Metagenomics analysis for microbial ecology investigation on historical samples: negligible effect of host DNA and optimal analysis strategies

Siu-Kin Ng[1,*], Rafal M. Gutaker[1,*]

1 – Royal Botanic Gardens, Kew, Richmond, Surrey, United Kingdom

* Corresponding author

Abstract

Microbiome composition and function are strongly influenced by environmental factors, with major shifts driven by intensified anthropogenic pressures over the past centuries. This timeframe extends beyond the scope of traditional experimental or longitudinal studies commonly used to investigate microbiome dynamics. The historical samples might provide important insights into the mechanistic consequences of anthropogenic pressures and the lost microbial diversity prior to the shifts. Despite their vast potential, historical samples available in museums and herbaria worldwide remain underutilized for exploring host-microbiome interactions across broad temporal and spatial scales due to incompatibilities with standard analytical pipelines and limited understanding of optimal classification parameters. While host DNA removal has conventionally been considered essential for accurate taxonomic assignment of metagenomic reads, and might be of particular importance when processing degraded DNA, this step is impractical for many samples because the reference genomes of host species are often unavailable. Here, we show that host DNA content does not significantly impact microbial ecology analyses in simulation and published datasets. Since, DNA molecules from historical samples are highly fragmented and uneven in length, we additionally analysed the impact of k-mer size on the accuracy of metagenomic assignments in historical samples. We propose a simple two-step approach in which reads are classified with two annotation databases constructed with a long and a short k-mer values. Through a simulation study, we demonstrated that this approach outperforms conventional workflows in effectively recovering microbial signals from a wide range of read lengths. Together, this study provides a solid foundation for

incorporating natural history collections into host-associated microbiome research, offering valuable insights into the long-term effects of anthropogenic change on microbial communities.

Introduction

The microbiome refers to a collection of microorganisms, which comprise bacteria, archaea, viruses, microbial eukaryotes, and the surrounding environmental conditions (Marchesi and Ravel 2015). Emerging evidence suggests the importance of resident microbiome in enhancing the fitness of host species, supporting adaptation to internal and external changes (Ley, Hamady et al. 2008, Xiao, Liang et al. 2022, Debray, Conover et al. 2023, Petersen, Hamerich et al. 2023). For example, the soil microbiome contributes to plant growth through biogeochemical cycling of nutrients and organic matter (Jansson and Hofmockel 2020, Banerjee and van der Heijden 2023), the endophytic microbiome contributes to plant survival under drought (Naylor and Coleman-Derr 2017, Trivedi, Leach et al. 2020) and the gut microbiome to the fitness and lifespan of drosophila (Gould, Zhang et al. 2018).

The field of microbial ecology aims to identify and characterize microbial community composition, function and their interactions with changing environment. Samples from controlled experiments (e.g. conventional case-control or longitudinal cross-section studies) are often sufficient to provide useful insights into the effects of short-term perturbations, such as dietary or seasonal climate shifts. The study of long-term disturbances or environmental changes (Tiao, Lee et al. 2012, Gancz, Farrer et al. 2023, Rout, Tripathy et al. 2023, Morais, Winkler et al. 2024), however, cannot be attained with these samples. Historical samples including those in herbarium or museum collections, have been preserved for centuries and offer an exclusive opportunity to examine the effects of long-term disturbance on host and their associated microbiome using metagenomic approaches (Warinner, Herbig et al. 2017, Weyrich, Duchene et

al. 2017). Examples of fundamental changes in the last two centuries include increased antimicrobial resistance in microbes in response to introduction of antibiotics, and decrease in soil nitrogen fixing activity in response to widespread application of synthetic fertilizers (Naamala, Jaiswal et al. 2016, Aeron, Khare et al. 2020, Rout, Dixit et al. 2024). Historical samples provide a global and extended temporal perspective that conventional studies simply cannot achieve (Burbano and Gutaker 2023). They could be an important source to uncover the "native" microbes predated major anthropogenic activities.

Host DNA and low endogenous microbial biomass (Eisenhofer, Minich et al. 2019) make metagenomic analysis with historical samples challenging. Additionally, nucleic acid material preserved in historical samples is prone to long-term degradation and environmental contamination. The impact of host DNA content on taxonomic classification of microbial communities has recently been investigated (Pereira-Marques et al. 2019) and host DNA removal prior to data analysis has been seen as a necessary step. Several commercial DNA extraction kits are designed to perform differential lysis and digestion of host cells to deplete their DNA, however these kits are only effective for human or animal host cells but not for plant nor host cells with complex cell wall structure (either fresh or historical). Alternatively, removal of host DNA can be performed *in silico* with the reference genomes of host species or closely related species. The breadth of microbe-host-environment interactions is best studied with diverse species including non-model organisms (e.g. diverse angiosperm plants (Zuntini, Carruthers et al. 2024)), where representative reference genomes are often absent.

Several critical questions regarding microbial analyses in historical samples remain unaddressed. First, is host DNA removal necessary when profiling the host-associated microbiome in historical samples, particularly those from non-model organisms? Although host DNA content may reduce the sensitivity for detecting rare taxa (Pereira-Marques et al. 2019), it remains unclear whether it significantly impacts microbiome analyses such as microbial diversity metrics. Second, the nucleic acids (either from host or microbes) preserving in historical samples are typically highly fragmented due to degradation processes such as cytosine deamination (Gutaker and Burbano 2017). Previous studies have developed methods to recover very short DNA fragments (e.g. 25bp) from historical samples (Glocke and Meyer 2017, Harney, Cheronet et al. 2021) but it remains unclear how reliable result these short metagenomic reads could provide. Without removing host DNA, short reads derived from host may be misclassified as microbial origin or vice versa. However, discarding these short reads from downstream analyses is not ideal, particularly when working with precious museum or herbarium collections. Microbiome profiling often relies on alignment-free tools such as Kraken2 (Lu, Rincon et al. 2022), where classification sensitivity and specificity are influenced by parameters such as k-mer size or seed length. By fine-tuning these parameters, short reads can still contribute to important insight. In this study, we explored the impact of host DNA on host-associated microbiomes analyses using both simulations and published datasets from historical collections. We also evaluated parameters and strategies for analyzing the metagenomic reads obtained from samples lacking host reference genomes.

Materials and Methods

Sampling, sequencing and collection of museum samples

A total of 864 metagenomes of contemporary and historical specimens was included in this study to investigate the impact of host DNA content on microbial analyses and to evaluate appropriate analytical strategies. For the contemporary samples, 330 publicly available metagenomes of leaf specimens from 110 rice cultivars across multiple time points with bacterial enrichment achieved through multiple washing steps were analyzed (Su, Wicaksono et al. 2022). For the historical samples, a total of 534 metagenomes were included, comprising 90 Asian rice specimens (Alam, Gutaker et al. 2025), 37 common ragweed (Bieker, Battlay et al. 2022), and 402 bumblebee (Mullin, Stephen et al. 2023). In addition, five phyllosphere metagenomes from herbarium rice leaf were generated in this study, and the raw sequencing data were deposited in the EBI nucleotide archive (PRJEB87646). Raw sequencing data for the public datasets were retrieved from the NCBI Sequence Read Archive (SRA): contemporary rice phyllosphere metagenomes (PRJEB45634), historical common ragweed (PRJEB48563), Asian rice (PRJNA1282210) and museum bumblebee (PRJEB52125). Genome assemblies for the host species were obtained from the NCBI GenBank or the EBI ENA: rice (*Oryza sativa* AGIS1.0; GCA_034140825.1 for reference genome; MZ424305.1 for chloroplast), common ragweed (*Ambrosia artemisiifolia*; GCA_024762085.1 for reference genome; MG019037.1 for chloroplast) and bumblebee assembly (*Bombus lapidaries; Bombus_lapidarius_EIv1*; PRJEB51891). In the simulation study, the catalogue of soil metagenome-assembled genomes (SMAGs; PRJNA983538) was used to compare the unique and shared k-mer sequences between

host species and common soil microbes. The SMAGs catalogue contains 21,077 genomic bins reconstructed from 3,304 soil metagenomes and resolved at the species level (Ma, Lu et al. 2023). The datasets and genomic sequences included in this study were summarized in Supplementary Table 1.

Collection and processing of herbarium rice samples

Metagenomics of five rice herbarium specimens were generated in this study. The rice specimens were collected between 1906 and 1914 in Philippines and stored at the herbarium collection of Smithsonian Institution. Briefly, a small fragment of leaf (approximately 20mm x 5mm) was removed and processed in the semi-clean room facilities (without air filtering) at the New York University Centre for Genomics and Systems Biology. Standard precautions for handling historical and ancient DNA (aDNA) were taken (disposable PPE, use of dedicated safety cabinets, frequent cleaning regimes and UV-sterilization). Sampled tissue was ground 'wet' with PTB lysis buffer in 37°C. Subsequently, DNA isolation and genomic library preparation protocols (without uracil-DNA glycosylase treatment) for historical plant tissues were adopted from our previous work (Latorre, Lang et al. 2020) with the following modifications: indexing PCR was carried out with Agilent Pfu Cx polymerase. Libraries were amplified with 10 polymerase chain reaction cycles. quantified using RT-qPCR, normalized and pooled to equimolar concentrations. Libraries were then sent to Macrogen for sequencing on a single lane of S4 flow cell with NovaSeq6000 instrument and demultiplexed.

Preprocessing of metagenomic data

Metagenomic analysis of the raw sequencing data was carried out with the nf-core/mag pipeline, an open-source and community curated best-practice bioinformatics pipeline for preprocessing, assembly, binning and annotation of metagenomes based on the Nextflow nf-core framework (Ewels, Peltzer et al. 2020, Krakau, Straub et al. 2022). Quality control of the raw metagenomic sequencing reads in FASTQ format was performed using FastQC v.0.12.0 (Andrews 2015). Adapters and low-quality reads were trimmed, and overlapping paired-end sequences were merged with an all-in-one fastp preprocessing tool using default settings (Chen, Zhou et al. 2018). Reads mapped to PhiX sequence using Bowtie2 v.2.3.4 were discarded (Langmead and Salzberg 2012).

Removal of host DNA

We evaluated the impact of host DNA content on metagenomic analyses by comparing the microbial community metrics in the same samples processed with (i) host DNA removal or (ii) without this step (Figure 1). For the host removal step (i), merged reads were aligned to host genome using Bowtie2 with default settings. Reads that did not align to the host genome were proceeded with taxonomic classification. For the workflow without host removal step, merged reads were taxonomically annotated immediately after raw read preprocessing.

