Bioinformatics Advances, 2026, **6**(1), vbag150
https://doi.org/10.1093/bioadv/vbag150
Published: 3 June 2026
Application Note | Genetics and population analysis

OXFORD

PopMAG: a Nextflow pipeline for population genetics analysis based on metagenome-assembled genomes

Daniel Sabogal-Rodriguez[1,2,] and Alejandro Caro-Quintero[2,3,*,]

[1]Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, 111321 Bogotá, Colombia
[2]Max Planck Tandem Group in Holobionts, Universidad Nacional de Colombia, 111321 Bogotá Colombia
[3]Departamento de Biología, Facultad de Ciencias, Universidad Nacional de Colombia, 111321 Bogotá, Colombia
*Corresponding author. Departamento de Biología, Facultad de Ciencias, Universidad Nacional de Colombia, 111321 Bogotá, Colombia. E-mail: acaroq@unal.edu.co
Associate Editor: Michael DeGiorgio

Abstract

Motivation: **Metagenome-assembled genomes (MAGs) are routinely recovered from metagenomic studies, yet the population genetic information embedded within these datasets remains largely underutilized. Analyzing within-species genetic variation can reveal adaptive evolution, selection pressures, and ecological dynamics that are hidden when MAGs are treated as homogeneous entities. Existing tools address individual analysis steps in isolation, requiring manual integration and creating barriers for researchers without extensive bioinformatics expertise.**

Results: **Here we present PopMAG, a Nextflow pipeline and interactive Shiny application that automates population genetics analysis of MAGs. PopMAG integrates quality control, community profiling, competitive read mapping, functional annotation, and microdiversity estimation into a single reproducible workflow. The pipeline calculates key population genetics metrics including nucleotide diversity (π), pN/pS ratios, fixation index (F_{ST}), Levins' index and SNVs counts with results consolidated into an interactive visualization platform for metadata-driven exploration. We demonstrate PopMAG's utility through analysis of longitudinal cystic fibrosis lung metagenomes, where we identify patterns consistent with antibiotic-driven selection in *Pseudomonas aeruginosa* efflux pump genes coinciding with treatment intervention.**

Availability and implementation: **PopMAG and corresponding documentation are publicly available at https://github.com/daasabogalro/PopMAG.**

1 Introduction

Metagenomic studies routinely recover metagenome-assembled genomes (MAGs) to characterize the taxonomic and functional composition of microbial communities (Sun *et al.* 2025). However, the population genetic information embedded within metagenomic datasets remains largely unexplored and with increasing interest of being characterized (Zhao *et al.* 2023). A MAG represents a consensus sequence derived from multiple cells of closely related organisms, potentially distinct subpopulations or strains coexisting within a sample (Van Rossum *et al.* 2020). This population structure possesses critical information about microbial adaptation, selection pressures, and ecological dynamics that is hidden when MAGs are treated as single, homogeneous entities.

Analysis of within-species genetic variation (microdiversity) in metagenomic data can reveal adaptive evolution in response to environmental change, host-microbe co-evolution, antibiotic resistance emergence, and niche differentiation among co-occurring strains (Ghazi *et al.* 2022). Population genetics metrics such as nucleotide diversity (π), fixation indices (F_{ST}), and pN/pS ratios provide quantitative frameworks for detecting selection, inferring demographic history, and identifying genes under adaptive pressure. Despite the biological and clinical relevance of these insights, few metagenomic studies exploit the full population genetic potential of their data, often limiting analyses to MAG presence/absence or gene content comparisons.

Several bioinformatics tools enable population genetics analyses of metagenomic data, including inStrain (Olm *et al.* 2021), MetaPop (Gregory *et al.* 2022), POGENOM (Sjoqvist *et al.* 2021), Anvi'o (Eren *et al.* 2015), MIDAS (Nayfach *et al.* 2016), and metaSNV (Costea *et al.* 2017). However, these tools typically address individual steps in isolation, such as SNV calling, diversity estimation, or visualization, requiring users to manually integrate outputs, format data across platforms, and navigate complex installation dependencies. No existing solution provides a reproducible, end-to-end pipeline that combines MAG quality control, competitive read mapping, comprehensive population genetics analyses, and interactive visualization in a single automated workflow. This fragmentation creates barriers for

Received: 22 December 2025. **Revised:** 11 May 2026. **Accepted:** 14 May 2026

researchers without extensive bioinformatics expertise and hinders reproducibility across studies.

