Received: 23 February 2022 | Accepted: 6 September 2022

DOI: 10.1111/2041-210X.14006

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

APPLICATION

*ngs*LCA—A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data

Yucheng Wang[1,2,3,4] | **Thorfinn Sand Korneliussen**[2] | **Luke E. Holman**[5,6] | **Andrea Manica**[1] | **Mikkel Winther Pedersen**[2]

[1]Department of Zoology, University of Cambridge, Cambridge, UK; [2]Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen K, Denmark; [3]ALPHA, State Key Laboratory of Tibetan Plateau Earth System, Environment and Resources (TPESER), Institute of Tibetan Plateau Research (ITPCAS), Chinese Academy of Sciences (CAS), Beijing, China; [4]BGI, BGI-Shenzhen, Shanghai, China; [5]School of Ocean and Earth Science, National Oceanography Centre Southampton, University of Southampton, Southampton, UK and [6]Section for Evolutionary Genomics, Faculty of Health and Medical Sciences, Globe Institute, University of Copenhagen, Copenhagen, Denmark

Correspondence
Mikkel Winther Pedersen
Email: mwpedersen@sund.ku.dk
Yucheng Wang
Email: yw502@cam.ac.uk

Funding information
Carlsbergfondet, Grant/Award Number: CF16-0728, CF16-0913, CF18-0024 and CF19-0712; Natural Environmental Research Council, Grant/Award Number: NE/L002531/1

Handling Editor: Antonino Malacrinò

Abstract

1. Metagenomic data generated from environmental samples is increasingly common in the analysis of modern and ancient biological communities. To obtain taxonomic profiles from this type of data, DNA sequences are aligned against large genomic reference databases and the lowest common ancestor (LCA) needs to be inferred for each sequence with multiple alignments. To date, efforts have mainly focused on improving the speed, sensitivity and specificity of alignment tools, and little effort has been applied to the LCA algorithm that generates the taxonomic profiles from alignments.
2. We present *ngs*LCA, a command-line toolkit with two separate modules: the main program (in C/C++) performing LCA inference, and an R package for generating tables and visualisations of the taxonomic profiles.
3. *ngs*LCA processed large datasets in BAM/SAM alignment format 4–11 times faster and used less memory compared to other available programs. It is compatible with the NCBI taxonomy and has flexible parameter settings. Furthermore, the toolkit offers functions for filtering, contamination removal, taxonomic clustering, and multiple ways of visualising the generated taxonomic profiles.
4. *ngs*LCA bridges a gap in current metagenomic analyses by supplying a computationally light, easy-to-use, accurate, fast and flexible LCA algorithm with R functions for processing and illustrating the taxonomic profiles

KEYWORDS
environmental DNA (eDNA), lowest common ancestor (LCA), metagenomics, next-generation sequencing, sedimentary ancient DNA (sedaDNA), shotgun sequencing, taxonomic profiling, toolkit

Yucheng Wang and Thorfinn Sand Korneliussen contributed equally to this work.

1 | INTRODUCTION

Understanding the distribution of species in an ecosystem through time and space is essential for a wide range of disciplines, such as ecology, conservation biology, microbiology, palaeontology, archaeology, and environmental sciences. Species incidence data have traditionally been generated using laboratory cultures for microbiota (Barer & Harwood, 1999), collected during field investigation or with Geographic Information System (GIS) tools for extant fauna and flora (Muller, 1997; Ryser-Degiorgis, 2013), and through the analysis of micro- and macro-fossils to learn about biota change through time (Jackson et al., 1997). More recently, DNA collected from environmental samples has been combined with innovations in high-throughput DNA sequencing to produce accurate and standardised biodiversity information (Thomsen & Willerslev, 2015). One such method is environmental DNA (eDNA) metabarcoding, which uses universal PCR primers to amplify a short marker gene for a specific group of organisms (such as vertebrate, vascular plant, bacteria) (Deiner et al., 2017). This non-invasive genetic approach allows for fast, efficient, sensitive, and comprehensive detection of many taxa in parallel to reconstruct biodiversity without a requirement to culture, observe or count organisms and fossils, and is therefore becoming frequently applied across disciplines (Deiner et al., 2017; Ruppert et al., 2019; Thomsen & Willerslev, 2015).

The continuously decreasing cost and expanding output of DNA sequencing technologies in the past few years have made it increasingly common to sequence the entire DNA pool isolated from environmental samples, without target species PCR amplification, to reconstruct ecosystem composition (Cowart et al., 2018; Garlapati et al., 2019; Pedersen et al., 2016). This method, known as eDNA shotgun metagenomics, has several advantages compared to metabarcoding. It enables genome-wide and functional analyses with greater sensitivity (Chua et al., 2021; Pedersen et al., 2021), can accurately recover the relative composition of DNA in a mixed sample (Bell et al., 2021; Wang et al., 2021), and permits the estimation of DNA damage for ancient DNA authentication (Pedersen et al., 2016). Shotgun metagenomics involves randomly sequencing a pool of DNA, either directly extracted from environmental samples (Cowart et al., 2018; Pedersen et al., 2016), or from a sample enriched in the laboratory targeting a specific taxon or group of taxa through hybridization capture (Jensen et al., 2021; Vernot et al., 2021). In both methods, the generated DNA sequences are bioinformatically aligned to a database containing reference genomes of multiple species, which in some cases can comprise millions of different taxa to minimise taxonomic bias (Garlapati et al., 2019; Wang et al., 2021). For each sequence with multiple possible alignments to the database, a lowest common ancestor (LCA) of the aligned taxa needs to be calculated to obtain taxonomic inference and classification.

