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 its 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. Despite their vast potential, historical samples available in museums and herbaria worldwide remains underutilized for studying host-microbiome interactions across broad temporal and spatial scales. It is due to incompatibilities with standard analytical pipelines and limited understanding of optimal classification parameters. While host DNA removal has traditionally been considered essential for accurate taxonomic assignment of metagenomic reads, this step is impractical for many historical samples because host reference genomes are unavailable for their species. Here, we show that host DNA content does not significantly affect microbial ecological analyses based on contemporary and historical samples. Additionally, DNA molecules from historical samples are highly fragmented and uneven in length. Conventional analysis workflows may be inefficient in this situation. To address this, we carried out detailed analyses on 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 using two annotation databases constructed with k=24 and k=31. Through a simulation study, we demonstrated that this approach outperforms conventional workflows in effectively recovering microbial signals from a wide range of read lengths, including fragments as short as 24 bp (21 bp if k=21 is used). 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).

High abundance of host DNA and low endogenous microbial biomass (Eisenhofer, Minich et al. 2019) make metagenomic analysis with historical samples challenging. Additionally, nucleic acid extracted from historical samples is prone to damage 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 the removal of host DNA 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. Alternatively, removal of host DNA can be performed *in silico* with the help of reference genomes from the host species or closely related species. The breadth of microbe-host-environment interactions, however, 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 the removal of host DNA necessary when analyzing 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 ecological metrics such as alpha and beta diversity. Second, the nucleic acids (either from host or microbes) surviving 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 uncertain how short metagenomic reads can provide reliable results. Without host DNA removal, shorter reads may increase the risk of misclassifying host-derived sequences as microbial in origin. However, discarding these short reads is not ideal, especially when working with precious samples from 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 like k-mer size or seed length. By adjusting these parameters, shorter reads can be incorporated into analyses, albeit at the cost of reduced specificity (lower taxonomic resolution). In this study, we investigated the impact of host DNA on profiling host-associated microbiomes in historical samples and evaluated parameters for analyzing the short reads typically recovered from these samples. To validate our findings, we conducted extensive analyses on simulated datasets containing damaged DNA patterns. Our ultimate goal is to assess whether metagenomic approaches can be applied broadly across samples lacking host reference genomes, both contemporary and historical.

Materials and Methods

Sampling, sequencing and collection of museum samples

In this study, we included 330 shotgun metagenomes of 110 contemporary rice cultivars by pooling leaf samples from multiple timepoints, followed by bacterial enrichment through multiple washing steps (Su, Wicaksono et al. 2022), and metagenomes from 17 museum samples (Table 1). Of the 17 museum samples, 5 phyllosphere metagenomes from herbarium samples of rice leaf were generated in this study and raw sequencing data was deposited to the EBI nucleotide archive (PRJEB87646). The raw sequencing data for other datasets can be found under the NCBI Sequence Read Archive (SRA): 330 rice phyllosphere metagenomes (*Oryza sativa)* (PRJEB45634), 5 historical common ragweed (*Ambrosia artemisiifolia*) (PRJEB48563) (Bieker, Battlay et al. 2022), 5 museum specimens of presumably extinct species of the rhinoceros beetle genus *Oryctes* (PRJEB34604) (Latorre, Herrmann et al. 2020) and 2 museum skin specimen of eastern gorilla (*Gorilla beringei*) (PRJEB29503) (van der Valk, Diez-Del-Molino et al. 2019). The genome assemblies for host (or closely related) species were obtained from the NCBI GenBank: rice (*Oryza sativa* AGIS1.0; GCA_034140825.1 for reference genome; MZ424305.1 for chloroplast; NC_011033.1 for mitochondrion), common ragweed (GCA_024762085.1 for reference genome; MG019037.1 for chloroplast), beetle (*Oryctes borbonicus*; GCA_001443705.1) and western gorilla (*Gorilla gorilla gorilla*; GCF_000151905; mitochondrial DNA). For the comparisons between host species and their associated microbiome, we included 2 catalogues of metagenome-assembled genome bins (MAGs) for the soil and skin microbiomes. The soil MAGs catalogue (SMAG catalogue) contains 21,077 MAGs

reconstructed from 3,304 soil metagenomes and resolved at the species level, were obtained under the accession PRJNA983538 (Ma, Lu et al. 2023). The skin microbial genome collection (SMGC) comprises 251 isolate genomes and 7,535 MAGs reconstructed from 4,802 skin metagenomes were retrieved from the EBI FTP server (http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/skin_microbiome/01_allprok/) (Saheb Kashaf, Proctor et al. 2022). The datasets included in this study were summarized in Supplementary Table 1.

