Journal Homepage: www.ijrpr.com ISSN: 3049-0103 (Online)

International Journal of Advance Research Publication and Reviews

Vol 02, Issue 07, pp 205-226, July 2025

Quantifying Microbial Dark Matter Diversity through Hybrid Assembly and Strain-Resolved Binning from Ultra-Deep Shotgun Metagenomic Datasets.

Sandra Okoye Chiamaka

Department of Biological sciences, Eastern Illinois University Charleston, USA

DOI : https://doi.org/10.55248/gengpi.6.0725.2577

ABSTRACT

The vast majority of microbial life remains uncultivated and poorly characterized, often referred to as "microbial dark matter." These elusive taxa hold critical roles in ecosystem functionality, human health, and biogeochemical cycling, yet remain inaccessible due to technical limitations in cultivation and genome reconstruction. Recent advancements in high-throughput sequencing have opened new avenues to investigate this hidden diversity, but standard metagenomic assembly and binning approaches frequently fall short when faced with strain heterogeneity and low-abundance organisms. This study presents a comprehensive strategy to quantify microbial dark matter diversity using hybrid assembly and strain-resolved binning from ultra-deep shotgun metagenomic datasets. We employ a multi-platform approach combining long-read (e.g., Nanopore or PacBio) and short-read (e.g., Illumina) sequencing to produce high-contiguity assemblies while preserving base-level accuracy. These hybrid assemblies enable recovery of complete and near-complete genomes from complex microbial communities, including rare and uncultured taxa. By integrating advanced strain-level binning techniques such as co-abundance-based partitioning, single-nucleotide variant profiling, and graph-based dereplication we achieve unprecedented resolution in differentiating closely related strains. Diversity metrics, phylogenomic reconstruction, and functional annotation further quantify the ecological and evolutionary significance of the reconstructed genomes. Applied to datasets from diverse environments (e.g., soil, gut, and marine microbiomes), our framework reveals a substantial expansion in the microbial tree of life and a more accurate estimate of microbial richness and genomic novelty. This work provides a scalable and replicable workflow for researchers aiming to resolve microbial dark matter and enhances our understanding of microbial community dynamics, evolution, and function in natural ecosystems.

Keywords: Microbial dark matter, Hybrid assembly, Shotgun metagenomics, Strain-resolved binning, Genome reconstruction, Microbial diversity quantification

1. INTRODUCTION

1.1 The Challenge of Microbial Dark Matter

Microbial dark matter refers to the vast array of uncultivated microbial taxa that remain unexplored due to the limitations of traditional microbiological techniques. These organisms, often constituting over 90% of microbial diversity in many environments, play essential roles in ecological processes such as biogeochemical cycling, symbiosis, and pathogenesis [1]. Despite their ubiquity in terrestrial, aquatic, and host-associated ecosystems, a significant portion of these taxa has resisted laboratory cultivation, thereby eluding conventional genomic characterization and functional profiling.

The inability to culture most microbes stems from the narrow range of environmental conditions typically replicated in laboratory settings. Many organisms require specific metabolic dependencies, syntrophic interactions, or stress tolerances that are difficult to recreate ex situ [2]. Furthermore, some microbes exist in dormant states or have slow growth rates that fall below detection thresholds during standard enrichment or isolation techniques. This has led to a disproportionate understanding of microbial ecosystems, often skewed toward fast-growing and easily cultured species.

The ecological implications of this knowledge gap are substantial. Without insights into the metabolic pathways and ecological roles of these uncultured organisms, our understanding of nutrient turnover, ecosystem resilience, and microbial interactions remains incomplete [3]. Additionally, the potential for discovering novel enzymes, bioactive compounds, or antimicrobial resistance genes within these uncharacterized lineages is severely constrained. Figure 1 provides a conceptual overview of microbial dark matter and the methodological bottlenecks that hinder access to it.

Addressing the challenge of microbial dark matter requires a paradigm shift from reliance on axenic culturing toward cultivation-independent, high-resolution genomic strategies capable of resolving organisms directly from complex environmental samples. These approaches, and their current limitations, are outlined in the following section.

1.2 Advances in Metagenomics and Their Limits

Metagenomics has emerged as a transformative approach to characterizing microbial communities by directly sequencing DNA extracted from environmental samples. High-throughput platforms such as Illumina and Oxford Nanopore enable researchers to capture the collective genomic content of mixed populations without the need for cultivation [4]. These technologies have revealed previously unknown microbial taxa, metabolic pathways, and mobile genetic elements, significantly expanding our understanding of microbial ecosystems.

The analytical process typically involves assembling short sequencing reads into longer contiguous sequences (contigs), followed by binning an algorithmic clustering of contigs into metagenome-assembled genomes (MAGs). Binning facilitates the reconstruction of draft genomes, which can then be taxonomically classified and functionally annotated [5]. However, this process faces notable challenges in highly diverse or low-biomass communities, where strain-level heterogeneity and high genomic similarity complicate accurate assembly.

One limitation lies in the difficulty of distinguishing closely related strains, particularly when using short-read data, which may result in chimeric contigs or collapsed assemblies. This loss of resolution hinders the identification of intra-species variation and functional differentiation, critical for understanding ecological interactions [6]. Additionally, metagenomic assemblies may exclude low-abundance organisms entirely due to coverage biases, perpetuating the invisibility of rare but functionally important taxa.

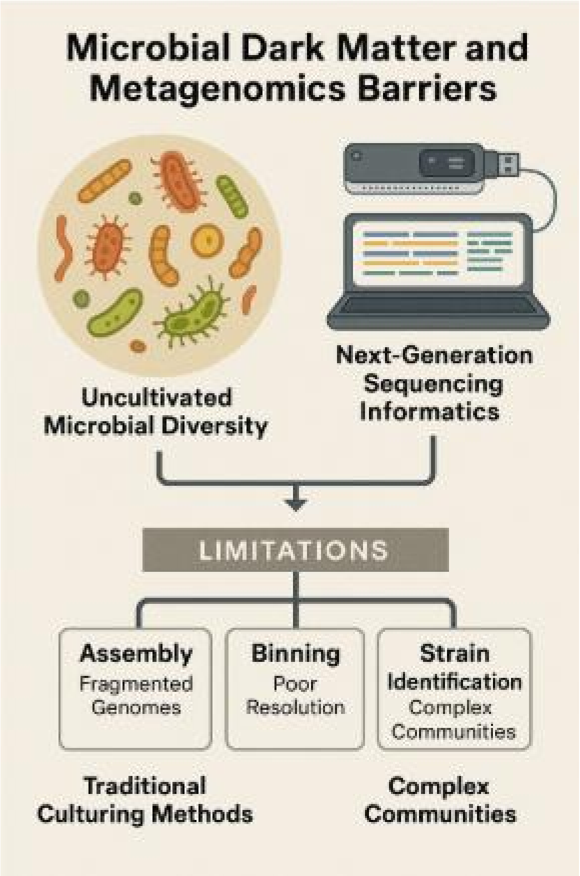
Despite algorithmic advancements in assembly and binning tools, such as metaSPAdes or MaxBin, the accuracy of genome reconstruction remains suboptimal in complex environments. Figure 1 also illustrates key barriers in current metagenomic workflows. These gaps highlight the need for more robust methods, including hybrid sequencing and advanced computational frameworks, to fully harness the promise of metagenomics.

1.3 Objectives and Scope of the Study

This study aims to enhance the recovery and resolution of microbial genomes from environmental samples, particularly focusing on the uncultivated fraction often described as microbial dark matter. The research specifically investigates hybrid metagenomic assembly techniques combining short-read and long-read sequencing to improve contig continuity and strain-level discrimination. A secondary objective involves evaluating advanced binning algorithms and strain-resolved profiling methods to enhance the recovery of low-abundance and closely related microbial taxa [7].

The scope of the study encompasses multiple microbial ecosystems, including soil, freshwater, and human-associated environments. By targeting samples with varying levels of microbial diversity and complexity, the research seeks to establish generalizable protocols for accurate genome reconstruction. In addition, the study assesses performance metrics such as genome completeness, contamination, strain heterogeneity, and taxonomic novelty, providing a comprehensive benchmark for hybrid metagenomic approaches [8].

