columns report the number of samples exceeding each threshold. While 108 metagenomes exhibit signals from multiple hosts at >1%, this number decreases substantially at higher thresholds (>2% and >5%), indicating that multi-host signals are primarily low-level and become rare at more stringent cutoffs.

1.2 FMG Filtering and Reference-based Assignability analysis

Building on the preliminary host-derived DNA analysis above, we implemented a filtering pipeline, aiming to better quantify the assignability of metagenomic data across diverse biomes accurately. This pipeline is designed to systematically exclude >10% host-driven datasets rather

than host-reads removal, to ensure that downstream k-mer matching reflects microbial genomic content rather than host-derived sequences.

1.2.1 FMG sample filtering

Because MGnify includes many host-associated and engineered samples that may contain substantial non-microbial DNA, filtering based on k-mer matching to eukaryotic reference databases was performed prior to microbial reference assignment (Figure 1.4). Each sample was first screened against an animal-plant (AP) reference database (db) using FMG at k-mer size of 51, implemented via a RocksDB-backed k-mer index. The larger k-mer size was selected to reduce spurious matches and minimize false-positive detection of conserved eukaryotic sequences. This step identified samples with substantial host genomic content.

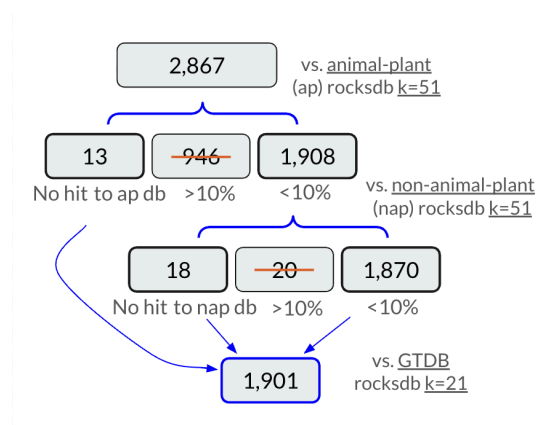


Figure 1.4: Workflow for host and non-microbial sample filtering.

Of the 2,867 samples, 13 samples showed no detectable k-mer matches to the AP database, suggesting minimal eukaryotic signal. In contrast, 946 samples exhibited more than 10% of k-mers matching the AP database and were excluded from further analysis. These excluded samples were primarily drawn from host-associated (785), engineered (83), and environmental (78) biomes, indicating substantial host or non-microbial signal. The remaining 1,908 samples, each with less than 10% AP-matching reads, were retained for further filtering.

The 1,908 retained samples were subsequently screened against a non-animal-plant (NAP) eukaryotic reference database, again using $k = 51$, to remove datasets dominated by other non-microbial eukaryotic content. In this step, 18 samples exhibited no detectable matches to the NAP database, while 20 samples exceeded the 10% NAP-matching threshold and were excluded. These excluded samples were primarily host-associated (14) or engineered (6). The remaining 1,870 samples, each with less than 10% NAP-matching k-mers, were retained.

To construct the final dataset for microbial reference assignment analysis, we combined the 1,870 samples passing both AP and NAP filters with the 13 AP-no-hit samples and 18 NAP-no-hit samples, yielding a total of 1,901 metagenomic samples. This final collection represents datasets with minimal host-derived or other non-microbial sequences and forms the basis for subsequent k-mer based comparison against the GTDB using k21, selected to maximize sensitivity for microbial reference coverage.

At the highest biome level, filtering removed approximately one-third of the original dataset, with the largest reduction occurring in host-associated datasets ($1,928 \rightarrow 1,129$), followed by engineered ($471 \rightarrow 382$) and environmental samples ($468 \rightarrow 390$). This pattern reflects the greater prevalence of host-derived and other non-microbial sequences in host-associated and engineered metagenomes. Despite these reductions, all three biome categories remained well represented after filtering, preserving broad biome coverage for downstream analysis.

1.2.2 Impact of FMG filtering on biome composition and reference assignability

Host-associated	b/f cutoff	std	count	a/f cutoff	std	count
Bacteria	65.65%	0.05	6	65.65%	0.05	6
Circulatory system	74.61%	0.23	11			
Cnidaria	38.35%	0.13	23	62.39%	0.06	5
Digestive system_Human	88.84%	0.13	195	92.19%	0.04	155
Digestive system_Mammals	87.23%	0.12	60	89.44%	0.10	52
Digestive system_Birds	85.27%	0.13	146	89.32%	0.07	124
Digestive system_Insecta	63.36%	0.22	311	85.27%	0.15	122
Digestive system_Animal	76.10%	0.14	279	77.42%	0.11	239
Digestive system_Fish	52.30%	0.24	103	64.38%	0.19	35
Gastrointestinal tract	85.22%	0.13	439	88.57%	0.09	289
Phylloplane	60.22%	0.18	31	68.56%	0.19	13
Respiratory system_Mammals	95.92%	0.03	30	94.83%	0.00	2
Respiratory system_Human	55.94%	0.30	2			
Rhizosphere	55.51%	0.06	23	55.01%	0.07	20
Root	62.51%	0.03	2	62.51%	0.03	2
Skin	77.22%	0.15	267	84.90%	0.08	65
Average	70.27%			77.17%		
Total counts			1928			1129

Engineered	b/f cutoff	std	count	a/f cutoff	std	count
Continuous culture	59.28%	0.12	13	61.47%	0.09	12
Dairy products	90.50%	0.07	48	91.13%	0.09	27
Fermented beverages	48.73%	0.23	33	90.30%	0.01	6
Fermented vegetables	57.59%	0.17	9	59.00%	0.17	8
Defined media	65.88%	0.23	23	63.00%	0.22	21
Genetic cross	60.36%	0.18	2	47.81%		1
Simulated communities (microbial mixture)	92.82%	0.07	18	96.00%	0.03	14
Simulated communities (sequence read mixture)	69.51%	0.19	9	56.89%	0.01	6
Composting	61.47%	0.12	25	61.47%	0.12	25
Landfill	56.69%		1	56.69%		1
Activated Sludge	73.47%	0.08	32	73.47%	0.08	32
Industrial wastewater	65.25%	0.05	8	65.25%	0.05	8
Water and sludge	71.86%	0.07	250	71.04%	0.07	221
Average	67.18%			68.73%		
Total counts			471			382

Environment	b/f cutoff	std	count	a/f cutoff	std	count
Aquaculture	58.30%	0.09	13	58.30%	0.09	13
Estuary	44.42%	0.04	36	44.42%	0.04	36
Freshwater	49.55%	0.08	64	49.55%	0.08	64
Lentic	76.04%	0.23	77			
Marine	51.35%	0.14	101	51.35%	0.14	101
Non-marine Saline and Alkaline	60.83%	0.09	52	60.83%	0.09	52
Sediment	42.41%	0.07	3	42.41%	0.07	3
Soil	50.87%	0.05	17	50.87%	0.05	17
Thermal springs	71.45%	0.18	105	71.44%	0.18	104
Average	56.14%			53.65%		
Total counts			468			390

Table 1.3: Aggregated Sourmash fastmultigather results before and after the filtering.

To evaluate the impact of FMG-based filtering on both biome composition and estimates of microbial reference assignability, we compared sample counts and average weighted match percentages before and after applying the >10% animal-plant (AP) and non-animal-plant (NAP) cutoff.

Filtering resulted in a systematic increase in average weighted reference match for host-associated samples, which increased from 70.27% to 77.17%, indicating enrichment for microbial genome content after removal of host-dominated datasets. Engineered samples showed a more modest increase (67.18% to 68.73%), while environmental samples exhibited little change (56.14% to 53.65%), suggesting that environmental metagenomes generally contain lower levels of host-derived and other non-microbial sequences and that filtering primarily removed a small subset of samples rather than broadly altering assignability.

At finer biome resolution, several host-associated subcategories exhibited substantial increases in reference match following filtering. Digestive system samples showed increases exceeding 20% points in multiple subgroups, and Cnidaria-associated samples increased from 38.35% to 62.39%, consistent with preferential removal of samples with elevated non-microbial content. These changes occurred alongside reductions in sample counts, indicating that filtering enriched for microbe-dominated datasets rather than inflating match percentages uniformly.

Engineered biomes displayed heterogeneous responses to filtering. Notably, fermented beverages showed a dramatic increase in reference assignability (48.73% to 90.30%) while sample counts dropped sharply (33 → 6), suggesting that filtering removed samples with substantial plant or other eukaryotic contributions, enriching for microbial fermentation communities. In contrast, categories such as activated sludge, industrial wastewater, and composting showed minimal changes in both assignability and sample count, indicating relatively stable microbial composition with limited host or eukaryotic interference.

Environmental sub-biomes were largely unaffected by filtering, with most categories showing negligible changes in average weighted match. The primary exception was lentic systems, where 77 samples were eliminated, yet average assignability remained stable, reinforcing the conclusion that environmental samples are less impacted by host-derived and non-microbial sequences and that FMG filtering does not substantially alter environmental biome profiles.

Together, these results demonstrate that FMG-based filtering selectively removes host- and eukaryote-rich datasets while preserving biome structure and improving microbial signals where host-derived and non-microbial sequences are prevalent. This filtering step therefore enhances the interpretability of downstream cross-biome comparisons without introducing major distortions in biome-level reference assignability.

1.2.3 Genome content assignability across biomes (GTDB rs220, $k = 21$)

Using the filtered dataset of 1,901 metagenomic samples, microbial genome content assignability was quantified via k-mer containment against GTDB release rs220, using $k = 21$ to increase sensitivity to divergent microbial lineages while retaining taxonomic resolution. Average weighted match percentages were calculated for each sample and summarized across biome categories and subcategories.

We grouped metagenomes by biome type and calculated the average weighted match percentage, standard deviation, and sample count per group. Results were visualized in a color-coded bar chart, showcasing the average assignable k-mer percentage for each biome type (Figure 1.5).

At the highest biome level, clear and consistent differences in reference assignability were observed (Figure 1.5). Host-associated metagenomes exhibited the highest average weighted match percentage, followed by engineered and then environmental samples. Host-associated samples ($n = 1,129$) showed mean assignability approaching ~75-80%,

reflecting the extensive representation of host-associated microbes in current reference databases. Engineered systems (n = 382) showed intermediate assignability (~65-70%), while environmental samples (n = 390) had the lowest average match (~50-60%), highlighting persistent gaps in reference coverage for natural ecosystems.

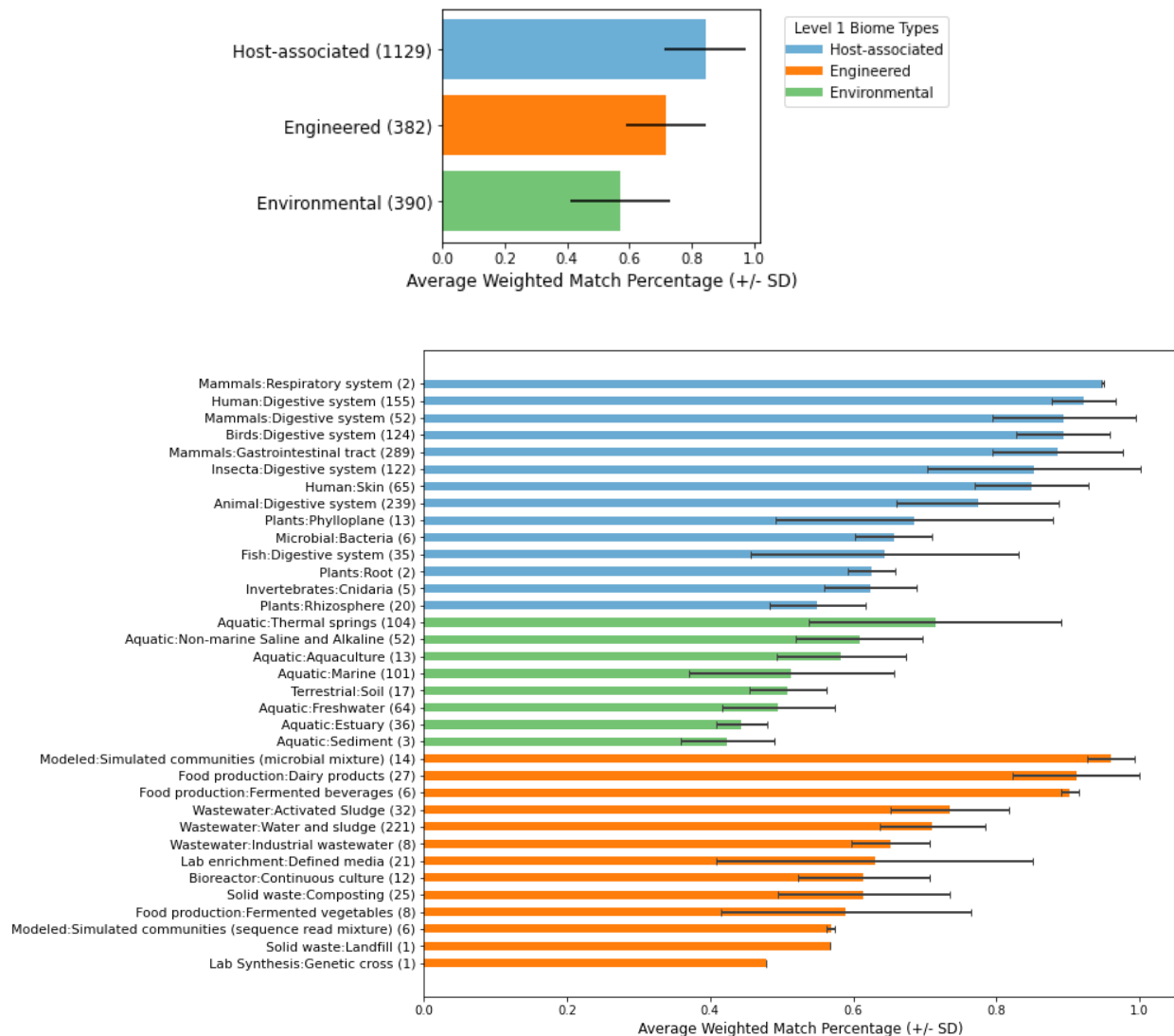


Figure 1.5: Biome-specific genome content assignability using GTDB (rs220, k = 21).

Average weighted k -mer match percentages ($\pm$ SD) are shown for 1,901 filtered MGnify metagenomes at the biome (top) and sub-biome (bottom) levels. The x-axis is the percentage of weighted metagenome k -mers that match to a known genome, indicating the percentage that

belong to a known host. The y-axis lists the type of biome and corresponding number for each different type. The standard deviation and sample count per group are also included. Error bars indicated substantial within-category variability, emphasizing that reference assignability is influenced not only by biome classification but also by finer-scale ecological context and sampling history.

Sub-biome analysis revealed substantial heterogeneity within each major biome category. Among host-associated samples, digestive system microbiomes, particularly those associated with humans, mammals, birds, and insects, consistently showed high assignability, often exceeding 80%. This pattern is consistent with the extensive representation of these microbial communities in existing reference databases, reflecting the greater investment and focus in research (counts of samples), where reference genome coverage is comparatively extensive, resulting in depth of reference genome coverage. In contrast, host-associated plant environments (e.g., rhizosphere, phylloplane, root) and invertebrate-associated samples (e.g., *Cnidaria*) displayed lower average match percentages, indicating reduced representation of these communities in GTDB. These lower assignability values may also partly reflect increased contributions from fungi and other eukaryotes, which are not represented in GTDB and are known to be more prevalent in plant-associated microbiomes. Despite filtering, variance remained high in some subcategories, reflecting biological diversity and uneven reference coverage.

In contrast, environmental samples exhibit substantially lower assignability overall. None of the samples exceed 80% weighted k-mer matching. Soil, sediment, freshwater, estuarine, and marine biomes consistently show low assignable fractions (generally <55%), whereas thermal spring and non-marine saline/alkaline environments display comparatively higher assignability. Although environmental sub-biomes vary in magnitude, they are uniformly less assignable than host-associated and many engineered systems. These patterns likely reflect the high microbial diversity and ecological complexity of environmental systems, which harbor

numerous unique and low-abundance taxa that remain underrepresented in current reference databases. In particular, soil and aquatic environments continue to show limited reference coverage despite ongoing genome recovery efforts.