Sample collection and preparation of rice metagenomics

In this study, we generated data for rice to compare the impact of host DNA removal in contemporary and historical samples. We identified five herbarium specimens of rice at the herbarium collection of Smithsonian Institution, collected between 1906 and 1914 in the Philippines. 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 established for historical plant tissues by Latorre et al. (2020) were followed. 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 shotgun metagenomic data

For reproducibility, 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). This pipeline streamlines high throughput metagenomic data analysis on large-scale computing facilities. To assess the effects of host genome removal on non-model organisms, we utilized the quality control and taxonomic classification modules of the pipeline. 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 on default settings (Chen, Zhou et al. 2018). PhiX derived reads were discarded by mapping the quality filtered reads to the PhiX genome using Bowtie2 (v.2.3.4) with the option --sensitive (Langmead and Salzberg 2012).

Removal of host derived sequences

We evaluated the effect of host DNA content on metagenomic analyses of contemporary and historical samples by comparing the microbial community metrics between the same dataset obtained by nf-core/mag pipeline with (i) host DNA removal or (ii) without this step (Figure 1). For the host removal step (i), filtered reads were aligned to host genome using Bowtie2 with default settings. Reads that did not align to the host genome were then subject to taxonomic classification. For the workflow without host removal step, filtered reads were taxonomically annotated after quality check and adaptor removal.

Taxonomic classification

Taxonomic classification of DNA reads processed with and without host removal step, was performed using Kraken2 (v2.0.9) (Wood, Lu et al. 2019) with a confidence value of 0.4 against the PlusPFP database (k-mer=35, version 2024.06.05), which contains archaea, bacteria, viral, plasmid, human, UniVec_Core along with protozoa, fungi and plants. Kraken2 classifies individual 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 parsed and converted to TSV format using the kraken-biom software (Dabdoub 2016). Non-microbial taxa (e.g. those assigned to eukaryotic species) were discarded from further microbiome analyses. To normalize sequencing depth, samples were rarefied (down-sampled) to the minimum value of bacterial read sums in their respective dataset

(contemporary rice=510,147; rice herbarium=4,409; museum ragweed=69,042; museum beetle=625; museum gorilla skin=1,632,030). OTUs were further filtered with a minimum read count of 50 and prevalence of 10% (for the contemporary rice samples).

K-mer analysis of host and resident genomes

To evaluate the similarity in DNA between host and resident microbiome, k-mer analysis of genomic sequences was performed with JellyFish (Marcais and Kingsford 2011). For the soil microbiome, a single representative OTU was shortlisted for each of 107 pangenomes in the SMAG catalogue based on 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 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 them from the NCBI GenBank using "download" command of the NCBI datasets tool. For the skin microbiome, 837 genomic sequences (622 prokaryotic species derived from 7,535 MAGs and 251 isolate genomes) representing core skin microbes in the SMGC were included (Saheb Kashaf, Proctor et al. 2022). Counting of k-mer shared between host and microbial genome sequences was carried out with JellyFish "count" command in 5 different k-mer size (--mer-len=11, 15, 18, 21 and 24), with Bloom filter disabled and both strands considered (--canonical). K-mer substrings were then

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 the window size, e.g. window.size=15 for k=15.