However, very few tools have been designed for performing LCA inference. Instead, efforts have been applied to develop faster and more accurate alternatives for sequence alignment such as DIAMOND (Buchfink et al., 2021) and PIA (Cribdon et al., 2020), tools for microbial alignments such as MALT (Herbig et al., 2016) and MetaPhlAn (Beghini et al., 2021), and faster k-mer based approaches such as kraken2 (Wood et al., 2019), krakenuniq (Breitwieser et al., 2018), Kaiju (Menzel et al., 2016), and PuffAligner (Almodaresi et al., 2021). Some of these tools (Kraken2, Diamond) can handle large databases, provided they are given enough memory on the computer, and KrakenUniq can do partial loading of the database. However, many of these tools are otherwise not compatible with large reference databases (e.g. NCBI-RefSeq) or do not accept custom genomes as reference, hindering the incorporation of constantly growing genome references. Some offer LCA functions for processing alignment results (such as PIA) or pre-clustering the reference database to assign the shared gnomonic regions to LCA (such as krakenuniq), but none of these programs process files in Sequence Alignment Format (i.e. BAM/SAM/CRAM) (Li et al., 2009) generated by commonly used aligners such as BWA (Li & Durbin, 2009) and bowtie2 (Langmead & Salzberg, 2012), limiting their broader application and often requiring a re-alignment of reads to reference genomes of interest for downstream analysis.

MEGAN6 (and the associated command-line tool sam2rma) (Huson et al., 2016) and sam2lca (Borry et al., 2022) are, to our knowledge, the only programs that can perform LCA taxonomic assignment on sequences from the standard alignment format. However, MEGAN6 is not directly compatible with the NCBI taxonomy format and relies on the release of a customised binary taxonomy file, which prevents the use of the latest updated NCBI databases. The custom taxonomy provided by MEGAN6 also does not allow for additional reference genomes from non-public databases. Furthermore, sam2lca does not supply functionalities needed for parsing, clustering, exploring, or illustrating of the generated taxonomic files. Finally, the LCA inference of both these programs is slower, particularly for larger datasets.

The size of both sequencing dataset and reference databases are rapidly increasing. A faster LCA inference toolkit is therefore needed to enable efficient processing of large alignment datasets in sequence alignment format, allow for the optional addition of taxonomy and reference genomes, and be compatible with command-line workflows for batch processing and incorporation into analytical pipelines.

Here, we present *ngs*LCA (next generation sequence lowest common ancestor algorithm), a standalone and easy-to-use toolkit that achieves fast and accurate LCA inference. Compared with other programs, it has the following advantages: (i) directly compatible with BAM/SAM/CRAM format alignments, (ii) computationally lighter (the entire workflow can be carried out on a regular laptop), (iii) built on the NCBI taxonomy and able to directly read in the downloaded NCBI taxonomy database without any reformatting requirements, (iv) accepts custom reference databases containing specific reference genomes of interests and a custom taxonomy confined to the NCBI format, (v) compatible with inputs from short or long read alignments, single or pair ended sequencing data, and datasets targeting any organismal assemblages, (vi) accepts a series of simplified parameters to provide flexible LCA inference, (vii) provides R functions for taxonomic profiling, filtering, sorting, and data visualisation, as well as generating outputs that are compatible with other subsequent analysis and visualisation tools, and (viii) is available under the GNU General Public License v3.0 (GPL).

2 | DESIGN, IMPLEMENTATION AND FEATURES

The *ngs*LCA toolkit comprises two modules, the main program (written in C/C++) that performs the actual LCA inference, and an R package that processes the LCA results and outputs tables and visualisations of the taxonomic profiles and other summary statistics.

2.1 | *ngs*LCA main program

The main program uses alignment files in BAM/SAM/CRAM format as input to infer the LCA per sequence based on the NCBI taxonomy (Figure 1a). The input alignment files can be generated by mapping a metagenomic dataset against one or several nucleotide reference databases. The reference databases can be customised by the user, but they need to comply with the NCBI taxonomy format (see *ngs*LCA GitHub repository for further details). All NCBI genomic databases (e.g. NCBI-nt and all NCBI-RefSeq divisions) can be directly used as reference.

We developed a new implementation for the naive LCA algorithm (Figure 1b): all alignments of each read are retained using a user-defined similarity, and we seek to find the lowest (furthest from the root) internal node in the complete taxonomic tree that spans these alignments. The taxonomic information is supplied as a tree with a branching order defined by lower-level nodes referencing their parent node. Nodes are identified by the NCBI taxonomic ID (taxaID) and include ranking information such as subspecies, species, genus and family. The LCA for a given read that is aligned to *n* reference identifiers (terminal nodes) is then characterised as the node

