Protocol Biol217 “Microbial -Omics”

biol217-01a (Marine) Microbial Omics - from sample to function 25/26

26/01/2026 – 06/02/2026

Submitted by: Sadikshya Khanal

Registration Number: 1204169

Taught by: Cynthia, Katharina, Almut, Hendrik, and Vincent

AG Schmitz-Streit

Date of Submission: 22/2/2026

Contents

List of Abbreviations

- DNA – Deoxyribonucleic acid
- RNA – Ribonucleic acid
- RNA-seq – RNA sequencing
- MAG(s) – Metagenome-Assembled Genome(s)
- vMAG(s) – Viral Metagenome-Assembled Genome(s)
- vOTU(s) – Viral Operational Taxonomic Unit(s)
- AMG(s) – Auxiliary Metabolic Gene(s)
- ANI – Average Nucleotide Identity
- MIMAG – Minimum Information about a Metagenome-Assembled Genome
- ORF(s) – Open Reading Frame(s)
- CDS – Coding Sequence
- tRNA(s) – Transfer RNA(s)
- rRNA(s) – Ribosomal RNA(s)
- CSS – Clade Separation Score
- RPKM – Reads Per Kilobase Million
- kbp – Kilobase pairs
- Mbp – Megabase pairs
- GC – Guanine-Cytosine content
- N50 – Contig length at which 50% of assembly is contained in contigs of equal or greater length
- Q20 / Q30 – Phred quality scores
- log2FC – Log2 Fold Change
- SRA – Sequence Read Archive
- SRR – Sequence Read Run accession
- NCBI – National Center for Biotechnology Information
- FASTQ – Sequence format including quality scores
- FASTA (.fna) – Sequence format (nucleotide)
- FASTG – Graph-based FASTA format
- GFF – General Feature Format
- SAM – Sequence Alignment/Map
- BAM – Binary Alignment/Map

Introduction

The advancement of high-throughput sequencing technologies has transformed research into the structure, function, and dynamics of microbial communities across various environments[1]. This study combines multiple omics approaches, including metagenomics, genomics, pangenomics, transcriptomics, and viromics, to thoroughly characterize microbial and viral populations across different samples. Such integrative analyses produce complex datasets, making it crucial to document workflows, manage software environments, and track versions to ensure reproducibility and long-term usability of the research outputs[2] [3].

Documentation plays a vital role in the research process. Tools like Git and GitHub facilitate version control and collaborative development, enabling researchers to track code changes and share their work transparently [3]. Additionally, Conda environments offer a reliable way to manage software dependencies, ensuring that analyses can be reproduced years after publication[4]. Metagenomics involves directly sequencing DNA from environmental samples without the need for laboratory cultivation[5]. Sequencing DNA extracted from different samples provides a comprehensive view of microbial community composition and functional capacity[5]. Modern metagenomic workflows, such as FastQC, MEGAHIT, METABAT, and CheckM, offer essential assessments of genome completeness and contamination[6] [7] [8] [9]. These methods enable the recovery of nearly complete genomes from environmental samples, revealing the phylogenetic diversity and metabolic potential of uncultivated microorganisms [9].

For studies on individual microbial isolates, genome assembly remains a fundamental technique. The development of third-generation sequencing technologies, especially Oxford Nanopore Technologies (ONT) and PacBio High-Fidelity (HiFi) sequencing, has enhanced the quality and continuity of genome assemblies[10]. Hybrid assembly strategies, which combine the accuracy of short reads with long-range information from long reads, can produce nearly complete genome assemblies[11]. Pangenomics extends beyond a single reference genome to encompass the entire genetic repertoire of a species or taxonomic group[12] [13]. It consists of three main components: the core genome, the accessory genome, and singletons. In bacterial populations, analyzing the accessory genome provides insights into niche-specific adaptations, including substrate utilization, stress resistance, and virulence factors [13]. Transcriptomics is the method used to measure gene expression levels to identify which genes are actively transcribed under specific conditions. RNA sequencing (RNA-seq) is the standard technique for differential expression analysis[14]. Viromics applies metagenomic principles to viral communities. The viromic workflow involves physically separating viral particles, extracting viral DNA or RNA, sequencing, and using computational methods to identify viral sequences from metagenomic assemblies[15]. Quality assessment tools like CheckV evaluate genome completeness, while clustering approaches define viral operational taxonomic units (vOTUs)[16] [17]. Functional annotation of viral genes, including the detection of auxiliary metabolic genes (AMGs), reveals how viruses manipulate host metabolism during infection[18].

This study utilizes these complementary omics approaches metagenomics, genome assembly, pangenomics, transcriptomics, and viromics to analyze microbial and viral communities. By maintaining strict documentation and reproducibility standards throughout the analytical process, we ensure that our results are transparent, verifiable, and reusable by the wider scientific community. The findings here enhance our understanding of microbial diversity, genomic structure, gene expression patterns, and viral ecology in the environment examined.

Material and Methods

Part 1: Metagenomics

To ensure high-quality raw reads, FastQC (Version 0.12.1) was used for initial quality assessment, analyzing metrics such as Phred scores. Next, fastp (Version 0.24.1) was applied for trimming and filtering to remove bases with Phred scores below 20 from paired-end reads. The filtered FastQ files underwent a second quality check with FastQC. For assembly, MEGAHIT (Version 1.2.9) utilizing the "meta-large" preset was employed, excluding sequences shorter than 1000 base pairs. The Megahit_toolkit converted the final sequences from FASTA to FASTG format, which was then visualized using Bandage (Version 0.8.1). To evaluate assembly quality, MetaQUAST (Version 5.3.0) was used on the assembled files, and short contigs were filtered out.

Before mapping, sequence IDs were simplified and shorter contigs were removed with anvi-script-reformat-fasta (Anvi'o Version 9)[20]. Bowtie 2 (Version 2.5.2) was then used for mapping, producing SAM files that were converted into BAM files with SAMtools (Version 1.23) for more efficient processing[21] [22]. Then, the BAM files were sorted and indexed with anvi-init-bam, and the contigs were subsequently binned into MetaGenome Assembled Genomes (MAGs) to identify the microbes present. After that, anvi-gen-contigs-database was used to create a database from the FASTA format; it also calculated k-mer frequencies and identified open reading frames (ORFs). Using anvi-run-hmms, the biological functions of the predicted ORFs were searched, and a sample profile was created with anvi-profile. These profiles were merged into one database (PROFILE.db) for unified analysis.

