

# nf-core/coproID v2.0: An improved pipeline for the identification of (palaeo)faecal depositors

Meriam van Os <sup>1,2,3</sup>  and Maxime Borry <sup>3,4</sup>

1 Department of Anatomy, University of Otago, Dunedin, New Zealand 2 Archaeology, School of Social Sciences, University of Otago, Dunedin, New Zealand 3 Microbiome Sciences Group, Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany 4 Associated Research Group of Archaeogenetics, Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute, Jena, Germany  Corresponding author

DOI: [10.21105/joss.09324](https://doi.org/10.21105/joss.09324)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Abhishek Tiwari](#)  

## Reviewers:

- [@vinegang](#)
- [@johnvusich](#)

Submitted: 24 June 2025

Published: 15 March 2026

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

## Summary

With the advancement of next-generation sequencing technologies, (palaeo)faeces have become a unique and valuable source in the fields of archaeology ([Battillo, 2019](#)), microbiome studies ([Rifkin et al., 2020](#); [Wibowo et al., 2021](#)), species ecology and conservation ([Ang et al., 2020](#); [Taylor et al., 2022](#)), and even in forensic investigations ([Quaak et al., 2017, 2018](#)). To use faecal samples for such studies, often the first question is “who deposited the faeces?”, before doing additional analyses. The nf-core/coproID pipeline helps to answer this question by taking raw sequencing data and predicting the depositor’s species.

The raw sequencing data is first pre-processed to trim adapters and remove low quality and low complexity reads with fastp ([Chen et al., 2018](#)). Bowtie2 ([Langmead et al., 2018](#)) is then used to align the reads to multiple user-specified reference genomes of potential depositor (host) species. Next, these reads are processed with sam2lca ([Borry et al., 2022](#)) to retain only reads specific to one of the references, and normalised according to the size of the genome. Furthermore, a taxonomic profile is created with kraken2 ([Wood et al., 2019](#)), and compared to a user-supplied database of potential sources using sourcepredict to estimate the percentages of contributing sources ([Borry, 2019](#)). Both the normalised host DNA and the sourcepredict results are used to predict the most likely depositor of the faeces. The pipeline also incorporates ancient DNA damage estimates using pyDamage ([Borry et al., 2021](#)) and damageprofiler ([Neukamm et al., 2021](#)) for authenticating the ancient nature of the host DNA, when working with palaeofaeces. All results are collated into a Quarto notebook HTML report for an easy overview of all the results.

## Statement of need

As mentioned above, (palaeo)faeces are valuable resources to study the depositor’s DNA, diet, microbiome, health, and more. However, it is often difficult to identify the depositor based on the faeces morphology alone. For example, humans and dogs often overlap in their diets, and produce similarly sized faeces. In 2020, the pipeline nf-core/coproID v1.0 was published ([Borry et al., 2020](#)), which uses both host and microbial DNA to predict the depositor of faecal samples. The microbiome can be a crucial part for a host prediction, as the host DNA content in faeces can be very low in certain species and/or individuals ([Ang et al., 2020](#); [Perry et al., 2010](#)), including humans and modern dogs ([Borry et al., 2020](#)). Since its first release, new tools have become available that can improve the accuracy and usability of nf-core/coproID. Here we present the newest version of the pipeline, nf-core/coproID v2.0, rewritten in the newest Nextflow DSL2 language to enhance modularity, reusability, and scalability ([Di Tommaso et](#)

al., 2017), and with newly added features to improve accuracy and reporting.

## Materials and Methods

nf-core/coproID combines the analysis of the putative host (ancient) DNA with a machine learning prediction of the faeces source, based on microbiome taxonomic composition:

A. First, nf-core/coproID performs parallel mapping of all reads against two (or more) target genomes (genome1, genome2, ..., genomeX) using bowtie2 (Langmead et al., 2018), and computes a host-DNA species ratio (NormalisedProportion) using sam2lca (Borry et al., 2022).

B. Next, nf-core/coproID performs metagenomic taxonomic profiling with kraken2 (Wood et al., 2019), and compares the obtained profiles to user-supplied modern reference samples of the target species metagenomes. Using machine learning, sourcepredict (Borry, 2019) then estimates the host source from the metagenomic taxonomic composition (SourcepredictProportion).