Taxonomic classification

Taxonomic classification of metagenomic reads was performed using Kraken2 v2.0.9 (Wood, Lu et al. 2019) with a confidence value of 0.6 against the PlusPFP database (prebuilt with k= 35, version 2024.06.05), which contains archaea, bacteria, viral, plasmid, human, UniVec_Core along with protozoa, fungi and plants. After evaluating the confidence values from 0.0 to 1.0 (data not shown), we found that 0.4-0.6 could provide a good balance between classification rate and false positive. Kraken2 classifies metagenomic reads by mapping all k-mers to the lowest common ancestor (LCA) of all genomes present in the database. Following Kraken2 analysis, Bracken (Bayesian Re-estimation of Abundance with Kraken) was used to estimate the taxon abundances at the species level (Lu, Breitwieser et al. 2017). The taxonomic lineage information for each operational taxonomic unit (OTU) was processed and converted to TSV format with kraken-biom (Dabdoub 2016). Non-microbial taxa (e.g. those assigned to eukaryotic species or viruses) were excluded from further microbiome analyses. OTUs were further filtered using a minimum read count of 50 (100 was used for contemporary rice samples due to their higher average read depth) and a prevalence cutoff of 2% of sample.

K-mer analysis of host and resident genomes

To evaluate the similarity in genomic sequences between host and associated microbiome, k-mer analysis of genomic sequences was performed with JellyFish (Marcais and

Kingsford 2011). For the soil microbiome, a single representative taxon was shortlisted for each of 107 soil microbe pangenomes in the catalogue of soil metagenome-assembled genomes (SMAGs) by assessing their genome completeness (the number of expected single-copy marker genes typically detected in a microbial genome assembly) and contamination (the percentage of redundant single-copy marker genes) using CheckM (Parks, Imelfort et al. 2015). The 107 pangenomes were generated from 2,200 species-level genomic bins with >10 high quality metagenome-assembled genomes (MAGs) by clustering conspecific protein sequences at 90% amino acid identity (Ma, Lu et al. 2023). For the representative microbes with a reference genome available on NCBI (sequence accession with a GCF_* or GCA_* prefix), we retrieved the sequences from the NCBI GenBank using "download" command of the NCBI Datasets command-line tools. Counting of k-mers shared between host and microbial genome sequences was carried out with JellyFish "count" command in 5 different k-mer size (11, 15, 18, 21 and 24). K-mer size higher than 24 was not evaluated as the analysis of large k-mer values using JellyFish was computationally expensive given the computational resources available to us. Also, we observed that the number of shared k-mers was decreased dramatically if k-mer size increases larger than 21. The conclusion remains valid even though larger k-mer values are not evaluated. K-mer substrings were exported to FASTA format with JellyFish "dump" command. Overlap between host and any microbe, which is the proportion of shared k-mer substrings, was measured using Jaccard distance, $Jaccard(host, microbe) = \frac{|host \cap microbe|}{|host \cup microbe|}$. Sequence complexity of k-mers was estimated as Wootton-Federhen complexity score using "sequence_complexity" function of the R package universalmotif (Tremblay 2024). Selected k-mer size was used as window size, e.g. window.size=15 for k=15.

Construction and analyses of simulated ancient DNA datasets

No *a priori* information on microbial content in the historical samples is available. To evaluate the effect of search parameters such as k-mer values, complexity of host genome and its DNA percentage in a sample, we performed a simulation study with datasets generated using gargammel, a NGS read simulator for aDNA (Renaud, Hanghøj et al. 2017). The simulated dataset comprises of 3 components: endogenous DNA (***endo***), bacterial (***bact***) and contamination (***cont***). We considered 3 plant host species to evaluate the impact of host genome size and complexity (Figure 5): rice (*Oryza sativa* AGIS1.0), maize (*Zea mays*; GCF_902167145) and bread wheat (*Triticum aestivum*; Genome assembly IWGSC CS RefSeq v2.1; chromosomes only). For each host species, a custom sequence set that consists of host (***endo***), soil microbes (***bact***), and contaminant skin microbes commonly observed in museum collections and human (***cont***) was compiled as the input sequences to gargammel. For the soil microbes, 120 bacteria were shortlisted (Supplementary Table 6) to represent the core soil microbes in the SoilGenomeDB database (Edwin, Fitzpatrick et al. 2024). Human genomes (GRCh38.p14), *Cutibacterium acnes* (strain NBRC 107605; NZ_AP019723) and *Staphylococcus epidermidis* (strain ATCC 14990; NZ_CP035288) were included as the contemporary contamination. For each set of reference sequences, we created 100 input sequence sets by randomly assigning abundance values to 107 soil microbes (Supplementary Figure 7). For each input sequence set, 20 simulated datasets consisting of 1,000,000 paired-end reads with a read length of 75 bp (HiSeq 2500) were iteratively generated from different proportions of input sequence components ***bact*** (0.01, 0.05, 0.1, 0.3 and 0.45), ***cont*** (0.01, 0.05, 0.3 and 0.45) and ***endo*** (1.0 –

(***bact*** + ***cont***)) using gargammel. A fragment length profile of a herbarium rice sample was used to generate the length distribution of simulated reads. The damage pattern settings (nick frequency *v*=0.03; length of overhanging ends *l*=0.4; probability of deamination of Cs in double-stranded parts *d*=0.01; and probability of deamination of Cs in single-stranded parts *s*=0.3) were adopted from the DNA isolated from Pleistocene organisms (Briggs, Stenzel et al. 2007). These settings were used to simulate the damage patterns for 10,000 year or older samples and unlikely to be the case for herbarium samples (100-200 years old typically) but we adopted these settings to generate datasets in a worst-case scenario. The fragment length profile was obtained by profiling the length of merged reads from one of the herbarium rice samples included in this study (HRB0263L) using readlength.sh from BBMap package (Bushnell 2014) with the settings (bin=1 and n=1,000,000). In total, 6,000 simulated datasets were generated from 3 sets of custom sequences (rice, maize and wheat). To investigate the effect of k-mer size, the 3 sets of custom sequences were used to build reference databases with k-mer sizes of 18, 21, 24, 28, 31 and 35 using kraken2-build command (k=11 and 15 were found not informative based on the analysis in the previous section and therefore not included in the analyses in this section). The simulated datasets were processed and analyzed with the same workflow as described in the previous section. The classification results were also examined with the exact k-mer search algorithm Krakenuniq version 0.5.8 (Breitwieser, Baker et al. 2018). To evaluate the misclassification rate of host DNA, taxonomy ID assigned to each read was obtained by parsing the header lines of classified reads exported by kraken2. Any host derived reads classified to any prokaryotic clads were considered as misclassified. The reads assigned with a taxonomy ID at species or finer level (e.g. subspecies or strain level) were considered as species-level resolved.

Statistical analysis

All statistical methods and visualization were performed in R v.4.2.2 with the corresponding R-packages (phyloseq, vegan, ggplot2). Before diversity analyses, samples were normalized by rarefying (downsampling) their sequencing depth to the minimum value of bacterial read sums in their respective dataset (contemporary rice=60,713; rice herbarium=5,099; museum ragweed=7,306; museum bumblebee=7,006). Samples with fewer than 5,000 reads were excluded from the diversity analysis. Alpha diversity (within-sample) diversity (Chao1 and Shannon index) was calculated using estimateR and diversity functions from the vegan package to provide an overview on the richness and diversity of microbial communities. Pairwise comparisons between the two pipelines with and without host genome removal were assessed with non-parametric Wilcoxon rank sum test with continuity correction. All *p*-values were adjusted with the conservative Holm-Bonferroni method. Beta diversity (between samples) was evaluated by analysis of similarities (ANOSIM) to test the difference in the composition of microbial community using anosim function (999 permutations) from the vegan package. Bray-Curtis distance metrics at the species level were used to perform two-dimensional principal coordinate analysis (PCoA) plots to visualize the differences among microbial communities. Rarefaction analysis was performed by subsampling read counts from the taxon table using rarefy_even_depth function from the phyloseq package. Taxon with a read count less than 10 was considered absent. The precision and recall of taxonomic classification for each read in the simulation datasets were evaluated by comparing the assigned taxonomy IDs with the true taxonomy IDs of the microbial species from which the reads were generated. Only species level

assignment were considered; reads assigned to higher taxonomic ranks (e.g. genus level) were treated as false negative (FN). Reads whose assigned taxonomy IDs did not match the true taxonomy IDs were classified as false positive (FP), while only those with correctly matched taxonomy ID were considered as true positive (TP). Precision was calculated as TP / (TP + FP), and recall as TP / (TP + FN). For the 2-stage classification approach, reads were treated as FP if they were assigned an incorrect (species-level) taxonomy ID during the first stage. Reads were considered as TP only if they received the correct taxonomy ID in the first stage, or if they were initially unassigned (FN) but correctly classified during the second stage. Part of the analyses in this study was conducted on CropDiversity-HPC (Percival-Alwyn 2024).

Results

Effect of host DNA removal on taxonomic assignments

Removal of host DNA is generally considered as necessary for the purpose of accurate microbial profiling. To evaluate the effects of host DNA removal on taxonomic classification, metagenomic datasets generated from contemporary and historical specimens of 3 different host species (rice, common ragweed and bumblebee) were included in this study (Supplementary Table 1). Each dataset was processed in two separate workflows in parallel: with host DNA removal step and without it (Figure 1). During host removal step, host-derived reads were removed by aligning them to the host reference genome. The reference genomes for rice and common ragweed were publicly available whereas genomic sequences of bumblebee were

released along with the metagenomic datasets (Supplementary Table 1). To evaluate the impact of host DNA removal on metagenomics analyses, we considered the number of reads, OTUs identified and microbial community diversity between the workflows.

Overall, the samples included in this study covered a wide spectrum of host DNA content (reads mapped to non-prokaryotic species), spanning from less than 0.1% in some of the contemporary rice samples and museum bumblebee to over 80% in some of herbarium rice samples (Supplementary Table 2 and Supplementary Figure 1). As expected, host DNA content was markedly reduced in subsequent downstream analyses when the samples were preprocessed with host DNA removal step (Supplementary Figure 1). On average, host derived reads (either from plant or metazoa) had as much as hundred-folds difference between the workflows. Meanwhile, the number of reads assigned to bacterial taxa was consistent (~1.0-fold difference between workflows), regardless of host DNA removal (Figure 2 and Supplementary Table 2-3).