Engineered biomes exhibited the widest range of assignability values among the major biome categories. Modeled and simulated microbial communities and dairy production systems showed high assignability, approaching near-complete reference coverage (e.g., simulated communities at 92.8% and dairy products at 90.5%). In contrast, engineered fermentation beverage samples showed the lowest assignability within this category (48.7%), while bioreactor cultures (59.3%), genetic cross samples (60.4%), composting systems, and landfill samples exhibited moderate match percentages. Wastewater-associated systems occupied an intermediate range. Together, these patterns indicate substantial heterogeneity in reference coverage across engineered environments, reflecting differences in microbial composition and representation in existing reference databases.

1.3 Impact of GTDB Version Update on Reference Assignability

To assess the sensitivity of reference-based metagenome assignability to database composition, we compared k-mer based containment estimates derived using two releases of the GTDB rs220 (April 2024) and rs226 (April 2025). These releases represent incremental growth of the GTDB reference collection through the continued addition of isolated genomes and MAGs, while preserving a consistent phylogenomic taxonomy. The substantial differences in genome content, taxonomic coverage, and curation between rs220 and rs226 provide a natural experiment for evaluating whether continued reference database expansion translates into measurable gains in assignable metagenomic content.

1.3.1 Expansion of reference content between GTDB rs220 and rs226

Between rs220 and rs226, GTDB experienced substantial growth across both bacterial and archaeal lineages (Table 1.4). The number of bacterial genomes increased from 584,382 to 715,230, representing a 22.39% increase in rs226, while archaeal genomes grew from 12,477 to 17,245 (38.21%). Species-level clusters also expanded, with bacterial species clusters increasing by 27.43% and archaeal species clusters by 18.73%.

	rs220	rs226	Growth %
Bacterial genomes	584,382	715,230	22.39%
Archaeal genomes	12,477	17,245	38.21%
Bacterial species clusters	107,235	136,646	27.43%
Archaeal species clusters	5,869	6,968	18.73%

Table 1.4: Reference content comparison between GTDB rs220 and rs226

A key methodological change in rs226 is the adoption of CheckM2 v1.0.2 for genome quality assessment, replacing the CheckM1 criteria used in rs220⁵². This update increased quality-control stringency, resulting in 12,214 additional genomes (11.1%) failing QC and only 178 (0.023%) newly passing, leading to the re-evaluation or removal of some previously recognized genomes. In parallel, rs226 incorporates taxonomic revisions, including higher-rank updates such as the reclassification of *Firmicutes* as *Bacillota*. Consequently, although rs226 substantially increases the total number of reference genomes, much of this expansion reflects improved genome quality and densification within already represented phylogenetic space rather than the introduction of novel deep lineages.

To note, the growth between rs220 and rs226 is dominated by the continued addition of MAGs, rather than new cultured isolates or newly discovered major lineages, as indicated in the GTDB's release reports⁵². Starting with rs220, newly added bacterial MAGs exceeded new isolated genomes, a pattern that persisted in rs226, while archaeal genomes remained overwhelmingly MAG-derived. These MAGs are often derived from a broader range of

environments and capture greater phylogenetic diversity than cultured isolates, reflecting the ability of metagenomic approaches to recover genomes from previously uncultivated lineages.

1.3.2. Proportional increase in raw k-mer matches across biomes

When comparing raw k-mer counts ($n_matches$), total match counts increase consistently under rs226 than rs220 across all three biome categories (Figure 1.6). Most samples clustered tightly along a near-linear trajectory above the 1:1 reference line, indicating that rs226 uniformly increases detectable matches.

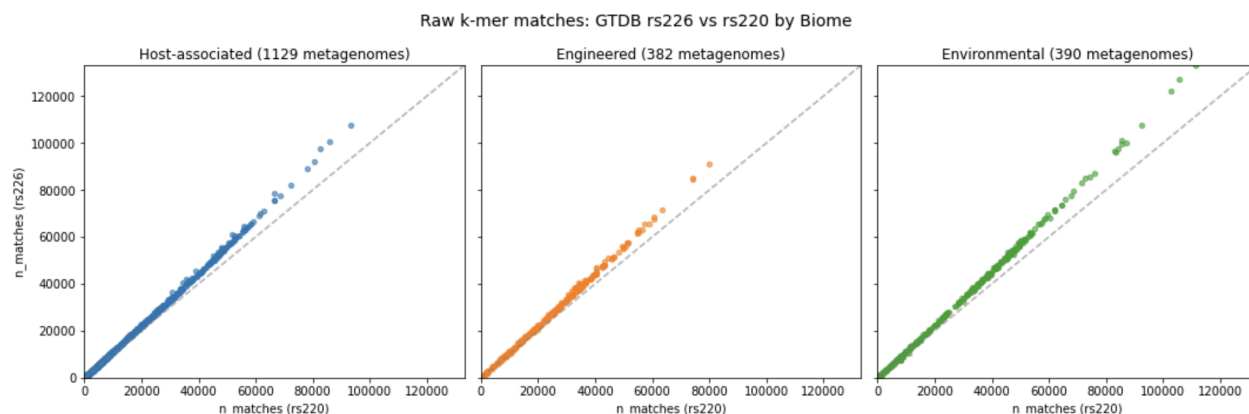


Figure 1.6: Comparison of total k-mer match counts with GTDB rs220 vs rs226.

Each point represents a single metagenome, and the dashed line indicates equality.

The relationship between rs220 and rs226 remained highly proportional within each biome, although environmental samples spanned the widest dynamic range and reached the highest absolute match counts, which likely reflects broader taxonomic spread and some highly-complexity samples. These results indicate that updating from rs220 to rs226 increases raw reference matches in a consistent and additive manner while preserving cross-biome structure. This increase is largely driven by the addition of MAG-derived genomes in newer GTDB releases, which expand coverage of previously uncultivated lineages. While this improves assignability, it may also influence the phylogenetic composition of detected matches.

1.3.3. Incremental gains by biome and where improvements plateau

Next, we examined the overall assignability changed from rs220 to rs226 for each metagenome, as measured by the FMG weighted reference fractions. Most samples cluster tightly along the 1:1 identity line, demonstrating strong concordance between database versions. The majority of points fall on or slightly above the identity line, indicating that assignability under rs226 is consistent with, or modestly greater than, that observed under rs220. Only a small, countable subset of samples falls below the identity line, and these deviations are limited in magnitude. For most metagenomes, absolute increases in weighted reference fraction under rs226 are small. This pattern indicates that while rs226 contains substantially more reference genomes, gains in assignability are generally incremental rather than large-scale shifts across the dataset.

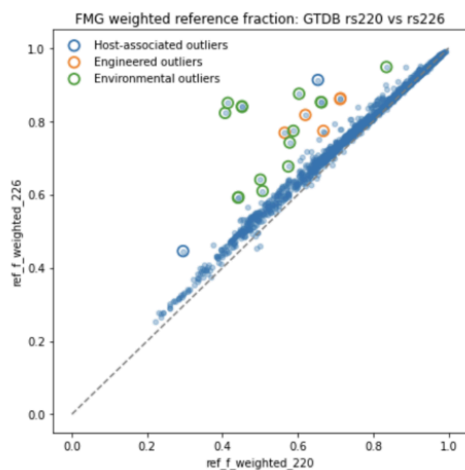


Figure 1.7. Comparison of FMG weighted reference fraction using GTDB rs220 and rs226.

Each point represents one metagenomic sample. The dashed line indicates equality between database releases (1:1). Points are colored by the top-level biome for identified outliers.

A distinct subset of (18) samples deviates markedly above the identity line (circled in Figure 1.7), exhibiting pronounced gains in weighted reference fraction under rs226. These

outliers are enriched in environmental systems, with additional representation from engineered and host-associated samples. In these cases, rs226 appears to capture previously underrepresented or absent taxa, likely contributing to these localized improvements in assignability.

1.4 Validation of k-mer based assignability via read mapping

1.4.1. Rationale for mapping-based analysis

While k-mer based containment of FMG provides a scalable estimate of the fraction of metagenomic genome content shared with existing reference databases, it does not require read-level alignment or genomic positional coherence and therefore represents an indirect measure of reference recoverability. To validate k-mer based assignability using an orthogonal approach, we employed genome-grist (<https://github.com/dib-lab/genome-grist>), a mapping-based workflow that quantifies the fraction of metagenomic reads that can be directly aligned to reference genomes.

Genome-grist performs read-level mapping to curated reference collections and summarizes alignment results at the genome and sample levels, providing an independent assessment of reference representation under stricter alignment constraints. Because both approaches use the same reference database, differences in assignability estimates can be attributed to methodological differences rather than database composition. Genome-grist was therefore used as an downstream validation framework to assess the robustness of FMG, k-mer based assignability estimates across biome types, establishing confidence in the containment-based metrics used throughout this chapter and providing a reference-based baseline for subsequent analyses.

1.4.2. Selection of samples for Genome-grist analysis

Because read-mapping based analysis is substantially more computationally intensive than k-mer containment, genome-grist was applied to a stratified subset of samples rather than the full dataset. From the filtered collection of 1,901 metagenomes analyzed with FMG, samples were subsampled by biome to preserve broad ecological representation while maintaining computational tractability (Figure 1.8).

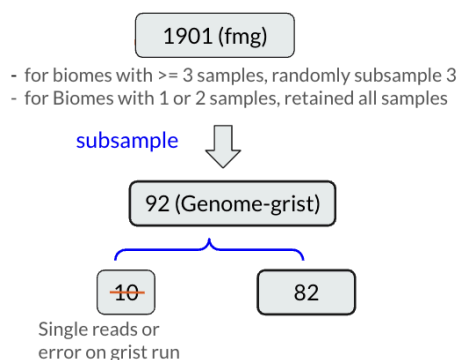


Figure 1.8: Workflow of sample selection for Genome-grist analysis.

For biome categories with three or more samples, three metagenomes were randomly selected; for those with one or two samples, all were retained. This approach yielded a final set of 92 metagenomes spanning host-associated, engineered, and environmental systems. Of these, 10 samples were excluded due to single-end reads or processing errors in genome-grist, resulting in a final dataset of 82 metagenomes for analysis.

This biome-stratified subsampling ensures that the genome-grist analysis captures diversity across ecological contexts while preventing overrepresentation of highly abundant biome categories. Importantly, no selection was based on FMG assignability values or expected mapping success. As a result, the subset provides an unbiased framework for evaluating concordance between k-mer based containment and read-mapping based reference assignability estimates.

1.4.3. Validation of Reference Assignability: FMG vs Genome-grist and Read Mapping

To validate k-mer based estimates of reference assignability, we compared weighted containment from standard genome-grist gather (k31), fastmulti-gather (FMG, k31), and genome-grist read mapping across representative host-associated, engineered, and environmental metagenomes (Figure 1.9).

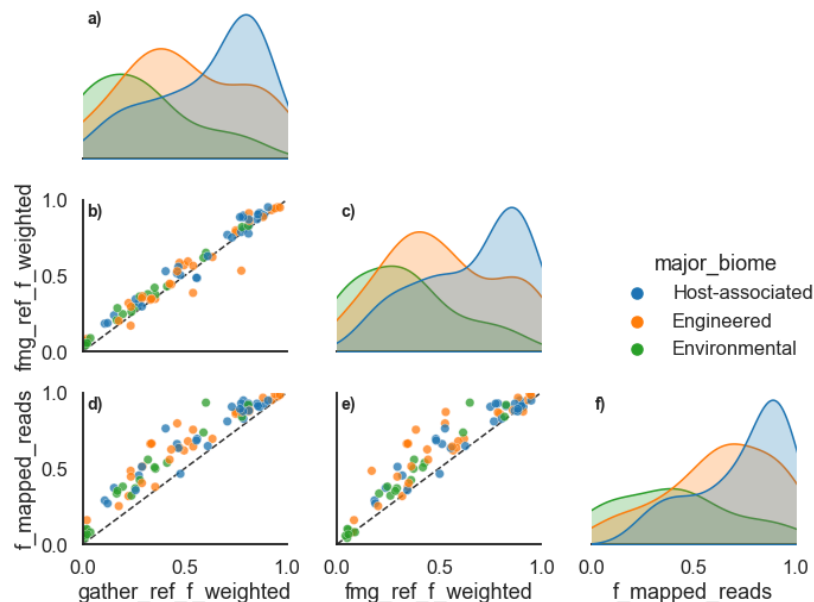


Figure 1.9: Pairwise comparisons of genome-grist, FMG, and mapping.

Genome-grist weighted reference fractions (gather_ref_f_weighted), FMG-derived weighted reference fractions (fmg_ref_f_weighted), and mapping-based fractions (f_mapped_reads), stratified by major biome categories. The dashed diagonal lines indicate perfect agreement (1:1) between methods.

The overall distributions of weighted containment are shown in panels (a) (gather) and (c) (FMG). Both methods display nearly identical biome-level structure: host-associated samples cluster at high assignability values, engineered systems occupy an intermediate range with broader dispersion, and environmental samples are concentrated at lower coverage fractions. The close similarity between panels (a) and (c) indicates that FMG preserves the quantitative

behavior of standard gather under the scaled workflow. Within the host-associated category, the bimodal distribution likely reflects differences between animal-associated and plant-associated microbiomes, with plant-associated samples exhibiting lower assignability, consistent with patterns observed in sub-biome analyses and their reduced representation in current reference databases.

Direct comparison between gather and FMG in panel (b) confirms this equivalence. Despite being executed under a substantially scaled workflow (10,000x versus 1,000x subsampling), FMG estimates closely tracked genome-grist gather values, with points tightly distributed along the 1:1 line across all biomes, with no systematic biome-specific bias observed. These subsampling levels were chosen to reflect typical parameter setting for standard gather and large-scale FMG workflows, indicating that FMG preserves weighted containment estimates under different sketch scaling.

To evaluate concordance between k-mer based containment and read-mapping based estimates, panels (d) and (e) compare genome-grist mapping fractions to standard gather and FMG, respectively. In both cases, assignability estimates are strongly positively correlated: samples with higher k-mer containment also exhibit higher proportions of mapped reads. The dispersion of FMG – mapping comparisons is comparable to that observed for gather – mapping comparisons, indicating that FMG does not introduce excess noise beyond the intrinsic differences between k-mer containment and alignment-based methods. Noticeably, for panels (d) and (e), most points lie above the 1:1 line. This offset is consistent with a methodological difference between the approaches: read mapping permits mismatches during alignment, allowing reads with close (but not exact) similarity to be counted as mapped, whereas k-mer containment requires exact k-mer matches, making it more stringent and therefore yielding lower fractions in these comparisons.

Panel (f) shows mapping-based assignability estimated by genome-grist. The overall biome ordering is preserved: host-associated samples remain highest, engineered samples

intermediate, and environmental samples lowest in reference coverage. Compared to k-mer based estimates, mapping-based fractions are generally shifted upward, consistent with the ability of read alignment to tolerate mismatches and capture near-reference similarity. This shift is most apparent in engineered systems, which move slightly toward higher coverage values relative to panels (a) and (c). For environmental samples, mapping-based estimates show a modest increase relative to strict k-mer containment, particularly among samples with very low k-mer assignability. However, the overall pattern remains unchanged: environmental metagenomes exhibit lowest reference coverage and greatest dispersion under both frameworks. The similarity in distributional structure across panels indicates that mapping increases sensitivity but does not fundamentally alter the biome-level hierarchy of reference representation.

Discussion

1) Host-derived sequence as a structural constraint on reference assignability

One of the central objectives of this project is to quantify the proportion of metagenomic sequence content that can be assigned to existing microbial reference genomes, and to determine whether assignability is biome dependent. The host-driven signal analysis presented in Result 1.1 establishes a critical prerequisite for this goal: accurate estimation of microbial assignability requires explicit separation of microbial and host-derived sequence space.

a) Evaluation of host-derived signal across biome categories

From the perspective of microbial reference assignability, the presence of host-driven signals has direct and potentially confounding consequences. K-mer based containment and

mapping approaches quantify the fraction of metagenomic sequence explained by reference databases. If a substantial proportion of k-mers originate from host genomes rather than microbial community members, the fraction of sequence unassigned to microbial references may be artificially inflated. In such cases, low assignability does not necessarily reflect novel microbial diversity but instead reflects the presence of non-microbial DNA. This effect may be further influenced by host genome size, as larger eukaryotic genomes contribute disproportionately more k-mers relative to microbial genomes, amplifying their impact on apparent unassigned sequence fractions.