Generation and analyses of simulated ancient DNA datasets

To directly evaluate the effect of various search parameters such as the size of k-mers, complexity of host genome and its DNA percentage in a sample, simulated datasets were generated using gargammel, a NGS read simulator for aDNA (Renaud, Hanghøj et al. 2017). A simulated aDNA dataset generated by gargammel comprises of 3 components: endogenous DNA (***endo***), bacterial (***bact***) and contemporary contamination (***cont***). We created 3 custom sequence sets, each of them comprised of the reference genomes of a plant species as endogenous DNA (***endo***), soil microbes (***bact***), and skin microbes commonly observed in museum collections and human as contamination (***cont***). The plant species were rice (*Oryza sativa* AGIS1.0), maize (*Zea mays*; GCF_902167145) and bread wheat (*Triticum aestivum*; Genome assembly IWGSC CS RefSeq v2.1; chromosomes only), featuring host genomes with increasing genome size and complexity (Figure 5). For the soil microbes, we shortlisted 120 bacteria (see Supplementary Table 6 for a complete list of the soil microbes included in the simulation study) representing 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 with a fragment length profile of a herbarium rice sample and a damage pattern setting (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) as observed in Pleistocene organisms (Briggs, Stenzel et al. 2007). The fragment length profile was obtained by profiling the length of merged reads from one of the herbarium rice samples included in this study (HRB0101R) 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. 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.

Computational cost assessment and statistical analysis

In order to assess the effects of host genome removal on computational resource usage and time cost for each workflow process of microbial community analysis, we used the nf-core/mag execution trace report files provided by nextflow. All statistical methods and visualization were performed in R (v.4.2.2) with the corresponding R-packages (phyloseq, vegan, ggplot2). Alpha diversity (within-sample) diversity (Chao1 and Shannon index) was calculated using estimateR and diversity functions from the vegan package. Pairwise comparisons using Wilcoxon rank sum test with continuity correction and the Holm-Bonferroni method for *P*-value adjustment were used to assess the differences between the two pipelines with and without host genome removal. Beta diversity (between samples) was assessed by analysis of similarities (ANOSIM) at 999 permutations to test for statistical significance (*P*-values) in microbial community assemblages among all the datasets using anosim function. Bray-Curtis distance metrics at the species level were used to construct 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 in subsampled taxon table was considered absent. Part of the analyses in this study was conducted on CropDiversity-HPC (Percival-Alwyn 2024).

Results

Effect of host DNA removal on taxonomic assignments

The host DNA removal 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 samples associated with 4 different host species (rice, ragweed, extinct rhinoceros beetle and eastern gorilla) 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, the reference genome for exact host species was used whenever possible. The reference genomes for rice and common ragweed were publicly available whereas genomic sequences for extinct rhinoceros beetle and eastern gorilla were unavailable. Instead, we used the reference genomes of two species closely related to them, scarab beetle (*Oryctes borbonicus*) and western lowland gorilla (*Gorilla gorilla gorilla*) (Supplementary Table 1). To evaluate the impact of host DNA removal on metagenomics analyses, we considered the number of reads, OTUs obtained and microbial community diversity between the two 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 5% in contemporary rice and museum beetle to over 50% in some of gorilla and herbarium rice samples (Supplementary Table 2). As expected, host DNA content was markedly reduced when the samples were processed

with the host DNA removal. The percentage of endogenous non-bacterial DNA content was decreased, especially in herbarium rice, museum gorilla and contemporary rice datasets (Supplementary Figure 1). On average, host derived reads (either from plant or metazoa) had as much as 25-fold difference between the workflows (Supplementary Table 3). 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 3).

The taxonomic assignment was performed using Kraken2 with the prebuilt 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 and eastern gorilla were common sunflower (*Helianthus annus*) and human (*homo sapiens*), respectively (Supplementary Table 4). Detected OTUs were then filtered based on their abundance and prevalence. Before filtering, the numbers of OTUs were between 8,200-10,600, 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). Of 330 contemporary rice samples (3 replicates for each rice variety), same number of OTUs (9,164) were detected in both workflows, and 21.6% of the OTUs were removed in subsequent filtering step with a filtering criterion applied (a minimum read count of 50 and prevalence of 10%). Only 4 OTUs (0.1%) failed the filtering step when the host removal step was applied, resulting in 7,184 OTUs. The number of OTUs with more than 5% change in read count (abundance proxy) was less than 0.1%. Host DNA removal exerted a negligible effect on the contemporary rice samples. For the historical samples, a significant portion of detected OTUs (>80%) were filtered out, except in herbarium ragweed samples. Museum beetle had the lowest number of filtered