The taxonomic assignment was performed using Kraken2 with the PlusPFP database (Figure 1). Except rice, genome sequences for other host species were not available in the database. The most closely related species in the database for common ragweed was common sunflower (*Helianthus annus*) while no comparative insect genomes for bumblebee dataset were included in the database (Supplementary Table 4). Detected OTUs were filtered based on their abundance and prevalence. Before filtering, the numbers of OTUs were between 4000-9000, regardless of host DNA removal (Table 1). Most of them were rare taxa with a relative abundance of 0.01% or below and removed in the subsequent filtering step (Supplementary Figure 2). Approximate 91-97% of OTUs were removed during the filtering step. For the

contemporary rice samples (3 replicates for each rice variety), same number of OTUs (390) were obtained in both workflows after filtering (a minimum read count of 100 and prevalence of 2%). Fifty-five (14.1%) of 390 OTUs was highly variable (more than 10% difference in read count) between workflows but their mean read count was significantly lower than the OTUs consistent between the two workflows (Supplementary Figure 2). For the historical samples, a significant portion of detected OTUs (>90%) were filtered out. Museum bumblebee had the lowest number of filtered OTUs (113) while 853 OTUs were retained in herbarium rice samples after filtering. The number and read count of filtered OTUs were highly consistent between the workflows. The inclusion of host DNA removal step only led to 0.5% difference in the number of filtered OTUs, of which almost all were low in abundance (mean read count), suggesting that these OTUs were rare (Table 1 and Supplementary Figure 2). Altogether, low abundance taxa were more prone to host DNA removal step.

Distinct taxonomic profiles were observed across different host-associated microbiota (Figure 3a and Supplementary Figure 3). The phylum Pseudomonadota was dominant in all datasets. Specifically, the proportion of genera with relative abundance lower than 2% were higher in the historical samples. Compared among contemporary and historical rice samples, taxa from the genus *Acinetobacter* were overrepresented in contemporary rice samples while a most diverse genera was observed in herbarium rice samples. Overall, relative abundance of taxa at phylum and genus level showed no marked difference between the workflows.

Effects of host DNA content on the metrics of microbial diversity

Host DNA might be misclassified as microbial origin (cross-domain), potentially leading to false detection of microbial taxa that do not exist in samples and affect diversity metrics. We sought to determine if host DNA removal step would alter the community diversity and richness metrics of the host-associated microbiome by measuring alpha and beta diversity. The rarefaction analysis indicated that there were no significant different in detected OTUs between workflows (Supplementary Figure 4). Alpha diversity was calculated using several metrics to account for different components of community diversity. Chao1 diversity metrics included an estimation of undetected species in the sample, and Shannon diversity index, which accounts for species presence/absence and evenness. For the contemporary rice samples, Chao1 diversity metrics were not significantly different (p>0.5) between the workflows with and without host removal, with average values of 59.3 ±23.9 and 57.7 ±23.4, respectively. A similar pattern was obtained for Shannon diversity index (with host removal=1.18 ±0.72 and without=1.17 ±0.72), where no significant differences (p=0.43) was observed (Figure 3a). The overall microbial diversity and community structure in historical, museum samples were highly consistent between workflows (Figure 3b-e).

To better understand the differences in the community diversity of the taxonomic profiles between the two workflows with and without host DNA removal, we assessed the differences in microbial community structure. Principal coordinate analysis (PCoA) plot of beta diversity for rice phyllosphere metagenome using Bray-Curtis dissimilarity index showed that host genome removal has minimally influenced the microbial community structure. Analysis of similarities (ANOSIM) showed no significant difference in microbial community structure between the

workflows (Figure 3). These results suggest that the community metrics were not sensitive to host DNA removal procedure.

K-mer analysis and optimal k-mer settings for studying historical samples

To evaluate the impacts of host DNA on downstream microbiome analysis, we decomposed the host genome and its associated microbiome into constituent k-mers, followed by assessing the proportion of overlap between these constituent k-mers (shared k-mers) using Jaccard index. This provided a theoretical assessment of the potential misclassification (the ratio of shared and distinct k-mers between host and microbes) of the host sequences as microbial signals purely by chance. The exact reference genomes were obtained for rice and ragweed while the genome of bumblebee generated from long-read sequencing were used. For the soil microbiome, representative soil microbes were compiled from either the pan-genomes in the catalogues of soil metagenome-assembled genome bins (SMAGs) or public available microbial reference genomes (microbial isolates). These pan-genomes were reconstructed from more than 3,000 metagenomes, suggesting that they robustly represent the core microbes in their respective biome. Distinct k-mers were estimated with two methods: (a) collective combination of all pan-genome MAGs, and (b) individually with each pan-genome MAG. The number of distinct k-mers depends on genome size and k-mer size (Supplementary Table 5). Rice genome is the smallest (391Mb) among the host genomes analyzed in this study. The number of distinct k-mers was 10-fold different between k=11 and k=24 (2,097,152 vs 282,788,287). In comparison, the number of distinct k-mers from the soil microbiome depends on the richness (the total genomic

size of all MAGs is 388Mb). Approximate 0.37 billion of distinct k-mers (k=24) were obtained from the soil microbiome (Supplementary Table 5).

The number of shared k-mers between host and its associated microbiome was measured using the Jaccard index, a metric that measures similarity based on the number of shared elements between two sets. At k=11, host and its associated microbiome shared nearly all k-mers, as reflected by a high Jaccard index regardless of host species (Figure 4a-4b; Supplementary Figure 5; Supplementary Table 5). As expected, the proportion of shared k-mers decreased significantly as the k-mer size increased, with Jaccard index dropping to 0.01 at k > 18 (Figure 4b). The number of observed k-mers was deviated from the expected distribution when longer k-mers was considered (Figure 4c). At k=24, the percent of high occurrence (5,000 or more) k-mers were markedly higher than the expected values. Furthermore, the sequence complexity of the observed k-mers was high (Figure 4d), indicating that low complexity sequences with high occurrence could be prone to be misclassified.

Analyses of simulated datasets for ancient DNA samples

Due to the limited availability of historical samples, a simulation study was performed to evaluate the impact of host DNA content on microbial analyses of the short DNA molecules (20-40bp) commonly recovered from historical samples. A total of 6,000 aDNA datasets were synthesized using *gargammel* for 3 plant species with different genomic complexity and size (rice, maize and bread wheat) and 107 representative soil microbes (Supplementary Table 6). To introduce the damage pattern, a length distribution obtained from one herbarium rice sample was

used (Supplementary Figure 6). Each dataset contained one million simulated paired-end reads, with host and microbial DNA proportions ranging from 0.01 to 0.45. An average of 0.98 million reads was obtained after quality filtering and merging. The merged read lengths followed a lognormal distribution, ranging from 21 to 150 bp with a peak around 28 bp (Supplementary Figure 6), closely resembling the fragment length profiles typically observed in historical samples. The merged reads were preprocessed with and without host DNA removal, followed by taxonomic classification against custom kraken2 databases built with k-mer sizes of 18, 21, 24, 28, 31 and 35. Overall, cross-domain misclassification of host reads (assigned to prokaryotes) was very low (Figure 5a and Supplementary Figure 7). Without host removal, the highest misclassification rate in the simulated rice datasets was 0.2% at k=21 and 24; this dropped to 0.1% or below when host DNA removal step was performed. The abundance of soil microbes between workflows was consistent across datasets (Supplementary Figure 8). As expected, the number of classified microbial reads correlated with their proportions in the simulated datasets, but also strongly influenced by k-mer value of the databases (Figure 5b; Supplementary Figure 9). The highest number of classified microbial reads (~300k for datasets with a bacterial fraction of 0.45) was obtained using databases built with k=24 or k=28, about one-third higher than k=35. In contrast, the number of classified reads dropped sharply to fewer than 100k when k-mer values below 24 were used.

Reads classified at the species level are generally more informative than those resolved only at higher taxonomic ranks. Among the simulation datasets, the number of microbial reads resolved at the species-level ranged from 7,890 to 375,850. The percent of reads with species-level classification was highest at k=28 (Figure 5c and Supplementary Figure 10). Higher percent

of reads were assigned to higher taxonomic ranks (e.g. genus or even phylum) at shorter k-mer size, indicating that the classification with databases built with smaller k-mer values may improve the overall classification rate but at the expense of taxonomic resolution.

Because DNA fragments from historical samples are highly variable in length (Supplementary Figure 6), we further examined read length distribution across different k-mer values. Regardless of k-mer values, the length distribution of classified microbial reads was lognormal with a peak length at 30-40 bp (Figure 5d), following the read length distribution of the herbarium rice sample used for the synthesis of simulation datasets. The shortest reads that could be annotated were constrained by the k-mer values (e.g. the lowest bound of read length for k=35 is 35bp), since reads shorter than the k-mer value were discarded by Kraken2. These observations suggested that annotation with two or more databases k-mer values could extend the coverage across a wider fragment length spectrum.

To determine the optimal annotation strategies, we evaluated the number of classified reads at species level among different combinations of k-mer values. In the single k-mer workflow (Figure 1), reads were annotated with a custom kraken2 database (prebuilt with k=21, 24, 28, 31 or 35). Only 47.7% of microbial reads were classified at the species level (Figure 6b) with a k=35 database. The publicly available Kraken2 databases (e.g. PlusPFP) were based on k=35, indicating that they might not be optimal for analyzing short reads in historical samples. The mean classification rate was boosted to 67.1% with a k=28 database, however the use of shorter k-mers might lead to a higher cross-domain misclassification, especially when the host DNA was not removed (Figure 5a). To address this, we evaluated a 2-step annotation workflow

(Figure 6a) in which reads were first searched against a database constructed with a longer k-mer value (k=31 for the simulated datasets). Unclassified reads or those resolved at genus or higher ranks were subsequently processed with another database generated with a shorter k-mer value (e.g. k=24). This strategy yielded a higher classification rate (71.7%), outperforming the single k-mer workflow. The recall rate was critically affected by the choice of k-mer values (single stage: 0.04-0.66; 2-stage: 0.49-0.74) while the precision of taxonomic classification was consistently maintained between 0.97-0.99 regardless of the choice of k-mer values (Figure 6c and Supplementary Table 7). Although host DNA removal is not strictly necessary, the removal of human or human-skin-associated microbes DNA prior to downstream analyses is still recommended.