To cluster contigs into MAGs, MetaBAT 2 (Version 2.18) and MaxBin 2.0 (Version 2.2.7) were used[23] [24]. The quality of the MAGs was assessed with anvi-estimate-genome-completeness, followed by manual inspection. To check for assembly errors, GUNC (Version 1.0.6) was run to identify chimeric genomes[25]. The best bins (over 70% completeness) were refined using anvi-refine and re-examined for their abundance with anvi-interactive. Taxonomic annotations were added to the MAGs by running anvi-run-scg-taxonomy with 20 threads, and taxonomy was estimated using anvi-estimate-scg-taxonomy in metagenome mode.

Part 2: Genomics/Genome Assembly

In this step, bacterial genomes were assembled from short and long reads. First, data was loaded into the working directory. FastQC (Version 0.12.1) was used to assess the quality of the raw short reads, followed by cleaning with fastp (Version 0.24.1) to trim adapters and filter out low-quality bases. FastQC was subsequently run again on the cleaned reads to confirm improvements in quality. For long reads, NanoPlot (Version 1.46.2) was used to visualize the quality and length distribution[26]. Filtlong (Version 0.3.1) was then applied to filter the reads, retaining only those above a specific length threshold (e.g., >1000 bp). This process ensured that both read types were of high quality prior to hybrid assembly[27].

Unicycler (Version 0.5.1) was used for hybrid assembly, integrating the cleaned short and long reads to generate a high-quality genome[28]. This approach combined the accuracy of short reads with the structural information provided by long reads, effectively addressing complex genomic regions. Quality assessment was performed using tools such as QUAST (Version 5.3.0) to provide metrics, including N50, while CheckM (Version 1.2.4) and CheckM2 (Version 1.1.0) evaluated genome completeness and contamination[29] [9] . For visual inspection of the assembly graph, Bandage (Version 0.8.1) was employed. Annotation of the genome was conducted using Prokka (Version 1.15.6) to identify genes, and GTDB-Tk (Version 2.6.1) was utilized for taxonomic classification against the Genome Taxonomy Database[30] [31] [32]. Finally, MultiQC (Version 1.33) aggregated the quality reports to verify the accuracy and completeness of the assembled genome[33].

Part 3: Pangenomics

The genomes of 52 *Vibrio jasicida* strains were downloaded and prepared for analysis using anvi-script-reformat-fasta (Anvi'o Version 9) to simplify sequence names and filter out short contigs. An Anvi'o contigs database was generated for each genome with anvi-gen-contigs-database. Contigs were annotated using Hidden Markov Models, NCBI COGs, tRNAs, and taxonomy with various anvi-run- commands within the Anvi'o environment, allowing for the attachment of essential functional information needed for genome comparison. An external genomes file was created for contamination checking and manual refinement, followed by anvi-estimate-genome-completeness to assess contamination and completeness. Contigs were visualized with anvi-interactive for manual inspection, separating high-quality "good" bins from "bad" bins. The cleaned selection was saved, and anvi-split was utilized to refine the genome, updating the external genomes file accordingly to ensure only high-quality genomic data was used for pangenome analysis.

A genome storage database was generated using anvi-gen-genomes-storage to consolidate information from all samples. The pangenome was calculated with anvi-pan-genome to identify gene clusters, while Average Nucleotide Identity (ANI) was assessed using anvi-compute-genome-similarity to evaluate genomic similarity between strains. Results were visualized with anvi-display-pan to compare the genomes, identify the core genome, and determine evolutionary relationships based on sequence similarity.

Part 4: Transcriptomics

First, raw RNA-Seq data for *Methanosarcina mazei* was retrieved from the Sequence Read Archive (SRA). To download the specific SRR files identified in the study by Prasse et al., grabseqs (Version 1.0.0) was used. The raw files were renamed to meaningful sample identifiers (wt_R1, mut_R1). Reference genome (fna) and gene annotation files (.gff) were sourced from NCBI to facilitate the interpretation of RNA reads. For the mapping of the raw reads to the reference genome, reademption align (READemption Version 2.0.4) was used[34]. Once aligned, the coverage was calculated using reademption coverage. Then, gene-wise quantification was performed using reademption gene_quanti, which counts the number of reads mapping to specific features such as CDS, tRNAs, and rRNAs. This was done to locate the origin of each read in the genome. Quantification converts these read locations into numerical data (counts), representing the expression level of each gene. This allows for the observation of which genes are active and to what degree. To calculate differential gene expression, reademption deseq, which acts as a wrapper for the DESeq2 package (Version 1.44.0), was used. The "mutant" condition was compared with the "wild-type" condition to identify genes with statistically significant expression changes[35]. Finally, visualizations were generated using various reademption viz commands to interpret the results. This step was done to identify biologically significant differences between conditions. It determines which genes are upregulated or downregulated in response to the experimental condition (mutation).

Part 5: Viromics

The Modular Viromics Pipeline (MVP) was utilized to automate the processing of viral metagenomic data. Within this pipeline, geNomad (Version 1.11.2) was used to identify viruses, proviruses, and plasmids from assembled data. Following identification, CheckV (Version 1.5.0) was used to assess the quality of viral sequences, determining completeness and filtering low-quality hits to establish a robust workflow for viromics[36] [16]. This process separated viral signals from background host data and classified viruses as "High-quality" or "Complete," ensuring that subsequent analyses relied on reliable sequences. For viral clustering, CheckV was used. It uses Average Nucleotide Identity to group similar viruses and identify cluster representatives, thereby reducing redundancy. Reads were subsequently mapped to these representatives using Bowtie 2 (Version 2.5.4), Minimap2 (Version 2.28), and CoverM (Version 0.7.0) to calculate abundance metrics, including RPKM[21] [37] [38]. The genes of viral cluster representatives were then annotated using Prodigal (Version -- 2.6.3), MMseqs2 (Release 18), and HMMER (Version 3.4) against databases such as PHROGs, facilitating the creation of Viral Operational Taxonomic Units (vOTUs) for simplified dataset management[39] [40] [41]. For genome reconstruction and host linking, vRhyme (Version 1.1.0) was utilized to bin viral contigs into Viral Metagenome-Assembled Genomes (vMAGs). iPHoP (integrated Phage-Host Prediction; Version 1.4.2) was executed manually to link identified viruses to their likely bacterial or archaeal hosts. This process reconstructed fragmented viral genomes by grouping contigs based on coverage and composition. Host prediction provides information on the ecological roles of viruses and the hosts they infect[42] [43].