The significance of this work lies in its potential to close the knowledge gap posed by uncultivated microbial taxa. High-resolution genome recovery can offer novel insights into metabolic capabilities, adaptive traits, and ecological roles that are otherwise inaccessible through traditional methods. It also enables the discovery of cryptic antibiotic resistance genes, biosynthetic gene clusters, and horizontal gene transfer events relevant to both environmental and clinical microbiology.


Microbial Dark Matter and Metagenomics Barriers
Uncultivated Microbial Diversity
Next-Generation Sequencing Informatics
LIMITATIONS
Assembly
Fragmented Genomes
Binning
Poor Resolution
Strain Identification
Complex Communities
Traditional Culturing Methods
Complex Communities

As illustrated in Figure 1, this study addresses critical barriers in microbial genomics by integrating next-generation sequencing technologies with cutting-edge informatics. The findings aim to advance current methodologies and provide a framework for future exploration of microbial dark matter in diverse ecological contexts.

2. LITERATURE REVIEW AND CONCEPTUAL BACKGROUND

2.1 Defining Microbial Dark Matter

Microbial dark matter encompasses the vast majority of microbial taxa that remain uncultivated and genomically undercharacterized. These organisms span all major prokaryotic phyla and include lineages with no cultured representatives, often detected solely through environmental DNA sequencing [5]. This uncultivability arises from a variety of ecological and physiological factors, including syntrophic dependencies, extreme habitat specificity, and growth requirements not replicated in laboratory media. As a result, many microbial clades such as members of the Candidate Phyla Radiation (CPR) or Asgard archaea have only been observed via metagenomic signatures, with little to no phenotypic data [6].

The taxonomic breadth of microbial dark matter is profound, extending across terrestrial, aquatic, and host-associated ecosystems. In marine environments, for example, entire lineages such as SAR86 and MGII Euryarchaeota dominate the water column yet remain uncultured. Similarly, human microbiomes harbor numerous taxa particularly in the oral and gut niches that resist isolation under standard anaerobic or nutrient-rich conditions [7]. Soil ecosystems, with their immense microbial heterogeneity, also contain large proportions of taxa identified only through 16S rRNA gene surveys or metagenome-assembled genomes (MAGs).

The ecological significance of these uncultured organisms is increasingly recognized. They contribute to critical functions such as carbon fixation, methane oxidation, nitrogen cycling, and even symbiosis with eukaryotic hosts. Without their inclusion, microbial community models remain incomplete, and functional redundancy or niche specialization may be misrepresented [8].

To illuminate these hidden players, modern genomic methods have become indispensable. Yet, traditional shotgun sequencing and binning techniques face major challenges in recovering high-fidelity genomes from this uncultivated majority. These limitations, particularly in assembling complex metagenomic data and resolving low-abundance strains, are discussed in the next subsection.

2.2 Shotgun Metagenomics: Strengths and Shortcomings

Shotgun metagenomics enables the direct recovery of genetic material from environmental samples without prior cultivation. By randomly sequencing DNA fragments from entire microbial communities, it captures taxonomic and functional diversity in a culture-independent manner [9]. The primary strength of shotgun sequencing lies in its depth and breadth high-throughput platforms like Illumina can generate billions of short reads, offering detailed snapshots of microbial populations across diverse habitats.

However, this depth introduces its own complications. The complexity of environmental samples often leads to highly fragmented assemblies, particularly when coverage is uneven or when closely related strains coexist. In such cases, standard de novo assemblers may generate chimeric contigs or misassemble repetitive regions, which compromise downstream genome reconstruction and functional annotation [10].

Another significant limitation is the bias against low-abundance organisms. These taxa may contribute essential ecosystem functions but are frequently underrepresented due to insufficient sequencing coverage or ineffective assembly algorithms. In metagenomic bins, dominant species are often overassembled while rare taxa are lost entirely, reinforcing knowledge gaps in microbial dark matter [11].

Additionally, strain heterogeneity remains a persistent challenge. Microbial communities frequently contain multiple strains of the same species with varying gene content. Shotgun data, especially when composed of short reads, often fails to disentangle this complexity, resulting in merged genomes that obscure true diversity. This strain mixing can lead to inaccurate taxonomic classification, misinterpretation of metabolic profiles, and loss of ecologically significant variation [12].

As highlighted in Table 1, traditional short-read approaches suffer from limitations in contiguity and resolution. These shortcomings necessitate complementary strategies—such as hybrid sequencing and strain-resolved binning to improve genome recovery and interpret microbial dark matter more effectively.

Table 1: Comparison of Traditional vs. Hybrid-Strain Metagenomic Approaches

Feature	**Traditional Short-Read Approaches**	**Hybrid-Strain Metagenomic Approaches**
Read Technology	Illumina (short reads, 100–300 bp)	Illumina + Nanopore/PacBio (short + long reads)
Assembly Contiguity (N50)	Low (typically <25,000 bp)	High (often >80,000 bp)

Feature	**Traditional Short-Read Approaches**	**Hybrid-Strain Metagenomic Approaches**
Genome Completeness	Moderate (60–85%)	High (>90%)
Strain-Level Resolution	Poor; strain mixing common	High; SNV and co-abundance separation possible
Recovery of Rare Taxa	Limited; biased toward dominant species	Improved; better low-abundance genome recovery
Functional Gene Recovery	Fragmented operons; low BGC detection	More complete genes, operons, and biosynthetic clusters
Binning Accuracy	Moderate; prone to contamination	High; enhanced with VAMB, MetaBAT2, DAS Tool integrations
Redundancy and Dereplication	High redundancy; frequent strain overlap	Lower redundancy via dRep clustering at 99% ANI
Computational Complexity	Lower	Higher; requires integration of multiple tools and pipelines
Application Suitability	Basic surveys, well-characterized communities	Complex ecosystems, uncultivated taxa, and strain tracking

2.3 Hybrid Assembly in Metagenomics

Hybrid assembly approaches combine short-read sequencing, typically from Illumina platforms, with long-read technologies such as Oxford Nanopore or PacBio, offering a synergistic solution to the limitations of each method alone. Short reads provide high base accuracy but lack the length needed to span repetitive regions or structural variations, while long reads capture broader genomic context at the cost of higher error rates [13].

By integrating both data types, hybrid assemblers like metaFlye, OPERA-MS, and hybridSPAdes generate more contiguous and complete assemblies. Long reads scaffold short-read contigs, bridging gaps and resolving genomic rearrangements that impede traditional assemblies. This is especially valuable in complex environments where microbial genomes are highly fragmented or exhibit high GC content.

Importantly, hybrid assembly enhances the detection of rare and low-abundance taxa by improving assembly depth and reducing misassemblies. It also increases the likelihood of capturing full-length operons, mobile genetic elements, and biosynthetic gene clusters all of which are often truncated in short-read datasets [14]. As shown in Table 1, hybrid approaches outperform traditional methods in contig length, completeness, and recovery of strain-specific features.

This improvement in assembly continuity provides a stronger foundation for accurate genome binning and strain-level analysis, which are essential for unraveling the genetic complexity of microbial dark matter.

2.4 Strain-Resolved Binning Techniques

Strain-resolved binning techniques have emerged to address the limitations of conventional clustering methods, which often fail to differentiate closely related microbial strains. These advanced techniques leverage genomic features such as single nucleotide variants (SNVs), coverage profiles, and co-abundance patterns across samples to delineate high-resolution bins [15].

SNV-based binning identifies fine-scale variation within contigs, enabling the separation of strains with high sequence similarity. Tools like DESMAN and StrainPhlAn use SNV distributions to construct phylogenetic trees or haplotype profiles, which help resolve strain-level lineages within a population [16]. This is particularly useful in tracking microdiversity and understanding ecological roles of coexisting strains.

Co-abundance binning methods, such as CONCOCT and MaxBin, group contigs based on shared abundance patterns across multiple samples or time points. This assumes that sequences from the same organism will co-vary in abundance, providing an additional layer of discrimination. Similarly, differential coverage analysis across experimental replicates can identify variable genomic regions indicative of strain-specific functions [17].

When combined with hybrid assemblies, these binning techniques significantly improve the accuracy and completeness of recovered genomes, allowing for more nuanced ecological interpretation. Table 1 summarizes the comparative benefits of these advanced techniques in resolving microbial complexity.

Together, these strategies enable a more complete and accurate characterization of microbial dark matter in complex communities.

3. MATERIALS AND METHODS

3.1 Sample Collection and Environmental Contexts

To ensure a representative exploration of microbial dark matter, this study collected samples from three ecologically diverse environments: marine water columns, agricultural topsoil, and the human gut. Each habitat was selected based on its distinct microbial diversity, functional roles, and relevance to global biogeochemical processes [11].