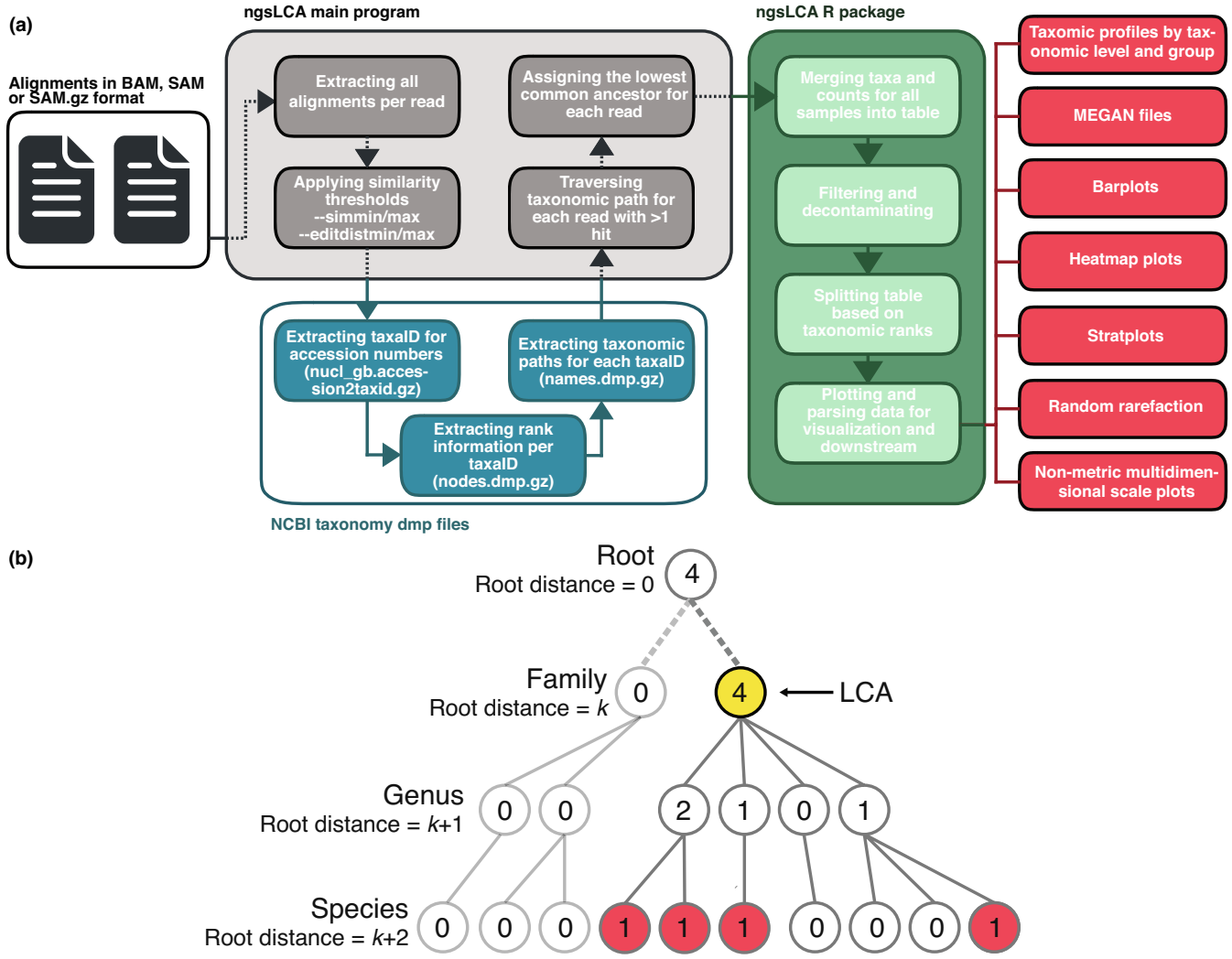

FIGURE 1 Workflow of *ngs*LCA toolkit and the algorithm for *ngs*LCA main program. The sketch map (a) shows the full *ngs*LCA workflow. The main program uses BAM/SAM/CRAM alignment file(s) as input to infer LCAs based on the NCBI taxonomy. *ngs*LCA R package supplies functions for parsing, processing, sorting and illustrating the LCA results. The schematic diagram (b) shows the algorithm for *ngs*LCA main program. An example is shown for an LCA inference where one read aligned to four different terminal nodes (i.e. $n = 4$). The number in each circle indicates the path multiplicity; red circles indicate the alignments (read alignments to terminal nodes). The LCA for this read is indicated by the yellow circle, which has the longest root distance (i.e. k) within the nodes for which the path multiplicity is equal to the number of alignments (i.e. n).

furthest from the root (i.e. to have the greatest root distance [*k*]) for which a direct path exists downwards in the tree for all the *n* aligned terminal nodes. The canonical solution for finding this furthest node in the presence of parent pointers is to tabulate the number of times (path multiplicity) the internal nodes are traversed from each of the terminal nodes toward the root. For internal nodes that have been traversed *n* times (i.e. for which path multiplicity is equal to *n*), the root distance is then calculated and the node with the longest root distance is reported as LCA for the read.

The LCA inference per read is reported as a flat text file (.lca file) per input file. At the same time, the LCA algorithm also outputs a look-up file (.bin file) that enables fast traversal of the reference database taxonomy. See *ngs*LCA GitHub repository for further details.

2.2 | *ngs*LCA R package

The *ngs*LCA R package supplies a series of easy-to-use functions to parse, filter, and sort the output files from the main program (Figure 1a). These R functions can generate taxonomic profiles in different table formats, split by taxonomic group and levels, which are compatible with downstream visualisation tools and subsequent analyses. It also offers functions that allow the user to visualise the taxonomic profiles in different ways (e.g. heatmaps, barplots, stratplots), and can implement non-metric multidimensional scaling (NMDS) and sample species richness tests (rarefaction analysis) (Dixon, 2003) for a quick assessment of the dataset.

The R function *ngsLCA_profile* first loads all ".lca" files from a user-defined directory, counts the number of reads assigned to each taxon, and subsequently merges the taxonomic profiles of all samples into a taxa-count matrix as a combined taxonomic profile. Low-abundance taxa that often derive from false-positives (due to sequencing errors, PCR errors or mis-annotations in the database [Steinegger & Salzberg, 2020]) can be removed by user-defined thresholds via the function *ngsLCA_filter*. All taxa that appear in the experimental controls (if supplied) will first be combined and filtered to form a "contamination" list, and the listed taxa will be subtracted from the combined taxonomic profile via the function *ngsLCA_deContam*. The *ngsLCA_rank* function sorts the taxonomic profiles into user-defined taxonomic levels (e.g. species, genus, family) by summing read counts from all lower taxonomic nodes of a given taxon. The *ngsLCA_group* function categorises taxa into user-defined groups (e.g. virus, bacteria, plant). Summary statistics for each filter, contamination removal, taxonomic level and group can be generated using the function *ngsLCA_count*.

Taxonomic profiles will all be written into the user-specified folder in tab-separated text files. All results from these functions can also be parsed into a MEGAN taxonomic profile format via *ngsLCA_meganFile* to be further analysed in MEGAN. Heatmaps, barplots, stratplots showing reads abundance of the taxonomic profiles can be generated using function *ngsLCA_heatmap* (Gu et al., 2016), *ngsLCA_barplot* and *ngsLCA_stratplot*, respectively. NMDS and rarefaction analysis on the generated taxonomic profiles can be performed using the functions *ngsLCA_NMDS* and *ngsLCA_rarefy*, respectively.

3 | COMPARISON WITH EQUIVALENT PROGRAMS

We tested *ngs*LCA by comparing its performance against the equivalent software packages discussed above. Specifically, we evaluated the specificity, speed and memory usage of *ngs*LCA through comparisons with MEGAN6 on a standard Mac laptop, and with sam2rma and sam2lca on a HPC (high performance computing) Linux server.