Discussion

Historical samples, such as those from museum or herbarium collections, are valuable to generate insights into how the diversity and community structure of host-associated microbiomes are shaped by environmental changes over time. Profiling host-associated microbiome in these samples presents challenges due to the DNA content from host species or other non-target organisms, such as humans or arthropods. These extraneous sequences have been suggested to interfere with downstream analyses, potentially leading to the false identification of taxa (primarily through the misclassification of host DNA as a microbial origin in this case) or misassemble of sequence contigs (Schmieder and Edwards 2011). The removal of non-target sequences prior to downstream analyses is a routine procedure. However, many historical samples come from non-model or extinct species, for which reference genomes are usually unavailable. Therefore, it raises the question to what extent the microbiome analyses are affected by host DNA. To this end, we profiled the host-associated microbiome in contemporary and historical samples along with theoretical analyses with k-mers. Our findings suggests that host DNA has a negligible effect on microbial analyses, regardless of their abundance in the samples. In contemporary rice, historical rice and ragweed, the host DNA content varied from less than 5% to nearly 90%. Our findings are not conditional on the presence of host genome in the database (the pre-built kraken2 database) for taxonomic assignment of sequencing reads. Notably, the rice genome is included in the database, whereas genomes of bumblebee or any closely related species is not available in the database. Moreover, plant species and arthropods are underrepresented in many biological databases, including but not limited to the prebuilt database used in this study. The database contains sequences from the common sunflower, which

is the closest possible species available to common ragweed within the Asteroideae subfamily (Supplementary Table 4). Nonetheless, it may be beneficial to adopt a more lenient mapping strategy e.g. reducing the seed length or lower mismatch and gap penalty when aligning reads to closely (or remotely) related species. The abundance of taxa detected was generally consistent between our two workflows (Table 1), indicating the impact of host DNA on downstream microbial analyses might not be as strong as previously suggested. In contrast to host DNA removal step, OTU filtering step has a strong impact on downstream microbiome analysis. A proper OTU filtering can reduce the number of falsely identified taxa (Table 1), which are typically in low abundance and prevalence (Cao, Sun et al. 2020). Other technical factors, such as sample size and sequencing depth, can also influence the identification of differential taxa (Nearing, Douglas et al. 2022). However, we did not examine these factors in this study, as they are often beyond the control of investigators when working with historical samples.

DNA isolated from historical samples is often degraded and highly fragmented, which complicates bioinformatic analysis (Burbano and Gutaker 2023). The typical workflow for profiling community composition of host-associated microbiome in historical samples relies on k-mer-based taxonomic classification tools such as Kraken2 and Centrifuge (Kim, Song et al. 2016). The sensitivity and specificity of these tools can be adjusted through k-mer size or seed length. The public available kraken2 databases were typically built with a k-mer value of 35 (Arizmendi Cardenas, Neuenschwander et al. 2022, Lu, Rincon et al. 2022). This setting is ideal for general use as it leverages between specificity and sensitivity but not optimal for degraded and heavily fragmented DNA sequences such as those from historical samples. Different values of k-mer have been proposed for various scenarios and types of datasets. A k-mer size of 29 was

suggested to built databases for filtering exogenous DNA in historical samples (Ravishankar, Perez et al. 2024). For metagenomic datasets, k-mer size of 21 was shown to be sufficient to obtain accurate results in most cases (Ondov, Treangen et al. 2016). However, it remained unclear if these values were suitable for profiling host-associated microbial DNA in historical samples. To this end, we conducted k-mer analyses on host genome and its associated microbiome to yield insights into the optimal k-mer size and supporting evidence on the negligible effects of host DNA on microbiome analysis in historical samples.

Microorganisms in the rhizosphere play vital roles in nutrient cycling and other biogeographical processes in soil (Philippot, Chenu et al. 2024). The composition of soil microbial communities is influenced by environmental perturbations; therefore, historical samples are useful in gaining insights into the microbial community dynamics and factors shaping them over long-time scales. We focused on core microbes in the soil microbiome, which were reconstructed from an extensive collection of soil metagenomes. The constituent k-mers shared between host genome and associated microbiome was regarded as an indicator of taxonomic misclassification of host-derived sequences. As expected, the number of k-mers shared between host genome and core microbes depends on the size of k-mers and their genomes. When a small k-mer value (e.g. k=11) was used, all constituent k-mers were shared simply by chance alone. With a larger k-mer size (e.g. k=18 or larger), the overlap in k-mer space was significantly reduced (Jaccard index < 0.1), suggesting that the misclassification of host DNA as microbial taxa become unlikely using a large k-mer size. Based on our observation, k-mer size of 21 is sufficient for profiling host-associated microbiome in historical samples from non-model species with a low level of misclassification (0.1% or less). This finding also explains

why host DNA has a negligible impact on microbial diversity metrices when taxonomic classification was performed using databases compiled with a large k-mer, such as k=35 in PlusPFP database. Furthermore, we observed that at k=24, 5.9% of the k-mers shared between host genome and associated microbiome (0.01% or less of total k-mers) were low-complexity k-mers with high occurrence (≥10) (Supplementary Table 5-6). It suggests that low complexity or repetitive elements in eukaryotic genomes may make taxonomic classification difficult. A sequence complexity filter may be useful in ameliorating the impact of low-complexity sequences prior to taxonomic classification. The size and complexity of eukaryotic hosts vary dramatically. To evaluate the impact of the genomic features of host genome, we performed a simulation study by synthesizing a sheer number of metagenomic datasets from 3 host plant species with varying genome size and complexity and soil microbes. DNA damage patterns commonly observed in historical samples were also incorporated into the synthesized datasets. The findings from the simulation study strongly suggested that while the microbiome analyses on historical samples are feasible, a thorough and careful considerations should be given to cope with the molecular features (e.g. short fragments) of DNA obtained from historical samples but not host DNA content. The two-step annotation workflow proposed in this study can effectively increase the number of classified short DNA fragments. For the simulated rice datasets, the proportion of reads annotated at the species level ranged from 4.5 to 67.1% using a single annotation database. This proportion increased to 71.1% (k=31 followed by k=24) with the two-step workflow without scarifying the precision. The optimal k-mer combination is subject to the length distribution of recovered DNA molecules from historical samples. This improvement has the potential to enable the recovery of more informative signals from precious historical samples that might otherwise be discarded.

In this study, we demonstrated that host DNA has a negligible impact on microbial analyses of host-associated microbiome, even in samples with high content of host DNA. Our findings further suggest that meaningful microbial analyses can still be performed on historical samples even though genomic sequences of the host species are unavailable. It may be applicable to explore the microbial ecology of non-model organisms. We also showed that microbial reads as short as 21 bp are still distinguishable from those derived from host genome and therefore informative results can still be obtained. To effectively annotate the short reads of this length, a 2-step annotation workflow was proposed (reads annotated with k=24 and k=31 databases). Although our findings suggest the genome size and complexity may not be critically affect downstream microbiome analyses, a comprehensive evaluation of the analytic workflow remains warranted. Finally, this study does not account for contamination arising from post-mortem colonization that accumulates over time. In this study, we did not separate the contamination from the reads derived from host-associated microbes. The findings were not substantially affected by the sources of the microbes (host-associated or contamination) as the host DNA content and optimal analysis strategies on microbial signals were the prime focus of this study. Future work should identify microbial contaminants in historical samples through DNA damage patterns (especially for recent contaminations) or a catalogue of post-mortem contaminants commonly detected in museum collections. Another aspect worth exploring, yet overlooked in this study, is the availability of microbial reference genomes in the annotation database. Historical samples may harbor microbial taxa that have become rare due to anthropogenic pressures or are entirely absent from contemporary samples. How to efficiently detect and characterize these poorly represented or unknown microbial signals remains an open question.

Overall, this study marks an important milestone in enabling metagenomic investigation of historical samples, contributing to our understanding of how anthropogenic changes have shaped host and associated microbiome.

Acknowledgement

SKN and RMG are both supported by UKRI Frontier Research Guarantee grant EP/X022404/1 to RMG. The authors acknowledge Research Computing at the James Hutton Institute for providing computational resources and technical support for the "UK's Crop Diversity Bioinformatics HPC" (BBSRC grants BB/S019669/1 and BB/X019683/1), use of which has contributed to the results reported within this paper. The authors also acknowledge help from Robin Becina Guevarra with early versions of the manuscript.

Data Accessibility and Benefit-sharing

Raw sequencing data generated in this study is available in EBI ENA under the SRA accession PRJEB87646.

Author Contributions

RMG conceptualized the study with inputs from SKN. SKN conceptualized the k-mer comparison approach. SKN carried out analyses and designed the figures with inputs from RMG. SKN has written the manuscript with inputs from RMG.

Tables and Figures

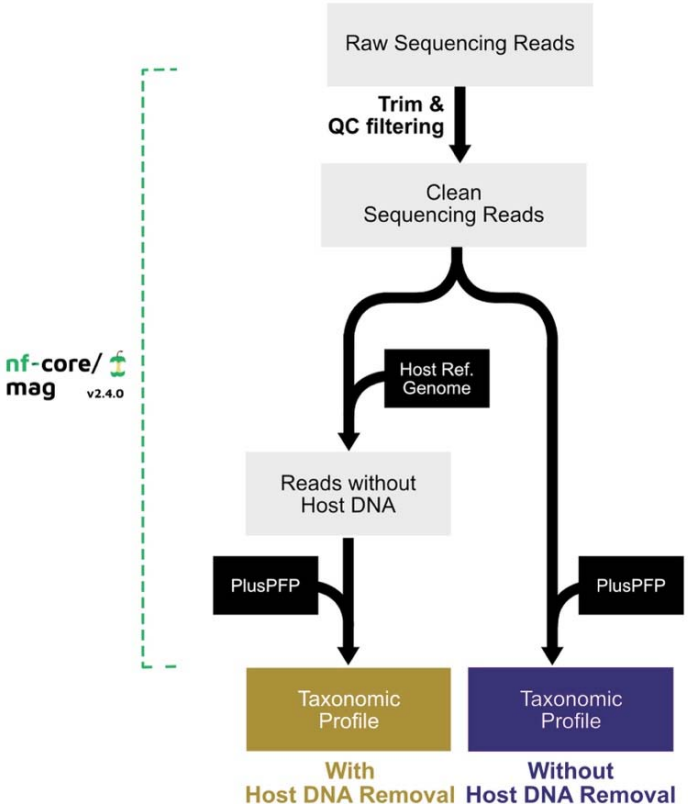

Figure 1. The bioinformatic workflow. The Nextflow metagenomic analysis workflow (nf-core/mag) was used in this study. Two separate workflows (gold for the workflow with host DNA removal and purple for the workflow without host DNA removal) were performed in parallel with identical steps except the host removal step prior to taxonomic classification.

