






1 nf-core/coproID v2.0: An improved pipeline for the 2 identification of (palaeo)faecal depositors

3 Meriam van Os ^{1,2}¶ and Maxime Borry ^{2,3}

4 1 Department of Anatomy, University of Otago, Dunedin, New Zealand 2 Microbiome Sciences Group,
5 Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
6 3 Associated Research Group of Archaeogenetics, Leibniz Institute for Natural Product Research and
7 Infection Biology Hans Knöll Institute, Jena, Germany ¶ Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

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

Editor: 

Submitted: 24 June 2025

Published: unpublished

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/)).

8 Summary

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

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

30 Statement of need

31 As mentioned above, (palaeo)faeces are valuable resources to study the depositor’s DNA, diet,
32 microbiome, health and more. However, it is often difficult to identify the depositor based
33 on the faeces morphology alone. For example, humans and dogs often overlap in their diets,
34 and produce similarly sized faeces. In 2020, the pipeline nf-core/coproID v1.0 was published
35 ([Borry et al., 2020](#)), which uses both host and microbial DNA to predict the depositor of faecal
36 samples. The microbiome can be a crucial part for a host prediction, as the host DNA content
37 in faeces can be very low in certain species and/or individuals ([Ang et al., 2020](#); [Perry et al.,
38 2010](#)), including humans and modern dogs ([Borry et al., 2020](#)). Since its first release, new tools
39 have become available that can improve the accuracy and usability of nf-core/coproID. Here
40 we present the newest version of the pipeline, nf-core/coproID v2.0.0, rewritten in the newest
41 Nextflow DLS2 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 (Borrry 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 (Borrry, 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 (Borrry 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 followed. First, a normalisation factor is calculated per reference, or source species (sp):

$$AverageReferenceLength = \sum_{sp} ReferenceLength_{sp} / NumberOfReferences$$

$$NormalisationFactor_{sp} = AverageReferenceLength / ReferenceLength_{sp}$$

Then, normalised read counts are calculated by:

$$NormalisedReads_{sp} = sam2lcaReads_{sp} * NormalisationFactor_{sp}$$

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 (Borrry, 2019) is then used to predict the source proportions, based on the kraken2 taxonomic profiles, and by using customer 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:

$$Probability_{sp} = NormalisedSam2lcaProportion_{sp} * SourcepredictProportion_{sp}$$

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

82 Output

83 The results are located in a nested folder architecture. Fourteen subfolders are created within
84 the customer identified output folder: - bowtie2 - create - damageprofiler - fastp - fastqc -
85 kraken - kraken2 - multiqc - pipeline_info - pydamage - quertonotebook - sam2lca - samtools -
86 sourcepredict

87 Discussion and conclusions

88 We present a new version of the nf-core/coproID pipeline, v2.0.0, designed to identify the true
89 depositor of (palaeo)faeces. Written in Nextflow DSL2, and adhering to the latest nf-core
90 standards and guidelines, nf-core/coproID v2.0.0 is more modular, reusable, and scalable. It
91 includes several new features, including fastp for faster pre-processing of the sequencing reads,
92 sam2lca to improve and generalise host DNA prediction, pyDamage to discriminate between
93 ancient and modern DNA, and the automated creation of a Quarto notebook html report.
94 The modular design also makes it easier for users to customise the pipeline, for example by
95 adding more modules and workflows.

96 Funding source declaration

97 MO was supported by a University of Otago Doctoral Scholarship.

98 Availability

99 The nf-core/coproID pipeline is freely available from the nf-core pipelines depository [https:](https://nf-co.re/coproID/2.0.0/)
100 [//nf-co.re/coproID/2.0.0/](https://nf-co.re/coproID/2.0.0/).

Figures

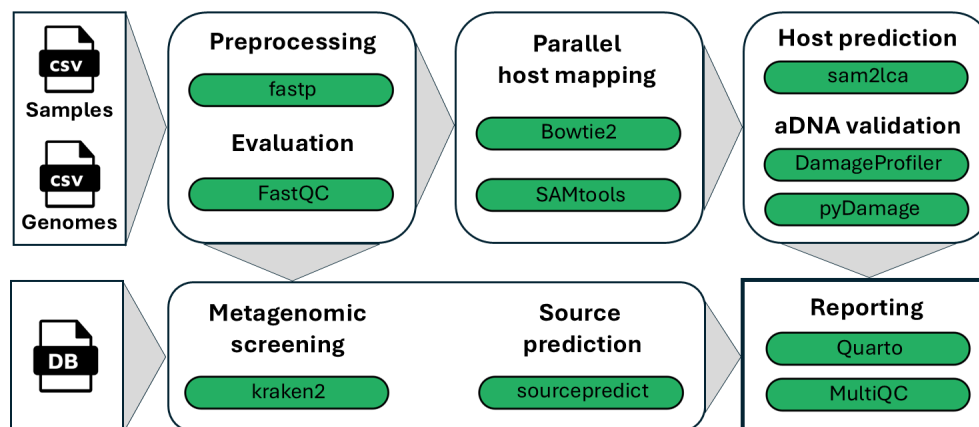


Figure 1: Figure 1

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 paleofecal 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 prepro-