To mimic a typical metagenomic experiment, we downloaded and indexed reference databases from two sources: (1) the NCBI-nt (downloaded on 21 November 2018), that, due to its large size, was divided into 9 equally sized FASTA files; (2) the NCBI-RefSeq database (release 90), which was also divided into equally sized FASTA files, including vertebrates (9 files), mammals (18 files) and invertebrates (3 files). We next selected 10 shotgun sequenced metagenomic samples (Spring Lake samples from Pedersen et al. (2016)) and aligned all reads against each of these reference databases using bowtie2 (version 2.3.2) end-to-end alignment following the default settings, allowing up to 5000 hits per read (–k 5000) to each indexed reference database file. The resulting BAM files for each sample were then merged and sorted using SAMtools (version 1.10).

3.1 | Comparison with MEGAN6

Generated BAM files were first converted to SAM format for compatibility with MEGAN6, and thereafter were analysed with *ngs*LCA and MEGAN6 (V6.18.6) using the same NCBI taxonomy (Oct-2019 version), allocating 14 GB memory (RAM) and using single-thread processing on macOS Catalina (V10.1.5.4), and requiring 100% similarity between reference and sequence. The generated taxonomic profiles, processing time, and RAM usage were recorded and compared.

3.1.1 | Specificity of *ngs*LCA and MEGAN6

We expected identical taxonomic profiles produced by *ngs*LCA and MEGAN6 using the same LCA parameters on identical input alignments. However, while this was the case, we found small differences between the profiles from between the two programs (Table 1). By manually checking the differences, we found the discrepancy was due to errors in the built-in taxonomic tree (NCBI taxdump) of MEGAN6, which did not match the corresponding Oct-2019 version accession2taxID archive (this archive aligns each accession ID of the reference database to the taxonomic ID). For example, in one tested sample (SPL_015_1444), MEGAN6 assigned 533 reads to taxaID 329,540. In the input SAM file, most of these reads were aligned to

TABLE 1 Differences in taxonomic profiles generated by *ngs*LCA and MEGAN6 for the 10 tested samples

Sample ID	Reads parsed by *ngs*LCA	Reads parsed by MEGAN6	Discrepancies between *ngs*LCA and MEGAN6
SPL_015_1444	238,868	238,868	965
SPL_055_3112	81,626	81,625	351
SPL_113_5430	207,252	207,252	809
SPL_153_7001	2335	2335	24
SPL_173_7782	5894	5894	31
SPL_191_8967	21,362	21,361	111
SPL_193_9133	14,425	14,425	101
SPL_195_9299	29,237	29,237	127
SPL_221_11440	6511	6511	77
SPL_235_11534	15,639	15,639	56

the reference LO018304.1 which is reported as representing taxaID 1160 in the accession2taxID archive. However, these reads were correctly parsed by *ngs*LCA, resulting in differences in the outputs between the two programs. This shows the importance of using an updated taxonomy database for accurate LCA inference, which can be easily handled by *ngs*LCA. MEGAN6, however, supplies no user option to update the NCBI taxonomy, but releases a binary-formatted taxonomy file once or twice a year.

3.2 | Speed and RAM usage of *ngs*LCA and MEGAN6

We also tracked the time consumed by both programs for processing the 10 SAM files and found that *ngs*LCA on average was 6.4 times faster and parsed 5.9 times more alignments per second than MEGAN6 (Figure 2a). For the three largest files with more than 100 million alignments, we found *ngs*LCA was 11.3 times faster. To further explore these observations, we merged all the 10 tested datasets into one large file containing 846.76 million alignments and parsed it using both programs. We found that *ngs*LCA was 11.6 times faster and used 686 s, while MEGAN6 used 7978 s in total.

Additionally, we logged RAM usage every 2 s during the processing of all files (Figure 2c). We found that *ngs*LCA consumed 22% less memory on average than MEGAN6. For the merged file combing all the 10 tested datasets, we found that *ngs*LCA consumed 30% less memory on average than MEGAN6.

3.3 | Comparison with SAM2RMA and SAM2LCA

To ensure compatibility, the BAM files were indexed using SAMtools to be used as input for sam2lca and converted to the SAM format to be used as input for sam2rma. Files were hereafter analysed with *ngs*LCA, sam2lca (version 1.0.0), and sam2rma (version 6.22.2) simultaneously on a Red Hat Enterprise Linux Server running 7.7 (Maipo) with 88 cores and 256 GB memory. The analyses were all performed with the same NCBI taxonomy (Feb-2022 version, the latest version that was available in the required binary format for sam2rma) for the 3 programs, using single thread and unlimited RAM access and requiring 100% similarity between reference and sequence. The generated taxonomic profiles, processing time, and RAM usage were compared using the same method as the comparison between *ngs*LCA and MEGAN6.

We found the three programs generated identical taxonomic profiles when based on the same taxonomy database. However, on average *ngs*LCA was 3.6 and 2.0 times faster than sam2lca and sam2rma, respectively (Figure 2d). Although sam2rma was marginally faster on the file with least alignments than *ngs*LCA (256 s versus 302 s for SPL_153_7001), the consumed time increased dramatically for sam2rma when analysing the big files. It took *ngs*LCA 2085 s to process the merged BAM file (all 10 tested BAM files combined), which was 4.4 and 4.2 times faster than sam2lca (9195 s) and sam2rma (8775 s) respectively. Considering the time consumed for indexing or converting the BAM files into the required format for sam2lca and sam2rma, *ngs*LCA provides the fastest LCA inference in real-world applications.

Without RAM limitations, we found all the three programs used approximately the same amount of RAM on average (Figure 2f). However, for the merged BAM file, *ngs*LCA used only 27.16% and 41.72% RAM compared to sam2lca and sam2rma, respectively. This again indicates that *ngs*LCA is superior in handling large datasets.