This distinction is critical for interpreting unexplained metagenomic content. For example, if 60% of k-mers are unmatched by microbial reference genomes but 50% align with host genomes, much of the “unexplained” sequence space is attributable to host-derived DNA rather than gaps in microbial reference coverage. Without systematic host screening, host sequence could therefore be misinterpreted as evidence of large reservoirs of unknown microbial diversity.

Across 2,867 MGnify metagenomes, the overall mean host-derived k-mer fraction was low (<2%), yet a nontrivial subset of samples exhibited extreme host dominance, in some cases exceeding 80%. These highly contaminated samples clustered within specific studies (e.g., murine gut datasets), indicating that host signal is strongly influenced by sampling context and experimental design rather than random noise. The heterogeneity pattern is expected, as host-associated microbiomes, particularly human and mouse samples, often contain abundant host DNA that is incompletely removed during library preparation. The predominance of human and mouse matches is therefore consistent with the strong representation of biomedical and model-organism studies in public metagenomic datasets.

When evaluated against biome annotations, most host signals were consistent with expected sample origins, with only occasional discrepancies observed. Potential explanations for these mismatches include conserved genomic regions shared across taxa, low-level

laboratory contamination, imprecision in broad biome classifications, mislabeling of samples, or engineered systems derived from host-associated material. Regardless of their specific origin, these observations reinforce the value of empirical, sequence-based host screening, even in samples not explicitly labeled as host-associated, to prevent host-derived sequence from influencing downstream microbial analyses. The host-detection workflow developed here provides a rapid and scalable approach to identify and filter out samples with substantial host-derived signals, ensuring that downstream analyses reflect microbial genomic content rather than non-microbial signal.

The observed multi-host signal patterns further highlight complexities in interpreting k-mer matches. While multi-host detection was uncommon at higher thresholds, low-level matches to multiple hosts likely reflect shared conserved k-mers, environmental mixtures (e.g., livestock-associated samples), or technical artifacts. These cases emphasize that k-mer matches must be contextualized biologically and analytically.

Overall, the host-driven signal study highlights the substantial presence of host-origin sequences in metagenomic datasets, with human and mouse genomes emerging as the most significant contributors. These results emphasize the importance of comprehensive identification and characterization of host-derived sequences to ensure accurate and reliable metagenomic data interpretation. By explicitly quantifying and filtering host-derived samples prior, we ensure that downstream analyses of microbial reference assignability reflect microbial genomic content rather than host signal. This step is essential for accurately assessing how much metagenomic sequence data from diverse biomes can be attributed to known microbial reference genomes.

b) Effectiveness and implications of FMG-based eukaryotic dataset filtering

The decision to implement broader eukaryotic filtering was motivated by the initial eight-host survey, which revealed that host-derived sequences could substantially influence assignability estimates in a subset of samples. Recognizing that limiting screening to a small set

of animal hosts might underestimate non-microbial contributions, we expanded the filtering strategy to include comprehensive animal and plant reference databases, as well as non–animal/plant eukaryotic genomes. This broader screening ensured that substantial eukaryotic sequence, whether from vertebrate hosts, plant material, or other eukaryotic sources, was identified prior to microbial reference assignment, thereby improving the accuracy and interpretability of cross-biome microbial comparisons.

Because this study aims to quantify microbial reference assignability across biomes at the sample level, rather than reconstruct individual microbial communities, a conservative sample-level exclusion strategy was adopted instead of read-level host removal. Samples containing substantial host-derived sequences are structurally different and can distort cross-biome comparisons of microbial assignability. A 10% host-match threshold was therefore selected as a pragmatic boundary at which host-derived k-mers begin to meaningfully influence overall sequence composition and bias reference coverage estimates.

The FMG-based filtering strategy removed host- and eukaryote-dominated samples while largely preserving overall biome structure. At finer biome resolution, the largest increases in assignability occurred in host-associated digestive system samples and in certain engineered fermentation contexts. These changes reflect the selective removal of samples with substantial non-microbial contributions, rather than shifts in underlying microbial composition. In contrast, the minimal effect of filtering on environmental biomes supports the interpretation that these datasets generally contain lower levels of host contamination and that their lower assignability reflects genuine reference gaps rather than technical artifacts. Together, these results demonstrate that conservative, sequence-based host filtering improves interpretability without altering core biome-level trends.

2) Assessing known and unknown microbial genome content across biomes

A central goal of this chapter was to quantify how much microbial genome content across diverse biomes can be assigned to existing reference databases, and to understand how this assignability varies as a function of biome, filtering strategy, and reference database composition. Using k-mer based containment as a scalable proxy for reference coverage, we observe consistent and biologically meaningful patterns that reflect both historical biases in microbial genomics and persistent gaps in reference representation.

a) Biome-dependent structure of reference assignability

Across all analyses, reference assignability exhibits a strong and consistent biome dependence. Host-associated microbiomes show the highest fraction of assignable genome content, followed by engineered systems, with environmental biomes consistently exhibiting the lowest assignability. This hierarchy reflects long-standing imbalances in microbial genome sequencing efforts. Human- and mammal-associated microbes, particularly those from digestive systems, have been extensively sampled, cultured, and sequenced, resulting in dense reference coverage that translates directly into high k-mer containment.

In contrast, environmental biomes such as soil, sediment, freshwater, estuarine, and marine systems show markedly lower assignability. These environments harbor immense microbial diversity, including many low-abundance and phylogenetically divergent taxa that remain underrepresented in reference databases. The persistence of low assignability even after aggressive filtering and across multiple GTDB releases suggests that current reference collections capture only a fraction of environmental microbial diversity. This gap is therefore biological and historical in origin, rather than methodological.

Engineered systems occupy an intermediate and highly heterogeneous position. Some engineered biomes, such as simulated communities and dairy fermentation systems, approach near-complete reference coverage, reflecting their construction from well-characterized taxa. Other engineered contexts, including composting systems, landfills, and certain bioreactor or laboratory synthesis environments, show substantially lower assignability. These patterns highlight that “engineered” does not imply uniform reference representation, but rather reflects a spectrum ranging from tightly controlled systems to environments that overlap strongly with natural microbial communities.

b) Heterogeneity within biomes and limits of broad classification

Sub-biome analysis reveals substantial heterogeneity within each major biome category. It is clear that biome-level labels alone are insufficient to explain reference coverage. For example, within host-associated microbiomes, plant-associated environments and invertebrate-associated samples show consistently lower assignability than vertebrate digestive systems. Similarly, environmental sub-biomes differ systematically, with thermal springs and saline systems exhibiting higher assignability than soil and aquatic systems. These patterns indicate that ecological context, evolutionary novelty, and sampling history interact to shape reference coverage at finer scales than broad biome categories.

The substantial within-category variance observed across nearly all biomes further suggests that reference assignability is influenced by factors beyond biome designation alone, including local diversity, strain-level heterogeneity, and uneven sampling effort across studies. This heterogeneity highlights the need for more granular analyses of taxonomic composition and clustering, which represent important directions for future work.

c) Implications for reference databases and downstream analyses

The Results 1.2 establish a quantitative baseline for how much microbial genome content is currently represented in reference databases across biomes. While host-associated and some engineered microbiomes are largely captured by existing references, a substantial fraction of environmental microbial content remains unassignable. This persistent gap motivates both continued expansion of reference databases, particularly through targeted recovery of environmental genomes, and careful interpretation of reference-based analyses in underrepresented biomes.

Importantly, these findings justify the use of k-mer based containment as a scalable metric for large-scale surveys, while also motivating the validation and complementary analyses introduced later in the chapter. The biome-dependent patterns observed here provide the foundation for examining how reference expansion over time alters assignability (Result 1.3) and for validating k-mer based estimates against read-mapping approaches (Result 1.4).

This result revealed the uneven representation of microbial communities across biome types in existing genomic databases, underscoring the challenges of microbial content identification in metagenomic studies. To address the significant gaps in reference genome coverage, efforts should focus on: 1) expanding genomic databases for underrepresented biomes, particularly environmental and non-mammalian host-associated systems; 2) increasing sequencing efforts in complex natural environments to capture rare and low-abundance taxa; 3) developing new bioinformatic approaches to improve genome recovery from complex and diverse metagenomic datasets. By bridging these gaps, future studies will enable a more comprehensive characterization of microbial communities.

Together, these results establish a biome-dependent baseline of microbial genome content assignability using current GTDB references. While host-associated and engineered microbiomes are largely captured by existing databases, a substantial fraction of environmental

metagenomic content remains unassigned, motivating further analyses of reference database expansion and orthogonal validation via mapping-based approaches.

3) Database expansion improves reference-based assignability

Compared to GTDB rs220, rs226 provides the addition of new genomes and refinement of genome quality, including the adoption of CheckM2-based quality control. These changes not only expanded the reference collection but also enabled curated genomes. The comparison between GTDB rs220 and rs226 provides direct evidence that these enhancements improved reference-based metagenome assignability. Across nearly all samples, rs226 increases total detectable k-mer matches.

The improvements in raw matches translate into stable or increased weighted reference fraction for the vast majority of samples. Most metagenomes fall on or above the 1:1 identity line when comparing weighted reference reflections. Although absolute gains are modest for some samples such as host-associated and engineered metagenomes, consistent with their already strong representation in earlier releases, the directionality of change is positive. Notably, several environmental metagenomes show substantial improvements, demonstrating that newly incorporated MAG collections previously underrepresented or absent taxa. In these cases, database expansion produces meaningful gains in the proportion of assignable content.

Although this analysis compares only two consecutive GTDB releases, the directional consistency of improvement supports the principle that expanding and refining high-quality reference databases enhances reference-based metagenome assignability. Future gains are likely to be maximized through strategic expansion, particularly by incorporating MAG recovery from underrepresented environments to improve overall metagenomic resolution, as highlighted by recent large-scale genome catalog studies⁴⁰.

4) Validation of k-mer based assignability using genome-grist read mapping

Because FMG weighted containment relies on exact k-mer matches rather than read-level alignment, it was essential to evaluate whether this scalable metric accurately reflects reference recoverability under stricter alignment constraints. The genome-grist validation analysis provides strong evidence that FMG-derived assignability estimates are both quantitatively reliable and biologically meaningful.

a) FMG recapitulates standard gather under computational scaling

Direct comparison between FMG and standard genome-grist gather ($k=31$) demonstrates near-identity across all sampled biomes. As shown in Figure 1.9 (panel b), weighted containment values closely follow the 1:1 diagonal, with no systematic deviation by biome category. The aggressive subsampling of FMG retains far fewer k-mers per genome than standard gather, yet produces nearly identical weighted containment estimates without introducing systematic bias.

This result confirms that FMG's scaled workflow does not distort reference assignability estimates. The preservation of quantitative behavior under scaling establishes FMG as a computationally efficient yet statistically faithful approximation of standard gather method. Given the large size of the full 1,901-sample dataset, this validation is critical: it demonstrates that the conclusions drawn throughout this chapter are not artifacts of computational acceleration.

b) Strong concordance between k-mer containment and read mapping

Comparison of k-mer based containment with genome-grist read mapping further validates FMG as a proxy for reference recoverability. Both standard gather and FMG show strong positive correlations with mapping-based fractions (Figure 1.9, panels d and e). This concordance indicates that k-mer containment captures the same underlying biological signal as

alignment-based approaches: the extent to which metagenomic reads derive from organisms represented in the reference database.

Across mapping comparisons, most points lie above the 1:1 line, indicating that mapping-based assignability estimates are generally higher than k-mer based containment. This offset is expected as read mapping allows mismatches and local alignment tolerance, enabling reads with near-reference similarity to be counted as assigned. In contrast, k-mer containment requires exact k-mer matches and is therefore inherently more stringent and conservative. This conservatism could be advantageous as it reduces ambiguity introduced by borderline similarity and ensures that assignability reflects strong genomic correspondence.

c) Preservation of biome-level structure across methods

When looking at the reference representation at biome-level, both containment- and mapping-based frameworks preserve the same hierarchy order (Figure 1.9, panels a, c, and f), although mapping increases slightly the overall assignability fractions. Host-associated metagenomes consistently exhibit the highest assignability, engineered systems occupy an intermediate position, and environmental biomes display the lowest coverage with broader dispersion. This stability across analytical frameworks indicates that the observed biome-dependent structure truly reflects differences in reference genome coverage rather than a methodological artifact.

The modest upward shift observed in environmental systems under mapping suggests that some sequences in this biome are near-reference but diverged enough to reduce exact k-mer overlap. However, this increased sensitivity does not erase the large gap between environmental and other two systems, reinforcing the core conclusion: environmental microbiomes remain substantially underrepresented in current reference collections.

Chapter 2. Metrics for measuring content loss in metagenome assembly

Abstract

Metagenome assembly plays a central role in contemporary microbiome analysis pipelines. Assemblies enable genome-centric analyses, facilitate the recovery of metagenome-assembled genomes (MAGs), and reduce sequencing errors by collapsing redundant reads into consensus contigs. At the same time, assembly necessarily removes sequence variation that does not conform well to linear contig representations, including low-abundance content, strain-level polymorphisms, and repetitive regions. Despite the widespread use of assembly, the magnitude and biological structure of the sequence content lost during this process remain incompletely quantified.

In this chapter, we develop and apply quantitative, k-mer based metrics to measure content loss during metagenome assembly across diverse biomes. By comparing raw metagenomes, assemblies, and reads that are able to map-back-to-assembly, we evaluate how much sequence information is removed, how this loss varies across biomes and abundance levels, and the extent to which lost content can be recovered.

Background

De novo assembly enables recovery of microbial genomes directly from sequencing reads without relying on existing reference genomes. In metagenomics, assembly reconstructs longer genomic fragments from short sequencing reads, allowing the identification of novel organisms and genomic regions that may not be represented in current reference

databases^{33,41}. This capability is particularly important for environmental microbiomes, where large fractions of microbial diversity remain uncharacterized^{26,27}.

Several metagenome assemblers are widely used today, each with strengths and trade-offs. MEGAHIT is one of the most popular for short-read metagenomics, offering high scalability and low memory usage, making it well suited for large, complex datasets⁵³. Benchmarking studies, including CAMI II, show that MEGAHIT achieves high genome recovery across complex datasets while remaining computationally efficiency²⁹. Alternatives like metaSPAdes often produce more contiguous assemblies with improved accuracy for smaller datasets but are computationally more demanding³². Earlier assemblers such as IDBA-UD were designed to handle uneven sequencing depth, many of these tools have been superseded by more recent algorithms with improved scalability and performance⁵⁴.

Despite their success, metagenome assemblers reconstruct consensus contigs from mixtures of closely related genomes and unevenly distributed sequence coverage. As a result, some sequence variation present in the original reads, particularly strain-level polymorphisms or low-abundance variants, may not be explicitly represented in the final assembly^{33,55}. Conversely, assembly can also reconstruct genomic segments from abundant organisms that are absent from reference databases, thereby expanding the observable genomic representation of microbial communities^{26,41}. These contrasting effects raise an important question: to what extent does assembly alter the representation of sequence information present in metagenomic datasets?

In this chapter, I will use MEGAHIT as the assembler, to evaluate how metagenome assembly reshapes sequence representation by comparing multiple representations of the same metagenomic datasets. Specifically, I quantify the fraction of read-level sequence signal retained in assembled contigs, the portion recoverable through read mapping to assemblies, and the extent to which assemblies capture abundant sequence content beyond reference-matched representations. By applying this framework across metagenomes from

diverse biomes, this analysis provides a quantitative assessment of information loss and recovery during the metagenome assembly process.

Methods

a) Experimental Design and Workflow Implementation

To evaluate assembly-associated sequence loss across diverse ecological contexts, we analyzed 82 metagenomic samples previously processed through the genome-grist workflow. These samples span host-associated, engineered, and environmental biomes and provide a controlled framework for comparing reference-based and assembly-based representations of the same underlying sequencing data.

For each sample, we constructed three representations of the same data sets: (1) raw trimmed metagenomic reads, representing maximal observed sequence diversity prior to reconstruction; (2) assembled contigs, generated using MEGAHIT (v. 1.2.9), representing consensus reconstructions derived from *de Bruijn* graph assembly; and (3) reads mapped back to assemblies, aligned using minimap2 (v. 2.17), representing the subset of read-level content explainable by the assembled contigs.

This three-tier representation enables systematic partitioning of read-level sequence content into (i) content retained in assemblies, (ii) content collapsed during assembly but recoverable via read alignment, and (iii) content absent from both assembly and mapping. By comparing these representations within each sample, we directly quantify the magnitude and structure of sequence compression introduced during the read-to-assembly transition.