OTUs (330) while 6,956 OTUs were retained in herbarium ragweed samples after filtering. The host DNA removal step led to 0.0%-16.4% decrease in the number of OTUs, of which almost all were in low-abundance with a read count of 100 or less, suggesting that these OTUs were rare or misclassified taxa (Table 1 and Supplementary Figure 2). For the retained OTUs, most (50.5% of OTUs in herbarium ragweed, 92.8% in museum beetle) had less than 5% difference in the number of assigned reads between workflows. Those with highly variable read counts (>5% difference) between the workflows had a mean read count much lower than those OTUs with a consistent read count. Altogether, low abundance or misclassified OTUs were more prone to the host DNA removal step (Table 1).

Distinct taxonomic profiles were observed across different host-associated microbiota (Figure 3a and Supplementary Figure 3). Genera with relative abundance lower than 2% were dominant in the samples of rice and ragweed, while taxa from the genus *Curtobacterium* and *Acinetobacter* were overrepresented in herbarium and contemporary rice samples, respectively. Overall, relative abundance of taxa at phylum and genus level showed no consistent difference between the workflows with and without host DNA removal. For example, the exclusion of the host DNA removal step increased share of metagenomic assignments to Actinomycetota in rice and gorilla, but decreased in ragweed and beetles.

Effects of host DNA content on the metrics of microbial diversity

The 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 the host DNA removal step would alter the community diversity and richness metrics of the host-associated microbiome by measuring alpha and beta diversity. 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=0.97) between the workflows with and without host DNA removal step, with average values of 1,667.7 ±1,200.1 and 1,660.2 ±1,201.7, respectively. A similar pattern was obtained for Shannon diversity index (with host removal=2.45 ±1.29 and without=2.46 ±1.29), where no significant differences (p=0.99) was observed (Figure 3a). The overall microbial diversity and community structure in historical, museum samples were consistent between workflows and non-significantly different, despite of a higher percentage of OTUs discarded after host DNA removal (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 two bioinformatic treatments (R=-0.003228, p=1), even in historical samples (Figure 3). These

results suggest that the community metrics were not sensitive to the host DNA removal procedure.

K-mer analysis and optimal k-mer size for 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 approach provided an overview of the extent to which sequence fragments from the host genome could be mistakenly assigned to its associated microbiota purely by chance. The exact reference genomes were obtained for rice and ragweed while the genomes of closely related species were used for beetle and western gorilla. For host-associated microbiomes, we focused on skin and soil microbiota whose comprehensive catalogues of pan-genomes of metagenome-assembled genome (MAG) bins and microbial reference genomes (microbial isolates) were available, 837 in skin and 107 in soil respectively. These pan-genomes were reconstructed from more than 3,000 metagenomes, suggesting that they robustly represent the core microbes in their respective biome. For each microbiome, we counted distinct k-mers 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 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, west gorilla genome (3.1Gb) produced 1000-

fold more distinct k-mers at k=11 compared to k=24 (2,097,081 vs 2,324,433,502). For skin and soil, the number of distinct k-mers depends on the richness of microbial community (the total genomic size of MAGs in this study). Roughly 1.51 billion of distinct k-mers were obtained from the skin microbiome (2.7Gb) at k-mer size of 24, compared to 0.37 billion in the soil microbiome (388Mb). We observed a similar pattern for k-mer numbers for individual pan-genome MAGs (Supplementary Table 3).