Results

Part 1: Metagenomics

The initial metagenomic assembly yielded 319 contigs totaling 2.2 Mbp and an N50 of 7.9 kbp. To reconstruct individual genomes from this mixed assembly, two binning strategies were applied. MetaBAT2 and MaxBin2 differed in their ability to recover high-quality metagenome-assembled genomes (MAGs). MaxBin2 binned a larger proportion of the total assembly (92.6%) but produced only five MAGs that met the stringent MIMAG criteria for high-quality drafts (>90% completeness, <5% redundancy). In contrast, MetaBAT2, which binned 46.9% of the assembly, generated nearly three times as many high-quality MAGs, recovering 13 distinct bins that satisfied these standards. Due to its superior performance, MetaBAT2 was selected for all subsequent analyses.

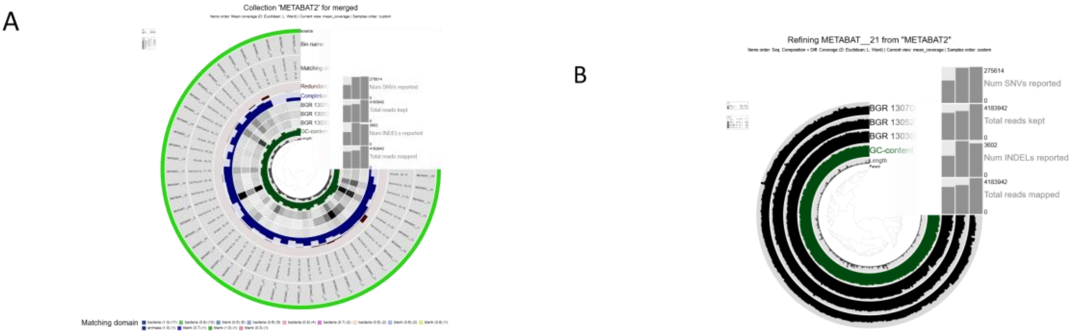

Figure 1: Visualization and manual refinement of MetaBAT2 metagenomic bins using Anvi'o. (A) Overview of the metagenomic bin collection generated by MetaBAT2. The central dendrogram organizes contigs based on sequence composition and differential coverage. Concentric circular layers display bin metrics, including GC-content, redundancy, completion, and mean coverage across various samples. (B) The radial plot illustrates the 262 contigs retained after manual pruning to eliminate chimeric fragments and improve bin purity.

The MetaBAT2 binning strategy recovered three distinct archaeal MAGs from the dataset. Evaluation of these bins indicated that only one, METABAT__21, met and exceeded the internationally recognized MIMAG standards. This MAG, taxonomically classified within the Archaea; *Halobacteriota*; *Methanomicrobia* lineage, demonstrated a completeness score of 98.68% and a redundancy of 3.95%, indicating near-complete genome recovery with minimal cross-species contamination. The remaining two archaeal bins, METABAT__10 and METABAT__13, did not meet the criteria for medium- or high-quality drafts, with completeness metrics of 48.7% and 38.2%, respectively. Following manual refinement of METABAT__21 to near-reference quality, a single chimeric contig was removed from its 263 contigs. Post-refinement validation using GUNC confirmed the elimination of this fragment, with the genome achieving a clade separation score (CSS) of 0.0 and a contamination proportion of 0.0% at the kingdom level. Additionally, 99% of its genes were confidently assigned to the target evolutionary lineage, confirming the bin's purity and phylogenetic consistency. The binning process was highly effective for the bacterial fraction of the community, yielding 12 high-quality bacterial MAGs. Taxonomic classification assigned 43 of the 46 recovered MAGs to Bacteria and three to Archaea. The bacterial community was dominated by the phylum Bacillota, particularly members of the class Clostridia. All three archaeal MAGs were affiliated with *Halobacteriota* and classified as methanogens within the orders *Methanomicrobiales* and *Methanosarcinales*, including the species *Methanoculleus thermohydrogenotrophicum* and *Methanosarcina flavescens*. These findings indicate a bacterial-dominated community with a distinct methanogenic archaeal component.

Part 2: Genomics/genome assembly

The initial paired-end short read dataset consisted of 3,279,098 reads, totaling 823.05 million bases. Read quality was high, with 94.10% of bases scoring above Q20 and 86.05% above Q30. The average sequence length was 251 bp, and the duplication rate was 4.12%. The peak insert size was 459 bp. After quality filtering and adapter trimming with fastp, 3,226,784 reads (790.25 million bases) were retained, accounting for over 98% of the original dataset. In the filtered dataset, the mean read length was 244 bp, bases scoring above Q20 increased to 94.74%, bases above Q30 increased to 86.96%, and GC content was 45.13%.

The initial Nanopore long-read dataset consisted of 16,000 sequences totaling 146.3 Mb. The raw reads had a median length of 3,270 bp, an average length of 9,166 bp, a read N50 of 21,971 bp, and a median quality score of 11.7. After filtering with filtlong, 12,400 reads (131.7 Mb) were kept. The filtered dataset showed a median length of 4,477 bp, an average length of 10,580 bp, an N50 of 22,747 bp, and a median quality score of 12.5.

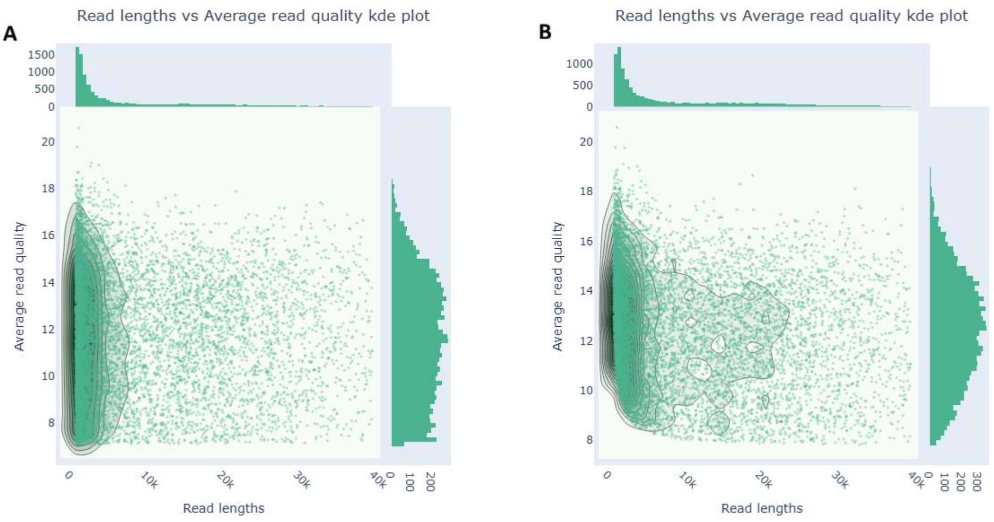

Figure 2: Distribution of Nanopore long-read lengths and quality scores before and after filtering with filtlong. Reads shorter than 1,000 bp and lower-quality sequences were effectively removed to optimize the hybrid assembly.