4 | BENCHMARK TEST

To assess the accuracy and specificity of taxonomic classification using the LCA approach supplied by *ngs*LCA, we simulated an ancient metagenomic dataset and tested how different *ngs*LCA parameters can affect the taxonomic profiles. First, we randomly selected 40 genomes from the NCBI genome-Refseq, including 19 bacteria (two genomes of *Escherichia coli* were included), 5 fungi, 3 invertebrate animals, 8 plants, and 4 vertebrate animals (Table 2). We then simulated divergence by applying mutation-simulator (Kuhl et al., 2021) to add mutations at a rate of 0.001. Thereafter single-end sequencing reads were generated using gargammel (Renaud et al., 2017) from

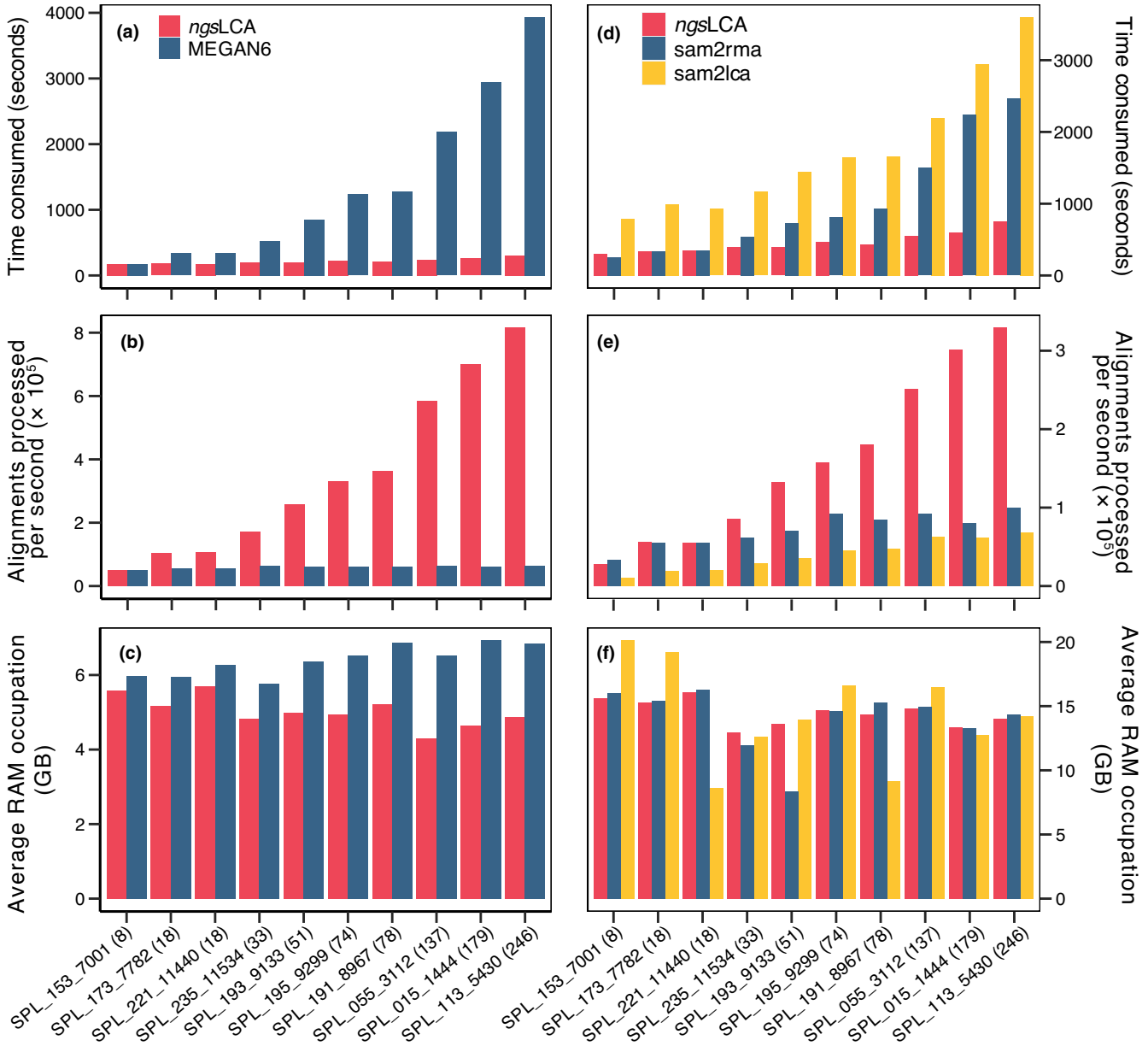

FIGURE 2 Comparisons of run time, alignments processed per second, and RAM usage between *ngs*LCA and MEGAN6 (a–c) on a laptop, and among *ngs*LCA, sam2rma, and sam2lca on a HPC cloud (d–f). Alignments of 10 shotgun sequenced dataset from (Pedersen et al., 2016) against NCBI databases were parsed by the four tested programs using the same parameters. The number in brackets after each sample name indicates the number of alignments (in millions) per dataset.

each genome, by adding (i) fragmentations (following suggested ancient DNA patterns in Fu et al., 2014), (ii) deamination damage patterns (–damagee 0.03, 0.4, 0.01, 0.3), and (iii) sequencing errors (–qs 30). We then randomly subset 1–500 simulated reads (reads number for each genome is shown in Table 2) from each genome and combined them into one FASTQ file as the simulated dataset. The dataset was mapped against the NCBI-nt (downloaded on 2 January 2022), NCBI-Refseq (all divisions of release 210), and the 40 template genomes, following the same method used in the comparison test (Section 3). The generated BAM file was processed by *ngs*LCA applying three different edit distances (i.e. mismatches between the quarry read and the reference genome) during the main program processing, nine different abundance filtering thresholds (i.e. threshold.1 in *ngsLCA_filter* of the R package), and three different taxonomic unit levels (species, genus and family). Results are shown in Figure 3.