The number of shared k-mers between host and its associated microbiome (e.g. gorilla vs the skin 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; Supplementary Figure 5; Supplementary Table 2). 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). Despite its large genome size, gorilla and the skin microbiome had a Jaccard index of 0.063 at k=18, which further declined to 0.02 at k=21. This pattern was consistent whether counting distinct k-mers from MAGs collectively or individually (Figure 4a-b; Supplementary Figure 6). When k-mer occurrence (i.e. the frequency of a given k-mer within a genome/metagenome) was considered, distinct density profiles were observed between k=11 and other larger k-mer values in rice genome and soil microbiota (Figure 4c). At k=11, k-mers with occurrence between 50 and 1,000 were predominant, whereas a vast proportion of k-mers was singleton (occurrence = 1) for k-mer size of 15 or greater. Since shared k-mers were a subset of the constituent k-mers detected in host genome, they were expected to have a similar distribution of density in occurrence. However, when k-mer sizes exceeded 18 (e.g. k=21 or 24),

the fraction of high occurrence k-mers (1 thousand times or more) were higher than expected (Figure 4c). Furthermore, the sequence complexity of k-mers shared between host and its associated microbiome were significantly lower with increasing k-mer size (Figure 4d), indicating that low complexity sequences could more easily be misclassified

Analyses of simulated datasets for ancient DNA samples

Due to the limited availability of historical samples, we performed a simulation study to evaluate the impact of host DNA content on microbial analyses and the short DNA molecules (20-70bp) commonly recovered from historical samples. A total of 6,000 aDNA datasets were generated using *gargammel* simulator from the genomes of three plant hosts (rice, maize and bread wheat), 107 representative soil microbes (Supplementary Table 6), human and two common human skin microbes (Supplementary Figure 7). 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 31 bp (Supplementary Figure 8), closely resembling the fragment length profiles typically observed in DNA from historical samples. The merged reads were processed with and without the removal of host DNA, followed by the taxonomic classification against 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 9). Without host removal, the highest misclassification rate in the simulated rice datasets was 0.2% when shorter k-mer values (k=21 and 24) were used; this dropped to below 0.1% after host DNA was

removed. It was reflected by the consistency in abundance of soil microbes between the datasets with and without host removal (Supplementary Figure 10). As expected, the number of identified microbial reads correlated with their proportions in the simulated datasets, but also strongly influenced by k-mer size (Figure 5B; Supplementary Figure 11). 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. Across our simulation, the number of microbial reads resolved at the species-level ranged from 7,890 to 375,850. Species-level classification was substantially higher with k=31 compared to k=35 (Figure 5C and Supplementary Figure 11) but only marginal improvement was observed with k-mers shorter than 31 (e.g. k=28). Instead, more reads were assigned to higher taxonomic ranks (e.g. genus or even phylum) with shorter k-mers, indicating that reducing k-mer size may boost the overall classification rate but at the expense of taxonomic resolution.

Because DNA fragments from historical samples are highly variable in length (Supplementary Figure 8), we further examined read length distribution across different k-mer values. Regardless of k-mer value, classified microbial reads followed a lognormal distribution with a peak length at 30-40 bp (Figure 5D). 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 are ignored by Kraken2. These profiles suggested that annotation with two or more k-mer values can extend the coverage across a wider fragment length spectrum.

To optimize metagenomic analyses of aDNA, we compared the single k-mer workflow with a two-step annotation strategy. In the single k-mer workflow, reads were annotated with a single kraken2 database (commonly built with k=35 database). Only 51% of microbial reads were classified at the species level (Figure 6B) with a k=35 database. The classification rate improved to 71.8% when a k=24 database was used, but shorter k-mers increased the risk of cross-domain misclassification, especially when the host DNA content was not removed. To address this, we implemented 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 as suggested by our simulations), and unclassified reads were subsequently processed with a database generated with a shorter k-mer value (e.g. k=24). The strategy achieved the highest classification rate (79.9%), outperforming the single k-mer workflow. Although host DNA removal is not strictly necessary, the removal of human-derived or skin microbes DNA prior to downstream analyses is still recommended.