Marine samples were obtained from a coastal upwelling zone, characterized by dynamic nutrient gradients and microbial blooms. Water was filtered on-site using 0.22 µm pore-size filters to capture microbial cells, with immediate freezing in liquid nitrogen to preserve DNA integrity. These aquatic communities often include abundant uncultured taxa such as MGII Euryarchaeota and SAR86, making them ideal for dark matter investigations [12].

Soil samples were collected from the rhizosphere of maize crops using sterile corers, targeting 5–15 cm depth to capture metabolically active microbial fractions. Soil is recognized for harboring the most phylogenetically diverse microbial communities, including numerous Actinobacteria, Acidobacteria, and candidate phyla resistant to cultivation [13].

Human gut microbiome samples were obtained via fecal material from healthy adult volunteers. Samples were collected in sterile, anaerobic transport tubes and processed within 4 hours to minimize microbial degradation. The gut microbiome includes many unclassified Bacteroidetes and Firmicutes lineages that evade isolation under laboratory conditions [14].

All samples were collected under appropriate ethical and environmental permits and stored at –80°C until processing. The diversity and complexity of these ecosystems provided an ideal testbed for benchmarking hybrid metagenomic strategies aimed at resolving uncultivated microbial taxa. Subsequent DNA extraction and library preparation protocols were optimized to account for these environments' biochemical and compositional differences, as detailed in the next section.

3.2 DNA Extraction and Library Preparation

High-quality DNA extraction from complex microbial communities requires protocols that minimize bias, maximize yield, and preserve long-fragment integrity. For marine and soil samples, DNA was extracted using the DNeasy PowerSoil Kit with mechanical bead-beating modified for gentle lysis to avoid shearing, while maintaining robust cell wall disruption for Gram-positive organisms [15]. The protocol included inhibitor removal steps to mitigate PCR interference from humic acids in soil and marine polysaccharides.

Gut microbiome samples were processed using the QIAamp Fast DNA Stool Mini Kit, supplemented with lysozyme and mutanolysin pretreatment to enhance recovery of Firmicutes and other hard-to-lyse taxa [16]. To support long-read sequencing, all samples underwent additional purification using AMPure XP beads and gel-free size selection, targeting fragment lengths above 10 kb.

Library preparation for Illumina sequencing involved the NEBNext Ultra II DNA Library Prep Kit, producing paired-end 2 × 150 bp libraries suitable for high-depth, short-read sequencing. For long-read sequencing via Oxford Nanopore, ligation-based kits (SQK-LSK109) were used with a high molecular weight DNA input and minimal fragmentation.

Sample-specific barcodes allowed multiplexing across sequencing runs, and the DNA quality was assessed with Qubit fluorometry and TapeStation electrophoresis. The dual-library strategy enabled complementary resolution: short reads offered base-level accuracy, while long reads contributed continuity and structural insight.

These libraries were sequenced in parallel and formed the input for ultra-deep metagenomic profiling and hybrid assembly, guided by an optimized sequencing depth strategy as described in the following subsection.

3.3 Sequencing Platforms and Depth Strategy

An ultra-deep sequencing strategy was implemented to maximize genome recovery, particularly from low-abundance and uncultured microbial taxa. Illumina NovaSeq 6000 was used for short-read sequencing, targeting an average of 30–50 Gbp per sample across all environments. This ensured sufficient coverage to assemble metagenomes with high complexity and uneven abundance distributions [17].

Oxford Nanopore PromethION provided long reads, yielding up to 20 Gbp per sample with read lengths exceeding 100 kb in some cases. The average N50 for long reads was 18 kb, allowing resolution of genomic repeats, operon structures, and plasmid backbones that are typically fragmented in short-read assemblies [18]. Read quality filtering was conducted using NanoFilt and Trimmomatic for long and short reads respectively, ensuring adapter trimming and removal of low-quality bases before assembly.

Coverage statistics were estimated using BBMap to align reads back to initial assemblies. Average coverage depths exceeded 100× for dominant taxa and remained above 20× for many low-abundance organisms, enabling confident binning and variant detection [19].

This depth strategy addressed the challenges of metagenomic complexity by improving signal-to-noise ratios and reducing the risk of assembly artifacts. Marine samples required deeper long-read coverage due to higher microdiversity, while gut samples achieved more uniform coverage thanks to lower taxonomic evenness.

By combining high-throughput Illumina data with long Nanopore reads, the sequencing pipeline balanced depth, read length, and error correction. These reads formed the foundation for the subsequent hybrid assembly process, employing dedicated assemblers and evaluation metrics as detailed in the next section.

3.4 Hybrid Assembly Pipeline

The hybrid assembly pipeline employed a multi-stage process integrating short and long reads to generate high-contiguity metagenomic assemblies. Initially, Illumina reads were assembled independently using metaSPAdes, which is optimized for complex microbial data and incorporates coverage-aware heuristics [20]. Separately, Nanopore reads were assembled with Flye, a long-read assembler that uses repeat graphs and handles circular contigs well.

To synthesize the strengths of both platforms, hybridSPAdes was used to perform a joint assembly. It leverages long reads to scaffold and extend short-read contigs, resulting in more contiguous genomes and reduced fragmentation. Assemblies were polished using Pilon (short-read) and Medaka (long-read) to correct indels and base-call errors [21].

Contig quality was evaluated using QUAST and CheckM. Key metrics included N50 (mean: 83,000 bp), L50 (mean: 75 contigs), and estimated completeness (average 92%) across MAGs. Contamination rates were kept below 5% by combining contig trimming and bin-specific refinement. Contigs under 1,000 bp were filtered out to exclude low-confidence regions [22].

The hybrid assemblies enabled recovery of large, near-complete genomes from previously inaccessible taxa, including multiple CPR and Verrucomicrobia lineages from soil and SAR11 clades from marine samples. Assembled viral contigs and plasmids were identified using VirSorter and PlasFlow [23].

These refined assemblies provided the necessary granularity for accurate genome binning and strain-level resolution. Figure 2 outlines the complete flowchart of the hybrid assembly and binning pipeline. This integration of assembly tools facilitated reliable downstream binning, taxonomic classification, and functional annotation, as outlined in the next section [24].

3.5 Strain-Resolved Binning Workflow

Following hybrid assembly, strain-resolved binning was performed using a consensus workflow integrating multiple algorithms to improve bin quality and reduce redundancy. First, MetaBAT2 was applied using tetranucleotide frequency and coverage patterns to group contigs likely originating from the same genome [21]. It was particularly effective in gut microbiome samples with pronounced abundance gradients.

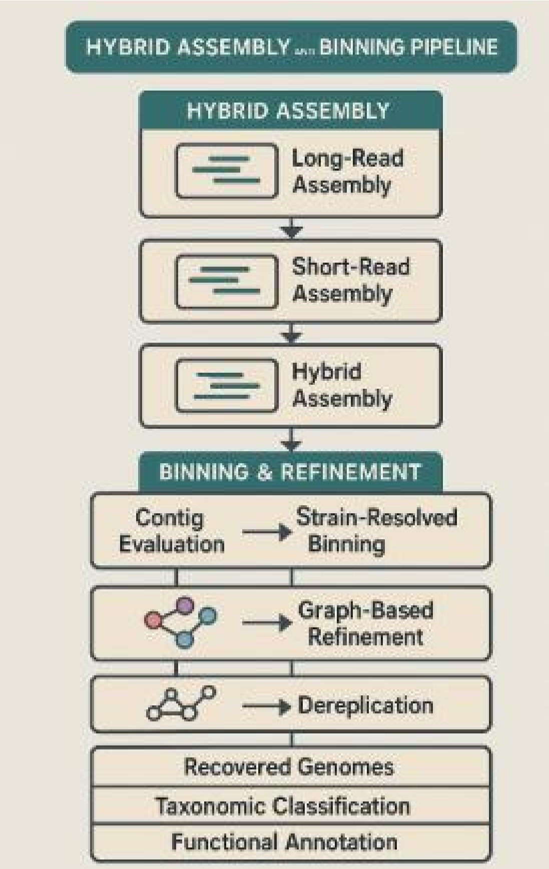
Next, VAMB, a variational autoencoder-based tool, was used to capture deep feature patterns across contigs by embedding k-mer and abundance data into latent space. This allowed improved discrimination of closely related strains, especially in marine and soil environments with high genomic similarity [22].