C. Finally, nf-core/coproID combines the A and B proportions to predict the likely host of the metagenomic sample.

## Workflow

Figure 1 describes the newest workflow:

1. Quality check of the input fastq reads with FastQC (Andrews, 2010).
2. Removal of adapters and low-complexity reads with fastp (Chen et al., 2018).
3. Mapping of adapter trimmed reads to multiple reference genomes with Bowtie2 (Langmead et al., 2018).
4. Lowest Common Ancestor analysis with sam2lca (Borry et al., 2022) to retain only genome-specific reads, i.e. reads that align equally well to multiple references are removed from the read counts. The sam2lca read counts are normalised by the size of the genome as follows. First, a normalisation factor is calculated per reference, or source species (sp):

$$\text{AverageReferenceLength} = \sum_{\text{sp}} \text{ReferenceLength}_{\text{sp}} / \text{NumberOfReferences}$$

$$\text{NormalisationFactor}_{\text{sp}} = \text{AverageReferenceLength} / \text{ReferenceLength}_{\text{sp}}$$

Then, normalised read counts are calculated by:

$$\text{NormalisedReads}_{\text{sp}} = \text{sam2lcaReads}_{\text{sp}} * \text{NormalisationFactor}_{\text{sp}}$$

5. Taxonomic profiling is performed on adapter-trimmed reads with kraken2 (Wood et al., 2019), and by using a custom supplied database. Kraken2 reports are parsed and merged into one table for all samples.
6. Sourcepredict (Borry, 2019) is then used to predict the source proportions, based on the kraken2 taxonomic profiles, and by using user-supplied reference sources (which should have been created with the same reference database).
7. Both the host DNA (NormalisedReads) and sourcepredict proportion are used to predict the most likely depositor of the (palaeo)faeces. The probability of each reference species is calculated by:

$$\text{Probability}_{\text{sp}} = \text{NormalisedSam2lcaProportion}_{\text{sp}} * \text{SourcepredictProportion}_{\text{sp}}$$

8. Ancient DNA damage patterns are estimated using pyDamage (Borry et al., 2021) and damageprofiler (Neukamm et al., 2021) to authenticate the ancient nature of the DNA when working on palaeofaecal samples.
9. MultiQC (Ewels et al., 2016) aggregates results of several individual nf-core modules.
10. Quarto notebook (Allaire et al., 2024) creates a report with an overview of all sample results (incl. tables and figures).

## Output

The results are located in a nested folder architecture. Fourteen subfolders are created within the user-identified output folder:

- bowtie2
- create
- damageprofiler
- fastp
- fastqc
- kraken
- kraken2
- multiqc
- pipeline\_info
- pydamage
- quartonotebook
- sam2lca
- samtools
- sourcepredict

## Funding source declaration

MO was supported by a University of Otago Doctoral Scholarship.

## Availability

The nf-core/coproID pipeline is freely available from the nf-core pipeline repository <https://nf-co.re/coproID/2.0.1/>.

## Figures

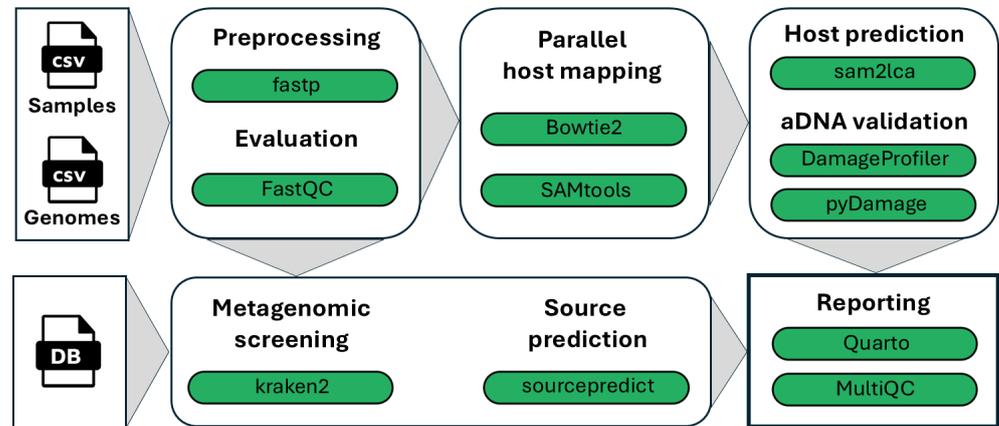


Figure 1: Workflow of nf-core/coproID version 2.0, showing the tools used for each step of the process.