Computational costs of host genome removal

We then investigated the influence of host genome removal on the computational resources for each of the workflow processes. To do this, we focused on the currently available

best-practice pipeline for performing each step of a typical metagenomic dataset analysis. In the box plots, we estimated the amount of memory in gigabytes (GB), CPU usage, and the time duration consumed by individual step in the workflow. Quality control check of the reads was performed twice using FASTQC before filtering and after filtering the raw sequencing reads using FASTP. Host DNA was removed using bowtie2, an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences (Langmead and Salzberg 2012). The average amount of memory used by all processes in the pipeline with host removal step enabled was 32.5 ± 56.9 GB, while it was reduced to 27.5 ± 52.9 GB when the removal step was disabled (Supplementary Figure 13). Among the steps in the pipeline, Kraken2 consumed the highest amount of system memory, with an average of 146 GB, regardless of host removal step enabled or not. All other steps had an average memory usage of 10 GB. The total runtime for the workflows with and without host removal was 8.5 ± 8.2 and 8.8 ± 7.4 minutes, respectively. The pipeline spent the longest CPU time on host removal step (bowties2). Besides, there was no significant difference between the two pipelines on CPU usage.

Discussion

Historical samples, such as those from museum or herbarium collections, offer valuable 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 false identification of taxa (primarily through the misclassification of host DNA as 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 considered to be an established 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 when host DNA is not removed. To this end, we profiled the host-associated microbiome in contemporary and historical samples and compared the microbial diversity and community structure between the workflows with host DNA removal and without it. Our findings suggests that host DNA has a negligible effect on microbial analyses, regardless of their content in the samples. In modern 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 human, the closet relatives of eastern gorilla which shares a high degree of genomic similarity, is the only primate species available in the database (Scally, Dutheil et al. 2012). 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 1). The abundance of taxa called differentially between our two tested approaches is generally low (except for the gorilla metagenome, Table 1). In contrast to the host DNA removal step, OTU filtering step also has a strong impact on downstream microbiome analysis, reducing the number of falsely identified taxa (Table 1), which are typically in low abundance and prevalence (Cao, Sun et al. 2020). It is important therefore to consider the abundance filtering criteria thoroughly, especially when excluding the host DNA removal step. 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). This was shown in our findings that the abundances of individual OTU were more stable between the workflows in contemporary rice samples, compared to historical rice samples, regardless of whether the host DNA removal step was performed (Figure 1). 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. By default, prebuilt kraken2 databases were compiled 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 size have been proposed for various scenarios and types of datasets. The databases used for filtering exogenous DNA in historical samples were built with a k-mer size of 29 (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 to validate 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). Similarly, skin-colonizing microbes help maintain cutaneous homeostasis and serve as an additional barrier against the invasion of exogenous pathogens (Byrd, Belkaid et al. 2018). The composition of both skin and 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 and skin microbiome, which were reconstructed from an extensive collection of soil and skin metagenomes. The constituent k-mers shared between host genome and associated microbiome was considered as an indicator of taxonomic misclassification due to host DNA contamination. As expected, the number of k-mers shared between host genome and core microbes in their

respective associated microbiome 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. Nonetheless, with a larger k-mer size (e.g. k=18 or above), the overlap in k-mer space was significantly reduced (Jaccard index < 0.1), suggesting that the misclassification of host DNA into microbial taxa become unlikely at a large k-mer size. Based on our observation, k-mer size of 18-21 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 strongly suggests that a sequence complexity filter, such as DUST algorithm (Morgulis, Gertz et al. 2006), should be applied prior to taxonomic classification to ameliorate the impact of low-complexity sequences. The findings were supported by an extensive analysis study of the simulated ancient DNA datasets using soil microbes and 3 host plant species with varying genome size and complexity. Our analyses strongly suggest that while the microbiome analyses on historical samples are feasible, a thorough and careful considerations must be given to cope with the unique features (e.g. short fragments) of DNA molecules retrieved from historical samples and to minimize the misclassification of host DNA. The 2-step annotation workflow proposed in this study can effectively improve the taxonomic annotation of short DNA fragments.