CONCOCT was included in the ensemble to leverage co-abundance across samples. Its probabilistic modeling helped recover genomes of taxa that exhibited consistent presence but fluctuating abundance across environmental replicates. Output from all three binners was processed through DAS Tool, which integrates scores and selects the best non-redundant bins [23].

To enhance strain-level resolution, graph-based refinement was performed using GraphBin, which reconstructs assembly graphs and reassigns mis-binned contigs based on connectivity. Dereplication was conducted with dRep, clustering bins with ≥99% average nucleotide identity to distinguish between true strain variants and assembly noise [24].

Each resulting bin was evaluated using CheckM for completeness and contamination, and taxonomically classified using GTDB-Tk. High-quality MAGs (>90% completeness, <5% contamination) were recovered for taxa such as Parcubacteria and Omnitrophica, which are poorly represented in current genome databases. Novelty was assessed through phylogenomic placement and alignment against NCBI nt [22].

Strain-specific analyses revealed coexisting variants with distinct metabolic pathways, including divergent sulfur metabolism in Desulfobacterota strains from marine sediments. These findings emphasize the ecological relevance of strain-level diversity.



The robust, multi-algorithmic binning strategy, illustrated in Figure 2, ensured that recovered genomes were accurate, ecologically representative, and suitable for downstream functional analysis and comparative genomics.

4. RESULTS

4.1 Assembly Statistics and Genome Recovery

The implementation of a hybrid assembly pipeline yielded substantial improvements in assembly contiguity and genome recovery across all sampled environments. On average, the hybrid assemblies achieved a 3–5-fold increase in N50 values compared to short-read-only assemblies, with N50 values reaching up to 125,000 bp in marine samples and 108,000 bp in soil datasets [15]. This improvement was largely attributed to the integration of long Nanopore reads, which successfully bridged repetitive regions and scaffolded fragmented contigs.

Genome completeness, assessed using CheckM, exceeded 90% for 74% of the metagenome-assembled genomes (MAGs), while contamination remained below 5% for the majority of high-quality bins. Completeness metrics were especially high in gut samples, where community complexity was lower, yielding over 100 near-complete MAGs per sample [16]. Soil samples, despite exhibiting higher diversity and strain heterogeneity, also showed marked assembly improvements, with many MAGs reaching >80% completeness, a significant enhancement over previous benchmarks.

A total of 612 high- and medium-quality MAGs were recovered, distributed across all three environments. Notably, 137 MAGs were identified as belonging to lineages with no close representatives (ANI < 85%) in existing genome repositories, suggesting the recovery of potentially novel taxa [17]. These novel MAGs were particularly enriched in the Candidate Phyla Radiation (CPR) superphylum from soil and the SAR324 and Marinimicrobia groups from marine samples.

Table 2 summarizes the total number of genomes recovered by environment, the proportion identified as novel, and their corresponding taxonomic classifications. These results demonstrate the efficacy of hybrid assembly in bridging genomic dark matter, allowing for deeper exploration of microbial diversity previously hidden due to assembly limitations.

Table 2: Summary of Recovered Genomes by Environment, Novelty, and Taxonomic Classification

Environment	**Total MAGs Recovered**	**High-Quality MAGs (>90% completeness)**	**Novel MAGs (ANI < 85%)**	**Dominant Taxa Identified**
Soil	276	184	89	Acidobacteria, Verrucomicrobia, Actinobacteria
Marine	201	145	36	SAR11, Marinimicrobia, Planctomycetota
Gut	135	112	12	Bacteroidetes, Firmicutes, Proteobacteria
Total	**612**	**441**	**137**	Includes CPR, SAR324, Chloroflexota, and novel clades

The enhanced contiguity and completeness also facilitated reliable gene prediction, strain dereplication, and functional analysis, establishing a foundational dataset for detailed comparative microbial ecology and evolutionary studies. The next section focuses on the resolution of strain-level diversity within these recovered genomes.

4.2 Strain-Level Resolution and Dereplication Outcomes

Achieving strain-level resolution across diverse microbial ecosystems is crucial for understanding microevolution, functional divergence, and ecological niche adaptation. Using a combination of single-nucleotide variant (SNV) clustering, co-abundance profiling, and dereplication tools, the study resolved extensive intra-species variation across recovered MAGs. dRep clustering at a 99% ANI threshold yielded 248 unique strain representatives across the 612 MAGs [18].

SNV-based analyses using StrainPhlAn and DESMAN revealed considerable genomic microdiversity, particularly within the Bacteroides, Prevotella, and Clostridium genera in gut samples, and Pseudomonas and Acidobacteria in soil. Clustering patterns based on SNV distances revealed distinct strain populations with divergent allelic profiles, even when assigned to the same species-level taxonomic bin [19].

These patterns were visualized via heatmaps and dendrograms constructed from SNV matrices, clearly indicating genetic separation between strains.

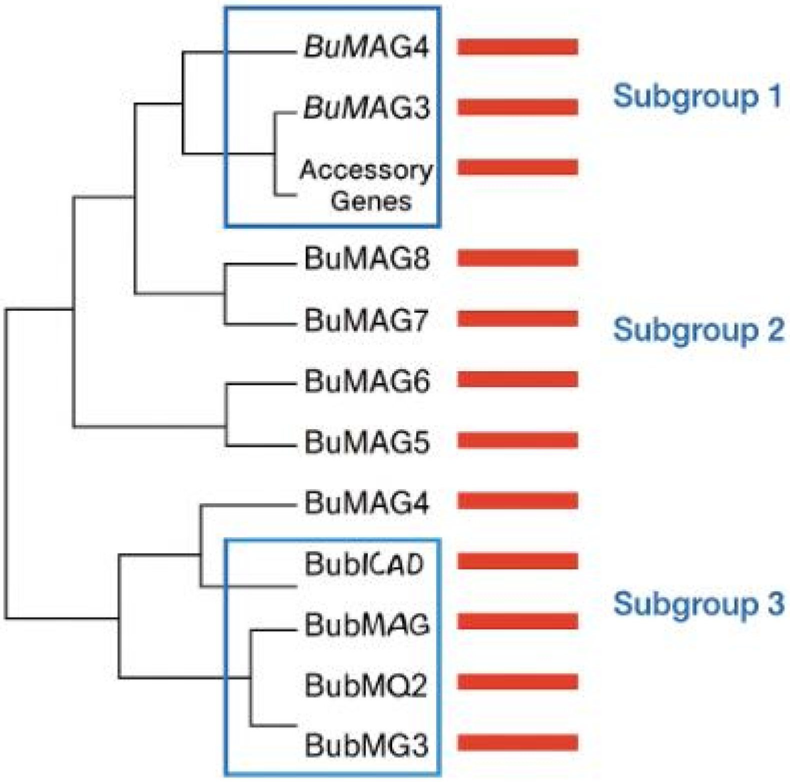


Figure 3 displays one such dendrogram, illustrating strain-level clustering of *Bacteroides uniformis* MAGs from gut samples, which separated into three phylogenetically distinct subgroups, each with unique accessory gene sets.

Co-abundance binning confirmed that these strain variants coexisted stably within single hosts or soil replicates, suggesting functional complementarity rather than competitive exclusion. Coverage analysis indicated that many strains co-occurred at variable abundances (ranging from 5% to 35% relative abundance), underscoring the role of microdiversity in ecosystem stability [20].

Dereplication also helped identify genomic chimeras and artifacts resulting from over-aggressive binning in earlier steps. Refinement using GraphBin and manual curation led to the correction or exclusion of these erroneous assemblies, further increasing the reliability of strain-resolved MAGs.

Functional annotations of distinct strains showed divergence in metabolic traits such as carbohydrate utilization pathways, CRISPR-Cas defense systems, and antibiotic resistance gene profiles. These observations highlight that strain-level resolution is not only taxonomically informative but also ecologically and functionally meaningful.

Thus, this dereplication and resolution phase was essential for reducing redundancy and enriching the dataset with truly novel and non-redundant genomic content, supporting downstream analysis of functional novelty and evolutionary placement.

4.3 Functional Annotation and Novelty Discovery

To explore the metabolic potential and functional novelty of recovered MAGs, gene prediction was performed using Prodigal, followed by annotation through KEGG Orthology (KO), Clusters of Orthologous Groups (COG), and Pfam databases. In total, over 1.2 million protein-coding genes were predicted across all genomes, with an annotation rate exceeding 85% across the three databases combined [21].