Here, we present PopMAG, an integrated scalable pipeline and Shiny application for population genetics analyses of MAGs that addresses these limitations. Our tool automates the complete workflow from quality-controlled MAGs and metagenomic reads to interactive visualizations, implementing competitive mapping strategies to ensure accurate strain-level resolution and providing interactive, metadata-driven exploration of evolutionary patterns. We demonstrate the pipeline's utility through a case study of antibiotic-driven selection in *Pseudomonas aeruginosa* populations from a cystic fibrosis patient, suggestive of adaptive evolution in a multidrug efflux pump gene coinciding with treatment intervention.

2 Implementation

We developed a Nextflow (Di Tommaso *et al*. 2017) pipeline for population genetics analysis of metagenome assembled-genomes (MAGs) that integrates quality control, competitive mapping, functional annotation, and microdiversity estimation into a single automated workflow. PopMAG is designed as a downstream complement to MAG reconstruction pipelines such as nf-core/mag (Krakau *et al*. 2022). The pipeline accepts three primary inputs: (i) a samplesheet of assembled MAGs, (ii) a samplesheet of metagenomic reads, and (iii) a metadata file for downstream visualization. PopMAG currently supports paired-end short-read sequencing data.

The workflow consists of five major stages (Fig. 1). First, MAG quality control assesses genome completeness and contamination using CheckM2 (Chklovski *et al*. 2023), filters bins based on user-defined quality thresholds, and performs dereplication with dRep (Olm *et al*. 2017) to generate a non-redundant genome database. Second, community profiling uses SingleM (Woodcroft *et al*. 2025) to characterize the taxonomic composition and provide context for population-level analyses. Third, competitive mapping creates a dereplicated genome database, generates contig-to-bin files, performs competitive mapping with Bowtie2 (Langmead and Salzberg 2012), and calculates per MAG coverage using CoverM (Aroney *et al*. 2025). Fourth, annotation and microdiversity analysis uses Prodigal (Hyatt *et al*. 2010) for gene prediction, MetaCerberus (Figueroa III *et al*. 2024) for functional annotation, inStrain (Olm *et al*. 2021) for within-species diversity profiling (including nucleotide diversity π and *pN*/*pS* ratios), and POGENOM (Sjoqvist *et al*. 2021) for between-population differentiation (F_{ST}). Finally, visualization consolidates inStrain profiles, POGENOM results, annotation summaries, and coverage statistics into formatted outputs that feed directly into an interactive Shiny application.

The pipeline uses containerization (Docker/Singularity) (Merkel 2014, Kurtzer *et al*. 2017) for reproducibility and implements parallel processing of samples to ensure scalability across datasets of varying sizes. Unlike existing tools that address individual steps in isolation, our end-to-end implementation eliminates the need for manual data formatting and transfer between analysis stages. The integrated Shiny application enables interactive exploration of results through metadata-driven visualizations, allowing researchers to dynamically filter, compare, and export population genetics metrics across experimental conditions or sample groups. All intermediate and final outputs are organized in a standardized directory structure for easy interpretation and downstream analysis.

2.1 Technical architecture

The pipeline is implemented in Nextflow (version $\geq$ 24.10.0) using DSL2 syntax for modular process design. All bioinformatics tools are packaged in individual containers available via Quay.io, with support for both Docker and Singularity execution. Alternative dependency management through Conda/Mamba environments is available for users preferring non-containerized execution.