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 be performed on historical samples where genomic sequences of the host species are unavailable. We also showed that microbial reads as short as 21 bp are distinguishable from those from host species and 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). Additionally, a sequence complexity filter should be useful to limit the misclassification of low-complexity host DNA sequences, particularly when host DNA removal is not feasible. In addition, despite our findings suggest the genome size and complexity may not be a critical parameter impacting the downstream microbiome analyses, a thorough evaluation on analytic workflow is still merited. Finally, this study does not consider thoroughly contamination and post-mortem colonization that is common in historical samples. The host DNA content and optimal analysis strategies on microbiome analyses were the prime focus of this study and the findings were not affected by the origin of the microbes. Future work should identify microbial contaminants through constructing a post-mortem reference database or constructing relevant metagenomic classifiers, likely unique to each kingdom. This work 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

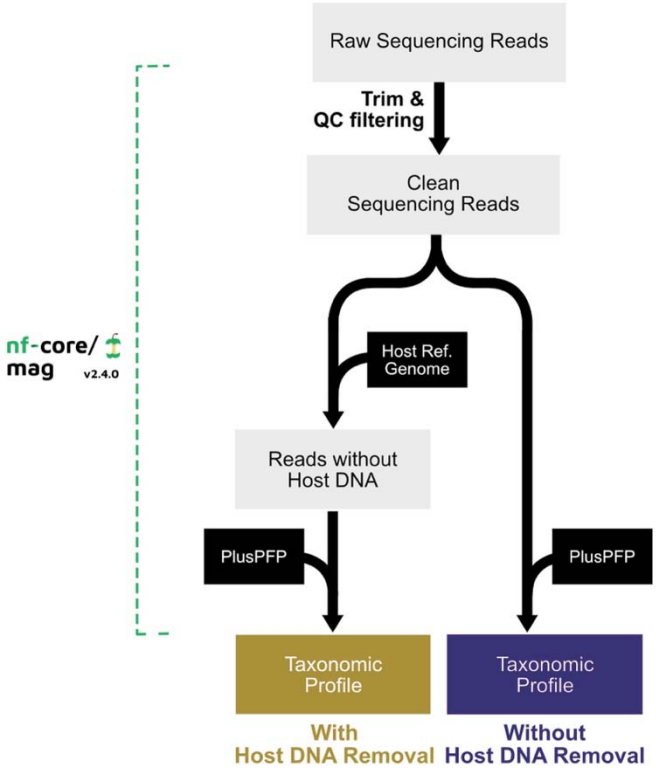


Figure 1. The bioinformatic workflow. The Nextflow metagenomic analysis workflow (nf-core/mag) was used in this study. Two separate workflows (gold and purple) were performed in parallel with identical steps except the host removal step prior to taxonomic classification.

