microPublication
BIOLOGY

2/19/2025 - *Open Access*

# BALROG-MON: a high-throughput pipeline for Bacterial AntimicrobiaL Resistance annOtation of Genomes-Metagenomic Oxford Nanopore

Edward Bird[1], Victoria Pickens[1], David Molik[2], Kristopher Silver[1], Dana Nayduch[2§]

[1]Entomology, Kansas State University, Manhattan, Kansas, United States

[2]Arthropod-Borne Animal Diseases Research Unit, Agricultural Research Service, United States Department of Agriculture, Manhattan, KS, United States

[§]To whom correspondence should be addressed: dana.nayduch@usda.gov

## Abstract

BALROG-MON is a Nextflow pipeline for automated analysis of metagenomic long-read data to detect pathogens, annotate antimicrobial resistance genes (ARGs), link ARGs to specific pathogens, predict ARG origin (e.g., plasmid, chromosomal) and optionally perform steps like community analysis. With both assembly-based and assembly-free workflows, BALROG-MON is applicable to a wide range of sample types with low or high coverage, varying complexities and origins. Optional genome binning provides a comprehensive overview of ARGs within the dataset. BALROG-MON additionally presents results in summarized reports, overall serving as a flexible analysis tool for exploring diverse metagenomic samples for pathogens and antibiotic resistance.

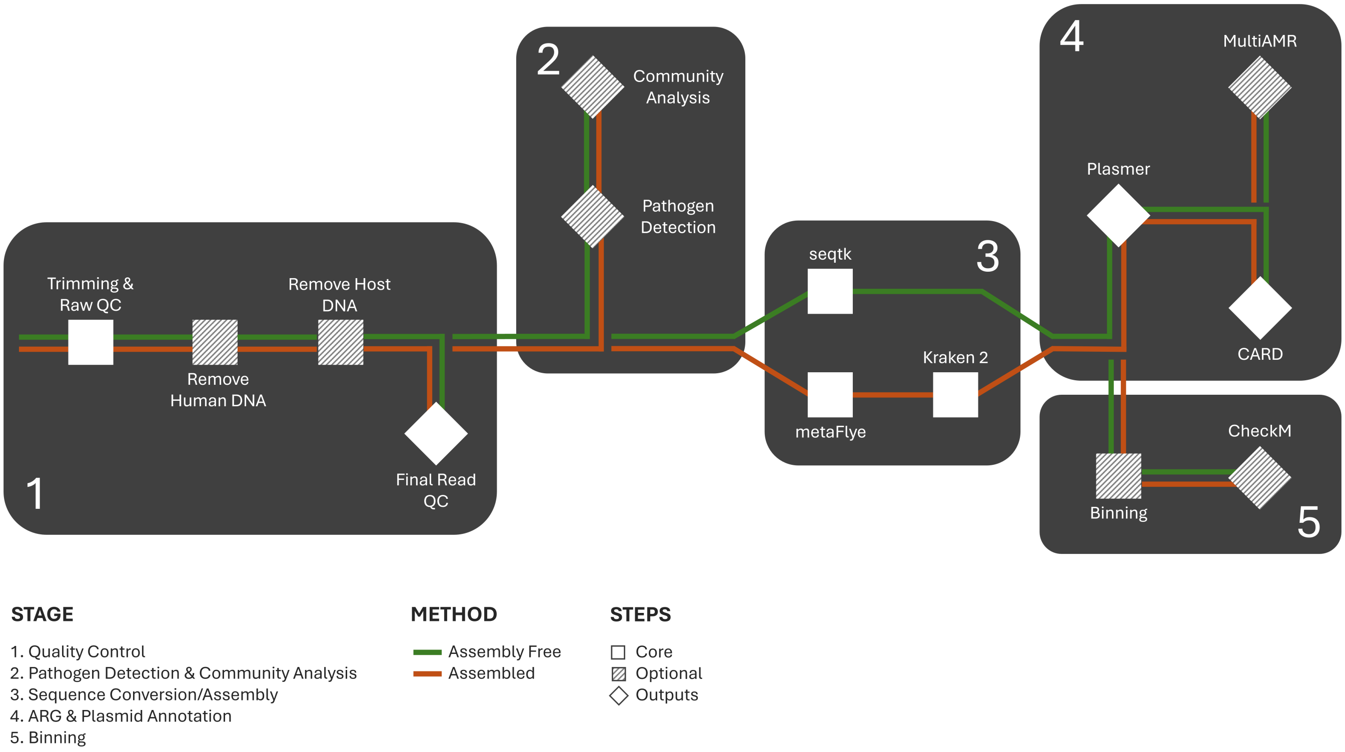

**Figure 1. BALROG-MON workflow diagram. :**

BALROG-MON enables the option of using an "assembled" or "assembly-free" method for the annotation of ARGs from metagenomic long read sequences.