Hybrid assembly with Unicycler resulted in 7 contigs totaling 4,454,332 bp. The N50 value was 4,454,332 bp (about 4.5 Mbp), indicating that over 97% of the bacterial genome was assembled into a single continuous fragment. Marker gene analysis using CheckM, which evaluated 492 markers across 269 marker sets for the order Bacteroidales, estimated a genome completeness of 98.88%, a contamination rate of 0.19%, and strain heterogeneity of 0.00%. Neural network-based assessment with CheckM2 predicted a completeness of 99.98% and a contamination rate of 0.29%. Visual inspection of the assembly using Bandage is shown in the figure below.

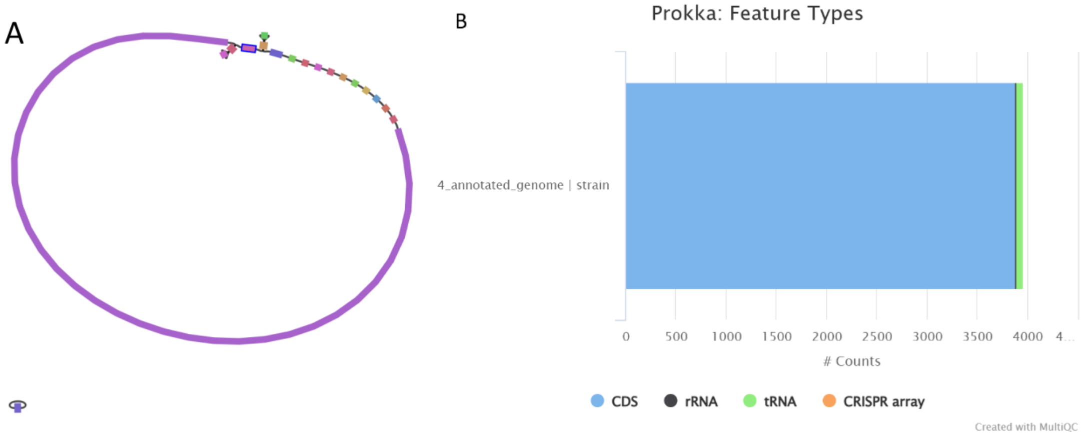

Figure 3: Structural and functional overview of the assembled bacterial genome. (A) Visual inspection of the de novo assembly graph using Bandage. The assembly resulted in a circular, unbroken bacterial genome with a total size of 4,454,332 bp. The graph represents a single connected component comprised of 7 nodes and 7 edges with no dead ends **(B)** Distribution of functional feature types annotated by Prokka. The horizontal stacked bar chart displays the total counts of genomic features identified in the annotated strain. The legend indicates colors corresponding to Coding Sequences (CDS, blue), rRNA (black), tRNA (green), and CRISPR arrays (orange).

Genome annotation using Prokka identified a total of 7,927 biological features, including 3,883 unique protein-coding sequences (CDS), 3,963 genes in total, 65 transfer RNAs (tRNAs), and 15 ribosomal RNAs (rRNAs). The overall coding density was 88.9%, with an average gene length of 338.7 base pairs. Taxonomic classification was performed with GTDB-Tk. Based on a GC content of 45.71% and the alignment of conserved marker genes, the assembly was classified as follows: Domain Bacteria, Phylum Bacteroidota, Class Bacteroidia, Order Bacteroidales, Family Bacteroidaceae, Genus Bacteroides, Species Bacteroides muris.

Part 3: Pangenomics:

Evaluation of the eight Vibrio jascida assemblies showed 100.00% completeness for all genomes. Redundancy was 5.63% in seven genomes, while V_jascida_52 had a higher redundancy of 21.13%. Visual inspection of the contig profile for V_jascida_52 revealed a structurally distinct cluster of contigs diverging from the main chromosome, distinguished by sequence composition and coverage metrics.

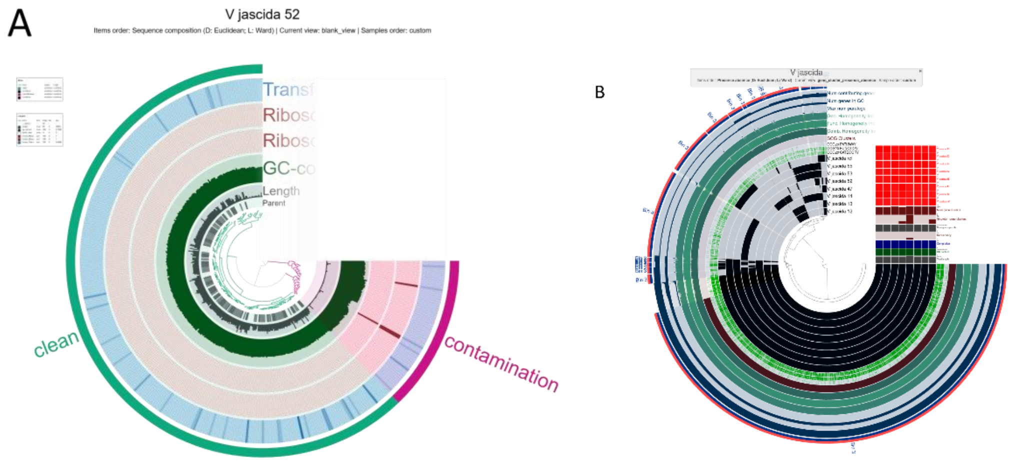

Figure 4. Genome refinement and pangenome analysis of Vibrio jascida lineages using Anvi'o. (A) Interactive contig profile for V. jascida strain 52. Contigs are clearly distinguished as contamination clusters or clean genome bins based on sequence composition and differential coverage metrics. (B) Pangenome presence/absence matrix of eight V. jascida genomes. Each central concentric ring represents an individual genome or MAG, with dark segments indicating the presence of specific gene clusters and light or blank segments indicating their absence.

Pangenome analysis of the V. jascida dataset revealed a highly organized presence/absence matrix. The core genome is represented by tightly packed, continuous inner layers of the hierarchical cluster. The accessory genome exhibits scattered presence/absence patterns in the outer layers of the plot. Distinct singleton gene clusters were identified, especially within strains V_jascida_47 and V_jascida_14. Pairwise comparisons among the eight genomes showed Average Nucleotide Identity (ANI) values above 97.5% for all pairs. The lowest identity, 97.50%, was observed between V_jascida_ref and V_jascida_13. Highly clonal sub-lineages were identified, with an ANI exceeding 99.99% between V_jascida_12 and V_jascida_13, and among V_jascida_14, V_jascida_53, and V_jascida_55.