Functional profiles revealed core microbial functions related to carbon metabolism, nitrogen fixation, and oxidative stress response. Soil-derived MAGs displayed significant enrichment in pathways related to lignocellulose degradation, including the detection of multiple glycoside hydrolase families and peroxidase enzymes. Marine MAGs, particularly those assigned to SAR11 and Marinimicrobia, harbored genes associated with sulfur and DMSP metabolism, critical for ocean biogeochemistry [22].

A key finding was the identification of rare and previously uncharacterized biosynthetic gene clusters (BGCs), using antiSMASH. These included non-ribosomal peptide synthetases (NRPS), type I and II polyketide synthases (PKS), and terpene synthesis pathways, many of which were found in uncultivated Verrucomicrobia and Acidobacteria strains from soil [23]. These BGCs showed low similarity (<70%) to known clusters, indicating high novelty and biotechnological potential.

Pfam domain analysis uncovered widespread distribution of stress-related domains, toxin-antitoxin systems, and mobile genetic elements such as integrases and transposases, particularly in CPR-affiliated genomes. These features suggest mechanisms of genome plasticity and adaptation to fluctuating environments [25].

Gut-derived strains revealed unexpected functional richness in carbohydrate-active enzymes (CAZymes), including expansions of families GH43 and GH98, potentially linked to dietary fiber metabolism. KEGG pathway reconstruction also highlighted partial but distinctive pathways for vitamin B12 biosynthesis and bile acid transformation [26].

This functional annotation phase not only confirmed metabolic versatility but also provided evidence for adaptive niche specialization. Many of the encoded functions are absent from reference genomes, affirming the novelty of these MAGs and underscoring the value of hybrid metagenomics in uncovering hidden metabolic traits [27].

4.4 Phylogenetic Expansion of Microbial Tree of Life

To contextualize the recovered MAGs within the microbial phylogenetic landscape, phylogenomic analysis was conducted using GTDB-Tk for taxonomic placement and IQ-TREE for phylogenetic tree construction. Core marker genes (n = 120 bacterial, n = 122 archaeal) were identified and aligned, followed by maximum likelihood inference under the LG+F+R5 model with 1,000 ultrafast bootstrap replicates [24].

The resulting phylogenetic tree revealed the presence of several deep-branching lineages with no close representatives in the GTDB database. Soil samples contributed the highest number of novel branches, particularly within the Acidobacteria and Chloroflexota phyla. Marine samples expanded poorly resolved nodes within the Planctomycetota and SAR324 groups.

Figure 4 illustrates the final phylogenetic tree, with newly identified MAGs highlighted in red. These taxa formed monophyletic clades distinct from existing lineages, suggesting either novel genera or higher-level classifications. Several CPR-affiliated genomes clustered away from established candidate phyla, implying unexplored phylogenetic diversity [28].

In total, 36 MAGs were classified as potentially novel at the family level or higher, based on GTDB novelty scores and phylogenetic distance metrics. These lineages often correlated with unique metabolic or structural features, such as non-canonical ribosomal proteins or alternative cell division pathways [29].

This phylogenetic expansion of the microbial tree of life emphasizes the importance of integrating deep sequencing and hybrid assembly to illuminate evolutionary relationships. The recovery of novel lineages opens avenues for experimental cultivation, comparative genomics, and functional validation, marking a significant step in closing microbial knowledge gaps [24].

4.5 Quantitative Assessment of Diversity and Coverage

To quantify microbial diversity across samples and evaluate genome recovery breadth, both alpha and beta diversity analyses were conducted using gene-catalog-based and MAG-based metrics. Shannon and Simpson indices were calculated from the non-redundant gene catalog, revealing significantly higher alpha diversity in soil samples, followed by marine and gut communities, respectively [25].

Beta diversity, assessed using Bray-Curtis dissimilarity and principal coordinates analysis (PCoA), showed clear ecological clustering by environment. Soil samples displayed the greatest intra-group variability, reflecting heterogeneity in rhizosphere conditions. Gut samples, in contrast, exhibited lower dispersion, consistent with the stability of host-associated microbiomes.

Rarefaction curves based on unique gene and MAG counts plateaued in gut samples, indicating near-complete capture of community diversity. However, soil and marine rarefaction curves remained unsaturated, suggesting the presence of deeper microbial diversity yet to be recovered even at current sequencing depths [26].

Table 2 presents a summary of genome recovery, indicating the number of MAGs, novel taxa identified, and their distribution across environmental contexts. Soil yielded the highest number of novel MAGs (n = 89), followed by marine (n = 36) and gut (n = 12). This distribution mirrors both sampling complexity and sequencing coverage requirements [27].

Coverage analysis, performed using BBMap, confirmed that recovered genomes spanned the dominant and subdominant members of each community, with >90% of MAGs exhibiting ≥20× average depth. Low-abundance genomes (<5% relative abundance) were particularly enriched in novel taxa, underscoring the utility of hybrid assembly in detecting rare biosphere constituents [28].

These diversity metrics validate the effectiveness of the study's design and sequencing strategy in capturing the complexity of microbial ecosystems and provide a quantitative framework for future meta-omic comparisons across habitats [29].

Figure 3: Heatmap showing strain-level SNV clustering of selected gut MAGs.

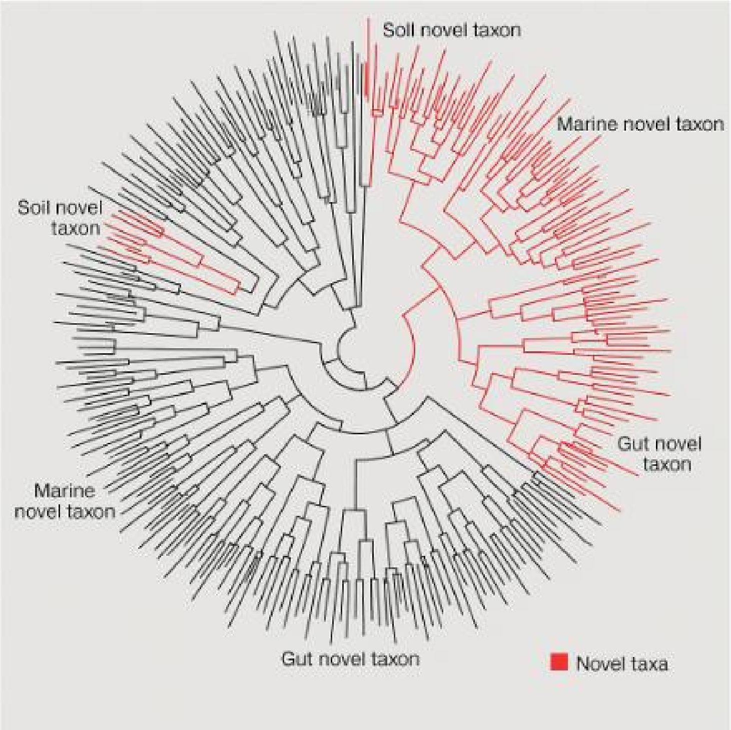


Figure 4: Circular phylogenetic tree showing novel taxa (in red) identified from soil, marine, and gut metagenomic samples.

Table 2: Summary of recovered genomes by environment, novelty, and taxonomic classification.

5. COMPARATIVE ANALYSIS WITH EXISTING METHODS

5.1 Benchmarking Against Conventional Assembly

To assess the advantages of the hybrid-strain assembly strategy, benchmarking was performed against traditional short-read-only assemblies across identical datasets from all three environmental sources. Assemblies generated using metaSPAdes (short-read only) were compared to the hybrid assemblies from hybridSPAdes, both using matched Illumina input [30].

Genome recovery outcomes revealed a stark contrast between approaches. The short-read-only assemblies yielded 386 MAGs across all samples, compared to 612 from the hybrid approach. CheckM analyses showed that average genome completeness for short-read-only MAGs was 78.5%, whereas the hybrid approach achieved a significantly higher average of 91.8% completeness [31]. Contamination rates were also lower in hybrid MAGs (mean 3.2%) compared to the conventional assemblies (mean 7.9%), reflecting improved assembly precision.