We found that classification at the genus level is more reliable than at species level, particularly when applying more relaxed abundance filtering. For example, when setting the minimum reads number required for confirming a taxon to 0, 2 or 5, averagely there were 6.9 false-positive identifications at species level, while only 2 were reported at genus level. We also found that read abundance is key when filtering out false-positives, particularly for the plant and animal taxa. By increasing the filtering threshold to 5, only two

TABLE 2 Genomes used for simulating the benchmark testing dataset

Genome accession ID	Taxa	NCBI taxaID	Simulated reads number
GCF_000091205.1	*Cyanidioschyzon merolae*	45,157	88
GCF_000149035.1	*Colletotrichum graminicola*	31,870	28
GCF_000150705.2	*Paracoccidioides lutzii*	1,048,829	467
GCF_000184155.1	*Fragaria vesca*	57,918	257
GCF_000250985.1	*Nematocida parisii*	586,133	104
GCF_000340665.1	*Cajanus cajan*	3821	300
GCF_000346465.2	*Prunus persica*	3760	332
GCF_000458825.1	*Escherichia coli*	562	52
GCF_000516155.1	*Acinetobacter baumannii*	470	383
GCF_000622305.1	*Nannospalax galili*	1,026,970	262
GCF_000686985.2	*Brassica napus*	3708	50
GCF_000717135.1	*Catenuloplanes japonicus*	33,876	100
GCF_000740945.1	*Brucella abortus*	235	13
GCF_000741045.1	*Vigna radiata*	157,791	427
GCF_000955945.1	*Cercocebus atys*	9531	441
GCF_000956235.1	*Wasmannia auropunctata*	64,793	346
GCF_001186125.1	*Sphaeroforma arctica*	72,019	309
GCF_001254135.1	*Shigella sonnei*	624	350
GCF_001590865.1	*Streptomyces* sp. NBRC 110611	1,621,259	187
GCF_001608675.1	*Vibrio parahaemolyticus*	670	344
GCF_001622625.1	*Staphylococcus epidermidis*	1282	84
GCF_001815405.1	*Peptoniphilus* sp. HMSC075B08	1,739,525	28
GCF_002096655.1	*Streptococcus oralis*	1303	358
GCF_002139215.1	*Campylobacter lanienae*	75,658	300
GCF_002303985.1	*Durio zibethinus*	66,656	281
GCF_003225955.1	*Enterobacter cloacae complex* sp.	2,027,919	379
GCF_003357145.1	*Venustampulla echinocandica*	2,656,787	44
GCF_003479435.1	*Eubacterium* sp. AF15-50	2,293,103	367
GCF_003586665.1	*Listeria monocytogenes*	1639	138
GCF_003713205.1	*Coffea eugenioides*	49,369	184
GCF_010015585.1	*Eremomyces bilateralis*	1,341,166	370
GCF_013073365.1	*Escherichia coli*	562	437
GCF_013340165.1	*Drosophila suzukii*	28,584	135
GCF_014824575.2	*Sturnira hondurensis*	192,404	299
GCF_016077325.2	*Equus asinus*	9793	90
GCF_016757015.1	*Marinobacter* sp. JB05H06	2,803,860	17
GCF_022811885.1	*Priestia megaterium*	1404	73
GCF_900136575.1	*Mycobacteroides abscessus*	36,809	68
GCF_900984035.1	*Streptococcus pyogenes*	1314	280
GCF_902162445.1	*Klebsiella michiganensis*	1,134,687	401

non-plant/animal false-positive identifications remained, while all the true taxa could be correctly identified, indicating that the *ngs*LCA approach is very reliable. We also noticed that more flexible mapping (i.e. greater edit distance) slightly improved the identification accuracy, likely because false-positives produced by reads with mismatches to a true reference genome are now able to be mapped onto the correct reference genome, and therefore will be classified appropriately by the LCA algorithm.

Overall, by classifying at genus level and requiring the read abundance greater than 5 for a given taxon, *ngs*LCA supplies