All analyses were implemented with a reproducible Snakemake workflow (Figure 2.1), which coordinated assembly, read mapping, k-mer sketching, and metric calculation under

consistent parameterization across samples. The workflow execution was parallelized execution on high-performance computing resources to ensure scalability and reproducibility.

b) Conceptual definition of content loss

Assembly-associated content loss was defined as sequence information present in raw metagenomic reads but absent from assembled contigs. Because *de novo* assembly intentionally removes sequencing artifacts and collapses redundant signals during graph simplification, not all loss reflects biological information reduction. However, loss becomes biologically consequential when genuine genomic variation, such as strain-specific alleles, accessory elements, or low-abundance taxa, is not represented in the final assembly.

To quantify content loss independently of assembly contiguity, gene annotation, or alignment-based bias, we adopted a k-mer centric framework. K-mers provide a uniform representation of sequence content across reads and assemblies, which enable direct comparison. This approach allows measurement of information retention and loss purely at the sequence level. Assembly-associated loss is therefore conceptualized as the subset of read-level k-mer signal not incorporated into assembled contigs.

c) Analytical Framework

Quantifying assembly-associated signal retention and loss

To better approximate the recovery of biologically supported sequences, we computed abundance-weighted k-mer retention (`assembly_f_weighted`) using Sourmash. Containment was calculated by comparing k-mer sets derived from assembled contigs against those derived from the corresponding raw trimmed reads. Specifically, each distinct k-mer was weighted by how often it appears in the reads before computing overlap with the assembly, so that common (high-abundance) k-mers have a larger influence on the final estimate than rare ones.

This abundance-weighted formulation provides an estimate of the fraction of total sequenced signal retained during assembly while downweighting rare and error-derived k-mers. Because each sequencing error introduces multiple low-abundance k-mers, metrics such as unweighted retention that treat all distinct k-mers equally are strongly influenced by sequencing noise in complex metagenomes. Abundance-weighted retention, in contrast, more accurately reflects recovery of well-supported genomic sequences.

Assembly-associated signal loss was then defined as: $1 - \text{assembly_f_weighted}$, representing the fraction of total sequenced k-mer signal not recovered in the assembled contigs. Samples were grouped by biome to evaluate whether assembly-associated signal loss differed systematically across ecological contexts (Figure 2.3).

Evaluating read-mapped recovery of abundance-weighted content

To determine how much assembly-associated content loss can be recovered by mapping reads back to assembled contigs, we compared two abundance-weighted retention measures for each metagenome. First, assembly weighted retention was calculated as the fraction of abundance-weighted read-level k-mers that are directly present in assembled contigs. This metric reflects the sequence content explicitly represented in the assembly.

Second, read-mapped weighted retention was computed by mapping reads back to the assembled contigs and measuring the fraction of abundance-weighted read content supported by alignments, including reads whose exact k-mers are not fully represented in contig sequences. This approach captures sequence variation and low-frequency differences that are collapsed during assembly but remain alignable to assembled consensus sequences.

For each sample, both metrics were derived from the same read set and assembly. Samples were ordered by assembly weighted retention to facilitate direct comparison of assembly-only representation and read-mapped recovery across the full spectrum of assembly performance (Figure 2.4).

Comparing reference-matched and total assembly content

To assess whether metagenome assembly recovers abundant sequence content beyond what is represented in reference databases, we compared reference-matched assembly retention to total assembly retention again using abundance-weighted k-mers. For each sample, abundance-weighted retention was computed in two ways: (i) considering only assembled k-mers that matched reference genomes (“assembly on reference”), and (ii) considering all assembled k-mers regardless of reference match (“total assembly”).

Reference matching was performed by intersecting assembled k-mers with those present in the reference database used for genome-grist analyses. Abundance-weighted retention reflects the fraction of total read-level k-mer abundance represented under each condition. The difference between total assembly retention and reference-matched assembly retention therefore quantifies abundant sequence content recovered by assembly that is not represented in the reference database.

Samples were grouped by biome (host-associated, engineered, environmental). Within each biome, samples were ordered by reference-based weighted retention to facilitate paired comparison between reference-matched and total assembly retention. Distributions across biomes were additionally summarized using boxplots to compare the magnitude and variability of novel assembly-recovered content (Figure 2.5).

d) Assessing the impact of sequencing depth and breadth on assembly performance

To evaluate whether poor assembly performance could be explained by undersampled or low-complexity metagenomes, we examined sequencing depth and breadth metrics (unpublished work) for each sample and tested their relationship with assembly retention.

Depth (Sequence Saturation). Sequencing depth was estimated using a signature-based saturation metric defined as: $\text{Saturation} = 1 - (\text{novel} / \text{total})$, where novel represents newly observed hashes and total represents the cumulative hashes observed in the signature. Higher saturation values indicate repeated observation of the same k-mers and therefore more complete sampling of the underlying sequence space. Samples with very low saturation (e.g., <25%) represent severely undersampled datasets.

Breadth (Signature Size). To assess coverage breadth, we calculated the total number of signature hashes for each metagenome and converted this value to approximate sequence representation. Signatures corresponding to <1 Mbp of sequence content were considered extremely small and likely insufficient to support robust assembly.

We examined whether samples failing the depth or breadth thresholds were associated with reduced assembly performance, to assess whether apparent assembly-associated content loss could be explained by insufficient sequencing depth or limited dataset breadth.

Results

2.1 Computational workflow for assembly loss and recovery analysis

To quantify how metagenome assembly alters the representation of sequence content, we developed a computational workflow that compares multiple representations of the same metagenomic dataset (Figure 2.1). This workflow extends the reference-based genome-grist presented in Chapter 1 by incorporating *de novo* assembly and read-mapping to assemblies, enabling direct comparison of sequence content retained, recoverable, or absent across different analytical strategies.

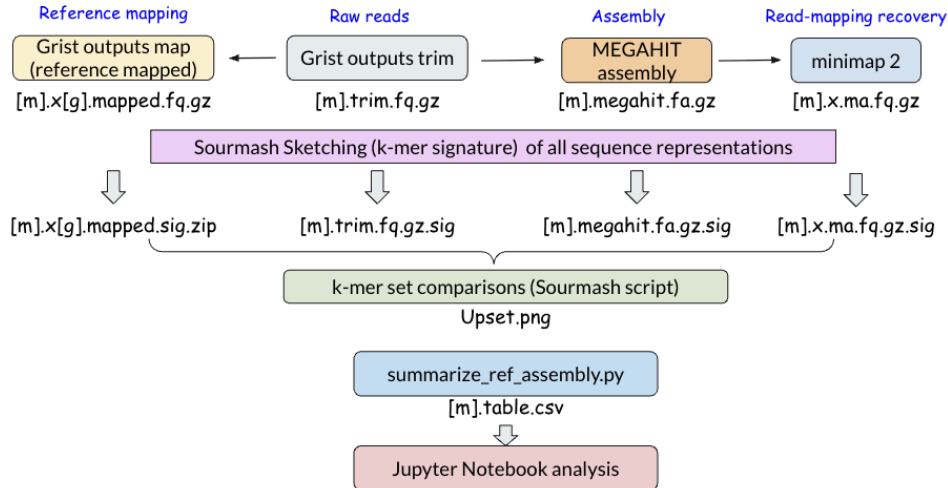


Figure 2.1: Computational framework for quantifying sequence retention, loss, and recovery during metagenome assembly.

For each metagenomic sample, four sequence representations were generated from the same set of trimmed reads:

- 1) raw trimmed reads represent the complete pool of sequenced fragments and therefore capture the maximal diversity present in the sample
- 2) reads mapping to reference genomes using genome-grist represent the fraction of sequence content assignable to known reference genomes
- 3) *de novo* assembly was performed using MEGAHIT, producing contigs that summarize read-level variation into consensus sequences
- 4) trimmed reads were mapped back to assembled contigs using minimap2 to identify read-level sequence content that is explainable by the assembly but not necessarily represented explicitly in assembled contigs.

To enable quantitative comparison among these representations, we used Sourmash to generate k-mer sketches from each dataset. These sketches summarize the k-mer composition of each sequence representation and allow efficient estimation of containment and overlap relationships among them. K-mer comparisons were performed using Sourmash scripts to

determine the extent of shared and unique k-mers across representations, and UpSet visualizations were generated to illustrate the structure of these overlaps (Supplementary Figure 2.1). Summary metrics were then calculated for each sample to quantify the fraction of read-level k-mer signal retained in assemblies, recoverable through read mapping to assemblies, or absent from both representations.

This framework allows the partitioning of read-level sequence content into three categories: k-mers retained in assembled contigs, k-mers absent from assemblies but recoverable through read mapping, and k-mers absent from both representations. The framework is applied in the following sections to characterize the magnitude and structure of assembly-associated sequence loss across metagenomes from diverse biomes.

2.2 Pilot analysis of metagenome representations

To illustrate how different analytical strategies capture metagenomic sequence content, we first conducted a pilot analysis using four representative metagenomes spanning diverse environments (Figure 2.2): human fecal (ERR505104), saltmarsh sediment (ERR257715), soil rhizosphere (SRR16117894), and Australian marine surface water (SRR8812665). For each sample, four sequence representations mentioned above (Result 2.1) were compared and overlaps among these representations were visualized using the Venn diagram.

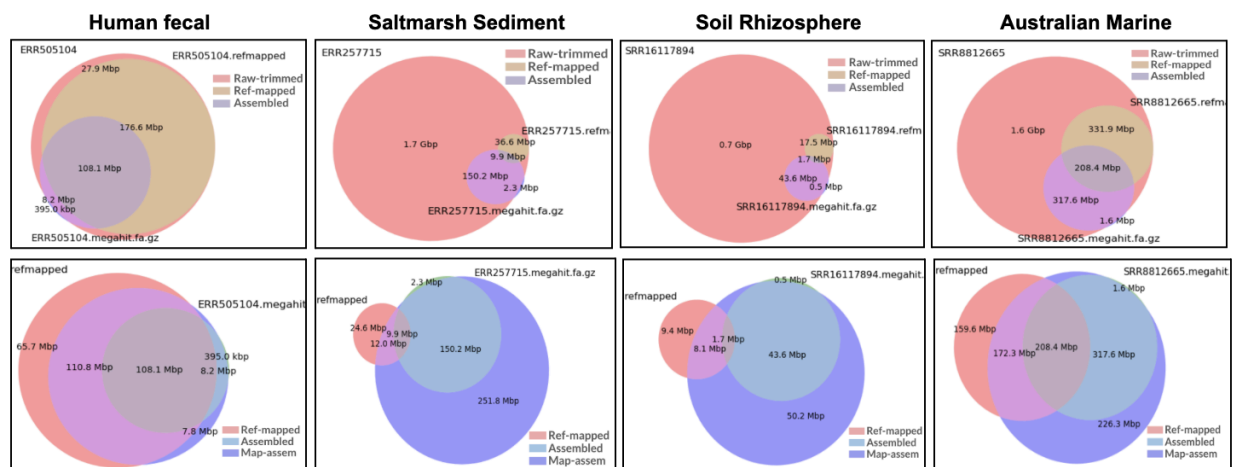


Figure 2.2: Pilot comparison of sequence representations across four different types of metagenomes.

Venn diagrams illustrate overlap among sequence representations. The top row compares raw reads, reads mapping to GTDB references (genome-grist), and MEGAHIT assemblies. The bottom row compares reference-mapped reads, assemblies, and reads mapping back to assemblies. Sequence amounts (bp) are shown for each region.

In the human fecal metagenome, where reference coverage is relatively comprehensive, reference-based mapping captured a large fraction of the sequence signal. However, assembly alone failed to represent a substantial portion of the read-level content, missing approximately 177 Mb of sequence relative to the read pool. Mapping reads back to the assembly recovered an additional 111 Mb beyond the assembled contigs, demonstrating that some sequence variation not explicitly represented in the assembly can still align to the consensus assembly.

In contrast, environmental samples with limited database representation showed markedly lower recovery through reference mapping. In the saltmarsh sediment and soil rhizosphere samples, MEGAHIT assemblies captured substantial amounts of sequence content absent from reference databases, recovering approximately 150 Mb and 44 Mb of novel genomic sequence, respectively. Mapping reads back to these assemblies further increased the amount of recoverable sequence signal, adding 252 Mb and 50 Mb of additional content.

The marine surface water sample exhibited intermediate behavior. Reference-based and assembly-based approaches each recovered comparable amounts of sequence content, while mapping reads back to the assembly contributed an additional 226 Mb of sequence beyond the assembly itself, including 172 Mb overlapping with the reference-mapped fraction.

These pilot analyses demonstrate that reference mapping, *de novo* assembly, and read mapping recover partially distinct sets of genomic content. Reference mapping performs well when representative genomes are present in databases, whereas assembly can reconstruct

genomic segments from organisms absent from references. Mapping reads back to assemblies further increases recovery by capturing read-level variation that is not explicitly represented in assembled contigs. These preliminary observations suggested that different analytical strategies capture complementary aspects of metagenomic sequence diversity, motivating the systematic analysis described below in which the same workflow was applied to a larger dataset of 82 metagenomes.

2.3 Assembly-associated content loss of abundance-weighted k-mer signal

For the large metagenome dataset, we first quantified assembly-associated content loss by measuring the fraction of abundance-weighted read-level k-mer signals not recovered in assembled contigs ($1 - \text{assembly_f_weighted}$). This metric estimates the proportion of total sequenced signal absent from the final assembly, downweighting low-abundance and error-derived k-mers, thereby emphasizing biologically meaningful content (Figure 2.2).

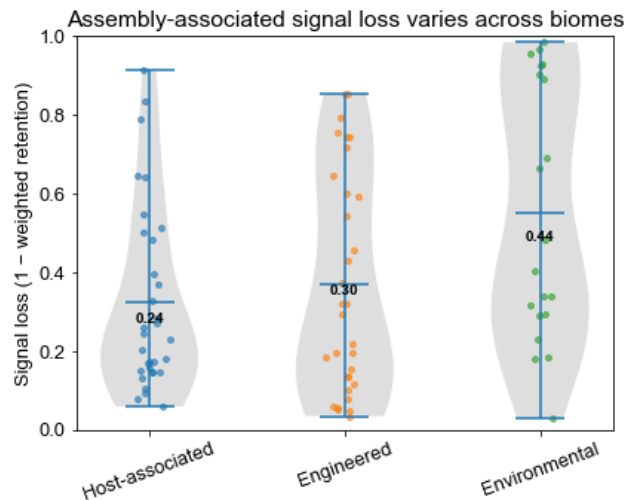


Figure 2.3: Assembly-associated signal loss across biomes.

Distribution of abundance-weighted k-mer loss for metagenomic assemblies grouped by top-level biome. Loss reflects the fraction of total sequenced k-mer signal not recovered in assembled contigs.

Across all samples, assembly-associated signal loss was not uniform. Host-associated metagenomes exhibited relatively low abundance-weighted loss (median = 0.24) with limited variability, indicating efficient recovery of dominant sequence content. Engineered systems showed intermediate loss (median = 0.30) and broader dispersion. In contrast, environmental metagenomes displayed substantially higher and more heterogeneous loss (median = 0.44), with several approaching near-complete loss of abundance-weighted signal. These results indicate that assembly-associated loss of abundance-weighted k-mer signal increases systematically with ecological complexity. Assemblies from complex environmental communities lose a larger fraction of sequenced signal than those derived from host-associated systems, consistent with differences in community complexity and coverage structure.

2.4 Recovery and expansion of sequence representation through assembly

Metagenome assemblies reconstruct genomes from raw sequencing reads. Although assembly collapses a fraction of read-level k-mer signal (Result 2.2), some missing sequence content may still be recoverable through read mapping to assemblies. To evaluate these complementary effects, we next examine the extent to which read mapping recovers assembly-associated signal loss and the degree to which assemblies capture abundant sequence content beyond reference-matched representations.

2.4.1 Read mapping consistently recovers additional abundance-weighted content

To assess recoverable assembly-associated content loss, we compared abundance-weighted retention from map-to-assembly (MA) to abundance-weighted retention represented directly in the assembly (Figure 2.4). Across all samples, the fraction of MA (orange points) consistently exceeds the fraction of abundance-weighted k-mers directly retained in the assembly (blue points). This indicates that a substantial portion of biologically real sequence

content is not explicitly represented in assembled contigs but can nonetheless be recovered through read mapping.