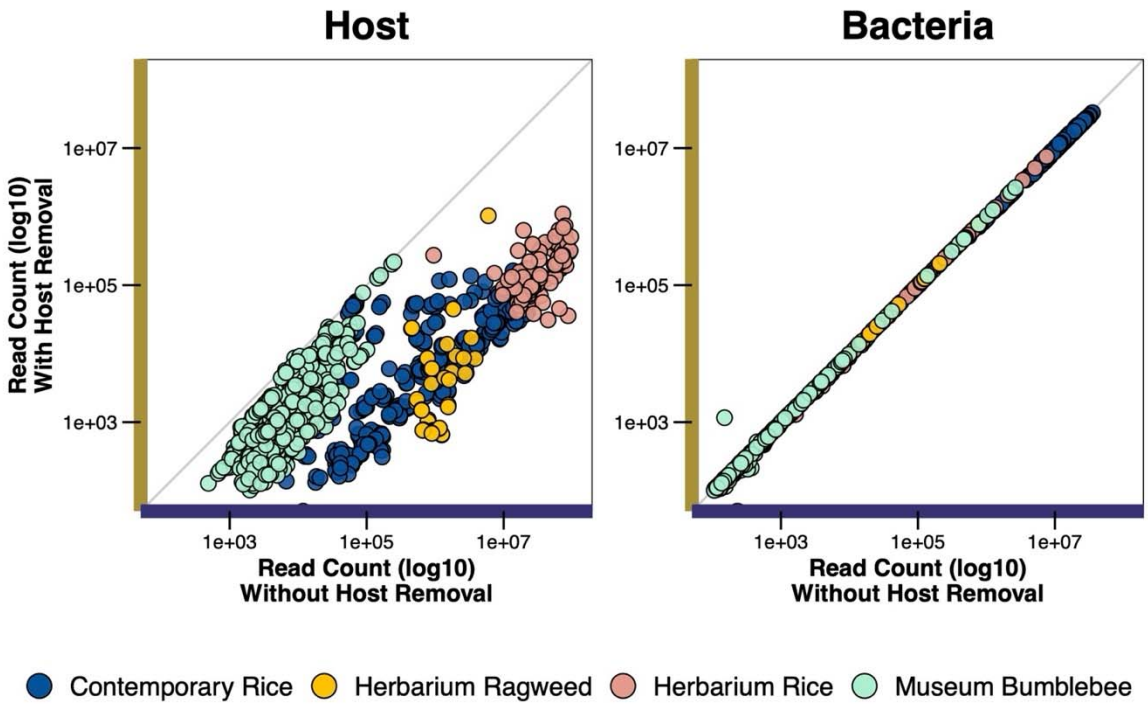

Figure 2. The number of sequencing reads (log10) assigned to host (left panel; plant or metazoa) and microbial taxa (right panel) with (y-axis; gold) and without (x-axis; purple) host DNA removal. Each circle represents a sample and is color-coded with their host species. Historical rice samples (Rice in pink circles) and contemporary rice samples (Cont. Rice in blue circles) were displayed separately despite the same host rice reference was used.

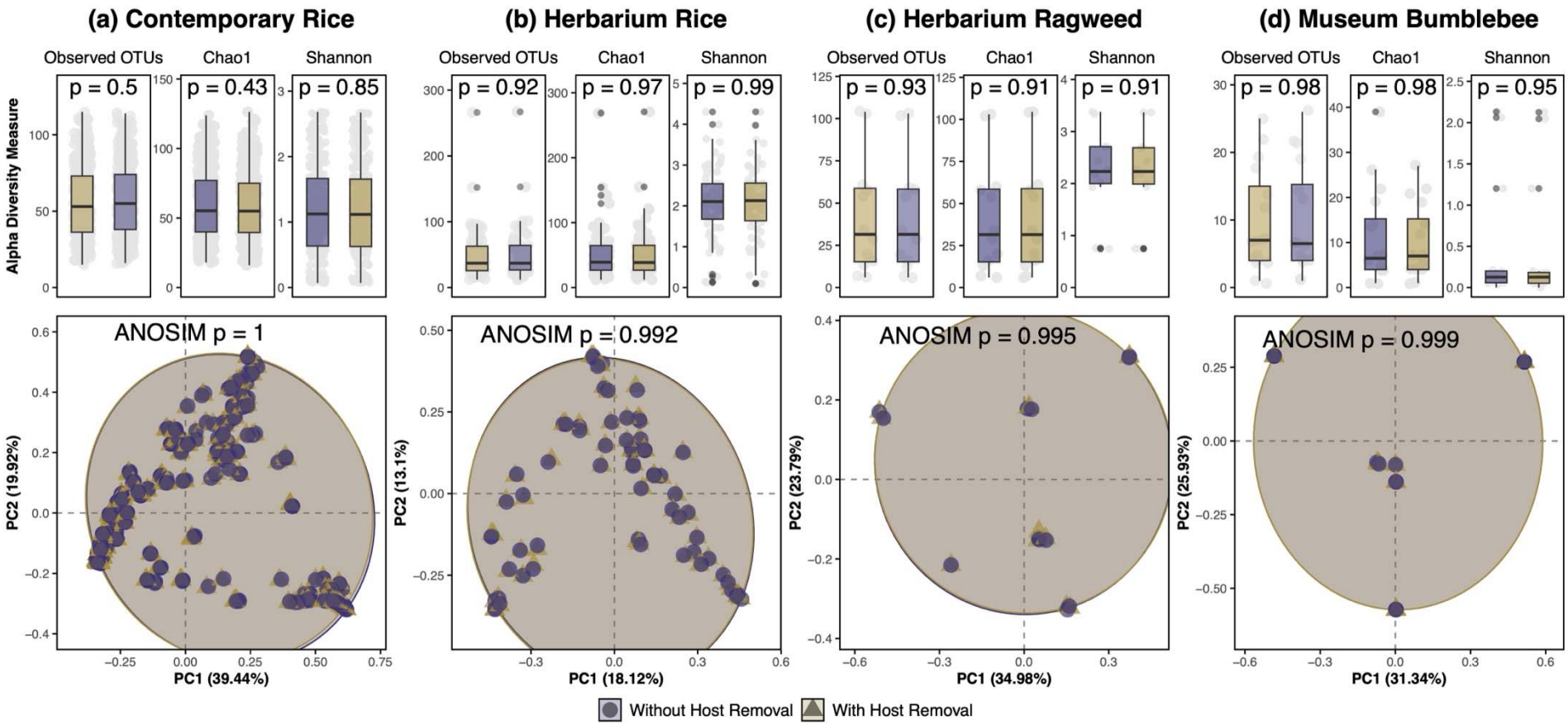

Figure 3. Alpha and beta diversity analysis of microbial community in the samples of (a) contemporary rice, (b) herbarium rice, (c) herbarium common ragweed and (d) museum bumblebee between the workflow with (gold) and without (purple) host DNA removal. Observed OTUs, Chao1 and Shannon diversity indices (top panels) were used to evaluate the richness of taxa and diversity. The structure of microbial community obtained (bottom panels) from both workflows (gold triangles for the dataset with host removal; purple circles for the same dataset without host removal) were compared using principal coordinates analysis (PCoA) based on Bray-Curtis dissimilarity index. *P*-value was obtained from the Wilcox-rank sum test for microbial diversity. The *p*-value of ANOSIM was based on 999 permutations.

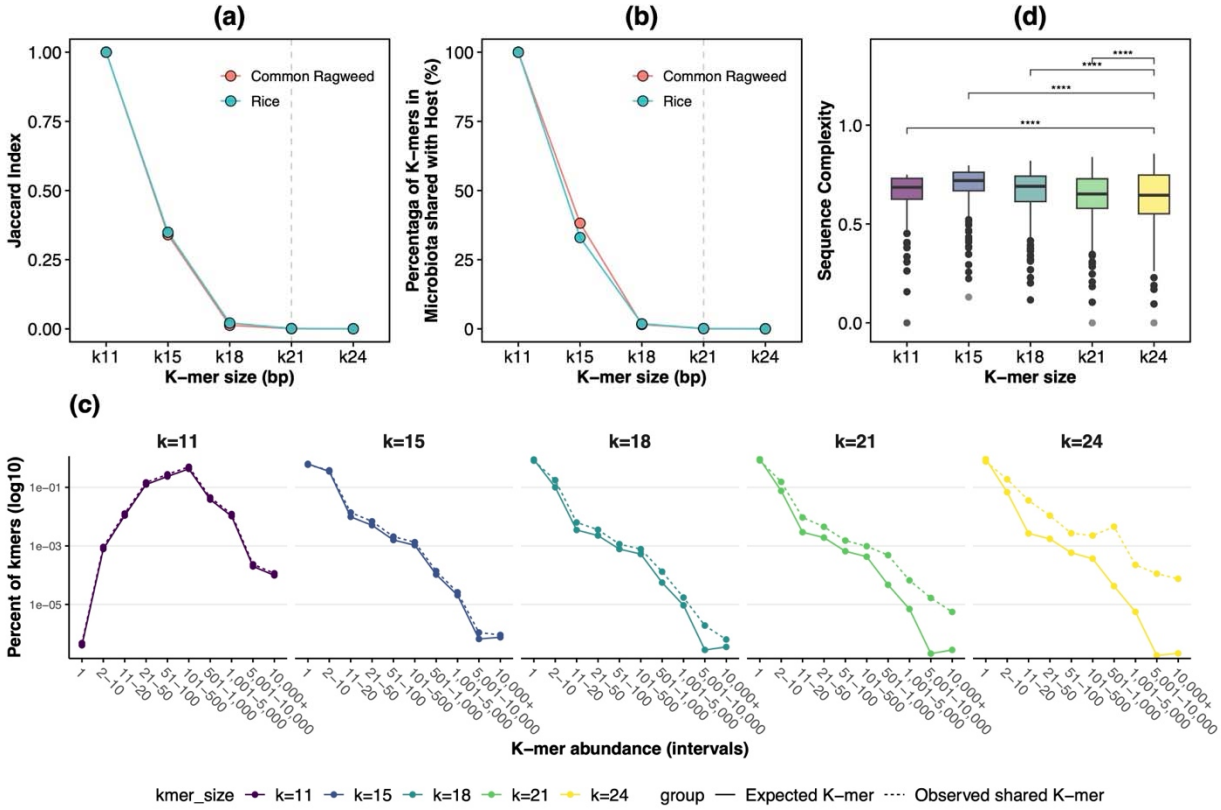

Figure 4. Shared k-mers between host species (rice and common ragweed) and representative MAGs of soil microbiome. (a) Jaccard index and (b) the percentage of k-mers shared between host and soil microbiome across different k-mer values using JellyFish "count" command with the option --mer-len=11, 15, 18, 21 and 24 and with Bloom filter disabled and both strands considered. Jaccard index reduced to 0.02 or below at k=21 (dashed line) was consistent across host species (color-coded lines). (c) The distribution of k-mers in rice (dashed line) and those shared between rice and the soil microbiome (solid line). K-mers were grouped into intervals (x-axis) based on their occurrence in host genome (1, 2-10, 11-20, etc). (d) Sequence complexity (Wootton-Federhen complexity score) of shared k-mers in rice and the soil microbiome across k-mer size.