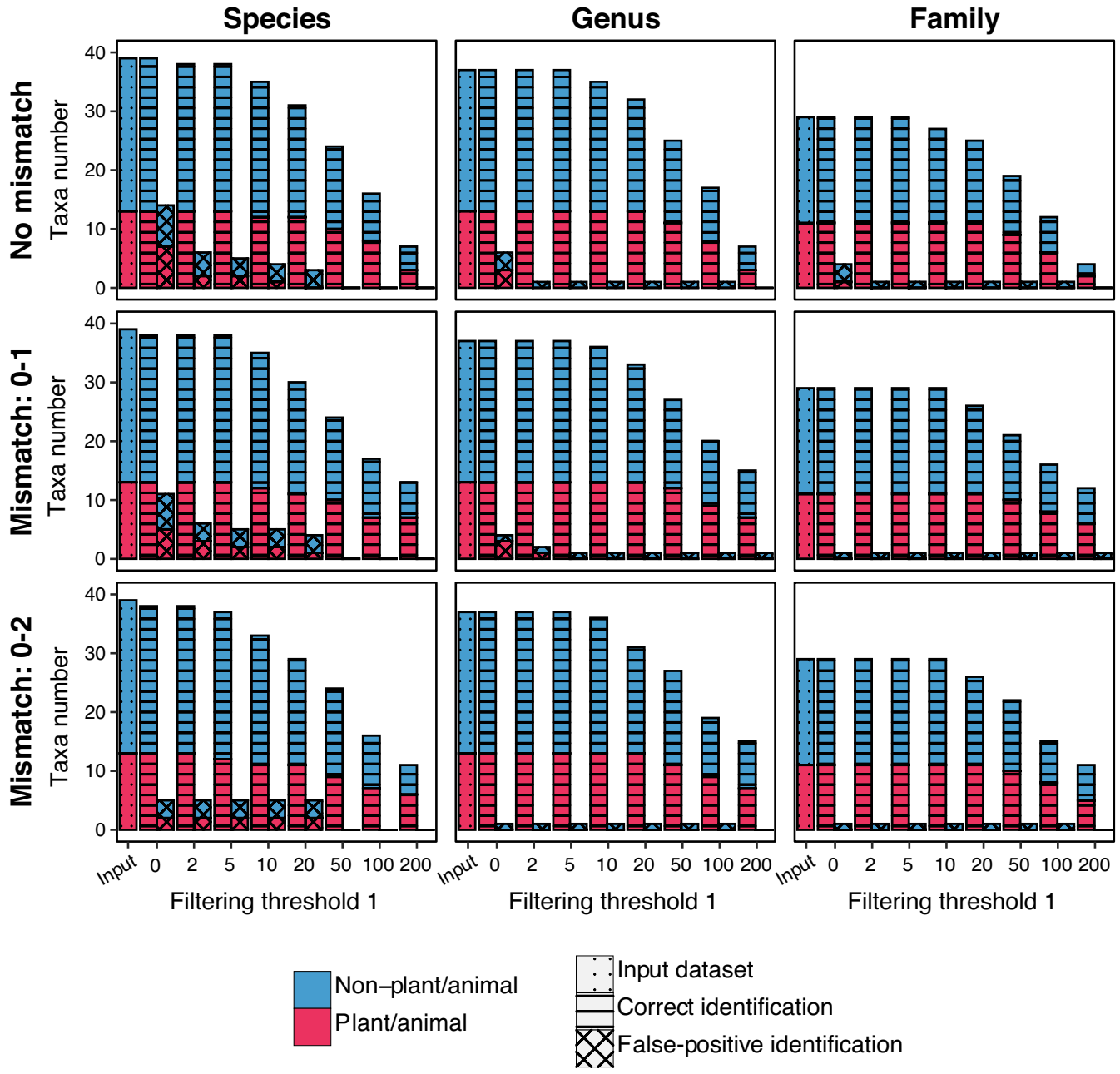

FIGURE 3 Accuracy and specificity of *ngs*LCA on a simulated ancient metagenomic dataset. A simulated dataset with known taxa composition with divergence, DNA fragmentation, deamination damage patterns, and sequencing errors was parsed by *ngs*LCA. The identified taxa are categorised into two different groups for each run: Plant/animal versus non-plant/animal and correct identifications versus false-positives. *X*-axis indicates the minimum reads number for authenticating the identification of a taxon.

accurate taxonomic classifications with all taxa correctly identified (i.e. no false-negative) and only 2 non-plant/animal false-positive identifications, for a dataset that has high divergence rate and is highly fragmented and damaged. This recommendation is not, however, universally applicable, as appropriate parameters can vary according to the sampling context and environment, DNA extraction and sequencing protocols, sequencing data quality controls, and importantly the reference database quality and completeness.

5 | CONCLUSION

*ngs*LCA is a fast, flexible, accurate and easy-to-use toolkit for lowest common ancestor inference and taxonomic profiling of metagenomic datasets. It contains 2 modules, the *ngs*LCA main program and the *ngs*LCA R package. The main program supplies a new implementation of the naive LCA algorithm which enables fast LCA inference, particularly for large datasets. A comparison test shows that it is more than ~11 times faster for large datasets and consumes less

memory than the commonly used MEGAN6, and ~4 times faster for large datasets than other available tools. By benchmarking on a simulated ancient metagenomic dataset, we show that *ngs*LCA supplies a very reliable approach for taxonomic classification with all true taxa correctly identified and only two prokaryote false positives. *ngs*LCA handles alignment in BAM/SAM/CRAM format derived from a vast range of sequenced material and is therefore compatible with a large number of bioinformatic analyses. It is computationally light, enabling the entire workflow to be carried out on a laptop. *ngs*LCA uses the NCBI taxonomy, for which the newest release can be downloaded directly from the NCBI ftp server without reformatting (reformatting is typically time consuming due to the large size). This makes maintenance and database updates simple and enables compatibility with the latest version of the NCBI genomic databases. *ngs*LCA is built upon the readily available NCBI taxonomic phylogeny, which makes it easy to add additional genomes not included in the NCBI databases. The *ngs*LCA R package supplies a series of functions for easy and fast taxonomic profiling, as well as filtering, decontaminating, sorting, grouping, illustrating, and clustering the taxonomic profiles, which collectively is of broad use to molecular ecologists and biologists.

With the rapid rise in metagenomic studies and the constant increase in the size of metagenomic data and genomic databases, metagenomic tools that improve speed and handling capabilities, such as *ngs*LCA, are necessary to maintain the current pace and scope of discovery.

AUTHOR CONTRIBUTIONS

Mikkel Winther Pedersen initiated and led the project. Thorfinn Sand Korneliussen implemented the LCA algorithm and developed the main program. Yucheng Wang developed the R package; performed the benchmark test; and drafted the manuscript with inputs from all other authors.

ACKNOWLEDGEMENTS

Y.W., M.W.P. and T.S.K. were funded by the Carlsberg Foundation (CF16-0728, CF16-0913, CF18-0024, and CF19-0712). L.H. was supported by the Natural Environmental Research Council (NE/L002531/1). We thank Eske Willerslev and Antonio Fernandez-Guerra for the discussions and inputs on this project.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.14006.

DATA AVAILABILITY STATEMENT

*ngs*LCA toolkit, as well as the document and tutorial are freely and publicly available at the GitHub repository (https://github.com/miwipe/ngsLCA) and Zenodo (Wang et al., 2022). Codes for the benchmark test are also archived at the GitHub repository under the folder "benchmark test". NCBI taxonomy is available at https://www.ncbi.nlm.nih.gov/taxonomy. The metagenomic dataset used for benchmark test can be downloaded from the European Nucleotide Archive under accession number PRJEB14494 and sample accessions SAMEA4374745–SAMEA4374754. The NCBI reference databases used for benchmark test are available at NCBI ftp servers: NCBI-nt (https://ftp.ncbi.nlm.nih.gov/blast/db/) and NCBI-RefSeq (https://ftp.ncbi.nih.gov/refseq/).

ORCID

Yucheng Wang https://orcid.org/0000-0002-7838-226X
Thorfinn Sand Korneliussen https://orcid.org/0000-0001-7576-5380
Luke E. Holman https://orcid.org/0000-0002-8139-3760
Andrea Manica https://orcid.org/0000-0003-1895-450X
Mikkel Winther Pedersen https://orcid.org/0000-0002-7291-8887

REFERENCES

Almodaresi, F., Zakeri, M., & Patro, R. (2021). Puffaligner: A fast, efficient, and accurate aligner based on the pufferfish index. *Bioinformatics*, *37*, 4048–4055.

Barer, M. R., & Harwood, C. R. (1999). Bacterial viability and culturability. *Advances in Microbial Physiology*, *41*, 93–137.

Beghini, F., McIver, L. J., Blanco-Miguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, *10*, e65088.

Bell, K. L., Petit, R. A., 3rd, Cutler, A., Dobbs, E. K., Macpherson, J. M., Read, T. D., Burgess, K. S., & Brosi, B. J. (2021). Comparing whole-genome shotgun sequencing and DNA metabarcoding approaches for species identification and quantification of pollen species mixtures. *Ecology and Evolution*, *11*, 16082–16098.

