The Journal of Open Source Software

nowayout: An automated pipeline for taxonomic classification of Eukaryotic mitochondrial reads

Kranti Konganti [1¶], **Monica Pava-Ripoll** [1], **Amanda Windsor** [1], **Christopher Grim** [1], **Mark Mammel** [1], **and Padmini Ramachandran** [1]

1 Human Foods Program, U.S. Food and Drug Administration, United States ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 15 January 2026
Published: unpublished

Summary

nowayout is an ultra-fast automated software pipeline for taxonomic classification of eukaryotic mitochondrial reads. The workflow is implemented in Nextflow and employs a custom database to first identify mitochondrial reads, then performs taxonomic classification on those reads. The pipeline has been specifically developed and evaluated for identifying arthropod contaminants in food matrices, enabling species-level assignment for research applications, potentially supporting routine monitoring. Additionally, nowayout can be used to detect eukaryotic DNA in shotgun metagenomic datasets, supporting verification of labeling claims when insects are used as food ingredients and extending to other eukaryotic taxa that may be present as food ingredients, making it a versatile tool for food safety research, regulatory monitoring, and other applications where eukaryotic composition is of interest.

Statement of Need

Food safety and regulatory programs are increasingly using sequencing and reference databases to improve monitoring and the identification of organisms in the food chain, including evaluation of metagenomic methods to taxonomically resolve arthropod material from unavoidable crop-associated pests versus avoidable stored-product pests, an important regulatory distinction. In most conventional foods, DNA extracts are dominated by the food matrix (plant/animal), while arthropod DNA is often rare and fragmented; consequently, recoverable arthropod DNA is well suited to targeted capture using mitochondrial probe (bait) panels. This strategy is effective because mitochondrial DNA occurs at high copy number, often persists when nuclear DNA is degraded, and is supported by extensive public reference databases for arthropod identification (Foran, 2006; Sujeevan & Hebert, 2007). By contrast, in insect-based foods, the focus shifts from detecting trace arthropod DNA to distinguishing among arthropod species and identifying non-declared components. In both settings, potential regulatory use requires taxonomic classification of arthropods, whether present as contaminants or ingredients.

nowayout addresses the resulting need for a standardized, end-to-end workflow that produces reproducible mitochondrial-based taxonomic identifications and reporting from arthropod metagenomic sequencing. The pipeline's focus on eukaryotic mitochondrial reads allows for broad applicability across different arthropod species and other eukaryotic organisms that may be present in foods. To our knowledge, nowayout is one of the first software tools for fully automated analysis of mitochondrial read identification and classification of eukaryotic species from shotgun metagenomics data.

The Journal of Open Source Software

Methods and Materials

A brief overview of the nowayout pipeline is presented in Figure 1.