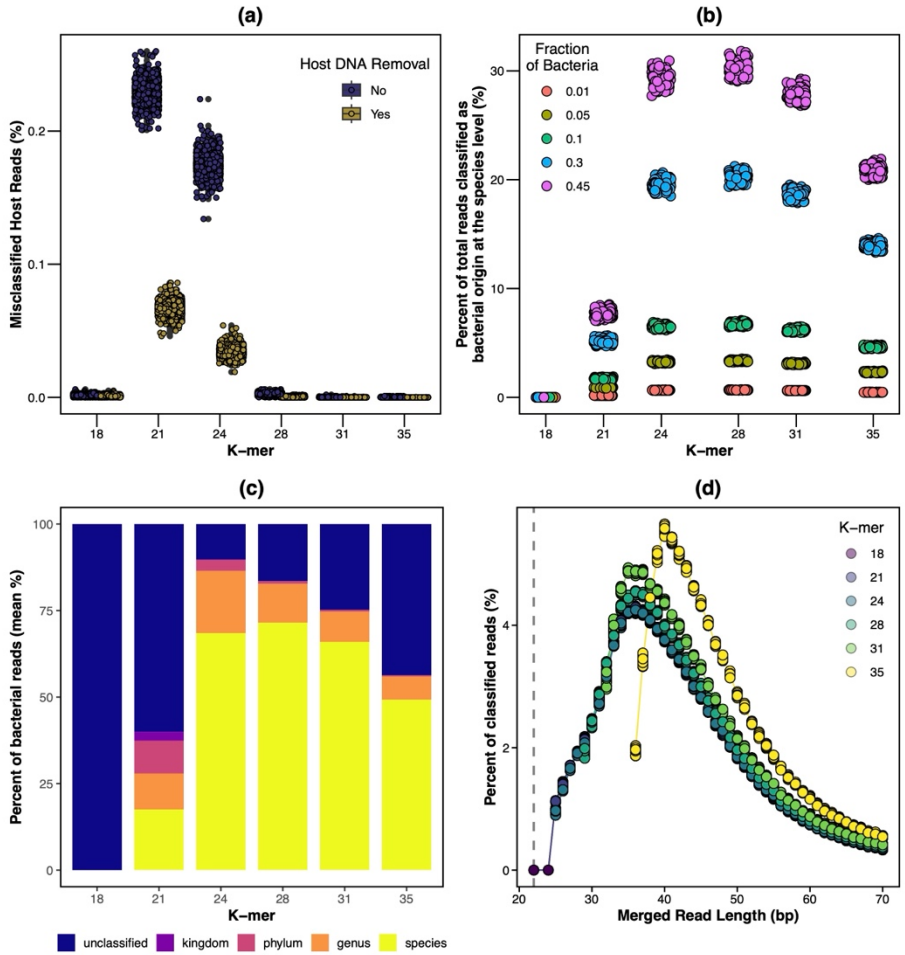

Figure 5. The analyses of the simulated rice datasets. **(a)** The percentage of host reads misclassified as microbial taxa. The simulated datasets were processed with (gold) and without (purple) host DNA removal step by comparing the mapping results against custom Kraken2 databases built with k=18 to k=35. **(b)** The number of classified bacterial reads in the simulated rice aDNA datasets after host DNA content removal. Merged reads from 2,000 simulated datasets were annotated with 5 databases created with different k-mer sizes (k=18, 21, 24, 28, 31 and 35). **(c)** The mean percent of microbial reads detected in the simulated rice aDNA datasets. Classified reads were color-coded to their lowest taxonomic rank (species, genus, phylum and kingdom were shown). Unclassified reads were shown in blue color. **(d)** The read length distribution of classified reads resolved at the species level across k-mer values.

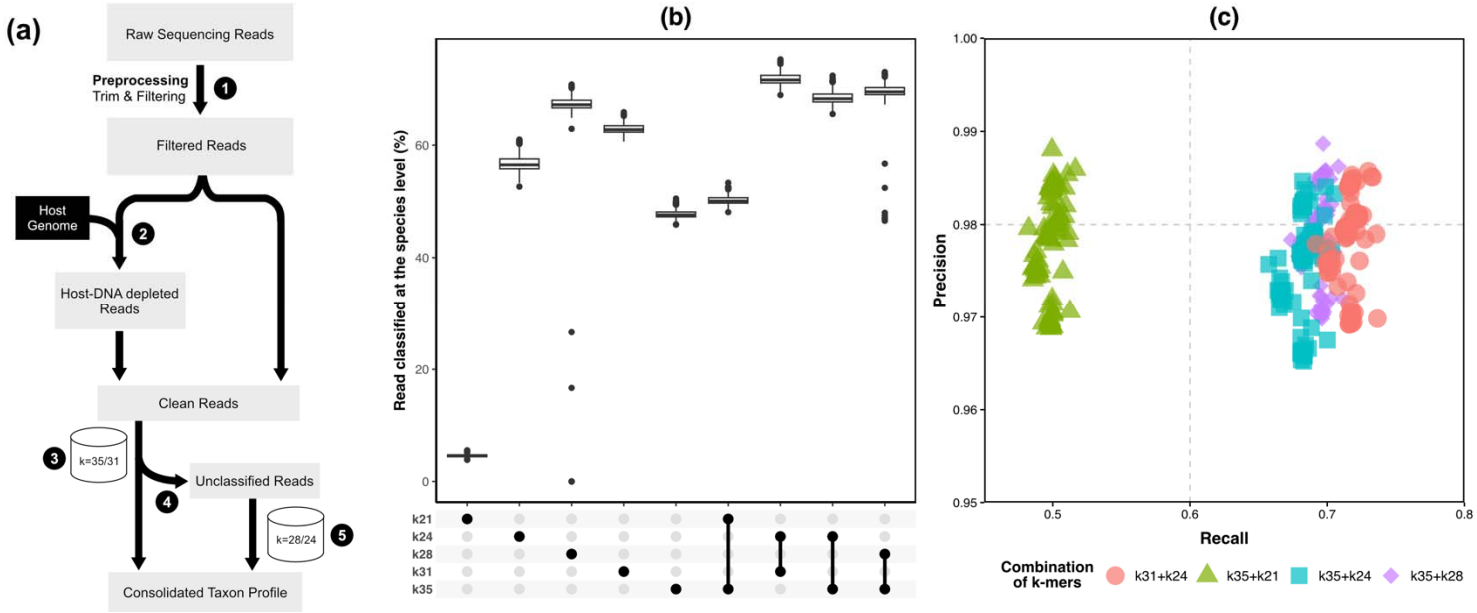

Figure 6 **(a)** Proposed 2-step analysis workflow for the taxonomic annotation of metagenomic reads in historical samples. 1. raw sequencing preprocessing: trimming, QC filtering and removal of low-complexity reads; 2. Removal of host and contaminant DNA if relevant genomic sequences are available. Low-complexity or repetitive sequences can be removed during this step; 3. The first stage of classification with an annotation database built with a large k-mer value (k35/k31). Hists on the optimal k-mer combination can be obtained by assessing the length distribution of the merged reads obtained in step 1 or 2; 4. Unclassified reads or those assigned with a non-species level taxonomy ID can be shortlisted for the second stage classification; 5. Classification of shortlisted reads using an annotated database built with a shorter k-mer value (k28/k24) **(b)** The percentage of microbial reads identified by one- or two-kmer search strategies in the simulated rice datasets. The k-mer value(s) used for the annotation was highlighted with a circle in black in the bottom upset plot. (c) The precision and recall of 100 randomly selected simulated rice datasets (out of 2,000 datasets) analyzed in 2-step analysis for each combination of k-mer values (k31+k24, k35+k21, k35+k24 and k35+k28) were shown. The precision and recall for all 2,000 simulated rice datasets were included in Supplementary Table 7.

Dataset	Sample Size (N) [§]	With Host DNA Removal			Without Host Removal			OTU Removed by Host DNA removal [#]	OTU# with ≥10% change in mean read count between workflows	Mean read count	
		Before filtering [¶]	After filtering	% filtered	Before filtering	After filtering	% filtered			OTU with ≥10% change	OTU with <10% change
Contemporary Rice	330	4,460	390	91.3%	4,614	390	91.6%	0 (0.0%)	55 (14.1%)	1,951.6	18,157.6
Herbarium Rice	95	8,057	853	89.4%	8,731	855	90.2%	3 (0.3%)	7 (0.8%)	90.6	467.0
Herbarium Ragweed	37	4,049	184	95.5%	4,315	185	95.7%	1 (0.5%)	0 (0.0%)	-	616.2
Museum Bumblebee	402	3,978	113	97.2%	4,497	113	97.0%	0 (0.0%)	0 (0.0%)	-	12,543.8

[§] Sample Size (N): no sample failed in host DNA removal step

[¶] Filtering cutoff: read count = 50 (100 for contemporary rice dataset), prevalence = 2% of samples

[#] OTU removed: the number of OTUs detected between the workflow with and without host DNA removal step before filtering

Table 1. Summary of the number of detected taxa (at the species level) between datasets with and with host DNA removal

References

Aeron, A., E. Khare, C. K. Jha, V. S. Meena, S. M. A. Aziz, M. T. Islam, K. Kim, S. K. Meena, A. Pattanayak, H. Rajashekara, R. C. Dubey, B. R. Maurya, D. K. Maheshwari, M. Saraf, M. Choudhary, R. Verma, H. N. Meena, A. Subbanna, M. Parihar, S. Shukla, G. Muthusamy, R. S. Bana, V. K. Bajpai, Y. K. Han, M. Rahman, D. Kumar, N. P. Singh and R. K. Meena (2020). "Revisiting the plant growth-promoting rhizobacteria: lessons from the past and objectives for the future." Arch Microbiol **202**(4): 665-676.