The pipeline implements a six-level resource labelling system ranging from lightweight single-CPU processes (6 GB RAM) to high-memory operations (200 GB RAM), with automatic error

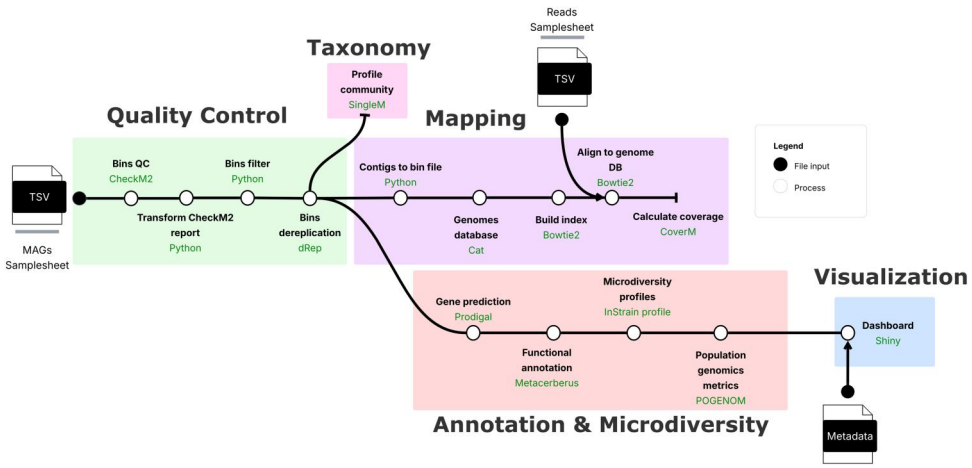

Figure 1 Workflow overview. Schematic representation of PopMAG. The workflow consists of five major stages: (i) MAG quality control, (ii) community profiling, (iii) competitive read mapping, (iv) functional annotation and microdiversity analyses, and (v) data integration and visualization through an interactive Shiny application. Each one of the five steps is represented by a different colour. The required input files, including the metadata file, are represented as filled circles, while the different processes are represented as empty circles.

recovery through dynamic resource scaling, failed processes retry with doubled CPU and memory allocations. Pre-configured execution profiles support local workstations, SLURM-managed HPC clusters, and AWS cloud infrastructure. Users customize pipeline behaviour through a documented configuration file or command-line parameters.

Nextflow automatically parallelizes processes across samples and within-sample per-MAG analyses (annotation, diversity profiling) based on available resources and workflow dependencies, enabling efficient scaling from small studies to large metagenomic surveys.

2.2 Core functionalities

2.2.1 MAG quality control and dereplication

The pipeline filters MAGs using stringent quality thresholds ($\geq$90% completeness, $\leq$5% contamination) as assessed by CheckM2, ensuring high-confidence genomes for population-level inference. These thresholds can be configured by the user to accommodate different study requirements. dRep performs genome dereplication using default parameters to generate a non-redundant genome set, eliminating near-identical genomes that would confound competitive mapping.

2.2.2 Competitive mapping and coverage estimation

To minimize read mis-mapping in complex metagenomic samples, the pipeline implements competitive read alignment where reads map simultaneously against all dereplicated genomes. This approach ensures that each genome in the database is sufficiently distinct to avoid mapping confusion and significantly reduces false-positive microdiversity signals. Bowtie2 performs the competitive alignment, and CoverM calculates genome coverage using trimmed mean (10–90th percentile) with a minimum covered fraction threshold of 0.5, providing robust coverage estimates that are resilient to coverage outliers.

2.2.3 Microdiversity and population genetics analyses

The pipeline integrates complementary tools for comprehensive evolutionary inference. inStrain profiles within-population diversity, calculating nucleotide diversity (π) as the average pairwise nucleotide difference per site across mapped reads (Nei and Li 1979) computed at positions meeting minimum coverage thresholds (5x by default), pN/pS ratios, SNV counts, allele frequencies, and synonymous/nonsynonymous mutations per gene. POGENOM performs between-population analyses, computing π within each sample and pairwise F_{ST} values across all sample pairs when multiple populations are available. When genomic annotations are provided, POGENOM calculates gene-wise π and F_{ST} at both nucleotide and amino acid levels, enabling detection of selection at the gene level. The pipeline also computes Levins' niche breadth index using the microniche R package, quantifying the ecological specialization of each population across samples.