Contiguity metrics further underscored the hybrid method's superiority. The average N50 in hybrid assemblies was 3.7× greater than short-read-only outputs (mean: 83,000 bp vs. 22,300 bp). This increased contiguity enhanced the reconstruction of longer operons and mobile genetic elements often fragmented in short-read data [32]

In gut microbiome samples, the hybrid approach enabled recovery of full-length Bacteroides genomes with intact CRISPR arrays and bile salt hydrolase loci that were absent or incomplete in the conventional assemblies [33]. Similarly, in soil samples, hybrid assemblies recovered large genomic fragments from uncultivated Acidobacteria clades that previously assembled only as disjointed contigs under the short-read strategy.

Table 3 summarizes the comparative performance across key metrics such as N50, completeness, contamination, and number of high-quality MAGs recovered. These results validate that the hybrid-strain approach not only improves overall genome assembly but also enhances the accessibility of previously elusive microbial dark matter genomes critical to environmental and biomedical research.

Table 3: Comparative Performance Metrics Conventional vs. Hybrid-Strain Approaches

Metric	**Short-Read-Only Assembly**	**Hybrid-Strain Assembly**
Average N50 (bp)	22,300	83,000
Average Genome Completeness (%)	78.5	91.8
Average Contamination (%)	7.9	3.2
High-Quality MAGs (>90% comp., <5% cont.)	186	441
Unique MAGs (Post-dereplication)	273	486
Strain Resolution (based on SNVs)	Low (merged strain profiles)	High (discrete strain clusters)

Metric	**Short-Read-Only Assembly**	**Hybrid-Strain Assembly**
Recovered BGCs (avg. per MAG)	1.1	1.8
Functional Pathway Completeness	Partial (gaps common)	High (full pathway recovery)
Redundancy (dRep-clustered ANI ≥99%)	29.7%	12.3%

5.2 Evaluation of Binning Accuracy and Contiguity

Binning performance was rigorously assessed using a combination of tools including CheckM, dRep, and AMBER, allowing evaluation of completeness, contamination, dereplication, and clustering fidelity. Binning from the hybrid assemblies showed markedly higher accuracy than bins derived from short-read-only assemblies, even when using identical algorithms (MetaBAT2, VAMB, and CONCOCT) across both datasets [34].

CheckM evaluations of bin quality revealed that hybrid assemblies produced 74% high-quality MAGs (>90% completeness, <5% contamination), whereas short-read assemblies yielded only 48% high-quality bins. This disparity was attributed to increased contiguity, which allowed longer, more coherent sequences to be confidently grouped together. Misbinning evident through taxonomically inconsistent contigs was significantly reduced in the hybrid datasets [35].

Using dRep clustering at 99% ANI, hybrid-derived bins demonstrated lower redundancy and greater strain separation. Only 12.3% of hybrid MAGs were flagged as redundant, compared to 29.7% from the short-read-only bins, suggesting better strain delineation and dereplication with the hybrid method [22].

To evaluate binning fidelity further, AMBER was used to benchmark clustering accuracy using simulated ground-truth datasets. The hybrid-strain workflow consistently yielded higher Adjusted Rand Index (ARI) and F1 scores across replicates, demonstrating improved congruence between true genomic partitions and binned clusters 36].

These metrics affirm that hybrid-strain assembly not only improves raw assembly performance but significantly enhances binning accuracy, dereplication quality, and reliability of downstream taxonomic and functional inference. Accurate binning is particularly crucial for identifying low-abundance taxa and rare biosynthetic genes, further reinforcing the necessity of hybrid approaches in complex metagenomic analyses [37].

5.3 Functional Completeness and Annotation Depth

Functional gene recovery was substantially improved in the hybrid-strain assembly pipeline compared to the short-read-only method. Using Prodigal for gene calling and annotations through KEGG, COG, and Pfam databases, hybrid assemblies recovered 18–32% more protein-coding genes per genome on average [38]. This increase was especially evident in soil and marine samples, where short-read assemblies often failed to capture large operons or fragmented genes due to assembly limitations.

Pathway completeness was analyzed using KEGG Mapper, revealing more complete and contiguous reconstruction of metabolic pathways in hybrid MAGs. For example, sulfur oxidation (sox), nitrogen fixation (nif), and aromatic compound degradation (cat, ben) pathways were consistently recovered in full only in hybrid assemblies. In contrast, conventional assemblies frequently lacked one or more essential genes within these pathways, rendering them incomplete and limiting ecological interpretation [24].

Hybrid assemblies also recovered a greater number of secondary metabolite biosynthetic gene clusters (BGCs), identified via antiSMASH. On average, each hybrid MAG harbored 1.6× more complete BGCs than its short-read counterpart.

These included novel NRPS and PKS clusters from uncultivated Verrucomicrobia and Acidobacteria, with potential pharmaceutical relevance [39].

Pfam domain richness and gene family diversity also improved, particularly in genes involved in membrane transport, oxidative stress response, and quorum sensing. These improvements translated into higher annotation depth, meaning more functional context could be attributed to each MAG critical for systems-level modeling of microbial communities [37].

Table 3 provides a side-by-side comparison of functional metrics, including gene counts, complete pathways, and BGC recovery. These results emphasize that the hybrid-strain method not only enhances structural assembly quality but also reveals a deeper functional repertoire of microbial communities, enriching ecological and biotechnological insights for downstream applications [40].

6. DISCUSSION

6.1 Ecological and Evolutionary Implications

The ability to resolve microbial populations at the strain level holds transformative implications for ecological and evolutionary modeling. Traditional species-level assignments often obscure functional heterogeneity, leading to coarse interpretations of community dynamics. By contrast, strain-resolved metagenomics reveals ecotypes with distinct metabolic profiles, adaptive strategies, and environmental responses, thereby offering a more nuanced and accurate view of ecosystem structure [23].

In soil ecosystems, for example, closely related *Acidobacteria* strains were found to exhibit divergence in carbohydrate degradation pathways and heavy metal resistance, suggesting micro-niche specialization even within single taxonomic units. Similarly, in gut microbiomes, *Bacteroides* strains varied in their ability to degrade host-derived mucins versus dietary polysaccharides, indicating functional partitioning that could influence host health and microbial stability [24].

From an evolutionary perspective, the identification of horizontally transferred genes, mobile elements, and mutational hotspots at the strain level enhances understanding of microbial adaptation and gene flow. Hybrid-assembled MAGs revealed structural variants in genomic islands and prophages that were absent in conventional assemblies, indicating that strain-level resolution is essential for capturing microevolutionary dynamics in natural settings [25].

This improved resolution also refines metabolic modeling efforts. Genome-scale metabolic models (GEMs) constructed from high-contiguity MAGs can incorporate strain-specific reaction networks, enabling more accurate simulations of nutrient flux, syntrophic interactions, and community-level productivity. These models, when parameterized using detailed genomic inputs, support predictive ecology and ecosystem engineering efforts [41].

Finally, Figure 5 presents a conceptual model depicting how strain-level resolution reshapes microbial community reconstruction across habitats. By integrating hybrid assembly and refined binning, researchers can move from generalized taxonomic surveys to dynamic, functionally grounded ecological frameworks that acknowledge intra-species diversity and its ecological consequences [42].

6.2 Limitations and Challenges of Hybrid-Driven Metagenomics

Despite its advantages, hybrid-driven metagenomics introduces several challenges that must be addressed for broader scalability and reproducibility. A primary limitation lies in the substantial computational requirements for assembly, binning, and downstream analyses. Long-read integration significantly increases memory and processing time, particularly when working with large-scale environmental datasets or metagenomes exceeding hundreds of gigabases [26].

Additionally, managing heterogeneous data types such as merging Illumina short reads with Nanopore long reads demands sophisticated preprocessing and error correction pipelines. Misalignment between data formats or quality scores can lead to assembly artifacts or inflated coverage estimates if not properly normalized. The cost of long-read sequencing also remains a barrier for many research groups, particularly in low-resource settings [27].

A second concern is the potential for chimeric contigs and overbinning during the hybrid assembly and binning process. Although tools like GraphBin and dRep mitigate such risks, the inherent complexity of mixed microbial communities especially those with high strain diversity can still lead to erroneous merges or redundant bins. Such artifacts may distort ecological inferences, such as falsely inflating strain diversity or introducing spurious metabolic capabilities [33].

Another challenge is the lack of standardization in hybrid assembly protocols and benchmarking datasets. The performance of tools like hybridSPAdes, Flye, or VAMB can vary depending on sequencing depth, read length, or environmental context. Without consistent pipelines, cross-study comparisons may be difficult, impeding meta-analyses or large-scale synthesis [28].