Part 4: Transcriptomics

The stacked species box plot shown in the figure below indicates that all mapped reads aligned exclusively to the reference genome of *Methanosarcina mazei* Gö1. Additionally, the aligned reads box plots show that most reads aligned uniquely to the reference genome, with little evidence of split or cross-alignment.

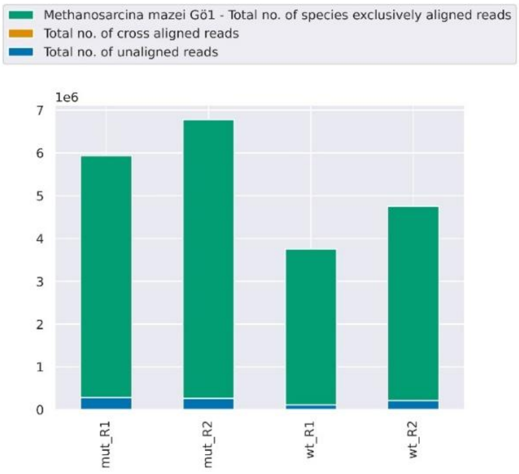

Figure 5: Comparative read alignment distribution across mutant (mut) and wild-type (wt) replicates. The stacked bar chart illustrates the total number of reads exclusively aligned to the *Methanosarcina mazei* Gö1 reference genome (green) versus unaligned reads (blue).

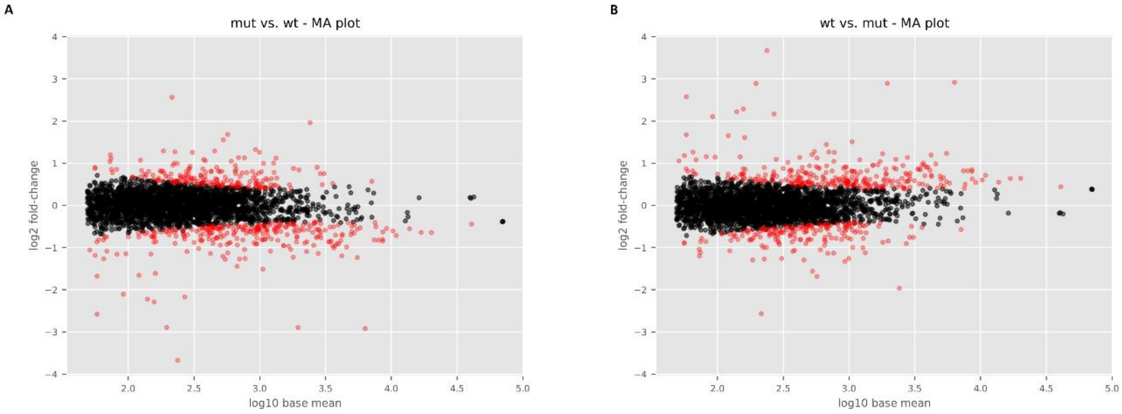

Figure 6. MA plots of differential gene expression between wild-type and mutant conditions. (A) The reciprocal MA plot shows the mutant (mut) versus wild-type (wt). (B) MA plot showing the comparison of wild-type (wt) versus mutant (mut). In both panels, the x-axis displays the log10-transformed mean expression (base mean), and the y-axis displays the log2 fold-change. Each point represents a gene. Red points indicate significantly differentially expressed genes (DEGs), while black points denote genes with non-significant changes in expression.

MA and volcano plots were created to visualize differential expression between conditions. Plots were generated for both raw and adjusted p-values. Adjusted p-values, which are more conservative, help reduce false positives. In both types of plots, the x-axis shows log2 fold changes, with positive values indicating upregulation in the first condition compared to the second. The y-axis displays the negative log10 of the p-value, so higher points indicate lower p-values. Green dotted lines mark significance thresholds: p-value less than 0.05 and log2 fold change greater than 1 or less than -1. Scatter plots of gene expression were created to compare levels across different conditions, showing sample correlations and R-values. Box plots for RNA classes were also produced. These plots display the total read counts for coding sequences (CDS), transfer RNA (tRNA), and ribosomal RNA (rRNA) for each sample.

Part 5: Viromics

After co-assembly and viral identification with geNomad, a total of 10,909 viral sequences were identified across all samples. Analysis of a representative sample, BGR_140717, revealed 849 free viral contigs and 11 proviruses, indicating a predominantly lytic or active viral population. Taxonomic classification of the filtered viral sequences showed a strong dominance of the class Caudoviricetes (tailed dsDNA viruses), which accounted for 10,764 of the identified viruses. A smaller subset of 145 sequences remained unclassified at this taxonomic level. Other less common viral groups were also detected, representing diverse genome types and hosts, including dsDNA viruses infecting eukaryotes (e.g., Mimiviridae, Phycodnaviridae), ssDNA viruses infecting prokaryotes (*Microviridae*), and ssRNA-RT viruses (Retroviridae). After co-assembly and viral identification with geNomad, a total of 10,909 viral sequences were identified across all samples. Analysis of a representative sample, BGR_140717, revealed 849 free viral contigs and 11 proviruses, indicating a predominantly lytic or active viral population. Taxonomic classification of the filtered viral sequences showed strong dominance of the class Caudoviricetes (tailed dsDNA viruses), accounting for 10,764 of the identified viruses. A smaller subset of 145 sequences remained unclassified at this taxonomic level. Other less common viral groups were also detected, representing diverse genome types and hosts, including dsDNA viruses infecting eukaryotes (e.g., Mimiviridae, Phycodnaviridae), ssDNA viruses infecting prokaryotes (Microviridae), and ssRNA-RT viruses (Retroviridae).