It is important to note that pN/pS estimated from within-sample metagenomic data reflects polymorphism pooled across potentially mixed time scales: it captures both recent mutations accumulating within a strain lineage and older fixed differences between co-occurring strains. These two sources of variation are not distinguishable from metagenomic read data alone.

2.2.4 Functional annotation

Prodigal predicts protein-coding genes, which are functionally annotated by MetaCerberus against comprehensive databases including KEGG (KO), COGs, CAZy, FOAM, VOGs, and PHROGs. Annotation results are integrated with microdiversity metrics, enabling users to link evolutionary patterns (pN/pS ratios, gene-wise F_{ST}) to functional categories and metabolic pathways.

2.2.5 Data integration

All outputs, inStrain profiles, POGENOM statistics, functional annotations, and coverage data are automatically merged into standardized tab-separated value (TSV) tables formatted for direct import into the Shiny app visualization, eliminating manual data wrangling steps.

2.3 Shiny application

The pipeline automatically launches an interactive Shiny application for exploratory data analysis and visualization of population genetics results. The application is configured directly from the pipeline, with customizable port settings and session duration to accommodate different deployment scenarios (local workstations, shared servers, or cloud instances).

The Shiny App features a sidebar for MAG selection and dynamic plot customization, allowing users to adjust visual parameters (colours, point sizes, axis scales) in real-time. A top navigation bar organizes visualizations into two main modules: metadata-driven analyses and MAG-specific population genetics metrics. The metadata section enables cross-sample comparisons through a correlation heatmap visualizing relationships between environmental or experimental variables, and principal component analysis (PCA) for dimensionality reduction and sample clustering. Users can interactively filter samples based on metadata criteria to explore condition-specific patterns. For each selected genome, the application provides five complementary visualizations: (i) Levins' niche breadth index indicating ecological specialization across samples, (ii) nucleotide diversity (π) barplots comparing within-population variation, (iii) scatter plots relating nucleotide diversity (π) to individual metadata variables with metadata-based filtering, (iv) SNVs and SNSs count distributions across samples, and (v) histograms of gene-wise pN/pS ratios with an interactive, filterable table displaying gene annotations and corresponding selection metrics (Supplementary Data). This integrated view enables users to identify genes under selection and their functional roles.

Each visualization is accompanied by its source data table, providing transparency and enabling users to inspect underlying values. Users can download plots and use filtered data tables for custom downstream analyses or reporting. The integration between pipeline outputs and visualization ensures that results are immediately accessible through an intuitive interface, lowering barriers for researchers without extensive bioinformatics expertise.

3 Case study

To demonstrate the pipeline's utility for detecting adaptive evolution in clinical metagenomes, we analysed an open access longitudinal dataset of sputum samples from a cystic fibrosis (CF) patient undergoing antibiotic therapy. The dataset consisted of

10 metagenomic samples from patient CFR11 (Table 1, available as supplementary data at *Bioinformatics Advances* online) collected over 19 months, with sequencing depths ranging from 11.9 to 43.6 million human-filtered reads per sample. The patient received continuous treatment with azithromycin, colistin, and insulin throughout the study period, with the CFTR modulator lumacaftor/ivacaftor added to the regimen at month 13 (Dmitrijeva *et al*. 2021).

The complete workflow (PopMAG v0.1.0-alpha) processed all 10 samples in 8.5 h on a local workstation with 12 CPUs and 128 Gb of RAM, with inStrain profiling accounting for approximately 5.6 h of runtime. BAM files ranged from 300 MB to 3.3 GB in size. Quality control identified six high-quality MAGs ($\geq$90% completeness, $\leq$5% contamination) representing key CF associated taxa: *Pseudomonas aeruginosa*, *Achromobacter* sp., *Fusobacterium* sp., *Parvimonas* sp., *Prevotella* sp., and *Streptococcus* sp. All six MAGs passed quality filters and were included in downstream analyses.