## Description

BALROG-MON (Bacterial AntimicrobiaL Resistance annOtation of Genomes – Metagenomic Oxford Nanopore) (https://github.com/edwardbirdlab/BALROG-MON/) automates the analysis of antimicrobial resistance genes (ARGs) from complex metagenomes. Antimicrobial resistance (AMR) is a significant global health challenge exacerbated by the increasing prevalence of resistance to and the declining discovery of new antibiotics. Over the past decade, multiple bioinformatic tools and databases have been developed to identify, characterize and understand the evolution of ARGs. While many tools focus on isolated bacterial genomes, BALROG-MON analyzes antimicrobial resistance metagenomically, offering a high-throughput approach to study ARGs across entire environments. It also provides insights into AMR in pathogens without the need for culture-based methods, making it a powerful tool for understanding resistance in complex microbial communities and what threats may be present.

Metagenomic sequencing is a recently feasible method that can generate data from which ARGs, microbiomes, and pathogens in the sample can be characterized simultaneously, serving as a more effective and comprehensive method than isolating and culturing bacteria. Typical analysis of metagenomic data involves either an assembly-based approach or a read-based approach, with each having its own benefits and limitations. Metagenomic assembly allows for upstream or downstream investigation of ARGs and provides accurate identification of their origin. However, this approach may lead to information loss, as low-coverage genomes are often not assembled. In contrast, read-based approaches enable mapping of all available data but lack the capability to explore surrounding genomic context or provide accurate taxonomic classifications. To address these challenges, we developed BALROG-MON, a versatile and reproducible Nextflow pipeline for surveying pathogens and ARGs from metagenomic long-read sequencing, offering both “assembled” and “assembly-free” workflow options.

BALROG-MON v1.0 (DOI: 10.5281/zenodo.14850876) consists of six major steps: (1) Quality Control, (2) Pathogen Detection and Community Analysis, (3) Sequence Conversion/Assembly, (4) ARG and Plasmid Annotation, (5) Binning (*optional*) and (6) Output Collection and Summary. Written in Nextflow, BALROG-MON compartmentalizes each process, allowing users to easily modify or customize steps to suit their analysis needs. All processes are executed with a single command, simplifying the workflow while ensuring scalability. Docker containers manage all dependencies, ensuring reproducibility across different computing environments. During data quality control host sequences and human sequences can be removed, and low-quality bases and adapters are trimmed. Quality-controlled sequences are then either assembled in assembly mode or converted to FASTA format in assembly-free mode. For high-coverage metagenomes, binning can be enabled, allowing reads or contigs to be grouped for generating metagenomically reconstructed genomes. The resulting sequences are classified as either plasmid or chromosomal in origin. ARGs are annotated using multiple tools, and the outputs are standardized. ARGs, taxonomic classifications, and plasmid classification results are integrated to generate a comprehensive report for each sample. Additionally, all quality control metrics (trimming, host and human sequence removal, final read QC) are summarized in a final report.

Input data and, if also provided, reference genome(s) are first run through data_validator v1.0 (https://github.com/edwardbirdlab/nextflow_input_std), a custom Python script that checks input data for a valid format and, if necessary, reformats FASTQ and FASTA files to the expected format. FastQC v0.12.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is run on raw reads to provide statistics on the quality of sequencing, and Porechop v0.2.4 (https://github.com/rrwick/Porechop) and chopper v0.7.0 (https://github.com/wdecoster/chopper) quality control reads. Optionally, reads can be mapped to the human genome (GCA_000001405.15_GRCh38) using minimap2 v2.26 (https://github.com/lh3/minimap2), and then non-human reads extracted using Samtools v1.17 (https://www.htslib.org/). Reads can also be mapped to one or more provided reference genomes, and host sequences removed as described above. Final quality-controlled host- and/or human-depleted reads are run through FastqQC v0.12.1 for final quality metrics.