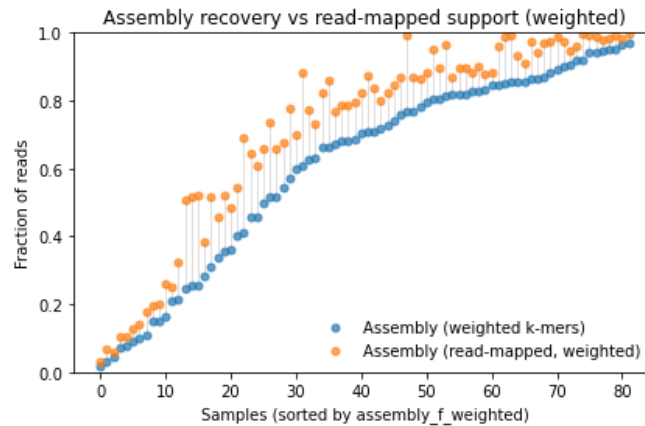


Figure 2.4: Read mapping partially recovers abundance-weighted sequence content lost during assembly.

Samples are ordered by abundance-weighted assembly retention (*assembly_f_weighted*). Blue points represent the fraction of abundance-weighted k-mers directly retained in assembled contigs, while orange points represent abundance-weighted retention after mapping reads back to the assembly (*assembly_f_readmapped_w*). Vertical segments indicate the magnitude of recoverable content for each metagenome.

The magnitude of recovery varies across samples and follows a non-linear pattern. When assembly weighted retention is very low ($< \sim 20\%$), read-mapped recovery provides only modest gains, suggesting that limited assembly representation constrains the ability of reads to map back effectively. Gains are largest among samples with intermediate assembly retention, consistent with strain-level variation being collapsed in contigs but still recoverable through read alignment to assembled consensus sequences. At the high end ($> \sim 90\%$), the difference between assembly and MA retention diminishes, as most abundance-weighted content is already captured in the assembly.

Across biomes, read mapping consistently recovers additional fraction of abundance-weighted signal beyond assembled contigs (Supplementary Figure 1.1). The magnitude of this difference is modest and broadly similar among host-associated, engineered, and environmental categories, with greater variability observed in engineered and environmental metagenomes compared to host-associated ones. Although mapping reads back to assembly can partially mitigate assembly-associated loss, MA retention remains below complete recovery for all samples and some abundance-weighted content is irrecoverably lost during assembly.

2.4.2 Assembly recovers abundant sequence content beyond reference-matched representations

To evaluate how much sequence content is captured by assembly relative to reference-based representations, we compared four quantities across samples: the fraction of reads mapping to reference genomes (Total Reference), the fraction of abundance-weighted k-mers retained in assembled contigs (Total Assembly), the overlap between assembly and reference-matched content (Assembly $\cap$ Reference), and the fraction of reads that are able to map back to assembled contigs (Map-to-Assembly, MA) (Figure 2.4).

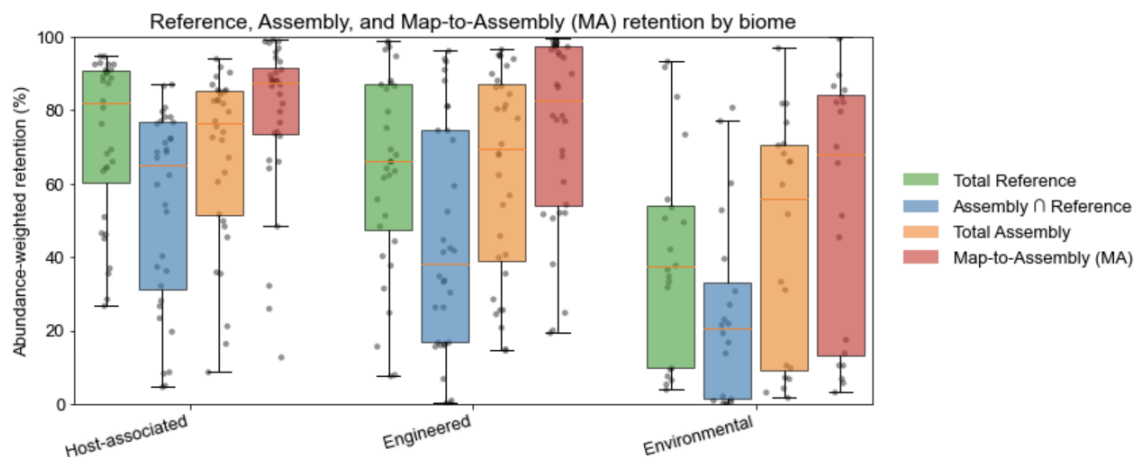


Figure 2.5: Novel abundant sequence recovered by assembly beyond reference-matched content.

Blue boxes indicated the K-mers used in the assembly that are also captured in reference-matching. Orange boxes indicated the k-mers captured by assembly. Green boxes represent k-mers matched by reference-based methods, while red boxes represent abundance-weighted signals recovered by mapping reads back to assembled contigs (MA).

In host-associated microbiomes, reference-based mapping (green) captures a large fraction of sequenced reads. Total assembly retention (orange) is similar to, but slightly lower than, reference-based retention for many samples, suggesting that assembly reconstructs most dominant genomic content while missing a small fraction of reads that map to known references. Because some reference-mapped reads are not represented in assembled contigs, and some assembled sequences lack reference matches, the overlap (blue) between assembly and reference representations is correspondingly reduced and cannot exceed the fraction represented in either dataset. Notably, mapping reads back to assemblies (red) increases the recoverable signal and approaches an upper bound of retention across samples. This increase likely reflects the recovery of sequence variation or read-level content that is not explicitly represented in assembled consensus contigs.

In engineered microbiomes, reference-based mapping (green) captures a moderate fraction of the abundance-weighted sequence signal, with substantial variability across samples. Assembly retention (orange) is broadly comparable to reference-based retention, although the median assembly fraction is slightly higher. Similar to host-associated systems, the overlap between assembly and reference-matched sequence is reduced in engineered microbiomes.

Unlike the other two biomes, assembly retention (orange) in environmental microbiomes is generally higher than reference-based (green) retention in most samples, with a substantially higher median value. This pattern again is consistent with the fact that a large portion of

abundant genomic content in these communities is not represented in current reference databases. Accordingly, the overlap (blue) between assembly and reference-matched sequence is relatively small. As observed in the other biomes, mapping reads back to assemblies (red) further increases the recoverable signal.

A small number of samples in the engineered and environmental categories show near-zero retention values. Examination of these outliers revealed that they contain very low numbers of reads and abundance-weighted k-mers (data not shown). These samples likely reflect technical limitations associated with very low sequencing depth and do not affect the overall biome-level trends.

2.5 Assembly retention is tightly constrained by sampling completeness

When plotting sequence saturation against assembly weighted retention, a strong and positive correlation was observed ($R^2 = 0.875$, $p = 9.6 \times 10^{-39}$, Figure 2.5). The relationship is monotonic and approximately linear across samples regardless of biome categories. For example, high saturation (> 0.8) consistently exhibits high retention ($> 0.8 - 0.95$), whereas low-saturation samples show uniformly poor retention. On the other hand, neither total nor novel number of signature hashes show correlation to the assembly retention (Supplementary Figure 2.2). The result implies that sequencing depth could be a dominant predictor of abundance-weighted assembly retention.

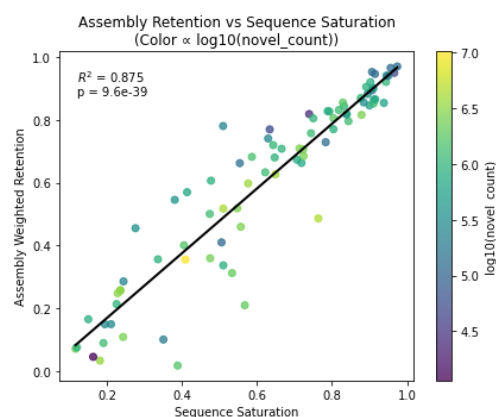


Figure 2.6: Relationship between sequencing saturation and assembly retention across metagenomes.

Each point represents a metagenome sample. The x-axis shows the sequencing saturation of the read-level signature. The y-axis shows abundance-weighted assembly retention. The solid line indicates the linear regression fit.

The pattern also has a mechanistic explanation. When sequence space is incompletely sampled, many k-mers appear only once and cannot be reliably incorporated into contigs, leading to reduced retention. As saturation increases, repeated observation of k-mers improves assembly continuity and representation of abundant genomic content. Insufficient sequencing depth therefore appears to be a major driver of assembly-associated content loss, suggesting that depth-based filtering may help separate undersampling effects from genuine biological complexity.

Discussion

1) Assembly preserves most abundance-weighted genomic signal despite measurable content loss

While *de novo* metagenome assembly enables recovery of genomic content absent from reference databases, it can also discard a fraction of the original sequencing signal during contig construction^{29,33,35}. Our analysis across the 82 metagenomes shows that metagenome assembly results in some loss of sequence information relative to the original sequencing reads, but most of the abundant sequence signal remains represented in the assembled contigs. Across all samples, assemblies retained the majority of abundance-weighted signal, with a median loss of approximately 24% in host-associated samples, 30% in engineered systems, and 44% in environmental metagenomes. These observations are consistent with previous

studies showing that metagenome assembly compresses read-level sequence diversity while still recovering most of the dominant genomic signal present in microbial communities^{36,55}.

The magnitude and variability of assembly-associated loss differed across environments. Environmental metagenomes exhibited both higher median loss and greater dispersion compared with host-associated and engineered systems. This pattern is consistent with the expectation that environmentally derived microbial communities contain greater strain-level heterogeneity and broader genomic diversity. Such diversity can challenge assembly algorithms, which reconstruct consensus sequences from mixtures of closely related genomes and may collapse or fragment highly variable regions. Consequently, some read-level sequence variation, particularly from low-abundance strains or polymorphic regions, may not be represented in the final assembled contigs. The results, however, also indicate that the majority of abundance-weighted sequence signal remains represented within assemblies, suggesting that the dominant genomic structure of microbial communities is generally recoverable through assembly.

In addition to biological factors, sequencing depth also influences the apparent magnitude of assembly-associated content loss. Our analysis shows a strong relationship between sequence saturation and abundance-weighted assembly retention. These observations highlight sequencing depth as an important technical contributor to assembly-associated content loss and emphasize the value of depth-based metrics for distinguishing undersampling effects from genuine biological complexity.

2) Read mapping reveals that much read-level content remains explainable by assemblies

Mapping reads back to assembled contigs consistently increased the fraction of abundance-weighted signal associated with assemblies. Across nearly all samples, the fraction

of reads mapping to assemblies exceeded the fraction of k-mers directly contained in the assembled contigs themselves. This result indicates that some sequence content absent from assembled contigs remains explainable by the assembly when reads are allowed to align to consensus regions.

This difference between assembly containment and read-mapped support likely reflects the consensus-building nature of assembly. Assemblers collapse closely related sequence variants into representative contigs, thereby removing some k-mer variation present in the raw reads. Although these variants may not appear explicitly in assembled sequences, reads derived from them can still align to the consensus assembly. As a result, read mapping can partially recover sequence signals that appear absent when assemblies are evaluated strictly at the k-mer level.

Notably, the difference between read-mapped support and direct assembly containment was somewhat larger in environmental samples, suggesting slightly greater strain-level variation in these communities. However, the magnitude of this difference remained modest across most samples, indicating that the additional diversity collapsed during assembly is limited relative to the total abundance-weighted signal present in the data.

3) Assembly complements reference-based representation of microbial communities

When compared to reference genome mapping, metagenome assemblies did capture genomic content that is not explained by reference genome mapping, although the extent of this expansion varies across environments. In host-associated and engineered microbiomes, assembly retention is broadly comparable to reference-based retention to reference genomes. This difference was particularly pronounced in environmental samples, where reference coverage remains limited.

These results highlight the complementary roles of reference-based and assembly-based approaches. While reference mapping efficiently captures sequence content corresponding to previously characterized genomes, assembly reconstructs genomic segments from abundant organisms lacking close representatives in reference databases. Consequently, assemblies provide an important mechanism for expanding the observable genomic representation, particularly in ecosystems where reference coverage remains incomplete.

4) Implications for genome-centric metagenomic analysis

Previous studies comparing metagenomic analysis strategies have typically focused either on benchmarking individual tools or evaluating specific pipelines under simulated conditions. Large community efforts such as the CAMI challenges provided valuable insights into the performance of taxonomic classifiers and assembly pipelines, but these studies generally examined a limited set of workflows or controlled datasets rather than systematic view of real biome diversity^{29,30}. As a result, the broader question of how different representations of metagenomic data: raw reads, reference-mapped reads, and assemblies, relate to one another across biomes has remained largely unexplored.

In this chapter, we address this gap by developing a unified analytical framework that systematically compares multiple representations of metagenomic sequence data across a large dataset spanning host-associated, engineered, and environmental microbiomes. The pipeline integrates genome-grist based reference mapping, *de novo* assembly, and read mapping back to assembled contigs within a common k-mer containment framework. By applying whole-genome sketching using Sourmash rather than marker-gene based profiling, such as SingleM²⁴, this approach evaluates genomic representation at the scale of entire genomes rather than a limited set of conserved markers. This enables a quantitative assessment of how much sequence content is captured by different analysis strategies and provides a scalable way to evaluate metagenomic information recovery across diverse datasets.

Our results re-stated that metagenome assembly plays a dual role in genome-centric metagenomic analysis. As expected from the consensus nature of assembly algorithms, assembly compresses read-level variation by collapsing closely related sequence variants into representative contigs. This process results in measurable reductions in k-mer diversity relative to the original read pool, reflecting both the removal of sequencing errors and the loss of biologically meaningful variation such as strain-level polymorphisms. At the same time, assemblies consistently recover substantial genomic content that is absent from existing reference databases, particularly in environmental microbiomes where reference coverage remains limited. These findings highlight the complementary roles of reference-based and assembly-based approaches for capturing microbial genomic diversity.

A key observation emerging from this analysis is the consistently high fraction of reads that map back to assembled contigs across many samples. Although assemblies do not explicitly represent all read-level k-mers, read mapping demonstrates that much of the underlying sequence signal remains explainable by the consensus assembly. In practice, this indicated that some sequence information lost during assembly can still be recovered when reads are aligned back to assembled contigs. This integrated representation therefore provides a more complete view of metagenomic sequence content than either reference mapping or assembly alone.

Importantly, by applying this framework across multiple biomes, our results suggest that the primary limitation of genome-resolved metagenomics may not lie in the assembly step itself. In many samples, assemblies capture a large fraction of the dominant genomic signal present in the reads. Instead, a major bottleneck likely arises downstream when assembled contigs must be partitioned into discrete genomes. Binning algorithms must separate sequences derived from mixtures of closely related organisms, a task that becomes increasingly difficult in communities with substantial strain-level diversity.

Beyond methodological comparisons, this work also provides a quantitative foundation for improving reference databases. By identifying biome-specific gaps in reference coverage and measuring the fraction of sequence content assignable to existing genomes, the framework developed here can guide targeted expansion of reference collections through incorporation of high-quality MAGs. Such targeted expansion may be particularly important for environmental microbiomes, where large fractions of microbial diversity remain poorly represented in current databases.

Together, these results demonstrate that different representations of metagenomic sequence data capture complementary aspects of microbial community structure. Reference mapping identifies sequences corresponding to known genomes, assembly reconstructs genomic content from abundant organisms regardless of reference availability, and read mapping back to assemblies recovers additional metagenomic content not explicitly represented in assembled contigs. By integrating these representations within a scalable k-mer based framework and applying it across diverse biomes, this study provides a quantitative foundation for evaluating metagenomic workflows and improving genome-resolved analysis of complex microbial communities.

Chapter 3. Quantifying Information Loss During Genome Binning

Abstract

Genome binning is widely used to reconstruct microbial genomes from metagenomic assemblies, enabling genome-resolved studies of microbial communities. However, it remains unclear how well genome bins represent the sequence diversity and taxonomic composition present in the original metagenomes. In this chapter, I quantified changes in sequence content and taxonomic representation across three stages of metagenomic processing, raw reads, assembled contigs, and genome bins, using k-mer based comparisons across 897 pig gut metagenomes. Assembly retained a substantial fraction of bacterial signal relative to reads, while genome binning dramatically reduced representation, with MAGs capturing only a small fraction of the original taxonomic signal. Archaeal representation was particularly limited, with near-zero recovery in genome bins across most samples. At the phylum level, many taxa detected in reads were either poorly represented in assemblies or entirely absent from genome bins.