Using the Shiny application's filterable and interactive gene-level *pN/pS* table, we identified a multidrug efflux pump subunit (MexA, membrane-fusion protein) in the *P.aeruginosa* MAG (mean coverage 155X) exhibiting a selection signature specifically in the post lumacaftor/ivacaftor treatment sample (pN/pS = 2.1). In the nine pre-treatment samples spanning months 0–13, this gene showed no evidence consistent with positive selection. The elevated pN/pS ratio is consistent with positive selection for nonsynonymous mutations in MexA coinciding with lumacaftor/ivacaftor administration, suggestive of adaptive resistance evolution. Notably, this gene was not previously reported among mutations associated with CFTR modulator treatment(Ledger *et al*. 2024) which identified mutations in genes such as *mutS*, *mutL*, *algG*, and *mucA*.

Additional genes in *P.aeruginosa* displayed signatures consistent with positive selection at various timepoints, including an AlgP family protein (pN/pS = 4.3), Cbb3-type cytochrome oxidase subunit 1 (2.5), fibronectin-binding autotransporter adhesin (2.2), and nitrate/nitrite transporter NarK (2.0), revealing temporally dynamic evolutionary responses throughout the treatment course.

We note that per-gene pN/pS estimates are inherently noisy, particularly for short genes or those with few segregating sites, and individual gene-level values should be interpreted cautiously. We recommend using the genome-wide pN/pS distribution as empirical background context when evaluating candidate genes.

Integration of clinical metadata revealed an inverse relationship between *P.aeruginosa* nucleotide diversity (π) and C-reactive protein (CRP) concentration, with microdiversity decreasing as inflammation markers increased. This pattern, readily visualized through the Shiny app's metadata-linked diversity plots (Fig. 2), suggests that host immune responses or treatment-associated pressures constrain within-population genetic variation during inflammatory episodes.

This case study demonstrates the pipeline's capability to potentially identify relevant adaptive evolution in pathogen populations, link selection signatures to treatment interventions through metadata integration, and show biologically interpretable patterns through intuitive visualizations, all within a single automated workflow.

4 Conclusions

We present an end-to-end Nextflow pipeline and interactive Shiny application for population genetics analyses of metagenome-assembled genomes, addressing the growing