Quality-controlled reads are classified utilizing Kraken 2 v2.1.3 (https://github.com/DerrickWood/kraken2) and Kraken’s pre-built PlusPFP (https://benlangmead.github.io/aws-indexes/k2) database by default. Optionally, species-level composition of the metagenome can be estimated using Bracken v2.9 (https://github.com/jenniferlu717/Bracken) and the same database used in Kraken 2 sequence identification. KrakenTools v1.2 (https://github.com/jenniferlu717/KrakenTools) then calculates Shannon’s alpha diversity and Bray-Curtis dissimilarity (Beta Diversity), as well as a Krona chart and MPA style report.

The sequence processing module workflow differs between BALROG-MON run modes (e.g., “assembly” or “assembly-free”). In assembly-free mode, quality-controlled reads are simply converted from FASTQ to FASTA format with SeqTK v1.4 (https://github.com/lh3/seqtk). In assembly mode, a draft metagenome will be assembled with metaFlye v2.9.3 (https://github.com/mikolmogorov/Flye) and contigs re-identified with Kraken 2 and the PlusPFP database. Finally, in both assembly or assembly-free mode, QUAST v5.2.0 (https://github.com/ablab/quast) is run to gather statistics (N50, Total Length, etc.) on the output sequences.

Sequences are first predicted to be of plasmid or chromosomal origin using Plasmer v23.04.20 (https://github.com/nekokoe/Plasmer), and then renamed to reflect their origin. ARGs are then annotated using Resistance Gene Identifier (RGI) v6.0.3 (https://github.com/arpcard/rgi) and the Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al., 2023) protein homologue model. Optionally, sequences can also be annotated with AMRFinderPlus v4.0.19 (https://github.com/ncbi/amr) and Resfinder v4.4.2 (https://github.com/cadms/resfinder).

If binning is enabled, LRBinner v2.1 (https://github.com/anuradhawick/LRBinner) will bin reads in assembly-free mode, or COMEBin v1.0.3 (https://github.com/ziyewang/COMEBin) will bin contigs in assembly mode. Genome completeness and binning accuracy is then assessed with CheckM v1.2.3 (https://github.com/Ecogenomics/CheckM).

The deliverables from BALROG-MON are summaries that combine and/or visualize the pipeline results. The main output from BALROG-MON is the summary that creates a table based on the combination of Kraken2 sequence identities, Plasmer sequence classifications (Plasmid/Chromosome), and ARGs. This table allows the user to investigate each ARG and determine the putative origin or location (e.g., plasmid vs. chromosomal). Additionally, all quality control metrics, including those from raw sequence data, trimming, host/human sequence removal, and final read metrics, are summarized using MultiQC v1.22.3 (https://github.com/MultiQC/MultiQC). Lastly, if multiple ARG annotation tools are used, hAMRonization v1.0.2 (https://github.com/pha4ge/hAMRonization) standardizes outputs by harmonizing gene names, recording databases and software versions, and consolidating results into a single output. Additionally, an easy to use web-based report is generated that combines all samples for easy cross sample and cross database comparisons.

BALROG-MON is a robust and adaptable Nextflow pipeline developed for surveying pathogens and ARGs from metagenomic long-read sequencing data. By offering both "assembled" and "assembly-free" workflows, BALROG-MON overcomes the limitations of existing tools, providing researchers with a powerful solution to study ARGs in a wide range of samples, including methods to analyze host associated and low coverage metagenomes. The pipeline incorporates several key steps, including data quality control, read-based sequence identification, ARG and plasmid annotation, and optional metagenomic binning, ensuring a detailed and accessible analyses of complex metagenomic samples. Notable features include versatility in pipeline execution to handle difficult samples, the integration of multiple ARG databases, plasmid prediction capabilities, and optional binning for deeper analysis. BALROG-MON unifies these functionalities into a streamlined, user-friendly platform, simplifying the study of ARGs in metagenomic datasets. As the field of metagenomics advances, BALROG-MON stands poised to play a vital role in elucidating the dynamics of AMR and shaping strategies to address this global health challenge across human, animal, and environmental domains.