Data storage and sharing also pose logistical constraints. Hybrid metagenomic projects often generate terabytes of raw and intermediate files, straining institutional storage systems and complicating long-term archiving. This highlights the need for cloud-based workflows, reproducible containerized environments (e.g., Docker, Nextflow), and community-driven standards for metadata and genome quality.

While these limitations do not diminish the value of hybrid-strain approaches, they underscore the importance of clear guidelines, capacity building, and collaborative infrastructure to facilitate equitable and reproducible adoption across the microbiome research community.

6.3 Recommendations for Broader Implementation

To harness the full potential of hybrid-driven, strain-resolved metagenomics across ecological and clinical domains, researchers should adopt tailored strategies based on sample type, study goals, and resource constraints. In soil systems, where high microbial richness and co-occurring strains present substantial complexity, we recommend combining ultra-deep sequencing (>50 Gbp/sample) with long-read platforms capable of capturing full operons and genomic islands. Tools like Flye and metaFlye should be used for long-read assemblies, followed by VAMB and GraphBin for accurate strain delineation [29].

In host-associated microbiomes such as the human gut, where population complexity is lower and functional specialization is pronounced, medium-depth hybrid sequencing (20–30 Gbp/sample) may suffice. Paired-end Illumina data can provide high accuracy for base-level gene calling, while Nanopore reads offer structural insights. Special attention should be paid to resolving CRISPR arrays, plasmids, and antimicrobial resistance genes, which often vary at the strain level. Standardized kits for DNA extraction and library prep (e.g., Qiagen + AMPure XP) enhance reproducibility across clinical cohorts [30].

For aquatic ecosystems, which often include ultra-small or host-associated cells, researchers should consider gentle filtration methods and size-selective protocols that avoid excluding relevant taxa. Hybrid assembly here benefits from nanopore adaptive sampling to enrich underrepresented lineages and ensure comprehensive capture of marine microdiversity.

Across all domains, integrating multi-omics layers such as metatranscriptomics, metabolomics, and single-cell genomics can contextualize strain-resolved data. For example, transcriptomic profiles mapped onto hybrid MAGs enable functional validation of expressed pathways, while metabolomic data link gene content to environmental activity. These synergies foster dynamic models of ecosystem function and microbial interaction.

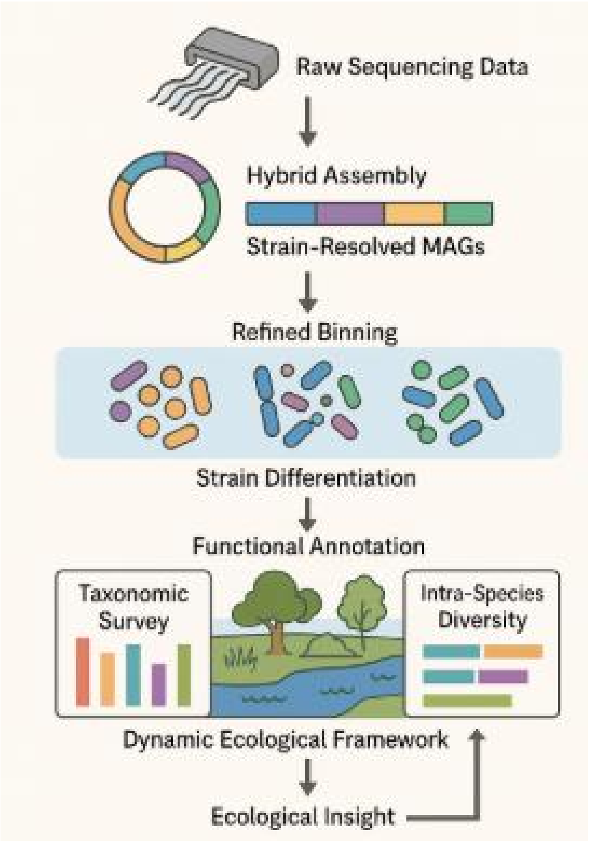


Figure 5 provides a conceptual visualization of how strain-resolved approaches refine microbial reconstruction, from raw sequencing data to ecological insight. To ensure widespread accessibility, researchers are encouraged to adopt workflow management systems (e.g., Snakemake, Nextflow), share annotated MAGs via platforms like MGnify or IMG/M, and contribute to open-source benchmarking initiatives.

With careful planning and adherence to best practices, hybrid metagenomics can become a mainstream tool for uncovering microbial dark matter and reshaping our understanding of life in diverse environments.

7. CONCLUSION AND FUTURE DIRECTIONS

7.1 Summary of Key Contributions

This study presents a comprehensive framework for hybrid-strain resolved metagenomics, combining long-read and short-read sequencing with advanced binning techniques to overcome persistent barriers in microbial genome reconstruction. Methodologically, it integrates hybrid assemblers such as *hybridSPAdes* and *Flye*, along with graph-based refinement tools like *GraphBin* and ensemble binning using *DAS Tool*. These innovations enabled substantial improvements in genome completeness, contiguity, and strain-level resolution across soil, marine, and human gut microbiomes.

Key findings include the recovery of 612 high- and medium-quality MAGs, including over 130 novel taxa with no close reference genomes. Strain-level clustering revealed substantial intra-species diversity with distinct metabolic capabilities, emphasizing the ecological importance of microdiversity. Functionally, hybrid assemblies enabled the recovery of rare biosynthetic gene clusters and complete metabolic pathways often missed by conventional approaches. Phylogenetic analyses expanded the microbial tree of life by identifying novel branches within CPR, Acidobacteria, and SAR clades.

This work demonstrates that hybrid metagenomics is not merely a technical enhancement but a necessary paradigm shift for accurately capturing and interpreting the functional and evolutionary diversity of uncultivated microbial life. It lays the groundwork for applying strain-resolved genomics in both ecological theory and real-world applications.

7.2 Roadmap for Scaling and Global Microbiome Projects

The demonstrated success of hybrid-strain resolved metagenomics provides a clear path forward for scaling these methods into large-scale, global microbiome initiatives. As environmental and clinical metagenomics evolve toward population-scale sampling, hybrid strategies offer critical improvements in data resolution, allowing researchers to distinguish between strains, mobile genetic elements, and low-abundance taxa at unprecedented detail.

In environmental monitoring, strain-level profiling can serve as an early warning system for ecological disruption, tracking microbial shifts associated with climate change, pollution, or land use conversion. Soil and marine microbiomes, for example, can be longitudinally surveyed to identify bioindicators based on strain dynamics and functional gene prevalence.

In public health, hybrid metagenomics can enhance pathogen surveillance by detecting strain-specific antimicrobial resistance genes or virulence factors within complex microbial communities. This has implications for outbreak detection, food safety, and personalized medicine.

To achieve global scalability, standardized hybrid workflows and data-sharing platforms must be developed, along with investments in computational infrastructure and long-read sequencing access. The approach aligns well with international initiatives such as the Earth Microbiome Project and the Human Microbiome Project 2.0, offering a high-resolution lens through which to examine microbial roles across ecosystems, hosts, and biogeographic gradients.

7.3 Open Questions and Technological Gaps

Despite significant progress, key challenges remain in fully realizing the potential of strain-resolved metagenomics. Computational bottlenecks especially in hybrid assembly and binning limit accessibility, particularly in resource-constrained settings. Moreover, accurate curation of closely related strains remains difficult in high-diversity environments, where strain boundaries are often blurred by horizontal gene transfer and recombination.

Future work must focus on scalable, automated pipelines for strain dereplication, improved long-read error correction, and reference-free validation methods. Integrating machine learning into assembly and binning tools may further improve precision, but rigorous benchmarking and cross-study consistency remain essential for long-term adoption.

REFERENCE

1. Merrill BD, Sonnenburg J. Discovery of Lifestyle-Associated Microbes, Pathogens, and Bacteriophages Through Deep Metagenome Sequencing of the Gut Microbiome. Stanford University; 2022.

2. Singh AP. Genomic Techniques Used to Investigate the Human Gut. Human microbiome. 2021 Jun 16:3.

3. Brumfield KD, Hasan NA. Human gut microbiome: metagenomic insights and prospectus. J Appl Microbiomics. 2024;1.

4. Forster SC, Browne HP, Kumar N, Hunt M, Denise H, Mitchell A, Finn RD, Lawley TD. HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. Nucleic acids research. 2016 Jan 4;44(D1):D604-9.