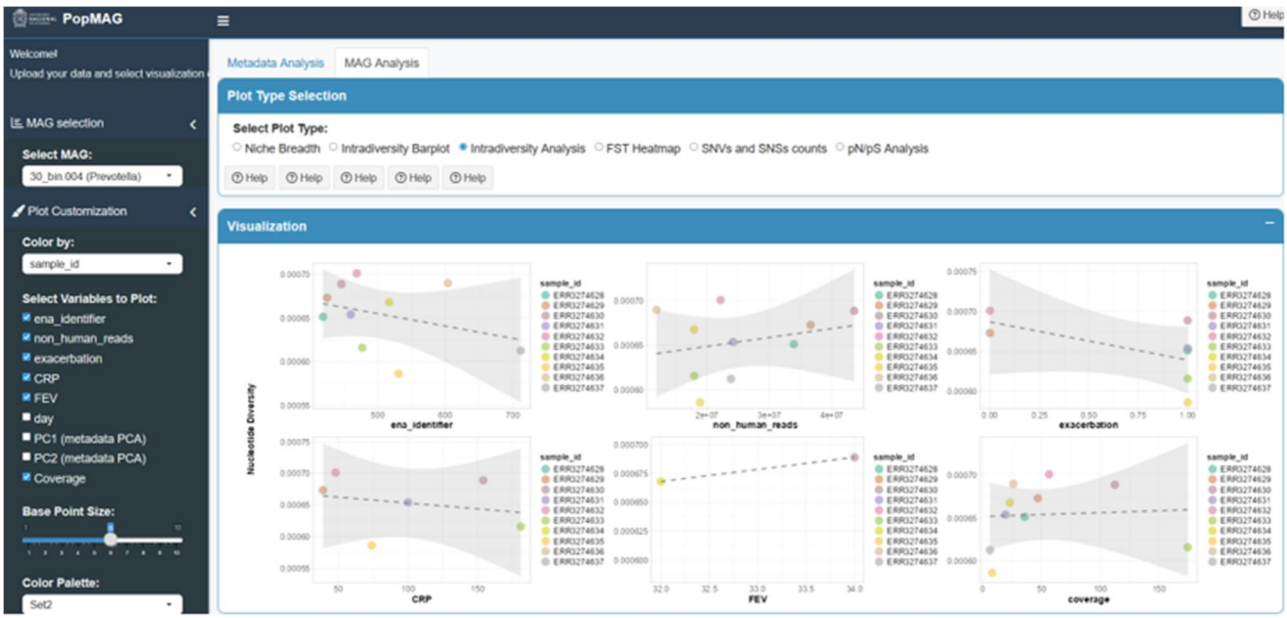

Figure 2 Interactive Shiny application interface for population genetics exploration. Structure of the shiny app generated after the PopMAG workflow execution. The MAG to be analysed can be selected using the sidebar on the left and the different plot types can be selected in the navigation bar at the top. The sidebar also features a 'Plot Customization' section to configure the visualizations. In this example we have selected the *P.aeruginosa* MAG and we're visualizing the 'Intradiversity Analysis' plots, where we can see the relationship between different metadata values (exacerbation, CRP and day, as selected in the sidebar) and the nucleotide diversity (π).

need for integrated tools to study microbial evolution in complex communities. By combining MAG quality control, competitive read mapping, microdiversity profiling, and functional annotation in a single automated workflow, our pipeline eliminates manual data formatting and streamlines the path from raw metagenomic data to interpretable evolutionary insights.

The pipeline's modular architecture and containerized implementation ensure reproducibility and portability across computing environments. The integrated Shiny application democratizes access to population genetics metrics from MAGs through metadata-driven visualizations, allowing researchers without extensive bioinformatics expertise to explore adaptive evolution, ecological specialization, and selection signatures in their metagenomic data.

Our case study of *Pseudomonas aeruginosa* during cystic fibrosis treatment illustrates the pipeline's utility for identifying candidate genes under selection and link selection signatures to treatment interventions. The identification of antibiotic-driven selection in a multidrug efflux pump gene highlights how this approach can reveal novel resistance mechanisms that complement traditional genomic surveillance efforts.

The pipeline and Shiny application are freely available at https://github.com/daasabogalro/popmag, along with documentation, to facilitate adoption by the microbiome research community. This tool can contribute to accelerate research into microbial adaptation and the ecological and evolutionary dynamics shaping microbiome function.

Author contributions

Daniel Sabogal-Rodriguez (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Methodology [equal], Software [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]) and Alejandro Caro-Quintero (Conceptualization [equal], Formal analysis [equal], Funding acquisition [lead], Methodology [equal], Resources [equal], Supervision [equal], Writing—review & editing [equal])

Supplementary material

Supplementary material is available at *Bioinformatics Advances* online.

Conflicts of interest

None declared.

Funding

This work was supported by the Universidad Nacional de Colombia and MINCIENCIAS through a grant awarded under the Max Planck Tandem Group initiative.

Data availability

The data used in this article was produced and published by (Dmitrijeva *et al*. 2021) and are available in the European Nucleotide Archive (PRJEB32062), at https://www.ebi.ac.uk/ena/browser/view/PRJEB32062.