**Acknowledgements:** Special thanks to Luke Brendel and Tanya Purvis (USDA-ARS) for collecting house flies, and Dr. Christian Young (USDA-APHIS) for providing sequencing data for validating the pipeline. The authors would also like to thank CDR Dr. Theron Hamilton, Dr. Jason Blanton, and Namratha Tarigopula of the United States Navy Entomology Center of Excellence for providing additional testing data for the pipeline. Some of the computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants CNS-1006860, EPS-1006860, EPS-0919443, ACI-1440548, CHE-1726332, and NIH P20GM113109. Mention of trade names or commercial products in this report is solely to provide specific information and does not imply recommendation or endorsement by the US Department of Agriculture. The US Department of Agriculture is an equal opportunity lender, provider and employer

## References

Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al., Aarestrup. 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy 75: 3491-3500. DOI: 10.1093/jac/dkaa345

Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, Wlodarski MA, et al., McArthur. 2022. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Research 51: D690-D699. DOI: 10.1093/nar/gkac920

Andrews S. 2010. FastQC. Available from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al., Li. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10: 10.1093/gigascience/giab008. DOI: 10.1093/gigascience/giab008

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

Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al., Klimke. 2021. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Scientific Reports 11: 10.1038/s41598-021-91456-0. DOI: 10.1038/s41598-021-91456-0

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

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al., Pevzner. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nature Methods 17: 1103-1110. DOI: 10.1038/s41592-020-00971-x

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094-3100. DOI: 10.1093/bioinformatics/bty191

Li H. Seqtk. Available from https://github.com/lh3/seqtk.

Lu J. Kraken Tools. Available from https://github.com/jenniferlu717/KrakenTools.

Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science 3: e104. DOI: 10.7717/peerj-cs.104

Mendes Is, Griffiths E, Manuele A, Fornika D, Tausch SH, Le-Viet T, et al., Maguire. 2024. hAMRonization: Enhancing antimicrobial resistance prediction using the PHA4GE AMR detection specification and tooling. : 10.1101/2024.03.07.583950. DOI: 10.1101/2024.03.07.583950

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

Wang Z, You R, Han H, Liu W, Sun F, Zhu S. 2024. Effective binning of metagenomic contigs using contrastive multi-view representation learning. Nature Communications 15: 10.1038/s41467-023-44290-z. DOI: 10.1038/s41467-023-44290-z

Wick R, Volkening J. Porechop. Available from https://github.com/rrwick/Porechop.

Wickramarachchi A, Lin Y. 2022. Binning long reads in metagenomics datasets using composition and coverage information. Algorithms for Molecular Biology 17: 10.1186/s13015-022-00221-z. DOI: 10.1186/s13015-022-00221-z

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biology 20: 10.1186/s13059-019-1891-0. DOI: 10.1186/s13059-019-1891-0

Zhu Q, Gao S, Xiao B, He Z, Hu S. 2023. Plasmer: an Accurate and Sensitive Bacterial Plasmid Prediction Tool Based on Machine Learning of Shared k-mers and Genomic Features. Microbiology Spectrum 11: 10.1128/spectrum.04645-22. DOI: 10.1128/spectrum.04645-22

**Funding:** A portion of this work was carried out by the Arthropod-Borne Animal Diseases Research Unit, ARS Project numbers 3020-32000-018-00D, 3020-32000-020-00D, and 3020- 32000-019-00D. This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project numbers 0201-88888-003-00D and 0201-88888-002-00D.

**Author Contributions:** Edward Bird: conceptualization, investigation, methodology, software, validation, visualization, writing - original draft, writing - review editing. Victoria Pickens: conceptualization, investigation, methodology, visualization, writing - original draft, writing - review editing. David Molik: conceptualization, investigation, methodology, writing - original draft, writing - review editing, software, validation. Kristopher Silver: supervision, resources, writing - original draft, writing - review editing. Dana Nayduch: funding acquisition, conceptualization, investigation, methodology, resources, supervision, writing - original draft, writing - review editing.

**Reviewed By:** Anonymous

**History: Received** November 21, 2024 **Revision Received** February 11, 2025 **Accepted** February 11, 2025 **Published Online** February 19, 2025 **Indexed** March 5, 2025

**Citation:** Bird E, Pickens V, Molik D, Silver K, Nayduch D. 2025. BALROG-MON: a high-throughput pipeline for Bacterial AntimicrobiaL Resistance annOtation of Genomes-Metagenomic Oxford Nanopore. microPublication Biology. 10.17912/micropub.biology.001427