Together, these results demonstrate that genome binning captures only a small and biased fraction of the biological information present in raw metagenomes, highlighting important limitations of genome-resolved metagenomic approaches.

Background

Genome binning serves as the cornerstone of genome-resolved metagenomics, facilitating the reconstruction of microbial genomes directly from complex environmental

assemblies. By clustering assembled contigs that share intrinsic biological signals, binning algorithms generate MAGs. These MAGs provide the necessary genomic context to investigate microbial physiology, evolution, and ecological interactions without the need for laboratory cultivation⁴². Over the past decade, advances in computational workflows have enabled the large-scale recovery of MAGs from diverse global environments, exponentially expanding microbial genome catalogs and revealing thousands of previously "dark" microbial lineages⁵⁶.

Genome binning methods, such as CONCOCT⁵⁷, MaxBin2⁵⁸, and MetaBAT2⁵⁹, typically rely on two primary signals: tetranucleotide frequencies (TNF), which serve as a genomic "signature," and coverage abundance profiles across multiple samples. Modern binning tools integrate these signals using statistical or machine-learning frameworks to improve clustering accuracy. For instance, VAMB employs deep variational autoencoders to encode k-mer and abundance features⁶⁰, while SemiBin2 utilizes self-supervised contrastive learning to extract feature embeddings⁶¹. Because different binning algorithms often recover partially overlapping genome sets, binning outputs are frequently refined using integration tools such as DASTool⁶², which integrates results from multiple algorithms and selects an optimized set of genome bins based on marker gene scoring.

Despite these advances, genome binning introduces important limitations. Binning methods depend on the availability of sufficiently long contigs, which themselves depend on successful assembly and adequate sequencing coverage. As a result, genomes from low-abundance organisms or highly fragmented assemblies are often excluded from bins^{29,42}. In addition, closely related strains may share similar sequence composition and coverage profiles, leading to collapsed bins that represent multiple strains or incomplete genome reconstructions^{36,63}. Mobile genetic elements (MGEs), plasmids, and antibiotic resistance genes, often lack the conserved k-mer signatures or coverage patterns of the host chromosome and are thus frequently excluded during the binning⁶⁴. These challenges can lead to biases in which genomes are recovered and how microbial diversity is represented.

Benchmarking efforts such as the CAMI have shown that binning accuracy varies widely depending on community complexity, sequencing depth, and strain heterogeneity^{29,30}. In complex communities with many closely related organisms, binning algorithms often struggle to separate genomes accurately, resulting in incomplete or chimeric bins⁶³. Even in relatively simple communities, only a subset of organisms present in the metagenome may ultimately be recovered as MAGs^{29,30}.

While previous studies have evaluated binning accuracy against reference genomes, there is a lack of direct quantification regarding how much of the original raw sequence information, the k-mer diversity and taxonomic breadth present in the initial reads, is actually captured in the final MAGs. In this chapter, I aim to quantify how genome binning alters the representation of metagenomic sequence content and taxonomic diversity in large pig metagenome datasets. Using k-mer based containment analysis and taxonomic profiling, I compare three representations of the same metagenomic datasets: raw reads, assembled contigs, and genome bins. This multi-stage comparison enables a direct measurement of the "information decay" that occurs during the binning process and provides insight into the extent to which MAG-based studies truly represent the underlying microbial community.

Methods

a) Dataset collection

A total of 897 pig gut metagenomes were analyzed in this study. These datasets were selected from the NCBI SRA using metadata queries performed in BigQuery⁶⁵. The initial SRA search was then subsequently filtered to include only datasets meeting the following criteria: whole-genome shotgun sequencing assay type, paired-end library layout, metagenomic library source, and dataset size greater than 1 GB. The size threshold was applied to ensure sufficient

sequencing depth for metagenome assembly and genome binning.

The selected datasets were previously processed by Anneliek ter Horst using the ATLAS metagenomic pipeline⁶⁶. Within this workflow, raw sequencing reads were first quality controlled and then assembled using MEGAHIT, with a minimum contig length threshold of 1,500 bp. Genome binning was performed using the standard ATLAS binning workflow, which applies multiple binning algorithms followed by refinement with DASTool to produce a final set of genome bins (MAGs).

b) K-mer content and taxonomic coverage comparison across representations

To quantify how taxonomic and sequence information changes across metagenomic processing steps, three representations of each metagenomic sample were analyzed independently: 1) the raw metagenome reads were aligned to the reference databases; 2) assembled contigs generated from those reads; 3) reads ended up in the genome bins. The design allowed direct comparison of the information retained at each stage of the metagenomic analysis pipeline.

All analyses were implemented in a reproducible Snakemake workflow (Figure 3.1). For each sample, FracMinHash sketches were first generated for the three data representations. Raw reads were sketched using the “read_sigs” rule, while assembled contigs and genome bins were sketched using the “contig_sigs”, and “bin_sigs” rules, respectively. Sketches were generated using sourmash with DNA k-mer sizes $k=21$, 31, 51, and abundance tracking enabled. For downstream analyses, $k=31$ sketches were used as the primary representation, as this k-mer size provides a balance between sensitivity to detect shared sequence and specificity to distinguish closely related genomes, and is commonly used in microbial comparative analyses.

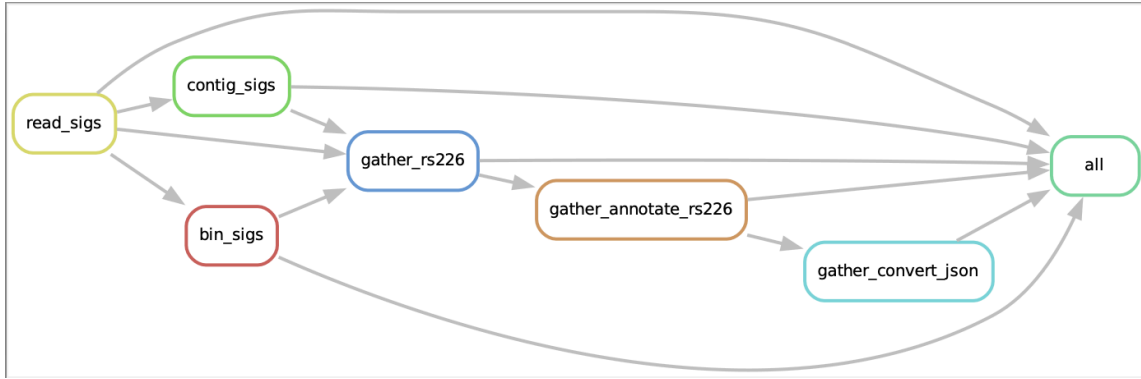


Figure 3.1: Snakemake workflow for k-mer and taxonomic comparison across metagenomic representations.

Raw metagenome reads, assembled contigs, and genome bins were independently converted into FracMinHash sketches and were queried against the reference database. Results were then annotated with taxonomic lineages using taxburst and used for downstream retention analyses and visualization.

The resulting sketches were subsequently queried against the same GTDB (rs226) using the “gather_rs226” rule, which applies the sourmash gather algorithm to identify the best matching reference genomes and estimate the fraction of query k-mers explained by each reference. Because identical reference database and sketch parameters were used for all representations, differences in gather results directly reflect changes in k-mer content introduced by assembly and binning rather than methodological artifacts. From the gather output, both the fraction of k-mer content explained by reference genomes and the associated taxonomic assignments were extracted for downstream comparisons.

c) Visualization of information flow across metagenomic representations

To visualize the flow of information across representations, we constructed boxplot and Sankey flow diagrams summarizing the transitions of k-mer content and taxonomy. Flows were aggregated at the superkingdom and phylum levels weighted by k-mer fraction. Each diagram

illustrates how much content from raw metagenomes is retained in assemblies and further retained in bins. These diagrams provide an intuitive, quantitative visualization of information bottlenecks introduced by assembly and binning, and the loss at each processing step.

Results

3.1 Overview of comparative framework and rationale

This chapter extends the read-level k-mer and taxonomic comparisons introduced in Chapter 2 by explicitly incorporating assembly and binning as intermediate representations of metagenomic data. Using a large cohort of previously assembled and binned metagenomes, we quantify how k-mer content and taxonomic signals propagate or fail to propagate, from raw reads, to assembled contigs, to genome bins.

To visualize and quantify information loss across these representations, we compare k-mer based taxonomic assignments using Sourmash gather across three data products and summarize transitions using Sankey flow diagrams. This approach enables a direct assessment of how much of the original metagenomic content is retained at each processing stage.

We compared k-mer content and taxonomic signal across three representations of the same 897 pig gut metagenomes: raw reads, assembled contigs, and genome bins. For each representation, k-mer based taxonomic profiles were generated using a consistent Sourmash sketching and gathering workflow against a fixed GTDB reference database. Comparisons focused on (i) k-mer content retention and (ii) taxonomic coverage across representations.

3.2 Superkingdom-level retention across assembly and genome binning

To quantify how genome-resolved metagenomic workflows affect microbial representation, we measured the fraction of taxonomic signal retained after assembly and

genome binning relative to the original metagenomic reads. We focused on the two major prokaryotic superkingdoms, Bacteria and Archaea, because they are the primary groups represented in current GTDB reference databases, enabling more reliable and comparable taxonomic assignment. In contrast, eukaryotic sequences are less comprehensively represented and more heterogeneous in metagenomic datasets, making retention estimates less consistent across samples. Retention was evaluated separately for Bacteria and Archaea, using abundance-weighted k-mer content assigned to each group. Metagenomic reads serve as the baseline (100% retention) and are not shown explicitly in the figure.

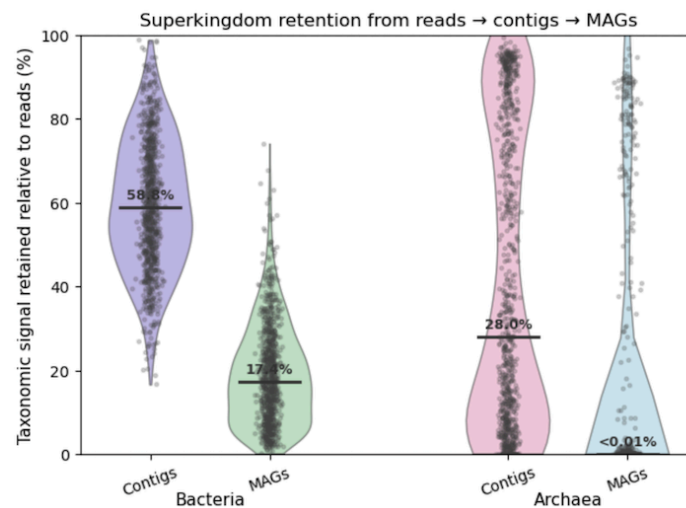


Figure 3.2: Superkingdom-level retention across 897 metagenome assembly and genome binning.

The fraction of taxonomic k-mer signal retained relative to reads is shown for Bacteria and Archaea after assembly (contigs) and genome binning (MAGs). Median retention values are indicated within each boxplot.

Across metagenomes, bacterial signals showed substantial but incomplete retention during assembly and were further reduced during genome binning (Figure 3.2). Assembled contigs retained a median of 58.8% of bacterial signal relative to reads. However, retention varied across samples, with some assemblies capturing 80 - 90% of bacterial signal, while others retained substantially less. Genome binning caused a further reduction in bacterial

representation. MAGs retained a median of 17.4% of bacterial signal, indicating that most bacterial genomic content present in assemblies is not recovered in bins. Although a subset of samples showed higher recovery, the majority exhibited low MAG retention.

Archaeal signals exhibited an even stronger reduction across the workflow. Assemblies retained a median of 28.0% of archaeal signal relative to reads, substantially lower than for bacteria. This likely reflects the generally lower abundance of archaeal taxa in many metagenomes, as assembly is sensitive to coverage. Recent large scale studies^{26,38,41} have shown that archaea typically comprise a small fraction of total community abundance across host-associated and environmental microbiomes. Consistent with this, archaeal signal is ~2x lower than bacterial signal at assembly stage across samples in our dataset. Following genome binning, archaeal representation declined dramatically, with near-zero median retention in MAGs, demonstrating that archaeal genomic content present in metagenomic reads and contigs is rarely recovered in genome bins.

Because taxonomic signals are derived from reference-matching k-mers, non-biological sequences such as sequencing errors or residual adapter contamination contribute minimally to these estimates. Thus, differences in retention primarily reflect biological and methodological effects rather than technical artifacts.

Our results here demonstrate that information loss accumulates across successive steps of genome-resolved metagenomic analysis. While assembly retains a substantial portion of bacterial signal and a smaller fraction of archaeal signal, genome binning dramatically reduces representation for both domains.

3.3 Progressive loss of taxonomic signal across metagenomes

To examine how microbial signal is partitioned across successive steps of genome-resolved metagenomic analysis, we quantified the fraction of phylum-level k-mer signal retained during assembly and genome binning. For each sample, the taxonomic signal was

divided into three components relative to the original reads: signal retained in genome bins (MAGs), signal present in assembled contigs but not incorporated into bins, and signal lost during assembly (Figure 3.3).

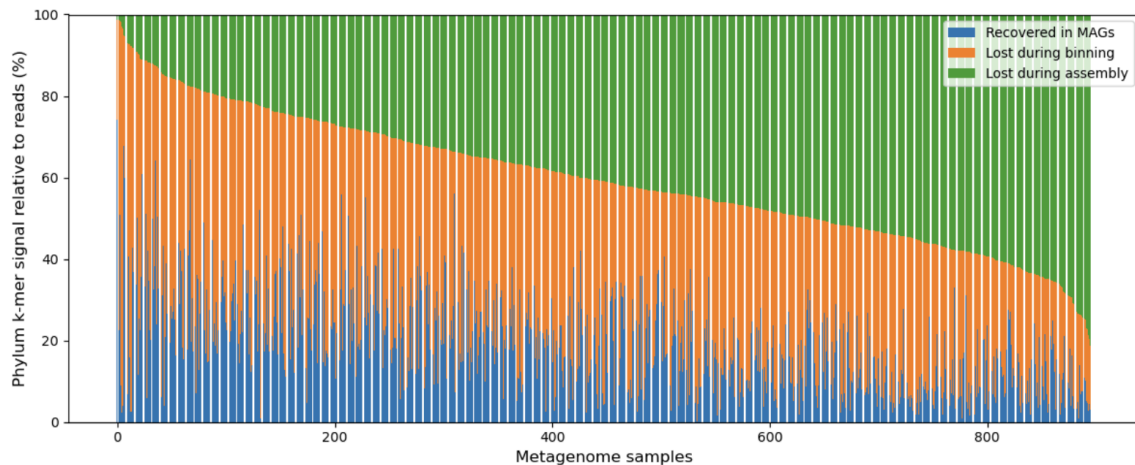


Figure 3.3: Partitioning of phylum-level taxonomic signals across genome-resolved metagenomic workflows.

For each metagenomic sample, the fraction of taxonomic k-mer signal retained relative to reads is shown for three categories: signal recovered in genome bins (MAGs), signal present in assembled contigs but not incorporated into bins, and signal lost during assembly. Samples are ordered by total signal retained in assembly.

Across metagenomes, microbial signal was progressively reduced across the workflow from reads to contigs to genome bins. Assemblies retained a substantial portion of the phylum-level signal present in reads, but a large fraction of this signal was subsequently lost during genome binning. In most samples, only a relatively small fraction of the original signal was recovered in MAGs, while a substantial proportion remained in assembled contigs that were not assigned to genome bins.

The extent of signal loss varied widely across samples. Some metagenomes retained a relatively large fraction of signal in genome bins, with up to approximately 60-70% of

phylum-level signal recovered in MAGs. However, most samples showed substantially lower recovery, with the majority of signal either remaining in unbinned contigs or lost during assembly. In many cases, signal loss during assembly represented a large component of the total loss, indicating that incomplete assembly contributes significantly to the overall reduction in microbial representation.

This wide variation across samples suggests that retention is not uniform. It could depend on properties of the underlying microbial communities. Samples with relatively high recovery may be consistent with lower-complexity communities dominated by a small number of abundant taxa, whereas samples with low recovery likely represent more diverse communities in which many taxa occur at low abundance and are therefore less recoverable through assembly and binning.