- 130 sor. *Bioinformatics*, 34(17), i884–i890. [https://doi.org/10.1093/BIOINFORMATICS/](https://doi.org/10.1093/BIOINFORMATICS/BTY560)
131 [BTY560](https://doi.org/10.1093/BIOINFORMATICS/BTY560)
- 132 DI Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C.
133 (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*
134 2017 35:4, 35(4), 316–319. <https://doi.org/10.1038/nbt.3820>
- 135 Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis
136 results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047–3048.
137 <https://doi.org/10.1093/bioinformatics/btw354>
- 138 Langmead, B., Wilks, C., Antonescu, V., & Charles, R. (2018). Scaling read aligners to
139 hundreds of threads on general-purpose processors. *Bioinformatics*, 35(3), 421–432.
140 <https://doi.org/10.1093/bioinformatics/bty648>
- 141 Neukamm, J., Peltzer, A., & Nieselt, K. (2021). DamageProfiler: Fast damage pattern
142 calculation for ancient DNA. *Bioinformatics*, 37(20), 3652–3653. [https://doi.org/10.1093/](https://doi.org/10.1093/bioinformatics/btab190)
143 [bioinformatics/btab190](https://doi.org/10.1093/bioinformatics/btab190)
- 144 Perry, G. H., Marioni, J. C., Melsted, P., & Gilad, Y. (2010). Genomic-scale capture and
145 sequencing of endogenous DNA from feces. *Molecular Ecology*, 19(24), 5332–5344.
146 <https://doi.org/10.1111/j.1365-294x.2010.04888.x>
- 147 Quaak, Frederike C. A., Graaf, M.-L. M. de, Weterings, R., & Kuiper, I. (2017). Microbial
148 population analysis improves the evidential value of faecal traces in forensic investiga-
149 tions. *International Journal of Legal Medicine*, 131(1), 45–51. [https://doi.org/10.1007/](https://doi.org/10.1007/s00414-016-1390-8)
150 [s00414-016-1390-8](https://doi.org/10.1007/s00414-016-1390-8)
- 151 Quaak, Frederike C. A., Wal, Y. van de, Maaskant-van Wijk, P. A., & Kuiper, I. (2018).
152 Combining human STR and microbial population profiling: Two case reports. *Forensic*
153 *Science International : Genetics*, 37, 196–199. [https://doi.org/10.1016/j.fsigen.2018.08.](https://doi.org/10.1016/j.fsigen.2018.08.018)
154 [018](https://doi.org/10.1016/j.fsigen.2018.08.018)
- 155 Rifkin, R. F., Vikram, S., Ramond, J.-B., Rey-Iglesia, A., Brand, T. B., Porraz, G., Val, A.,
156 Hall, G., Woodborne, S., Le Bailly, M., Potgieter, M., Underdown, S. J., Koopman, J.
157 E., Cowan, D. A., Van de Peer, Y., Willerslev, E., & Hansen, A. J. (2020). Multi-proxy
158 analyses of a mid-15th century middle iron age bantu-speaker palaeo-faecal specimen
159 elucidates the configuration of the “ancestral” sub-saharan african intestinal microbiome.
160 *Microbiome*, 8(1), 62–23. <https://doi.org/10.1186/s40168-020-00832-x>
- 161 Taylor, R. S., Manseau, M., Redquest, B., Keobouasone, S., Gagné, P., Martineau, C., &
162 Wilson, P. J. (2022). Whole genome sequences from non-invasively collected caribou
163 faecal samples. *Conservation Genetics Resources*, 14(1), 53–68. [https://doi.org/10.1007/](https://doi.org/10.1007/s12686-021-01235-2)
164 [s12686-021-01235-2](https://doi.org/10.1007/s12686-021-01235-2)
- 165 Wibowo, M. C., Yang, Z., Borry, M., Hübner, A., Huang, K. D., Tierney, B. T., Zimmerman, S.,
166 Barajas-Olmos, F., Contreras-Cubas, C., García-Ortiz, H., Martínez-Hernández, A., Luber,
167 J. M., Kirstahler, P., Blohm, T., Smiley, F. E., Arnold, R., Ballal, S. A., Pamp, S. J., Russ,
168 J., ... Kostic, A. D. (2021). Reconstruction of ancient microbial genomes from the human
169 gut. *Nature (London)*, 594(7862), 234–239. <https://doi.org/10.1038/s41586-021-03532-0>
- 170 Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2.
171 *Genome Biology*, 20(1), 1–13. <https://doi.org/10.1186/S13059-019-1891-0>