Borry, M., Hübner, A., & Warinner, C. (2022). sam2lca: Lowest common ancestor for SAM/BAM/CRAM alignment files. *Journal of Open Source Software*, *7*, 4360.

Breitwieser, F. P., Baker, D. N., & Salzberg, S. L. (2018). KrakenUniq: Confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, *19*, 198.

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*, 366–368.

Chua, P. Y. S., Crampton-Platt, A., Lammers, Y., Alsos, I. G., Boessenkool, S., & Bohmann, K. (2021). Metagenomics: A viable tool for reconstructing herbivore diet. *Molecular Ecology Resources*, *21*, 2249–2263.

Cowart, D. A., Murphy, K. R., & Cheng, C. C. (2018). Metagenomic sequencing of environmental DNA reveals marine faunal assemblages from the West Antarctic peninsula. *Marine Genomics*, *37*, 148–160.

Cribdon, B., Ware, R., Smith, O., Gaffney, V., & Allaby, R. G. (2020). PIA: More accurate taxonomic assignment of metagenomic data demonstrated on sedaDNA from the North Sea. *Frontiers in Ecology and Evolution*, *8*, 84.

Deiner, K., Bik, H. M., Machler, E., Seymour, M., Lacoursiere-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*, 5872–5895.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, *14*, 927–930.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., Johnson, P. L., Aximu-Petri, A., Prufer, K., de Filippo, C., Meyer, M., Zwyns, N., Salazar-Garcia, D. C., Kuzmin, Y. V., Keates, S. G., Kosintsev, P. A., Razhev, D. I., Richards, M. P., Peristov, N. V., … Paabo, S. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, *514*, 445–449.

Garlapati, D., Charankumar, B., Ramu, K., Madeswaran, P., & Ramana Murthy, M. V. (2019). A review on the applications and recent advances in environmental DNA (eDNA) metagenomics. *Reviews in Environmental Science and Bio/Technology*, *18*, 389–411.

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*, 2847–2849.

Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., & Huson, D. H. (2016). MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean iceman. *bioRxiv*, 050559.

Huson, D. H., Beier, S., Flade, I., Gorska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H. J., & Tappu, R. (2016). MEGAN Community edition—Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computational Biology*, *12*, e1004957.

Jackson, S. T., Overpeck, J. T., Webb, T., Keattch, S. E., & Anderson, K. H. (1997). Mapped plant-macrofossil and pollen records of late quaternary vegetation change in eastern North America. *Quaternary Science Reviews*, *16*, 1–70.

Jensen, M. R., Sigsgaard, E. E., Liu, S., Manica, A., Bach, S. S., Hansen, M. M., Moller, P. R., & Thomsen, P. F. (2021). Genome-scale target capture of mitochondrial and nuclear environmental DNA from water samples. *Molecular Ecology Resources*, *21*, 690–702.

Kuhl, M. A., Stich, B., & Ries, D. C. (2021). Mutation-simulator: Fine-grained simulation of random mutations in any genome. *Bioinformatics*, *37*, 568–569.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature Communications*, *7*, 11257.

Muller, E. (1997). Mapping riparian vegetation along rivers: Old concepts and new methods. *Aquatic Botany*, *58*, 411–437.

Pedersen, M. W., De Sanctis, B., Saremi, N. F., Sikora, M., Puckett, E. E., Gu, Z., Moon, K. L., Kapp, J. D., Vinner, L., Vardanyan, Z., Ardelean, C. F., Arroyo-Cabrales, J., Cahill, J. A., Heintzman, P. D., Zazula, G., MacPhee, R. D. E., Shapiro, B., Durbin, R., & Willerslev, E. (2021). Environmental genomics of late Pleistocene black bears and giant short-faced bears. *Current Biology*, *31*, 2728–2736.e8.

Pedersen, M. W., Ruter, A., Schweger, C., Friebe, H., Staff, R. A., Kjeldsen, K. K., Mendoza, M. L., Beaudoin, A. B., Zutter, C., Larsen, N. K., Potter, B. A., Nielsen, R., Rainville, R. A., Orlando, L., Meltzer, D. J., Kjaer, K. H., & Willerslev, E. (2016). Postglacial viability and colonization in North America's ice-free corridor. *Nature*, *537*, 45–49.

Renaud, G., Hanghoj, K., Willerslev, E., & Orlando, L. (2017). gargammel: A sequence simulator for ancient DNA. *Bioinformatics*, *33*, 577–579.

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, *17*, e00547.

Ryser-Degiorgis, M. P. (2013). Wildlife health investigations: Needs, challenges and recommendations. *BMC Veterinary Research*, *9*, 223.

Steinegger, M., & Salzberg, S. L. (2020). Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biology*, *21*, 115.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA—An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4–18.

Vernot, B., Zavala, E. I., Gomez-Olivencia, A., Jacobs, Z., Slon, V., Mafessoni, F., Romagne, F., Pearson, A., Petr, M., Sala, N., Pablos, A., Aranburu, A., de Castro, J. M. B., Carbonell, E., Li, B., Krajcarz, M. T., Krivoshapkin, A. I., Kolobova, K. A., Kozlikin, M. B., … Meyer, M. (2021). Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments. *Science*, *372*, eabf1667.

Wang, Y., Korneliussen, T. S., Holman, L. E., Manica, A., & Pedersen, M. W. (2022). miwipe/ngsLCA: v1.0.5 (v1.0.5). *Zenodo*.

Wang, Y., Pedersen, M. W., Alsos, I. G., De Sanctis, B., Racimo, F., Prohaska, A., Coissac, E., Owens, H. L., Merkel, M. K. F., Fernandez-Guerra, A., Rouillard, A., Lammers, Y., Alberti, A., Denoeud, F., Money, D., Ruter, A. H., McColl, H., Larsen, N. K., Cherezova, A. A., … Willerslev, E. (2021). Late quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature*, *600*, 86–92.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biology*, *20*, 257.

How to cite this article: Wang, Y., Korneliussen, T. S., Holman, L. E., Manica, A., & Pedersen, M. W. (2022). *ngs*LCA—A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data. *Methods in Ecology and Evolution*, *00*, 1–10. https://doi.org/10.1111/2041-210X.14006