Alam, O., R. Gutaker, N. Surya, C. Castillo, D. Fuller, J. D. Guedes and M. Purugganan (2025). "Evolutionary histories of functional mutations during the domestication and spread of japonica rice in Asia." Proc Natl Acad Sci U S A **122**(43): e2514614122.

Andrews, S. (2015). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

Arizmendi Cardenas, Y. O., S. Neuenschwander and A. S. Malaspinas (2022). "Benchmarking metagenomics classifiers on ancient viral DNA: a simulation study." PeerJ **10**: e12784.

Banerjee, S. and M. G. A. van der Heijden (2023). "Soil microbiomes and one health." Nat Rev Microbiol **21**(1): 6-20.

Bieker, V. C., P. Battlay, B. Petersen, X. Sun, J. Wilson, J. C. Brealey, F. Bretagnolle, K. Nurkowski, C. Lee, F. S. Barreiro, G. L. Owens, J. Y. Lee, F. L. Kellner, L. van Boheeman, S. Gopalakrishnan, M. Gaudeul, H. Mueller-Schaerer, S. Lommen, G. Karrer, B. Chauvel, Y. Sun, B. Kostantinovic, L. Dalen, P. Poczai, L. H. Rieseberg, M. T. P. Gilbert, K. A. Hodgins and M. D. Martin (2022). "Uncovering the genomic basis of an extraordinary plant invasion." Sci Adv **8**(34): eabo5115.

Breitwieser, F. P., D. N. Baker and S. L. Salzberg (2018). "KrakenUniq: confident and fast metagenomics classification using unique k-mer counts." Genome Biol **19**(1): 198.

Briggs, A. W., U. Stenzel, P. L. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann and S. Pääbo (2007). "Patterns of damage in genomic DNA sequences from a Neandertal." Proc Natl Acad Sci U S A **104**(37): 14616-14621.

Burbano, H. A. and R. M. Gutaker (2023). "Ancient DNA genomics and the renaissance of herbaria." Science **382**(6666): 59-63.

Bushnell, B. (2014). "BBMap: a fast, accurate, splice-aware aligner."

Cao, Q., X. Sun, K. Rajesh, N. Chalasani, K. Gelow, B. Katz, V. H. Shah, A. J. Sanyal and E. Smirnova (2020). "Effects of Rare Microbiome Taxa Filtering on Statistical Analysis." Front Microbiol **11**: 607325.

Chen, S., Y. Zhou, Y. Chen and J. Gu (2018). "fastp: an ultra-fast all-in-one FASTQ preprocessor." Bioinformatics **34**(17): i884-i890.

Dabdoub, S. (2016). kraken-biom: Enabling interoperative format conversion for Kraken results (Version 1.2) [Software].

Debray, R., A. Conover, X. Zhang, E. A. Dewald-Wang and B. Koskella (2023). "Within-host adaptation alters priority effects within the tomato phyllosphere microbiome." Nat Ecol Evol **7**(5): 725-731.

Edwin, N. R., A. H. Fitzpatrick, F. Brennan, F. Abram and O. O'Sullivan (2024). "An in-depth evaluation of metagenomic classifiers for soil microbiomes." Environ Microbiome **19**(1): 19.

Eisenhofer, R., J. J. Minich, C. Marotz, A. Cooper, R. Knight and L. S. Weyrich (2019). "Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations." Trends Microbiol **27**(2): 105-117.

Ewels, P. A., A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. Di Tommaso and S. Nahnsen (2020). "The nf-core framework for community-curated bioinformatics pipelines." Nat Biotechnol **38**(3): 276-278.
Gancz, A. S., A. G. Farrer, M. P. Nixon, S. Wright, L. Arriola, C. Adler, E. R. Davenport, N. Gully, A. Cooper, K. Britton, K. Dobney, J. D. Silverman and L. S. Weyrich (2023). "Ancient dental calculus reveals oral microbiome shifts associated with lifestyle and disease in Great Britain." Nat Microbiol **8**(12): 2315-2325.
Glocke, I. and M. Meyer (2017). "Extending the spectrum of DNA sequences retrieved from ancient bones and teeth." Genome Res **27**(7): 1230-1237.
Gould, A. L., V. Zhang, L. Lamberti, E. W. Jones, B. Obadia, N. Korasidis, A. Gavryushkin, J. M. Carlson, N. Beerenwinkel and W. B. Ludington (2018). "Microbiome interactions shape host fitness." Proc Natl Acad Sci U S A **115**(51): E11951-E11960.
Gutaker, R. M. and H. A. Burbano (2017). "Reinforcing plant evolutionary genomics using ancient DNA." Curr Opin Plant Biol **36**: 38-45.
Harney, E., O. Cheronet, D. M. Fernandes, K. Sirak, M. Mah, R. Bernardos, N. Adamski, N. Broomandkhoshbacht, K. Callan, A. M. Lawson, J. Oppenheimer, K. Stewardson, F. Zalzala, A. Anders, F. Candilio, M. Constantinescu, A. Coppa, I. Ciobanu, J. Dani, Z. Gallina, F. Genchi, E. G. Nagy, T. Hajdu, M. Hellebrandt, A. Horvath, A. Kiraly, K. Kiss, B. Kolozsi, P. Kovacs, K. Kohler, M. Lucci, I. Pap, S. Popovici, P. Raczky, A. Simalcsik, T. Szeniczey, S. Vasilyev, C. Virag, N. Rohland, D. Reich and R. Pinhasi (2021). "A minimally destructive protocol for DNA extraction from ancient teeth." Genome Res **31**(3): 472-483.
Jansson, J. K. and K. S. Hofmockel (2020). "Soil microbiomes and climate change." Nat Rev Microbiol **18**(1): 35-46.
Kim, D., L. Song, F. P. Breitwieser and S. L. Salzberg (2016). "Centrifuge: rapid and sensitive classification of metagenomic sequences." Genome Res **26**(12): 1721-1729.
Krakau, S., D. Straub, H. Gourle, G. Gabernet and S. Nahnsen (2022). "nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning." NAR Genom Bioinform **4**(1): lqac007.
Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.
Latorre, S. M., P. L. M. Lang, H. A. Burbano and R. M. Gutaker (2020). "Isolation, Library Preparation, and Bioinformatic Analysis of Historical and Ancient Plant DNA." Curr Protoc Plant Biol **5**(4): e20121.
Ley, R. E., M. Hamady, C. Lozupone, P. J. Turnbaugh, R. R. Ramey, J. S. Bircher, M. L. Schlegel, T. A. Tucker, M. D. Schrenzel, R. Knight and J. I. Gordon (2008). "Evolution of mammals and their gut microbes." Science **320**(5883): 1647-1651.
Lu, J., F. P. Breitwieser, P. Thielen and S. L. Salzberg (2017). "Bracken: estimating species abundance in metagenomics data." PeerJ Computer Science **3**: e104.
Lu, J., N. Rincon, D. E. Wood, F. P. Breitwieser, C. Pockrandt, B. Langmead, S. L. Salzberg and M. Steinegger (2022). "Metagenome analysis using the Kraken software suite." Nat Protoc **17**(12): 2815-2839.
Ma, B., C. Lu, Y. Wang, J. Yu, K. Zhao, R. Xue, H. Ren, X. Lv, R. Pan, J. Zhang, Y. Zhu and J. Xu (2023). "A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources." Nat Commun **14**(1): 7318.
Marcais, G. and C. Kingsford (2011). "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." Bioinformatics **27**(6): 764-770.
Marchesi, J. R. and J. Ravel (2015). "The vocabulary of microbiome research: a proposal." Microbiome **3**: 31.
Morais, S., S. Winkler, A. Zorea, L. Levin, F. S. P. Nagies, N. Kapust, E. Lamed, A. Artan-Furman, D. N. Bolam, M. P. Yadav, E. A. Bayer, W. F. Martin and I. Mizrahi (2024). "Cryptic