References

Aroney STN, Newell RJP, Nissen JN *et al.* CoverM: read alignment statistics for metagenomics. *Bioinformatics* 2025;**41**:btaf147. https://doi.org/10.1093/bioinformatics/btaf147

Chklovski A, Parks DH, Woodcroft BJ *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;**20**:1203–12. https://www.nature.com/articles/s41592-023-01940-w

Costea PI, Munch R, Coelho LP *et al.* metaSNV: A tool for metagenomic strain level analysis. *PLoS One*, 2017;**12**:e0182392. https://doi.org/10.1371/journal.pone.0182392

Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;**35**:316–9. https://doi.org/10.1038/nbt.3820

Dmitrijeva M, Kahlert CR, Feigelman R *et al.* Strain-Resolved dynamics of the lung microbiome in patients with Cystic Fibrosis. *mBio* 2021;**12**. https://doi.org/10.1128/mBio.02863-20

Eren AM, Esen OC, Quince C *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 2015;**3**:e1319. https://doi.org/10.7717/peerj.1319

Figueroa JL III, Dhungel E, Bellanger M *et al.* MetaCerberus: distributed highly parallelized HMM-based processing for robust functional annotation across the tree of life. *Bioinformatics* 2024;**40**:btae119. https://doi.org/10.1093/bioinformatics/btae119

Ghazi AR, Münch PC, Chen D *et al.* Strain identification and quantitative analysis in microbial communities. *J Mol Biol* 2022;**434**:167582. https://doi.org/10.1016/j.jmb.2022.167582

Gregory AC, Gerhardt K, Zhong Z-P *et al.* MetaPop: a pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations. *Microbiome* 2022;**10**:49. https://doi.org/10.1186/s40168-022-01231-0

Hyatt D, Chen G-L, LoCascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;**11**:119. https://doi.org/10.1186/1471-2105-11-119

Krakau S, Straub D, Gourlé H *et al.* nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning. *NAR Genomics Bioinf* 2022;**4**:lqac007. https://doi.org/10.1093/nargab/lqac007

Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. *PLoS One*, 2017;**12**:e0177459. https://doi.org/10.1371/journal.pone.0177459

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

Ledger EL, Smith DJ, Teh JJ *et al.* Impact of CFTR modulation on pseudomonas aeruginosa infection in people with cystic fibrosis. *J Infect Dis* 2024;**230**:e536–47. https://doi.org/10.1093/infdis/jiae051

Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J* 2014;**2014**:2.

Nayfach S, Rodriguez-Mueller B, Garud N *et al.* An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res* 2016;**26**:1612–25. https://doi.org/10.1101/gr.201863.115

Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA* 1979;**76**:5269–73. https://doi.org/10.1073/pnas.76.10.5269

Olm MR, Brown CT, Brooks B *et al.* dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 2017;**11**:2864–8. https://doi.org/10.1038/ismej.2017.126

Olm MR, Crits-Christoph A, Bouma-Gregson K *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* 2021;**39**:727–36. https://doi.org/10.1038/s41587-020-00797-0

Sjoqvist C, Delgado LF, Alneberg J *et al.* Ecologically coherent population structure of uncultivated bacterioplankton. *ISME J* 2021;**15**:3034–49. https://doi.org/10.1038/s41396-021-00985-z

Sun Y, Chen Q, Fan G, *et al.* Gcmeta. GcmetaA global repository of metagenome-assembled genomes enabling cross-ecosystem microbial discovery and function research. *Nucleic Acids Res* 2025;**54**:D724–33. https://doi.org/10.1093/nar/gkaf1115.

Van Rossum T, Ferretti P, Maistrenko OM *et al.* Diversity within species: interpreting strains in microbiomes. *Nat Rev Microbiol* 2020;**18**:491–506. https://doi.org/10.1038/s41579-020-0368-1

Woodcroft BJ, Aroney STN, Zhao R *et al.* Comprehensive taxonomic identification of microbial species in metagenomic data using SingleM and Sandpiper. *Nat Biotechnol* 2025;1–6. https://doi.org/10.1038/s41587-025-02738-1

Zhao C, Shi ZJ, Pollard KS. Pitfalls of genotyping microbial communities with rapidly growing genome collections. *Cell Syst* 2023;**14**:160–76.e3. https://doi.org/10.1016/j.cels.2022.12.007

Bioinformatics Advances, 2026, 6, 1–6
https://doi.org/10.1093/bioadv/vbag150
Application Note