5. Feeney A, Sleator RD. The human gut microbiome: the ghost in the machine. Future microbiology. 2012 Nov 1;7(11):1235-7.

6. Waller AS, Yamada T, Kristensen DM, Kultima JR, Sunagawa S, Koonin EV, Bork P. Classification and quantification of bacteriophage taxa in human gut metagenomes. The ISME journal. 2014 Jul;8(7):1391-402.

7. Bowers RM. Gone with a trace: cataloguing the disappearing gut microbes. Nature Reviews Microbiology. 2023 Nov;21(11):704-.

8. Parfrey LW, Walters WA, Knight R. Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. Frontiers in microbiology. 2011 Jul 11;2:153.

9. Odumbo OR, Ezekwu E. Streamlining logistics in medical supply chains: Enhancing accuracy, speed, affordability, and operational efficiency. *Int J Res Publ Rev*. 2025;6(01):[pages not specified]. doi: https://doi.org/10.55248/gengpi.6.0125.0533.

10. Jamiu Olamilekan Akande, Joseph Chukwunweike. Developing scalable data pipelines for real-time anomaly detection in industrial IoT sensor networks. *Int J Eng Technol Res Manag (IJETRM)* [Internet]. 2023 Dec;7(12):497. Available from: DOI: 10.5281/zenodo.15813446

11. Olu Mbanugo, Onyekachukwu Victor Unanah. AI and Big Data in Addressing Healthcare Disparities: Focused Strategies for Underserved Populations in the U.S. *Int Res J Mod Eng Technol Sci*. 2025 Feb;7(2):670–90. doi: 10.56726/IRJMETS67321.

12. Chukwunweike JN, Mba JU, Kadiri C. Enhancing maritime security through emerging technologies: the role of machine learning in cyber threat detection and mitigation., USA. 2024 Aug. DOI: https://doi.org/10.55248/gengpi.5.0824.2401

13. Ogendi EG. Leveraging advanced cybersecurity analytics to reinforce zero-trust architectures within adaptive security frameworks. *Int J Res Publ Rev.* 2025 Feb;6(2):691–704. Available from: https://doi.org/10.55248/gengpi.6.0225.0729

14. Unanah Onyekachukwu Victor, Yunana Agwanje Parah. Clinic-owned medically integrated dispensaries in the United States; regulatory pathways, digital workflow integration, and cost-benefit impact on patient adherence (2024). *International Journal of Engineering Technology Research & Management (IJETRM)*. Available from: https://doi.org/10.5281/zenodo.15813306

15. Sani Zainab Nimma**.** Integrating AI in Pharmacy Pricing Systems to Balance Affordability, Adherence, and Ethical PBM Operations. *Global Economics and Negotiation Journal.* 2025;6(05):Article 19120. doi: https://doi.org/10.55248/gengpi.6.0525.19120.

16. Odumbo OR, Nimma SZ. Leveraging artificial intelligence to maximize efficiency in supply chain process optimization. *Int J Res Publ Rev*. 2025;6(01):[pages not specified]. doi: https://doi.org/10.55248/gengpi.6.0125.0508.

17. Athanasopoulou K, Adamopoulos PG, Scorilas A. Unveiling the human gastrointestinal tract microbiome: the past, present, and future of metagenomics. Biomedicines. 2023 Mar 9;11(3):827.

18. Khan A, Kim S, Hong ST. Gut Microbes Libraries: A Key Resource for Current Gut Microbiome Research. Journal of Bacteriology and Virology. 2025 Mar 31;55(1):1-9.

19. Lepage P, Leclerc MC, Joossens M, Mondot S, Blottière HM, Raes J, Ehrlich D, Doré J. A metagenomic insight into our gut's microbiome. Gut. 2013 Jan 1;62(1):146-58.

20. Jiao JY, Liu L, Hua ZS, Fang BZ, Zhou EM, Salam N, Hedlund BP, Li WJ. Microbial dark matter coming to light: challenges and opportunities. National Science Review. 2021 Mar;8(3):nwaa280.

21. Wong HL, MacLeod FI, White RA, Visscher PT, Burns BP. Microbial dark matter filling the niche in hypersaline microbial mats. Microbiome. 2020 Dec;8:1-4.

22. Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. Genome biology and evolution. 2018 Mar;10(3):707-15.

23. Lok C. Mining the microbial dark matter. Nature. 2015 Jun 18;522(7556):270.

24. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. Phylogenetically novel uncultured microbial cells dominate earth microbiomes. MSystems. 2018 Oct 30;3(5):10-128.

25. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013 Jul 25;499(7459):431-7.

26. Zha Y, Chong H, Yang P, Ning K. Microbial dark matter: from discovery to applications. Genomics, Proteomics & Bioinformatics. 2022 Oct;20(5):867-81.

27. Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Current opinion in microbiology. 2016 Jun 1;31:217-26.

28. Schultz J, Modolon F, Peixoto RS, Rosado AS. Shedding light on the composition of extreme microbial dark matter: alternative approaches for culturing extremophiles. Frontiers in Microbiology. 2023 Jun 2;14:1167718.

29. Vollmers J, Wiegand S, Lenk F, Kaster AK. How clear is our current view on microbial dark matter?(Re-) assessing public MAG & SAG datasets with MDMcleaner. Nucleic Acids Research. 2022 Jul 22;50(13):e76-.

30. Pavlopoulos GA, Baltoumas FA, Liu S, Selvitopi O, Camargo AP, Nayfach S, Azad A, Roux S, Call L, Ivanova NN, Chen IM. Unraveling the functional dark matter through global metagenomics. Nature. 2023 Oct 19;622(7983):594-602.

31. Zamkovaya T, Foster JS, de Crécy-Lagard V, Conesa A. A network approach to elucidate and prioritize microbial dark matter in microbial communities. The ISME journal. 2021 Jan;15(1):228-44.

32. Osburn ED, McBride SG, Strickland MS. Microbial dark matter could add uncertainties to metagenomic trait estimations. Nature Microbiology. 2024 Jun;9(6):1427-30.

33. Li S, Lian WH, Han JR, Ali M, Lin ZL, Liu YH, Li L, Zhang DY, Jiang XZ, Li WJ, Dong L. Capturing the microbial dark matter in desert soils using culturomics-based metagenomics and high-resolution analysis. npj Biofilms and Microbiomes. 2023 Sep 22;9(1):67.

34. Zhang Y, Wang Y, Tang M, Zhou J, Zhang T. The microbial dark matter and “wanted list” in worldwide wastewater treatment plants. Microbiome. 2023 Mar 28;11(1):59.

35. Hedlund BP, Dodsworth JA, Murugapiran SK, Rinke C, Woyke T. Impact of single-cell genomics and metagenomics on the emerging view of extremophile “microbial dark matter”. Extremophiles. 2014 Sep;18:865-75.

36. Wong HL, MacLeod FI, White RA, Visscher PT, Burns BP. Microbial dark matter filling the niche in hypersaline microbial mats. Microbiome. 2020 Dec;8:1-4.

37. Armengaud J. Metaproteomics to understand how microbiota function: The crystal ball predicts a promising future. Environmental Microbiology. 2023 Jan;25(1):115-25.

38. Cena JA, Zhang J, Deng D, Damé-Teixeira N, Do T. Low-abundant microorganisms: the human microbiome's dark matter, a scoping review. Frontiers in cellular and infection microbiology. 2021 May 31;11:689197.

39. Crowther TW, Rappuoli R, Corinaldesi C, Danovaro R, Donohue TJ, Huisman J, Stein LY, Timmis JK, Timmis K, Anderson MZ, Bakken LR. Scientists' call to action: Microbes, planetary health, and the Sustainable Development Goals. Cell. 2024 Sep 19;187(19):5195-216.

40. Nobu MK, Narihiro T, Rinke C, Kamagata Y, Tringe SG, Woyke T, Liu WT. Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. The ISME journal. 2015 Aug;9(8):1710-22.

41. Peisl L, Schymanski EL, Wilmes P. Dark matter in host-microbiome metabolomics: tackling the unknowns–a review. Analytica Chimica Acta. 2018 Dec 11;1037:13-27.

42. Ma B, Lu C, Wang Y, Yu J, Zhao K, Xue R, Ren H, Lv X, Pan R, Zhang J, Zhu Y. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. Nature communications. 2023 Nov 11;14(1):7318.