After co- assembly and viral identification with geNomad, a total of 10, 909 viral sequences were found across all samples. Analysis of a representative sample, BGR _ 140717, showed 849 free viral contigs and 11 proviruses, indicating a mainly lytic or active viral population. Taxonomic classification of the filtered viral sequences revealed a strong dominance of the class Caudoviricetes (tailed dsDNA viruses), which made up 10, 10,764 of the identified viruses. A smaller group of 145 sequences remained unclassified at this taxonomic level. Other less common viral groups were also detected, representing diverse genome types and hosts, including dsDNA viruses infecting eukaryotes (e. g., Mimiviridae, Phycodnaviridae), ssDNA viruses infecting prokaryotes (Microviridae), and ssRNA-RT viruses (Retroviridae). Quality assessment showed most viral sequences were low- quality. Across all 18 time points, a total of 10, 884 low- quality, 21 medium- quality, 2 high- quality, and 3 complete viral genomes were identified. The complete viruses were only found in three samples: BGR _ 131021, BGR _ 140121, and BGR _ 140717. None were identified as integrated proviruses. Their genome lengths were 46, 113 bp (BGR _ 131021), 34, 34,619 bp (BGR _ 140121), and 31, 258 bp (BGR _ 140717). The number of viral hallmark genes varied, with BGR _ 131021 containing 12, BGR _ 140121 containing 24, and BGR _ 140717 containing 21. These hallmark genes comprised roughly 23%, 51%, and 48% of their total predicted genes, respectively. The remaining genes did not show significant similarity to known viral or host sequences in current databases. Clustering identified 5, 375 non-redundant vOTUs, including 91 (1. 7%) proviruses. Three complete viral genomes each represented a vOTU, with cluster sizes of 7, 25, and 11 for samples BGR _ 131021, BGR _ 140121, and BGR _ 140717. Functional annotation of the 5, 375 vOTUs revealed 7, 749 hits. The BGR _ 131021 virus (53 genes) mainly encoded proteins involved in DNA, RNA, and nucleotide metabolism. BGR _ 140121 (47 genes) showed a similar profile. BGR _ 140717 (44 genes) encoded a wider range of functions, including lysis, structural, and nucleic acid metabolism genes.

Key metabolic pathways among viral AMGs included sulfur and nucleotide metabolism, cofactor and vitamin biosynthesis, amino acid metabolism, and transcriptional regulation. Toxin-related and host-interaction genes were also identified, including zot, PAAR motif, lar, and homing endonucleases. Of the three complete viral genomes, only the BGR_140717 virus had a host prediction: it likely infects MAG BGR_130829_bin.14 (phylum Bacillota, class Clostridia, order Tissierellales, family Peptoniphilaceae). The other two viruses had no host predictions. Viruses with multiple host predictions showed two trends: some linked to closely related hosts (e.g., BGR_130305_NODE_1097 to several *Bacillus* A species with high confidence) and others to distantly related hosts (e.g., BGR_130305_NODE_1187 to *Clostridium_AE oryzae* and Fervidobacterium, based only on blast homology and lower confidence).

Discussion:

The metagenomic workflow was used to reconstruct microbial genomes from a complex community, resulting in a set of high-quality metagenome-assembled genomes (MAGs) that reveal the community's structure and function. The initial assembly produced 319 contigs totaling 2.2 Mbp. Although this assembly is relatively small, it provides a sufficient foundation for genome-resolved analysis. The choice of binning strategy was crucial in achieving these results. MaxBin2 binned a larger portion of the assembly (92.6%) but produced only five high-quality MAGs[44]. In contrast, MetaBAT2, which binned a smaller part of the data (46.9%), was much more effective at reconstructing nearly complete genomes, recovering 13 MAGs that meet the strict MIMAG standards for high-quality drafts (>90% complete, <5% redundant)[23] [45]. This difference shows that different binning algorithms, which use distinct features such as tetranucleotide frequencies or differential coverage, can yield complementary results. MetaBAT2's superior performance in generating complete MAGs from this dataset led to its selection for all future analyses, confirming its effectiveness for understanding community structure based on available coverage data[23]. The dominance of the phylum Bacillota, particularly the class Clostridia, aligns with expectations for anaerobic settings rich in organic material, where these groups play a key role in fermentation[46]. The recovery of three archaeal MAGs, all identified as methanogens within Halobacteriota (orders Methanomicrobiales and Methanosarcinales), indicates a functionally important, though less abundant, archaeal community.

A complete bacterial genome was assembled using a hybrid approach with Unicycler, which combines short and long sequencing reads[28]. Short Illumina reads, processed with fastp to ensure high base quality (Q20 >94%), provided the accuracy needed to resolve fine-scale sequence details[19]. Long Nanopore reads, filtered by Filtlong to remove fragments shorter than 1000 bp and to improve overall quality (N50 increased to 22.7 kbp), supplied the long-range information essential for bridging repetitive elements[27]. This combined method was crucial for assembling the circular chromosome into a single 4.33 Mbp contig, demonstrating that Unicycler leverages the strengths of both data types to overcome the limitations of either alone. The assembly's high quality was carefully validated with multiple complementary tools. QUAST confirmed excellent contiguity with an N50 of 4.45 Mbp, indicating the genome is essentially complete as a single piece[29]. Biological completeness was assessed with CheckM and CheckM2, which searched for lineage-specific single-copy marker genes. High completeness scores (98.88% from CheckM and 99.98% from CheckM2) and low contamination levels (0.19-0.29%) suggest that nearly all expected genes are present and the genome is free from significant foreign DNA[9]. Bandage visualization of the assembly graph revealed a single, unbroken circular structure[32]. This visual confirmation verifies that the assembly is not a linear misassembly but a truly complete, circular chromosome. Taxonomic assignment was performed using GTDB-Tk, which identified the isolate as *Bacteroides muris*[30]. This identification was further confirmed by comparing the genome to its closest relative using FastANI, which calculates genome-wide Average Nucleotide Identity (ANI)[47]. Manual refinement of METABAT_21 improved bin quality by removing chimeric fragments and nearly perfecting validation scores for lineage purity[23]. The other two archaeal bins were incomplete, indicating challenges in recovering genomes from low-abundance or fragmented populations. Dominant *Bacillota* and *Bacteroidota* are typical of anaerobic, organic-rich environments. Other detected phyla suggest a complex, multi-niche community[48]. Species-level taxonomy, including uncultivated lineages and identified methanogens, provides functional insights.

Pangenome analysis of eight *Vibrio jasicida* genomes provided a comprehensive overview of the species' genomic structure. Initial quality checks with anvi'o indicated all genomes were 100% complete[20] [49]. However, a redundancy of 21.13% in strain V_jasicida_52, compared to 5.63% in other strains, pointed to possible contamination. These findings emphasize the importance of manual correction, as automated metrics might overlook such issues. Using anvi-split to isolate and remove this contamination was essential, ensuring that only high-quality data were used in the subsequent pangenome analysis. Visualization of the pangenome as a presence/absence matrix revealed a densely packed inner layer representing the core genome, which includes genes shared by all eight strains. This core likely encodes essential functions such as replication, metabolism, and environmental adaptation, indicating strong purifying selection[50]. Conversely, the outer layers represented the variable accessory genome, with distinct singleton gene clusters observed only in V_jascida_47 and V_jascida_14. These singletons, identified through anvi-pan-genome analysis, may correspond to recent evolutionary innovations, such as mobile genetic elements or niche-specific adaptations, underscoring the dynamic nature of the V. jasicida pangenome[51]. Pairwise Average Nucleotide Identity (ANI), calculated using anvi-compute-genome-similarity, confirmed that all strains belong to the same species, with values exceeding the 97.5% threshold. The minimum identity of 97.5% between the reference and V_jascida_13 suggests possible geographic or ecological divergence[13]. Meanwhile, ANI values greater than 99.99% between several strains, such as V_jascida_12 and V_jascida_13, indicate highly clonal sub-lineages, reflecting recent common ancestry or strong selective pressures.