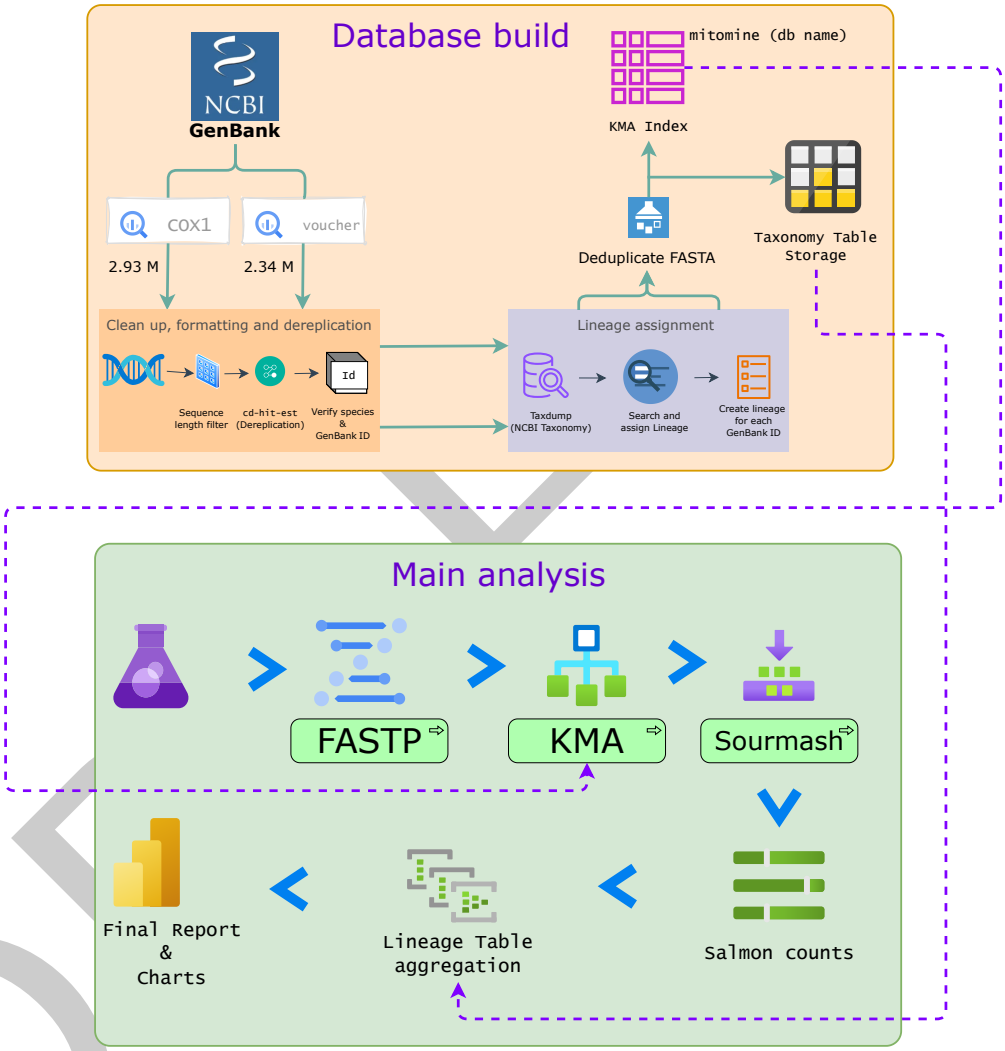

Figure 1: The database preparation step produces the `mitomine` database which is used during KMA alignment stage of the main analysis.

Database Generation

The nowayout pipeline uses a custom database generated from sequences downloaded from NCBI GenBank (Benson et al., 2012). The database construction process begins by downloading the NCBI Taxonomy dump (Schoch et al., 2020) and converting it to lineages using the `ncbitax2lin` tool (Xue & Cumbo, 2023). All sequences catalogued as voucher or cytochrome c oxidase subunit 1 (COX1) are then downloaded from NCBI GenBank (Benson et al., 2012; Pava-Ripoll et al., 2023) using a keyword search, with separate catalogs maintained for each sequence type.

In the next stage, sequences less than 200 base pairs in length are filtered out of the dataset. CD-HIT (Fu et al., 2012) is employed to perform sequence deduplication at 100% identity for each sequence catalog. For each filtered GenBank accession, the corresponding lineage is

identified in the NCBI Taxonomy dump and assigned to create a comprehensive taxonomic reference.

In the final stage of database preparation, the voucher and COX1 catalogs are merged and subjected to a final deduplication step at 100% sequence identity using CD-HIT (Fu et al., 2012). This final catalog of GenBank sequences is then indexed using KMA (Clausen et al., 2018), creating the main database for all subsequent taxonomic classification tasks. This custom database is named `mitomine`.

Taxonomic Read Classification

The main analysis steps in `nowayout` (Figure 1) begin with metagenomic sequencing and read preprocessing using fastp (Chen et al., 2018) for adapter trimming and quality filtering. Next, mitochondrial reads are identified by aligning to the `mitomine` database using KMA (Clausen et al., 2018). All mitochondrial reads are then extracted, and sketches are created for both the identified mitochondrial reads and accession hits. Reads identified as mtDNA are then classified using the `gather` command from sourmash (Irber et al., 2024). Additionally, Salmon (Patro et al., 2017) is used to bin the number of reads mapped to each lineage. Finally, a consolidated Krona (Ondov et al., 2011) chart and an aggregated MultiQC (Ewels et al., 2016) report is generated for all samples

`nowayout` offers three preset threshold filters (strict, mild, relax) for exploring results and optimizing taxonomic classification specificity and stringency trade-offs. The pipeline is also available on the HFP GalaxyTrakr platform (version >= 25.x), providing a user-friendly web interface for researchers without command-line expertise.