## References

- Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., Dervieux, C., & Woodhull, G. (2024). *Quarto* (Version 1.6). <https://doi.org/10.5281/zenodo.5960048>
- Andrews, S. (2010). *Babraham Bioinformatics - FastQC a quality control tool for high throughput sequence data*. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Ang, A., Roesma, D. I., Nijman, V., Meier, R., Srivathsan, A., & Rizaldi. (2020). Faecal DNA to the rescue: Shotgun sequencing of non-invasive samples reveals two subspecies of Southeast Asian primates to be critically endangered species. *Scientific Reports*, *10*(1), 9396–9396. <https://doi.org/10.1038/s41598-020-66007-8>
- Battillo, J. (2019). Farmers who forage: Interpreting paleofaecal evidence of wild resource use by early corn farmers in the North American Southwest. *Archaeological and Anthropological Sciences*, *11*(11), 5999–6016. <https://doi.org/10.1007/s12520-019-00944-y>
- Borry, M. (2019). Sourcepredict: Prediction of metagenomic sample sources using dimension reduction followed by machine learning classification. *Journal of Open Source Software*, *4*(41), 1540. <https://doi.org/10.21105/joss.01540>
- Borry, M., Cordova, B., Perri, A., Wibowo, M., Prasad Honap, T., Ko, J., Yu, J., Britton, K., Girdland-Flink, L., Power, R. C., Stuijts, I., Salazar-García, D. C., Hofman, C., Hagan, R., Samdapawindé Kagoné, T., Meda, N., Carabin, H., Jacobson, D., Reinhard, K., ... Warinner, C. (2020). CoproID predicts the source of coprolites and paleofeces using microbiome composition and host DNA content. *PeerJ (San Francisco, CA)*, *8*, e9001–e9001. <https://doi.org/10.7717/peerj.9001>
- Borry, M., Hübner, A., Rohrlach, A. B., & Warinner, C. (2021). PyDamage: Automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly. *PeerJ*, *9*, e11845. <https://doi.org/10.7717/peerj.11845>
- Borry, M., Hübner, A., & Warinner, C. (2022). sam2lca: Lowest common ancestor for SAM/BAM/CRAM alignment files. *Journal of Open Source Software*, *7*(74), 4360. <https://doi.org/10.21105/joss.04360>
- 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>
- Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology* 2017 35:4, 35(4), 316–319. <https://doi.org/10.1038/nbt.3820>
- Ewels, P., Magnusson, M., Lundin, S., & Källér, 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>
- Langmead, B., Wilks, C., Antonescu, V., & Charles, R. (2018). Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, 35(3), 421–432. <https://doi.org/10.1093/bioinformatics/bty648>
- Neukamm, J., Peltzer, A., & Nieselt, K. (2021). DamageProfiler: Fast damage pattern calculation for ancient DNA. *Bioinformatics*, 37(20), 3652–3653. <https://doi.org/10.1093/bioinformatics/btab190>
- Perry, G. H., Marioni, J. C., Melsted, P., & Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology*, 19(24), 5332–5344. <https://doi.org/10.1111/j.1365-294x.2010.04888.x>
- Quaak, F. C. A., Graaf, M.-L. M. de, Weterings, R., & Kuiper, I. (2017). Microbial population analysis improves the evidential value of faecal traces in forensic investigations. *International Journal of Legal Medicine*, 131(1), 45–51. <https://doi.org/10.1007/s00414-016-1390-8>
- Quaak, F. C. A., Wal, Y. van de, Maaskant-van Wijk, P. A., & Kuiper, I. (2018). Combining human STR and microbial population profiling: Two case reports. *Forensic Science International : Genetics*, 37, 196–199. <https://doi.org/10.1016/j.fsigen.2018.08.018>
- Rifkin, R. F., Vikram, S., Ramond, J.-B., Rey-Iglesia, A., Brand, T. B., Porraz, G., Val, A., Hall, G., Woodborne, S., Le Bailly, M., Potgieter, M., Underdown, S. J., Koopman, J. E., Cowan, D. A., Van de Peer, Y., Willerslev, E., & Hansen, A. J. (2020). Multi-proxy analyses of a mid-15th century Middle Iron Age Bantu-speaker palaeo-faecal specimen elucidates the configuration of the “ancestral” sub-Saharan African intestinal microbiome. *Microbiome*, 8(1), 62–23. <https://doi.org/10.1186/s40168-020-00832-x>
- Taylor, R. S., Manseau, M., Redquest, B., Keobouasone, S., Gagné, P., Martineau, C., & Wilson, P. J. (2022). Whole genome sequences from non-invasively collected caribou faecal samples. *Conservation Genetics Resources*, 14(1), 53–68. <https://doi.org/10.1007/s12686-021-01235-2>
- Wibowo, M. C., Yang, Z., Borry, M., Hübner, A., Huang, K. D., Tierney, B. T., Zimmerman, S., Barajas-Olmos, F., Contreras-Cubas, C., García-Ortiz, H., Martínez-Hernández, A., Lubber, J. M., Kirstahler, P., Blohm, T., Smiley, F. E., Arnold, R., Ballal, S. A., Pamp, S. J., Russ, J., ... Kostic, A. D. (2021). Reconstruction of ancient microbial genomes from the human gut. *Nature (London)*, 594(7862), 234–239. <https://doi.org/10.1038/s41586-021-03532-0>
- Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1), 1–13. <https://doi.org/10.1186/S13059-019-1891-0>