Transcriptomic analysis of *Methanosarcina mazei* , comparing a mutant to the wild type, revealed distinct patterns of differential gene expression. The stacked species box plot confirmed that all mapped reads aligned exclusively to the *M. mazei* Gö1 reference genome, validating the RNA-Seq experiment's specificity and ruling out cross-species contamination[34]. The MA plot offers a global overview of expression changes by plotting log2 fold change against mean expression. Genes deviating from the zero line are candidates for altered expression, and the magnitude of change reflects the strength of the transcriptional response[35]. In *Methanosarcina*, even moderate fold changes in metabolic genes can have significant pathway-level effects, especially if they involve rate-limiting enzymes. Volcano plots comparing raw and adjusted p-values provide a clearer statistical view. The adjusted plot shows fewer significant genes but with greater confidence, effectively controlling the false discovery rate. This conservative approach, used in DESeq2, helps reduce false positives common in RNA testing[35]. Box plots generated with Reademption display the distribution of reads across coding sequences (CDS), tRNA, and rRNA. Typically, rRNA dominates transcriptomic libraries, reflecting the cell's high demand for protein synthesis. Accurate mapping to tRNA and rRNA genes confirms that the annotation accurately captures these non-coding features and that library preparation preserved these RNA classes[14]. Scatter plots of expression correlations (R-values) between conditions are essential for assessing experimental reproducibility. A high correlation between biological replicates (>0.95) indicates low technical variability and reliable quantification, whereas a lower correlation between conditions suggests genuine biological differences[14] [35]. The observed R-values increase confidence that the differential expression signals are robust.

Viromic analysis of 18 samples using the MVP pipeline revealed a diverse viral community, mainly composed of the class Caudoviricetes (tailed dsDNA viruses), with 10, 10,764 sequences identified[15]. This dominance is consistent with global trends, as Caudoviricetes are repeatedly the most abundant viral group across various ecosystems. Quality assessment using CheckV classified most sequences as low-quality, a common finding due to the difficulty of assembling complete viral genomes from complex metagenomes[16]. The three high-quality viral genomes showed significant variation in hallmark gene densities, ranging from 23% in BGR _ 131021 to 51% in BGR _ 140121, indicating distinct genomic strategies. Clustering viral sequences into 5, 375 non- redundant viral operational taxonomic units (vOTUs) reduced redundancy and revealed population structure[16]. Functional annotation of vOTUs using Prodigal and MMseqs 2 against the PHROGs database yielded 7, 749 hits, including AMGs involved in sulfur and nucleotide metabolism, cofactor biosynthesis, and transcription regulation. Notably, the detection of sulfur metabolism genes (cysD, cysH) suggests viral manipulation of host sulfur assimilation pathways, possibly enhancing nucleotide synthesis or affecting biogeochemical cycling[15] [39] [41] [52]. The discovery of toxins such as zot, the PAAR motif, and lar indicates advanced viral counter- defense strategies, highlighting the ongoing evolutionary arms race with bacterial hosts. Host prediction with iPHoP established a direct ecological link between the complete virus BGR _ 140717 and a specific bacterial metagenome- assembled genome (MAG), BGR _ 130829 _ bin.14 (phylum Bacillota, family Peptoniphilaceae) [53]. The lack of host predictions for the other two complete viruses underscores current database limitations for less- studied viral lineages. Virus BGR _ 130305 _ NODE _ 1097 was predicted to infect multiple *Bacillus* A species with high confidence (>93%), indicating a polyvalent phage capable of targeting closely related hosts. In contrast, cross-phylum predictions for virus BGR _ 130305 _ NODE _ 1187, which link Firmicutes and Thermotogota, should be interpreted with caution. Genuine cross- phylum infections are extremely rare, and these predictions lack supporting CRISPR or k- mer signatures, relying solely on BLAST homology.

Conclusion

We used an integrated multi-omics approach to reconstruct microbial and viral community structures at the genomic level, yielding high-quality genome assemblies, transcriptomic profiles, and insights into viral populations from a complex anaerobic environment. The recovery of 13 high-quality bacterial metagenome-assembled genomes (MAGs), a nearly complete methanogenic archaeal genome, and a complete circular chromosome of *Bacteroides muris* assembled through hybrid sequencing highlights the importance of genome-resolved metagenomics combined with strict quality validation.

The integrated results reveal a functional cascade linking fermentative Bacillota and methanogenic archaea, mediated by viral community activity. Pangenomic analysis of Vibrio jasicida strains highlights the evolutionary potential within this species, identifying a conserved core genome along with strain-specific accessory elements and clonal sub-lineages that may facilitate niche adaptation. Transcriptomic profiling of Methanosarcina mazei shows consistent differential expression patterns between mutant and wild-type conditions, reflecting metabolic adjustments within the methanogenic population. Viromic analysis indicates that viral communities influence this system, with Caudoviricetes bacteriophages established as the dominant viral lineage and auxiliary metabolic genes involved in sulfur and nucleotide metabolism. The confirmed virus-host linkage suggests that viral manipulation of host metabolism could impact anaerobic digestion and methane cycling, implying that viral auxiliary metabolic genes (AMGs) might serve as an underrecognized control point in methanogenic environments.

Combining community composition, individual genome structure, gene expression dynamics, and viral ecology across biological scales within a reproducible framework enables a systems-level understanding that cannot be achieved with any single method. Limitations of this approach include incomplete recovery of some archaeal genomes, a high percentage of low-quality viral sequences, and uncertainties in host prediction for divergent viral lineages. Despite these challenges, the findings enhance understanding of microbial interactions in anaerobic environments and show that genome-resolved multi-omics is crucial for unraveling the ecological complexity of natural microbial communities. Future research should focus on improving host predictions for uncultivated viruses and experimentally confirming the functional roles of viral auxiliary metabolic genes in biogeochemical cycles.