diversity of cellulose-degrading gut bacteria in industrialized humans." Science **383**(6688): eadj9223.
Mullin, V. E., W. Stephen, A. N. Arce, W. Nash, C. Raine, D. G. Notton, A. Whiffin, V. Blagderov, K. Gharbi, J. Hogan, T. Hunter, N. Irish, S. Jackson, S. Judd, C. Watkins, W. Haerty, J. Ollerton, S. Brace, R. J. Gill and I. Barnes (2023). "First large-scale quantification study of DNA preservation in insects from natural history collections using genome-wide sequencing." Methods in Ecology and Evolution **14**(2): 360-371.
Naamala, J., S. K. Jaiswal and F. D. Dakora (2016). "Antibiotics Resistance in Rhizobium: Type, Process, Mechanism and Benefit for Agriculture." Curr Microbiol **72**(6): 804-816.
Naylor, D. and D. Coleman-Derr (2017). "Drought Stress and Root-Associated Bacterial Communities." Front Plant Sci **8**: 2223.
Nearing, J. T., G. M. Douglas, M. G. Hayes, J. MacDonald, D. K. Desai, N. Allward, C. M. A. Jones, R. J. Wright, A. S. Dhanani, A. M. Comeau and M. G. I. Langille (2022). "Microbiome differential abundance methods produce different results across 38 datasets." Nat Commun **13**(1): 342.
Ondov, B. D., T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren and A. M. Phillippy (2016). "Mash: fast genome and metagenome distance estimation using MinHash." Genome Biol **17**(1): 132.
Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz and G. W. Tyson (2015). "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." Genome Res **25**(7): 1043-1055.
Percival-Alwyn, L., Barnes, I., Clark, M.D., Cockram, J., Coffey, M.P., Jones, S., Kersey, P.J., Kidner, C.A., Kosiol, C., Li, B., Marsh, W.A., Zhou, J., Caccamo, M., Milne, I. (2024). "UKCropDiversity-HPC: A collaborative high-performance computing resource approach for sustainable agriculture and biodiversity conservation." Plants, People, Planet.
Petersen, C., I. K. Hamerich, K. L. Adair, H. Griem-Krey, M. Torres Oliva, M. P. Hoeppner, B. J. M. Bohannan and H. Schulenburg (2023). "Host and microbiome jointly contribute to environmental adaptation." ISME J **17**(11): 1953-1965.
Philippot, L., C. Chenu, A. Kappler, M. C. Rillig and N. Fierer (2024). "The interplay between microbial communities and soil properties." Nat Rev Microbiol **22**(4): 226-239.
Ravishankar, S., V. Perez, R. Davidson, X. Roca-Rada, D. Lan, Y. Souilmi and B. Llamas (2024). "Filtering out the noise: metagenomic classifiers optimize ancient DNA mapping." Brief Bioinform **26**(1).
Renaud, G., K. Hanghøj, E. Willerslev and L. Orlando (2017). "gargammel: a sequence simulator for ancient DNA." Bioinformatics **33**(4): 577-579.
Rout, A. K., S. Dixit, P. S. Tripathy, S. S. Rout, S. N. Parida, P. K. Parida, D. J. Sarkar, B. Kumar Das, A. K. Singh and B. K. Behera (2024). "Metagenomic landscape of sediments of river Ganga reveals microbial diversity, potential plastic and xenobiotic degradation enzymes." J Hazard Mater **471**: 134377.
Rout, A. K., P. S. Tripathy, S. Dixit, D. U. Behera, B. Behera, B. K. Das and B. K. Behera (2023). "Unveiling the Microbiome Landscape: A Metagenomic Study of Bacterial Diversity, Antibiotic Resistance, and Virulence Factors in the Sediments of the River Ganga, India." Antibiotics (Basel) **12**(12).
Schmieder, R. and R. Edwards (2011). "Fast identification and removal of sequence contamination from genomic and metagenomic datasets." PLoS One **6**(3): e17288.
Su, P., W. A. Wicaksono, C. Li, K. Michl, G. Berg, D. Wang, Y. Xiao, R. Huang, H. Kang, D. Zhang, T. Cernava and Y. Liu (2022). "Recovery of metagenome-assembled genomes from the phyllosphere of 110 rice genotypes." Sci Data **9**(1): 254.
Tiao, G., C. K. Lee, I. R. McDonald, D. A. Cowan and S. C. Cary (2012). "Rapid microbial response to the presence of an ancient relic in the Antarctic Dry Valleys." Nat Commun **3**: 660.

Tremblay, B. (2024). "universalmotif: An R package for biological motif analysis." Journal of Open Source Software **9**: 7012.
Trivedi, P., J. E. Leach, S. G. Tringe, T. Sa and B. K. Singh (2020). "Plant-microbiome interactions: from community assembly to plant health." Nat Rev Microbiol **18**(11): 607-621.
Warinner, C., A. Herbig, A. Mann, J. A. Fellows Yates, C. L. Weiss, H. A. Burbano, L. Orlando and J. Krause (2017). "A Robust Framework for Microbial Archaeology." Annu Rev Genomics Hum Genet **18**: 321-356.
Weyrich, L. S., S. Duchene, J. Soubrier, L. Arriola, B. Llamas, J. Breen, A. G. Morris, K. W. Alt, D. Caramelli, V. Dresely, M. Farrell, A. G. Farrer, M. Francken, N. Gully, W. Haak, K. Hardy, K. Harvati, P. Held, E. C. Holmes, J. Kaidonis, C. Lalueza-Fox, M. de la Rasilla, A. Rosas, P. Semal, A. Soltysiak, G. Townsend, D. Usai, J. Wahl, D. H. Huson, K. Dobney and A. Cooper (2017). "Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus." Nature **544**(7650): 357-361.
Wood, D. E., J. Lu and B. Langmead (2019). "Improved metagenomic analysis with Kraken 2." Genome Biol **20**(1): 257.
Xiao, K., X. Liang, H. Lu, X. Li, Z. Zhang, X. Lu, H. Wang, Y. Meng, A. Roy, W. Luo, X. Shen, D. M. Irwin and Y. Shen (2022). "Adaptation of gut microbiome and host metabolic systems to lignocellulosic degradation in bamboo rats." ISME J **16**(8): 1980-1992.
Zuntini, A. R., T. Carruthers, O. Maurin, P. C. Bailey, K. Leempoel, G. E. Brewer, N. Epitawalage, E. Francoso, B. Gallego-Paramo, C. McGinnie, R. Negrao, S. R. Roy, L. Simpson, E. Toledo Romero, V. M. A. Barber, L. Botigue, J. J. Clarkson, R. S. Cowan, S. Dodsworth, M. G. Johnson, J. T. Kim, L. Pokorny, N. J. Wickett, G. M. Antar, L. DeBolt, K. Gutierrez, K. P. Hendriks, A. Hoewener, A. Q. Hu, E. M. Joyce, I. Kikuchi, I. Larridon, D. A. Larson, E. J. de Lirio, J. X. Liu, P. Malakasi, N. A. S. Przelomska, T. Shah, J. Viruel, T. R. Allnutt, G. K. Ameka, R. L. Andrew, M. S. Appelhans, M. Arista, M. J. Ariza, J. Arroyo, W. Arthan, J. B. Bachelier, C. D. Bailey, H. F. Barnes, M. D. Barrett, R. L. Barrett, R. J. Bayer, M. J. Bayly, E. Biffin, N. Biggs, J. L. Birch, D. Bogarin, R. Borosova, A. M. C. Bowles, P. C. Boyce, G. L. C. Bramley, M. Briggs, L. Broadhurst, G. K. Brown, J. J. Bruhl, A. Bruneau, S. Buerki, E. Burns, M. Byrne, S. Cable, A. Calladine, M. W. Callmander, A. Cano, D. J. Cantrill, W. M. Cardinal-McTeague, M. M. Carlsen, A. J. A. Carruthers, A. de Castro Mateo, M. W. Chase, L. W. Chatrou, M. Cheek, S. Chen, M. J. M. Christenhusz, P. A. Christin, M. A. Clements, S. C. Coffey, J. G. Conran, X. Cornejo, T. L. P. Couvreur, I. D. Cowie, L. Csiba, I. Darbyshire, G. Davidse, N. M. J. Davies, A. P. Davis, K. J. van Dijk, S. R. Downie, M. F. Duretto, M. R. Duvall, S. L. Edwards, U. Eggli, R. H. J. Erkens, M. Escudero, M. de la Estrella, F. Fabriani, M. F. Fay, P. L. Ferreira, S. Z. Ficinski, R. M. Fowler, S. Frisby, L. Fu, T. Fulcher, M. Galbany-Casals, E. M. Gardner, D. A. German, A. Giaretta, M. Gibernau, L. J. Gillespie, C. C. Gonzalez, D. J. Goyder, S. W. Graham, A. Grall, L. Green, B. F. Gunn, D. G. Gutierrez, J. Hackel, T. Haevermans, A. Haigh, J. C. Hall, T. Hall, M. J. Harrison, S. A. Hatt, O. Hidalgo, T. R. Hodkinson, G. D. Holmes, H. C. F. Hopkins, C. J. Jackson, S. A. James, R. W. Jobson, G. Kadereit, I. M. Kahandawala, K. Kainulainen, M. Kato, E. A. Kellogg, G. J. King, B. Klejevskaja, B. B. Klitgaard, R. R. Klopper, S. Knapp, M. A. Koch, J. H. Leebens-Mack, F. Lens, C. J. Leon, E. Leveille-Bourret, G. P. Lewis, D. Z. Li, L. Li, S. Liede-Schumann, T. Livshultz, D. Lorence, M. Lu, P. Lu-Irving, J. Luber, E. J. Lucas, M. Lujan, M. Lum, T. D. Macfarlane, C. Magdalena, V. F. Mansano, L. E. Masters, S. J. Mayo, K. McColl, A. J. McDonnell, A. E. McDougall, T. G. B. McLay, H. McPherson, R. I. Meneses, V. Merckx, F. A. Michelangeli, J. D. Mitchell, A. K. Monro, M. J. Moore, T. L. Mueller, K. Mummenhoff, J. Munzinger, P. Muriel, D. J. Murphy, K. Nargar, L. Nauheimer, F. J. Nge, R. Nyffeler, A. Orejuela, E. M. Ortiz, L. Palazzesi, A. L. Peixoto, S. K. Pell, J. Pellicer, D. S. Penneys, O. A. Perez-Escobar, C. Persson, M. Pignal, Y. Pillon, J. R. Pirani, G. M. Plunkett, R. F. Powell, G. T. Prance, C. Puglisi, M. Qin, R. K. Rabeler, P. E. J. Rees, M. Renner, E. H. Roalson, M. Rodda, Z. S. Rogers, S. Rokni, R. Rutishauser, M. F. de Salas, H. Schaefer, R. J. Schley, A. Schmidt-Lebuhn, A. Shapcott, I. Al-Shehbaz, K. A. Shepherd, M. P. Simmons, A. O. Simoes, A. R. G.

Simoes, M. Siros, E. C. Smidt, J. F. Smith, N. Snow, D. E. Soltis, P. S. Soltis, R. J. Soreng, C. A. Sothers, J. R. Starr, P. F. Stevens, S. C. K. Straub, L. Struwe, J. M. Taylor, I. R. H. Telford, A. H. Thornhill, I. Tooth, A. Trias-Blasi, F. Udovicic, T. M. A. Utteridge, J. C. Del Valle, G. A. Verboom, H. P. Vonow, M. S. Vorontsova, J. M. de Vos, N. Al-Wattar, M. Waycott, C. A. D. Welker, A. J. White, J. J. Wieringa, L. T. Williamson, T. C. Wilson, S. Y. Wong, L. A. Woods, R. Woods, S. Worboys, M. Xanthos, Y. Yang, Y. X. Zhang, M. Y. Zhou, S. Zmarzty, F. O. Zuloaga, A. Antonelli, S. Bellot, D. M. Crayn, O. M. Grace, P. J. Kersey, I. J. Leitch, H. Sauquet, S. A. Smith, W. L. Eiserhardt, F. Forest and W. J. Baker (2024). "Phylogenomics and the rise of the angiosperms." Nature **629**(8013): 843-850.