The results here clearly demonstrate that genome-resolved metagenomic workflows capture only a subset of the taxonomic signal present in the original sequencing data. While assembly preserves a portion of microbial signal across samples, genome binning consistently accounts for only a small fraction of the original phylum-level diversity.

3.4 Phylum-level retention patterns across assembly and genome binning

3.4.1 Global patterns of phylum-level signal loss

While the per-sample analysis demonstrates that substantial microbial signal is lost during assembly and genome binning, it does not reveal whether this loss is evenly distributed across microbial taxa. Differences in genome abundance, sequence composition, and assembly complexity may influence how effectively different lineages are recovered. To investigate whether taxonomic biases contribute to these patterns, we next examined retention of taxonomic signals at the phylum level across the metagenomes.

	Reads --> Contigs	Contigs --> MAGs
Fraction with Zero recovery	60.10%	86.90%
Median retention (≥ 1000 reads)	26.30%	0%
Median retention (top 10 phyla)	35.20%	19%

Table 3.1: Phylum-level retention across assembly and binning.

Retention is measured relative to the read-level signal for each phylum. Assembly retains a moderate portion of taxonomic signal, while genome binning dramatically reduces representation.

Across the dataset, 45,743 phylum - sample observations were analyzed from 897 metagenomes, representing 186 distinct phyla. Assembly retained a subset of the original taxonomic signal, while genome binning further reduced representation substantially.

At the assembly stage, approximately 60% of phylum observations showed zero recovery in contigs, indicating that many phyla present in the reads were not represented in the assembly at all. The effect was even more pronounced during genome binning: ~87% of phyla showed zero recovery in MAGs, demonstrating that most detected phyla were not captured in genome bins.

To reduce the influence of extremely low-abundance taxa and avoid unstable retention estimates driven by very small read counts, we restricted analysis to phyla supported by at least 1,000 reads per sample. Even after this filtering, assemblies retained only a modest fraction of the original read-level signal. The median phylum-level retention in contigs was 26.3%, indicating that roughly one quarter of the taxonomic signal detected in reads was represented in assembled sequences. In contrast, median retention in MAGs was 0%, indicating that more than half of these phyla were still not recovered in genome bins.

When focusing on the ten most abundant phyla across the dataset, retention improved but remained incomplete. Assemblies retained a median of 35.2% of read-level signal, while MAGs retained only 19%. Despite representing a small subset of overall microbial diversity,

these dominant phyla still exhibit substantial loss, indicating that genome binning captures only a minority of the underlying metagenomic signals.

3.4.2 Retention patterns among the eight most abundant microbial phyla

To examine whether retention of taxonomic signals vary across phyla, we quantified retention of taxonomic k-mer signals at the phylum level for the eight most abundant microbial phyla observed across metagenomes. For each phylum, we calculated the fraction of signal retained in assembled contigs and genome bins relative to the corresponding read-level signal across samples. Retention distributions for individual samples are shown in Figure 3.4, while median retention values are summarized in Figure 3.5.

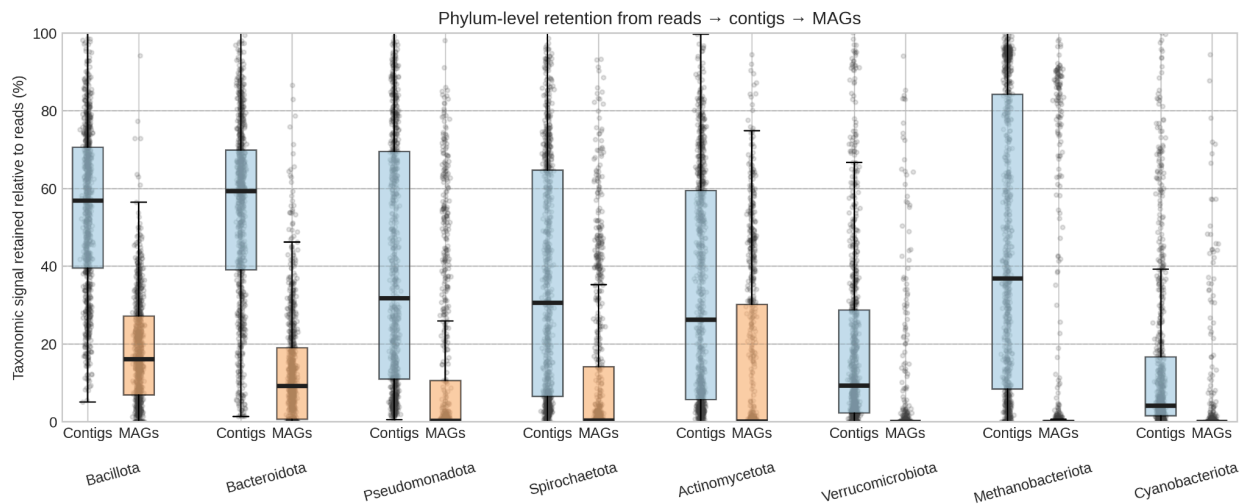


Figure 3.4: Taxonomic retention of top 8 phyla across assembly and binning.

Distributions of taxonomic k-mer signals retained relative to metagenomic reads are shown for the eight most abundant microbial phyla across samples. For each phylum, retention is calculated for assembled contigs and genome bins (MAGs). Each point represents a single metagenomic sample. The boxplots summarize the distribution of retention values across samples, with the median indicated by the central line and the interquartile range shown by the box.

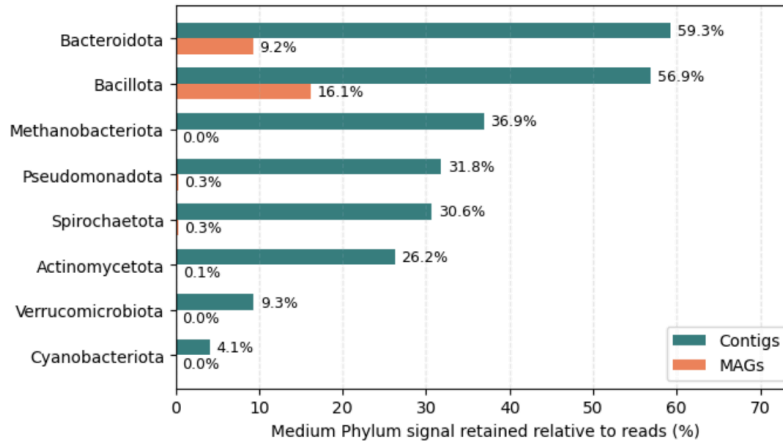


Figure 3.5: Median phylum-level retention relative to reads for contigs and bins.

Median fractions of taxonomic *k*-mer signal retained relative to metagenomic reads are shown for the eight most abundant microbial phyla. Retention is reported for assembled contigs and genome bins (MAGs).

Across phyla, assembly retained substantially more taxonomic signal than genome binning, although retention varied widely among groups (Figure 3.5). Bacteroidota and Bacillota, two dominant bacterial phyla in many host-associated microbiomes, showed the highest assembly retention, with median values of approximately 59.3% and 56.9%, respectively. Several additional phyla exhibited moderate assembly retention, including Methanobacteriota (36.9%), Pseudomonadota (31.8%), and Spirochaetota (30.6%). In contrast, Actinomycetota (26.2%), Verrucomicrobiota (9.3%), and Cyanobacteriota (4.1%) showed substantially lower assembly retention.

Genome binning resulted in a dramatic reduction in representation across all phyla. Even among phyla with relatively high assembly retention, only a small fraction of the signal was recovered in MAGs. For example, Bacillota retained a median of 16.1% of read-level signal in MAGs, while Bacteroidota retained 9.2%. For most other phyla, MAG retention was extremely limited, with median values close to zero. In particular, Methanobacteriota, Verrucomicrobiota, and Cyanobacteriota showed minimal recovery in genome bins across samples.

The distribution of retention values across samples (Figure 3.4) further highlights the variability of recovery across taxa. While some samples retained relatively large fractions of signal for certain phyla in assemblies, the majority showed substantially reduced representation in MAGs. This pattern indicates that although assembly may partially capture genomic content from many phyla, genome binning consistently recovers only a small subset of that signal, demonstrating the primary lineage-specific loss driven by binning rather than biological absence.

Together, these results show that signal loss increases from assembly to genome binning across all examined phyla. While assembly filters a portion of phylum-level diversity, binning imposes a much stronger constraint on taxonomic representation. As a result, many phyla present in metagenomic reads and assemblies are poorly represented or absent in genome bins, substantially narrowing the taxonomic breadth captured by MAG-based analyses.

Discussion

While metagenome-assembled genomes have greatly expanded microbial genome catalogs, the extent to which genome binning faithfully represents the full diversity present in metagenomic sequencing data remains poorly quantified. In this chapter, I compared k-mer content and taxonomic signals across three representations of metagenomic data: raw reads, assembled contigs, and genome bins, to directly measure how microbial information changes across successive steps of genome-resolved metagenomic analysis. Together, these results reveal that genome binning captures only a limited and biased subset of the microbial information present in raw metagenomic data.

1) Information loss accumulates across genome-resolved metagenomic workflows

Our results show that microbial signal progressively decreases from reads to assemblies to genome bins. While this general trend is expected, our analysis quantifies the magnitude of loss at each stage and identifies contributing factors. At the superkingdom level, assembled contigs retained a median of 58.8% of bacterial k-mer signal relative to raw reads, whereas genome bins dropped to a median of just 17.4%, indicating that approximately 70% of the genomic content successfully captured in assemblies fails to be incorporated into MAGs. Archaeal signals showed an even stronger reduction: assemblies retained a median of ~28% of raw signal, while genome bins exhibited near-zero median retention. These results indicate that although assembly preserves a substantial (though incomplete) portion of microbial sequence content, while genome binning introduces an additional and substantial reduction in taxonomic representation, likely driven by coverage-dependent assembly, uneven abundance distributions, and the inability of binning algorithms to recover fragmented or low-abundance genomes.

As indicated from previous studies, this loss reflects inherent constraints of current binning approaches, which rely on minimum contig lengths and consistent compositional and biological signals. Sequences that deviate from these patterns, such as fragmented assemblies from low-abundance, highly variable, or poorly assembled organisms, as well as mobile genetic elements, remain in the "unbinned fraction".

The results highlight both the scale and underlying drivers of information loss across genome-resolved metagenomic workflows, showing that these approaches do not simply reorganize metagenomic information into genomes but progressively filter the original sequence content. It confirms that this unbinned signal represents the majority of the original metagenomic diversity, suggesting that genome-resolved approaches may be overlooking the significant reservoir of genetic potential in microbial communities.

2) Taxonomic representation is uneven across microbial lineages

The loss of microbial signal is not evenly distributed across taxa. Phylum-level retention revealed strong variation in how different lineages are represented after assembly and binning. Dominant host-associated phyla such as Bacteroidota and Bacillota exhibited relatively high assembly retention (~ 59% and 57%, respectively), whereas other phyla such as Verrucomicrobiota and Cyanobacteriota showed substantially lower retention in assemblies.

Despite these differences at the assembly stage, genome binning consistently captured only a small fraction of the signal across nearly all phyla. Bacteroidota and Bacillota retained a median MAGs representation of only 9.2% and 16.1%, respectively. Many other phyla exhibited near-zero recovery in MAGs across most samples, added ~87% of phyla detected in the original reads showed zero recovery in the final bins. These results indicate that genome binning substantially narrows the taxonomic breadth represented in genome-resolved datasets, even for relatively abundant microbial groups.

3) Genome binning captures dominant taxa but misses community diversity

The progressive filtering observed in this study clearly shows that genome binning preferentially captures dominant and well-assembled organisms while excluding a large fraction of the broader microbial community. In many samples, a relatively small number of genome bins accounted for the majority of the recovered signal, indicating that binning tends to concentrate genomic representation into a limited set of abundant taxa. This observation is consistent with previous benchmarking studies showing that genome binning performance strongly depends on organism abundance and assembly contiguity^{29,30,63}. These results suggest that the transition from assemblies to genome bins substantially narrows the representation of microbial diversity present in metagenomic datasets. These patterns also provide a potential explanation for the substantial variation in signal retention observed across samples (Figure 3.3), where differences

in community structure may influence the extent to which microbial content can be recovered through assembly and binning.

Several characteristics of binning algorithms help explain this selective recovery. Most binning methods rely on sequence compositional signatures and coverage profiles to cluster contigs belonging to the same genome. Sequences that do not form clear coherent genome-scale grouping are therefore often excluded from bins. These sequences likely represent a mixture of low-abundance organisms, fragmented assemblies, strain variation, or genomic regions with atypical sequence composition. In addition, closely related strains sharing similar coverage patterns may collapse into incomplete or chimeric bins, further complicating genome reconstruction. As a result, even when assemblies contain fragments from taxonomically diverse organisms, the binning step may fail to recover them as coherent genomes.

These limitations highlight an important distinction between genome recovery and community representation. While MAGs provide valuable genomic context for dominant organisms, they do not necessarily represent the full taxonomic or functional diversity present in microbial communities. The strong reduction in archaeal representation observed here illustrates how some microbial groups may be systematically underrepresented in MAG-Based datasets. Thus, ecological interpretations based solely on genome bins may therefore disproportionately emphasize taxa that assemble well and form stable bins, while overlooking contributions from low-abundance lineages, mobile genetic elements, and antibiotic resistance genes that influence community function and adaptation.

Future work could explicitly test these hypotheses by correlating per-sample retention with community diversity metrics, such as Shannon diversity, richness, or coverage evenness, to determine how ecological complexity shapes genome recovery across metagenomes.

4) Limitations and future directions

The methodology of this Chapter enables a direct, quantitative comparison of k-mer content and taxonomic signal across reads, assemblies, and bins, which systematically evaluates how much metagenomic information survives common assembly and binning workflows. However, several limitations of this analysis should be considered. First, taxonomic assignments were derived from k-mer based comparisons to a reference genome database. Although this approach provides a consistent framework for comparing representations, it depends on the completeness and taxonomic coverage of the reference database. Some microbial sequences present in the metagenomes may therefore remain unclassified.

Second, the present analysis focused on k-mer containment and taxonomic representation rather than functional content. It remains possible that genome bins capture a larger fraction of key functional genes than suggested by taxonomic retention alone. Future work could extend this analysis to evaluate functional gene representation across reads, contigs, and genome bins.

Finally, the results presented here reflect a specific dataset and workflow, including assembly with MEGAHIT and binning using the ATLAS pipeline. Different assembly or binning strategies may alter the magnitude of information loss observed. Nevertheless, the consistent patterns observed across hundreds of metagenomes suggest that the progressive reduction in microbial signal is a general feature of genome-resolved metagenomic analysis.

5) Conclusions

Overall, this study provides a quantitative assessment of how microbial information changes across genome-resolved metagenomic workflows. By comparing k-mer content and taxonomic signals across reads, assemblies, and genome bins, we developed a scalable analytical framework for evaluating information retention across successive stages of

metagenomic analysis. Our results showed that while assembly preserves a substantial fraction of microbial signal, genome binning dramatically narrows the taxonomic representation of the community in raw metagenomic data.

These findings suggest that genome-resolved metagenomics should be interpreted as a powerful but selective view of microbial communities. Genome bins provide valuable genomic context for dominant organisms, but they do not fully represent the breadth of microbial diversity present in the underlying metagenome. Alternative analytical strategies based on k-mer analysis, read-level profiling, or reference-free sequence comparison may provide complementary perspectives that are not limited by the constraints of genome reconstruction.

Recent advances in machine learning have begun to address some of the limitations of traditional genome binning methods. Deep learning approaches such as VAMB⁶⁰ and SemiBin2⁶¹ use neural networks to learn representations of contigs that integrate sequence composition and coverage signals, improving genome recovery from complex communities. Graph-based methods such as GraphMB⁶⁷ further incorporate assembly graph information to better cluster fragmented contigs. In parallel, deep learning models for DNA sequence representation, including DNABERT⁶⁸, offer new opportunities for read-level classification and reference-free sequence analysis. Integrating these approaches with genome-resolved metagenomics may further help recover microbial diversity that is currently lost during assembly and binning.