Results

To evaluate nowayout, we analyzed sequencing data (FASTQ files) generated from three mock genomic DNA (gDNA) mixtures comprising 23 insect taxa across seven orders (*Coleoptera*, *Blattodea*, *Diptera*, *Hymenoptera*, *Orthoptera*, *Lepidoptera*, and *Hemiptera*) combined with whole wheat flour gDNA as the food matrix. Mixture 1 contained one insect taxon, mixture 2 six, and mixture 3 twenty-two with staggered DNA inputs to mimic an uneven community. In all mixtures, whole wheat flour gDNA was added at four times the total insect gDNA mass. For each mixture, libraries were prepared in parallel and subjected to hybridization capture using two arthropod bait panel versions (v1 and v2) to enable direct panel-to-panel comparisons on identical mixture inputs.

The sequencing libraries were generated using the KAPA HyperPlus Library Preparation Kit (Roche Diagnostics) following the manufacturer's instructions. For targeted hybridization capture, amplified libraries with similar concentrations were pooled and enriched for mitochondrial targets using custom arthropod bait panels, applying either panel v1 or panel v2 under the same capture workflow. Enriched libraries were sequenced on an Illumina MiSeq platform.

Using `nowayout`, all expected insect taxa were detected in Mixtures 1 (a) and 2 (b), and most taxa were recovered in Mixture 3 (c) (22 of 23), with complete removal of wheat-flour background signal using panel v2 (Table 1). The nowayout visualizations produced a more interpretable and informative taxonomic profile.

Software Design

The `nowayout` pipeline is implemented in Nexflow (Di Tommaso et al., 2017) following DSL2 principles and as such can be run on any UNIX based platform. All the individual steps are parallelized and run concurrently for all samples. `nowayout` is released under MIT license and comprehensive documentation is hosted on GitHub. Current efforts are being undertaken to develop custom algorithms and classification methods to better handle ambiguous read

assignments. Additionally, expanding to support Oxford Nanopore long reads in addition to the current Illumina short read capability is planned, enabling analysis of a wider range of sequencing data types.

AI Usage Disclosure

Generative AI was not used in any aspects of the development of the software, in writing the documentation or during any aspects of the paper authorship process.

Acknowledgements

We would like to thank the HPC team for providing systems administration support for the HFP Reedling HPC Cluster.

References

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42. https://doi.org/10.1093/nar/gks1195

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Clausen, P. T., Aarestrup, F. M., & Lund, O. (2018). Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics*, *19*(1), 307. https://doi.org/10.1186/s12859-018-2336-6

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Foran, D. R. (2006). Relative degradation of nuclear and mitochondrial DNA: An experimental approach. *Journal of Forensic Sciences*, *51*(4), 766–770. https://doi.org/10.1111/j.1556-4029.2006.00176.x

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Irber, L., Pierce-Ward, N. T., Abuelanin, M., Alexander, H., Anant, A., Barve, K., Baumler, C., Botvinnik, O., Brooks, P., Dsouza, D., Gautier, L., Hera, M. R., Houts, H. E., Johnson, L. K., Klötzl, F., Koslicki, D., Lim, M., Lim, R., Nelson, B., … Brown, C. T. (2024). Sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets. *Journal of Open Source Software*, *9*(98), 6830. https://doi.org/10.21105/joss.06830

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, *12*(1), 385. https://doi.org/10.1186/1471-2105-12-385

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Pava-Ripoll, M., Miller, A. K., & Ziobro, G. C. (2023). Development of a multiplex polymerase chain reaction (PCR) assay for the potential detection of insect contaminants in food.

Journal of Food Protection, *86*(8), 100120. https://doi.org/10.1016/j.jfp.2023.100120

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., & others. (2020). NCBI taxonomy: A comprehensive update on curation, resources and tools. *Database*, *2020*, baaa062. https://doi.org/10.1093/database/baaa062

Sujeevan, R., & Hebert, P. (2007). BOLD: The barcode of life data system. *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Xue, Z., & Cumbo, F. (2023). *ncbitax2lin: A tool to convert NCBI taxonomy dump into lineages*. https://github.com/zyxue/ncbitax2lin