Data Availability

Data for this work can be found at: khanalsadeeksha-hub/Biol_217_omics

References

1. Zhao Z, Gänzle MG. Sequence based characterization of microbial communities in food: The panacea for smart detection of food microbes or dirty deeds done dirt cheap? Trends in Food Science & Technology. 2025;162: 105113. doi:10.1016/j.tifs.2025.105113

2. Bretaudeau A, Corguillé GL. FAIR principles applied to genome assembly and annotation. 19 Jan 2024 [cited 21 Feb 2026]. Available: https://ifb-elixirfr.gitlab.io/training/fair-gaa/module-4-pipeline-et-conteneurisation/#/title-slide

3. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. PLOS Computational Biology. 2013;9: e1003285. doi:10.1371/journal.pcbi.1003285

4. Practical Power: Reproducibility, Automation, and Layering with Conda | conda.org. 11 Nov 2025 [cited 21 Feb 2026]. Available: https://conda.org/blog/conda-practical-power/

5. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Informatics Exp. 2012;2: 3. doi:10.1186/2042-5783-2-3

6. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. 1 Mar 2023 [cited 21 Feb 2026]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

7. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph | Bioinformatics | Oxford Academic. [cited 21 Feb 2026]. Available: https://academic.oup.com/bioinformatics/article/31/10/1674/177884

8. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities [PeerJ]. [cited 21 Feb 2026]. Available: https://peerj.com/articles/1165/

9. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25: 1043–1055. doi:10.1101/gr.186072.114

10. Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology | Applied and Environmental Microbiology. [cited 21 Feb 2026]. Available: https://journals.asm.org/doi/10.1128/aem.00626-21

11. High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats | mBio. [cited 21 Feb 2026]. Available: https://journals.asm.org/doi/10.1128/mbio.02288-14

12. Osmundson J, Dewell S, Darst SA. RNA-Seq Reveals Differential Gene Expression in Staphylococcus aureus with Single-Nucleotide Resolution. Vertessy BG, editor. PLoS ONE. 2013;8: e76572. doi:10.1371/journal.pone.0076572

13. Vernikos GS. A Review of Pangenome Tools and Recent Studies. In: Tettelin H, Medini D, editors. The Pangenome: Diversity, Dynamics and Evolution of Genomes. Cham: Springer International Publishing; 2020. pp. 89–112. doi:10.1007/978-3-030-38281-0_4

14. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17: 13. doi:10.1186/s13059-016-0881-8

15. Coclet C, Camargo AP, Roux S. MVP: a modular viromics pipeline to identify, filter, cluster, annotate, and bin viruses from metagenomes. mSystems. 2024;9: e00888-24. doi:10.1128/msystems.00888-24

16. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39: 578–585. doi:10.1038/s41587-020-00774-7

17. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol. 2021;6: 960–970. doi:10.1038/s41564-021-00928-6

18. Tian F, Wainaina JM, Howard-Varona C, Domínguez-Huerta G, Bolduc B, Gazitúa MC, et al. Prokaryotic-virus-encoded auxiliary metabolic genes throughout the global oceans. Microbiome. 2024;12: 159. doi:10.1186/s40168-024-01876-z

19. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34: i884–i890. doi:10.1093/bioinformatics/bty560

20. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol. 2021;6: 3–6. doi:10.1038/s41564-020-00834-3

21. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

22. Twelve years of SAMtools and BCFtools | GigaScience | Oxford Academic. [cited 21 Feb 2026]. Available: https://academic.oup.com/gigascience/article/10/2/giab008/6137722

23. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7: e7359. doi:10.7717/peerj.7359

24. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32: 605–607. doi:10.1093/bioinformatics/btv638

25. Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biol. 2021;22: 178. doi:10.1186/s13059-021-02393-0

26. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics. 2023;39: btad311. doi:10.1093/bioinformatics/btad311

27. rrwick/Filtlong: quality filtering tool for long reads. [cited 21 Feb 2026]. Available: https://github.com/rrwick/Filtlong

28. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads | PLOS Computational Biology. [cited 21 Feb 2026]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005595

29. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29: 1072–1075. doi:10.1093/bioinformatics/btt086

30. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022;38: 5315–5316. doi:10.1093/bioinformatics/btac672

31. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069. doi:10.1093/bioinformatics/btu153

32. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31: 3350–3352. doi:10.1093/bioinformatics/btv383

33. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048. doi:10.1093/bioinformatics/btw354

34. Förstner KU, Vogel J, Sharma CM. READemption—a tool for the computational analysis of deep-sequencing–based transcriptome data. Bioinformatics. 2014;30: 3421–3423. doi:10.1093/bioinformatics/btu533

35. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

36. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol. 2024;42: 1303–1312. doi:10.1038/s41587-023-01953-y

37. Minimap2: pairwise alignment for nucleotide sequences | Bioinformatics | Oxford Academic. [cited 21 Feb 2026]. Available: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778

38. Aroney STN, Newell RJP, Nissen J, Camargo AP, Tyson GW, Woodcroft BJ. CoverM: Read alignment statistics for metagenomics. Bioinformatics. 2025. p. btaf147. doi:10.1093/bioinformatics/btaf147

39. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119. doi:10.1186/1471-2105-11-119

40. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform. 2021;3: lqab067. doi:10.1093/nargab/lqab067

41. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35: 1026–1028. doi:10.1038/nbt.3988

42. vRhyme enables binning of viral genomes from metagenomes | Nucleic Acids Research | Oxford Academic. [cited 21 Feb 2026]. Available: https://academic.oup.com/nar/article/50/14/e83/6584432

43. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. PLOS Biology. 2023;21: e3002083. doi:10.1371/journal.pbio.3002083

44. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32: 605–607. doi:10.1093/bioinformatics/btv638

45. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35: 725–731. doi:10.1038/nbt.3893

46. Full article: Microbiological insights into anaerobic digestion for biogas, hydrogen or volatile fatty acids (VFAs): a review. [cited 21 Feb 2026]. Available: https://www.tandfonline.com/doi/full/10.1080/21655979.2022.2035986

47. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9: 5114. doi:10.1038/s41467-018-07641-9

48. Boutard M, Cerisy T, Nogue P-Y, Alberti A, Weissenbach J, Salanoubat M, et al. Functional Diversity of Carbohydrate-Active Enzymes Enabling a Bacterium to Ferment Plant Biomass. PLOS Genetics. 2014;10: e1004773. doi:10.1371/journal.pgen.1004773

49. Anvi’o: an advanced analysis and visualization platform for ‘omics data [PeerJ]. [cited 21 Feb 2026]. Available: https://peerj.com/articles/1319/

50. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome” | PNAS. [cited 21 Feb 2026]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0506758102

51. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Current Opinion in Microbiology. 2015;23: 148–154. doi:10.1016/j.mib.2014.11.016

52. PHROG: families of prokaryotic virus proteins clustered using remote homology | NAR Genomics and Bioinformatics | Oxford Academic. [cited 21 Feb 2026]. Available: https://academic.oup.com/nargab/article/3/3/lqab067/6342220

53. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria | PLOS Biology. [cited 21 Feb 2026]. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3002083