Supplementary material

query	p_genom	avg_abu	p_metag	metagenom	query	p_genome	avg_abun	p_metag	metagenom
mm39.fa.gz	89.80%	4.7	70.00%	SRR18439526	hg38.all	26.20%	1.6	87.80%	ERR2749187
mm39.fa.gz	89.40%	4.3	76.90%	SRR18439531	hg38.all	23.00%	1.6	81.30%	ERR2749913
mm39.fa.gz	88.60%	4.4	76.60%	SRR18439537	hg38.all	20.00%	1.6	77.80%	ERR2750373
mm39.fa.gz	88.50%	4.4	68.70%	SRR18439549	hg38.all	11.50%	1.5	51.30%	ERR2750405
mm39.fa.gz	87.60%	4.2	71.20%	SRR18439548	hg38.all	6.30%	1.4	88.60%	ERR2750515
mm39.fa.gz	87.60%	4.3	66.60%	SRR18439532	hg38.all	31.90%	1.7	86.90%	ERR2750639
mm39.fa.gz	86.40%	3.9	76.80%	SRR18439533	hg38.all	52.40%	2.7	58.90%	ERR3274631
mm39.fa.gz	85.40%	4	68.90%	SRR18439530	hg38.all	39.00%	2	85.40%	ERR3593541
mm39.fa.gz	84.50%	3.6	65.90%	SRR18439539	hg38.all	29.10%	3.2	57.10%	SRR17301062
mm39.fa.gz	83.80%	3.9	58.80%	SRR18439528	hg38.all	43.90%	2.6	71.90%	SRR1950750
mm39.fa.gz	83.40%	3.9	60.00%	SRR18439546	hg38.all	16.00%	2.1	75.40%	SRR1950765
mm39.fa.gz	82.70%	3.8	62.00%	SRR18439527	hg38.all	6.60%	2.6	70.40%	SRR1950771
mm39.fa.gz	82.60%	4	59.70%	SRR18439529	hg38.all	14.50%	2.1	85.70%	SRR1950773
mm39.fa.gz	82.20%	3.5	53.00%	SRR18439545	hg38.all	12.90%	3.4	86.60%	SRR1950776
mm39.fa.gz	79.00%	3.3	50.90%	SRR18439542	hg38.all	12.90%	1.9	66.30%	SRR1950787
mm39.fa.gz	78.70%	3.5	52.50%	SRR18439547	hg38.all	5.20%	24.3	59.80%	SRR4420316
mm39.fa.gz	77.60%	3.2	52.80%	SRR18439538	hg38.all	63.40%	6.4	84.10%	SRR4420318
mm39.fa.gz	77.20%	3.2	52.10%	SRR18439535	hg38.all	15.60%	12.6	66.00%	SRR4420319
mm39.fa.gz	77.20%	3.2	47.30%	SRR18439541	hg38.all	44.40%	2.4	54.30%	SRR6145077
mm39.fa.gz	75.00%	3.1	43.10%	SRR18439544	hg38.all	40.70%	2.3	62.60%	SRR6145084
mm39.fa.gz	73.90%	3	48.80%	SRR18439540	hg38.all	11.80%	2.1	51.40%	SRR6216751
mm39.fa.gz	72.40%	2.9	46.80%	SRR18439534	hg38.all	9.60%	2	62.80%	SRR6216758
mm39.fa.gz	70.10%	2.9	42.80%	SRR18439536	hg38.all	7.80%	2	52.20%	SRR6216779
mm39.fa.gz	69.20%	2.7	41.90%	SRR18439543	hg38.all	11.50%	1.9	60.80%	SRR6216796
mm39.fa.gz	50.60%	2.8	43.20%	SRR17257133	hg38.all	8.70%	2	75.90%	SRR6216867
					hg38.all	23.30%	8.1	85.10%	SRR6757611
					hg38.all	43.00%	2.4	69.50%	SRR6757613
					hg38.all	34.20%	5.4	82.20%	SRR6757641
					hg38.all	26.20%	1.9	82.80%	SRR9684990

Supplementary Table 1.1: Samples with the highest mouse- and human-derived signals and their study-level grouping

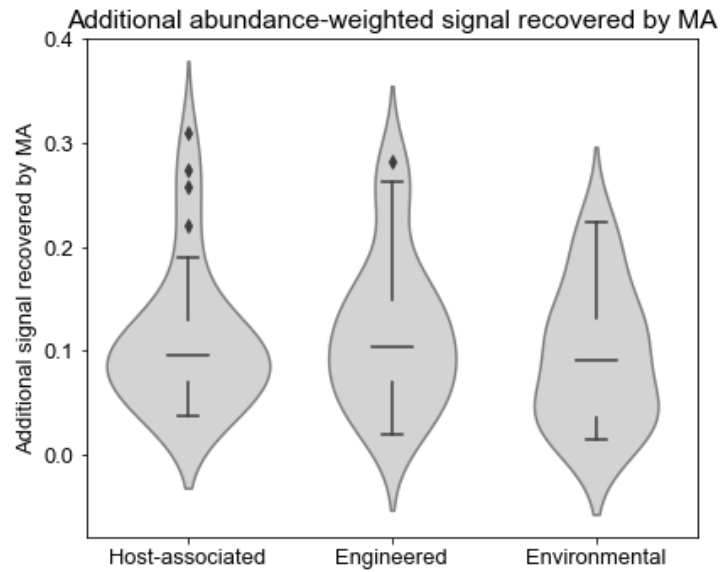
Query is the host genome identifier (e.g., mm39.fa.gz for mouse, hg38.all for human).

p_genome: Maximum percentage of the host genome detected in the metagenome. **avg_abun**:

Average abundance of the host genome across the sample. **p_metag**: Overall signal

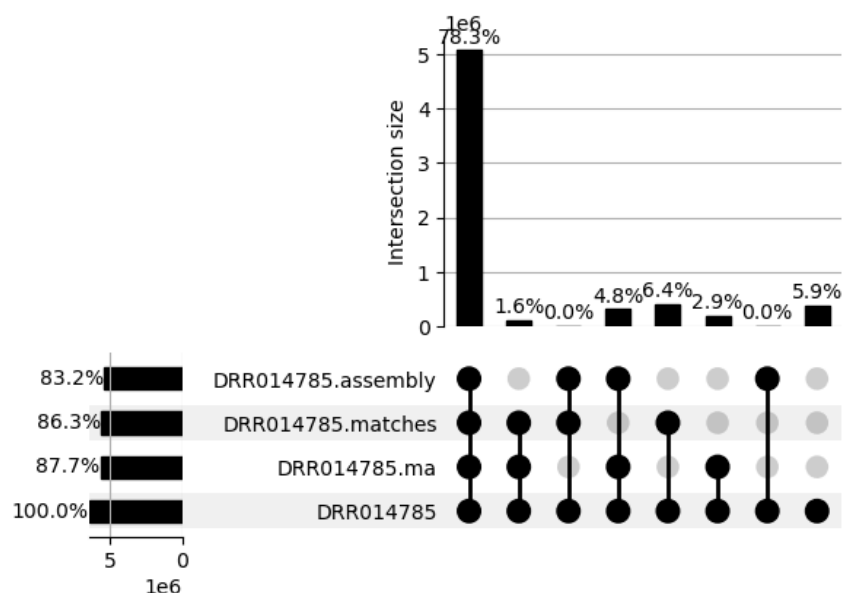
percentage in the metagenome. **Metagenome**: Sample ID, linking each sample to its study.

Among samples with the highest mouse-derived signal detected, 24 of the 25 (blue-highlighted) originate from a single study (BioProject: PRJNA815731). For human samples, highlighted entries indicate groups of samples belonging to the same BioProject. For example, four samples with accessions ERR2750xxx correspond to a shared BioProject study, while six samples with accessions ERR1950xxx belong to a different BioProject study.



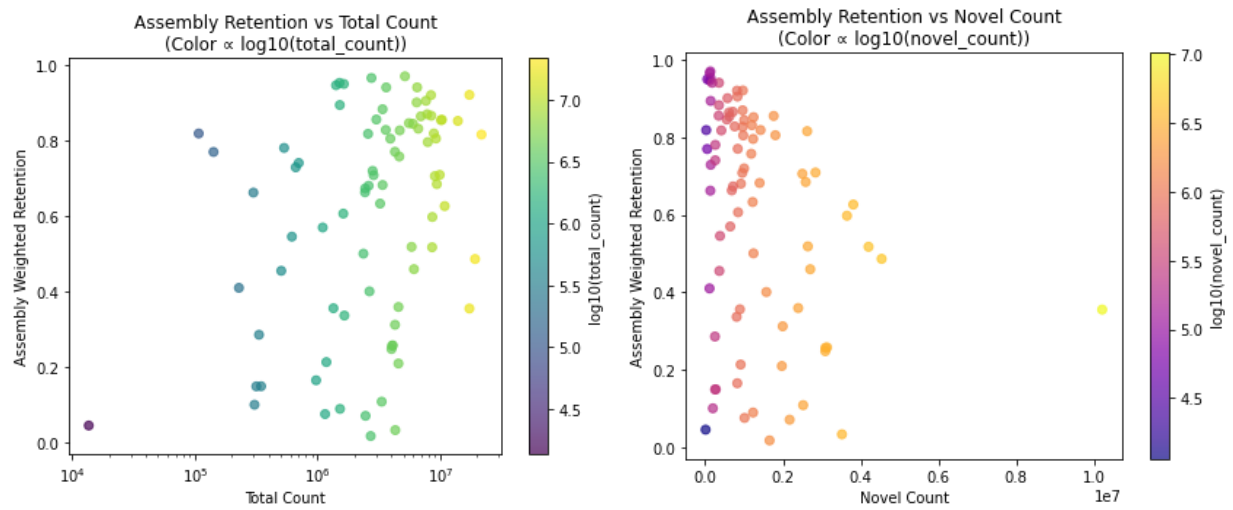
Supplementary Figure 1.1: Additional abundance-weighted signal recovered by read mapping across biomes

Violin plots show the distribution of the difference between map-to-assembly (MA) retention and assembly retention ($\Delta = \text{MA weighted} - \text{assembly weighted}$) for host-associated, engineered, and environmental metagenomes. Positive values indicate additional abundance-weighted k-mer signal recovered when reads are mapped back to assembled contigs relative to assembled contigs alone.



Supplementary Figure 2.1: Example of Atta Snakefile Upset diagram

The UpSet plot summarizes shared and unique k -mer content among four data representations of a single Host-associated:Mammals:Respiratory system metagenome sample (DRR014785): raw reads, assembled contigs (“assembly”), reads mapped back to assembly (“ma”), and reference-matched reads (“matches”). The horizontal bars (left) indicate the total fraction of abundance-weighted k -mers retained in each representation relative to raw reads, while the vertical bars (top) show the size of intersections between combinations of datasets. The largest intersection corresponds to k -mers shared across all representations, whereas smaller intersections represent content uniquely retained or lost at specific stages.



Supplementary Figure 2.2: Assembly Retention vs. Total/Novel Read Count

Scatter plots showing the relationship between assembly weighted retention and (left) total read count and (right) novel (non-reference-matching) read count across metagenomic samples. Each point represents one sample, with color indicating $\log_{10}$ -transformed total read count (left) or novel read count (right).

Bibliography

1. Falkowski, P. G., Fenchel, T. & Delong, E. F. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science* **320**, 1034–1039 (2008).
2. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12**, 69 (2014).
3. Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z. & Ettema, T. J. G. Innovations to culturing the uncultured microbial majority. *Nat. Rev. Microbiol.* **19**, 225–240 (2021).
4. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**, 10.1128/msystems.00055-18 (2018).
5. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
6. Handelsman, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiol. Mol. Biol. Rev.* **68**, 669–685 (2004).
7. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
8. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
9. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).
10. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**, 150023 (2015).
11. Tara Ocean Foundation *et al.* Priorities for ocean microbiome research. *Nat. Microbiol.* **7**, 937–947 (2022).

12. Wu, X. *et al.* Microbial life in 25-m-deep boreholes in ancient permafrost illuminated by metagenomics. *Environ. Microbiome* **18**, 33 (2023).
13. Varliero, G. *et al.* Biogeographic survey of soil bacterial communities across Antarctica. *Microbiome* **12**, 9 (2024).
14. Cao, S. *et al.* Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. *Microbiome* **8**, 47 (2020).
15. Herbold, C. W. *et al.* Nutritional niches of potentially endemic, facultatively anaerobic heterotrophs from an isolated Antarctic terrestrial hydrothermal refugium elucidated through metagenomics. *Environ. Microbiome* **19**, 104 (2024).
16. Tighe, S. *et al.* Genomic Methods and Microbiological Technologies for Profiling Novel and Extreme Environments for the Extreme Microbiome Project (XMP). *J. Biomol. Tech. JBT* **28**, 31–39 (2017).
17. Salido, R. A. *et al.* The International Space Station has a unique and extreme microbial and chemical environment driven by use patterns. *Cell* **188**, 2022–2041.e23 (2025).
18. Schultz, J. *et al.* Genomic insights into novel extremotolerant bacteria isolated from the NASA Phoenix mission spacecraft assembly cleanrooms. *Microbiome* **13**, 117 (2025).
19. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
20. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
21. Shaw, J. & Yu, Y. W. Rapid species-level metagenome profiling and containment estimation with sylph. *Nat. Biotechnol.* **43**, 1348–1359 (2025).
22. Chorlton, S. D. Ten common issues with reference sequence databases and how to mitigate them. *Front. Bioinforma.* **4**, 1278228 (2024).
23. Knight, R. *et al.* Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).

24. Woodcroft, B. J. *et al.* Comprehensive taxonomic identification of microbial species in metagenomic data using SingleM and Sandpiper. *Nat. Biotechnol.*
<https://doi.org/10.1038/s41587-025-02738-1> (2025) doi:10.1038/s41587-025-02738-1.
25. Yang, C. *et al.* A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* **19**, 6301–6314 (2021).
26. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
27. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).
28. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
29. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
30. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
31. Treichel, N. S., Hitch, T. C. A., Penders, J. & Clavel, T. Benchmarking of shotgun sequencing depth highlights strain-level limitations of metagenomic analysis. Preprint at <https://doi.org/10.1101/2025.03.27.645659> (2025).
32. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
33. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).

34. Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl. Environ. Microbiol.* **87**, e02593-20 (2021).
35. Vollmers, J., Wiegand, S. & Kaster, A.-K. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PLOS ONE* **12**, e0169662 (2017).
36. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
37. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
38. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
39. Liu, Y.-X. *et al.* A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell* **12**, 315–330 (2021).
40. Prasoodanan Pk, V. *et al.* Unbinned contigs expand known diversity in the global microbiome. *Nat. Microbiol.* <https://doi.org/10.1038/s41564-026-02314-6> (2026)
doi:10.1038/s41564-026-02314-6.
41. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
42. Liu, S. *et al.* Analysis of metagenomic data. *Nat. Rev. Methods Primer* **5**, 5 (2025).
43. Chen, I.-M. A. *et al.* The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.* **51**, D723–D732 (2023).
44. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

45. Irber, L. *et al.* Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. Preprint at <https://doi.org/10.1101/2022.01.11.475838> (2022).
46. Titus Brown, C. & Irber, L. sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.* **1**, 27 (2016).
47. Pierce, N. T., Irber, L., Reiter, T., Brooks, P. & Brown, C. T. Large-scale sequence comparisons with sourmash. *F1000Research* **8**, 1006 (2019).
48. Irber, L. *et al.* sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets. *J. Open Source Softw.* **9**, 6830 (2024).
49. Irber, L., Pierce-Ward, N. T. & Brown, C. T. Sourmash Branchwater Enables Lightweight Petabyte-Scale Sequence Search. Preprint at <https://doi.org/10.1101/2022.11.02.514947> (2022).
50. C Titus Brown, Tessa Pierce Ward, Mohamed Abuelanin, Marisa Lim. dib-lab/genome-grist: v0.7.4. (2021).
51. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
52. Parks, D. H. *et al.* GTDB release 10: a complete and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes. *Nucleic Acids Res.* **54**, D743–D754 (2026).
53. 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 (2015).
54. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
55. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).

56. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
57. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
58. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
59. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
60. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
61. Pan, S., Zhao, X.-M. & Coelho, L. P. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics* **39**, i21–i29 (2023).
62. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
63. Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5**, 10.1128/msystems.00731-19 (2020).
64. Maguire, F. *et al.* Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microb. Genomics* **6**, (2020).
65. Tessa Pierce Ward. BigQuery Metadata search. (2023).
66. Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M. & McCue, L. A. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics* **21**, 257 (2020).
67. Lamurias, A., Sereika, M., Albertsen, M., Hose, K. & Nielsen, T. D. Metagenomic binning with assembly graph embeddings. *Bioinformatics* **38**, 4481–4487 (2022).

68. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).