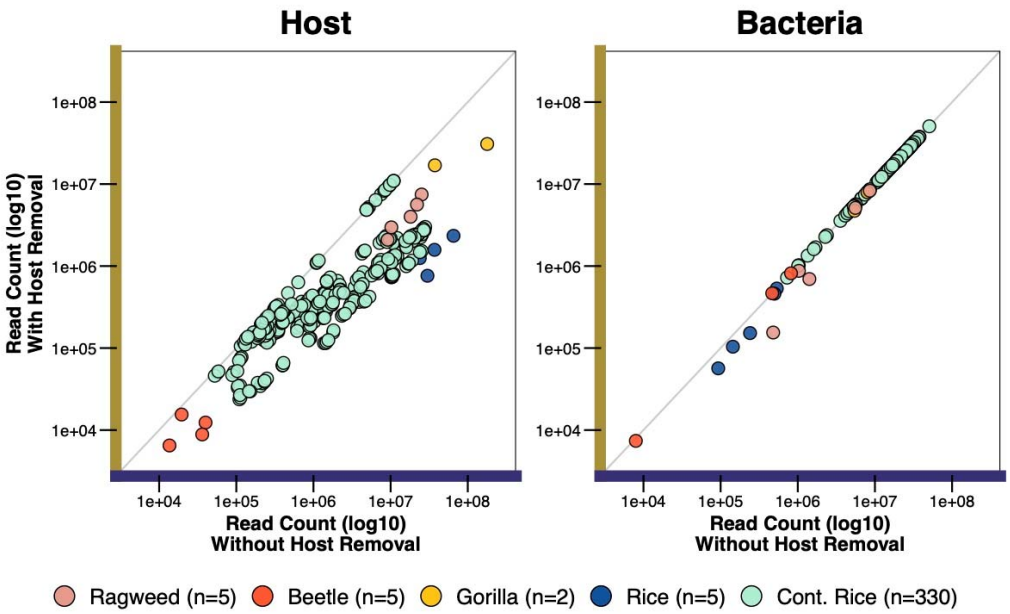


Figure 2. Proportion of sequencing reads (log10) assigned to host (plant or metazoa) and microbial taxa with (y-axis; gold) and without (x-axis; purple) host DNA removal. Each circle represents a sample and is color-coded with host species.

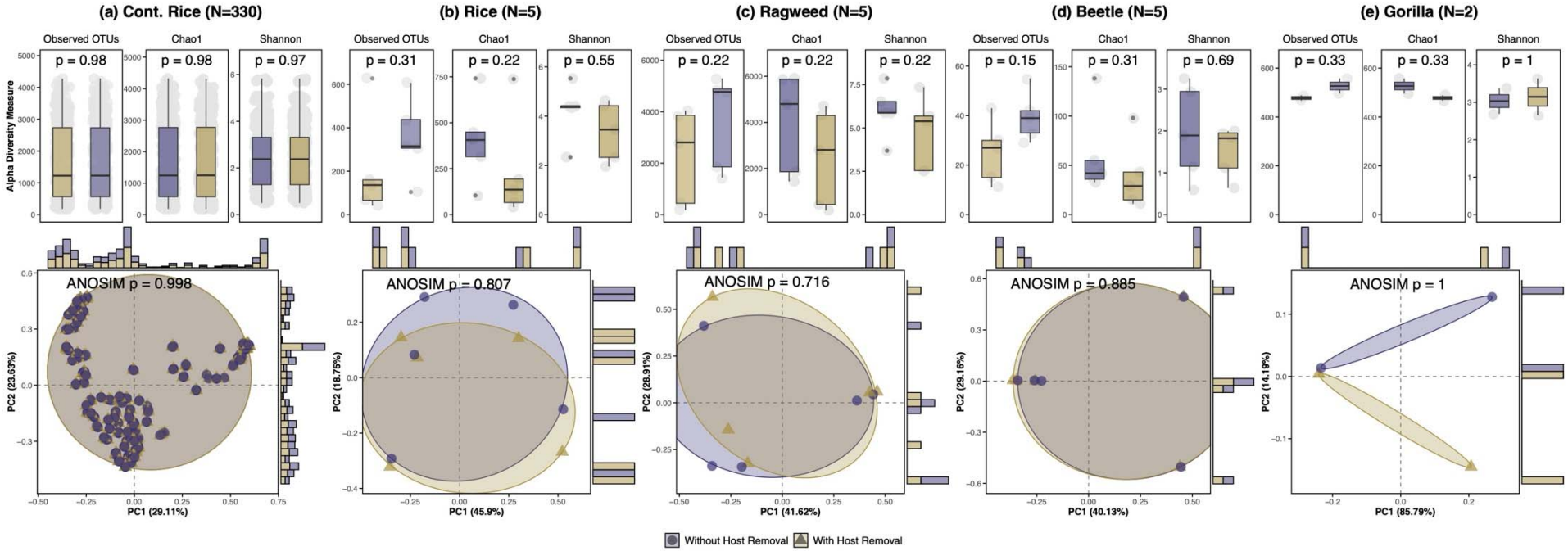


Figure 3. Alpha and beta diversity analysis of microbial community in the samples of contemporary rice (a), herbarium rice (b), herbarium ragweed (c), museum beetle (d) and museum gorilla (e). Microbial diversity analysis using Chao1 and Shannon diversity indices (top panels) and microbial community structure using principal coordinates analysis (PCoA) based on Bray-Curtis dissimilarity index (bottom panels) in pipelines with (gold) and without (purple) host DNA removal. *P*-value was obtained from the Wilcox-rank sum test for microbial diversity.

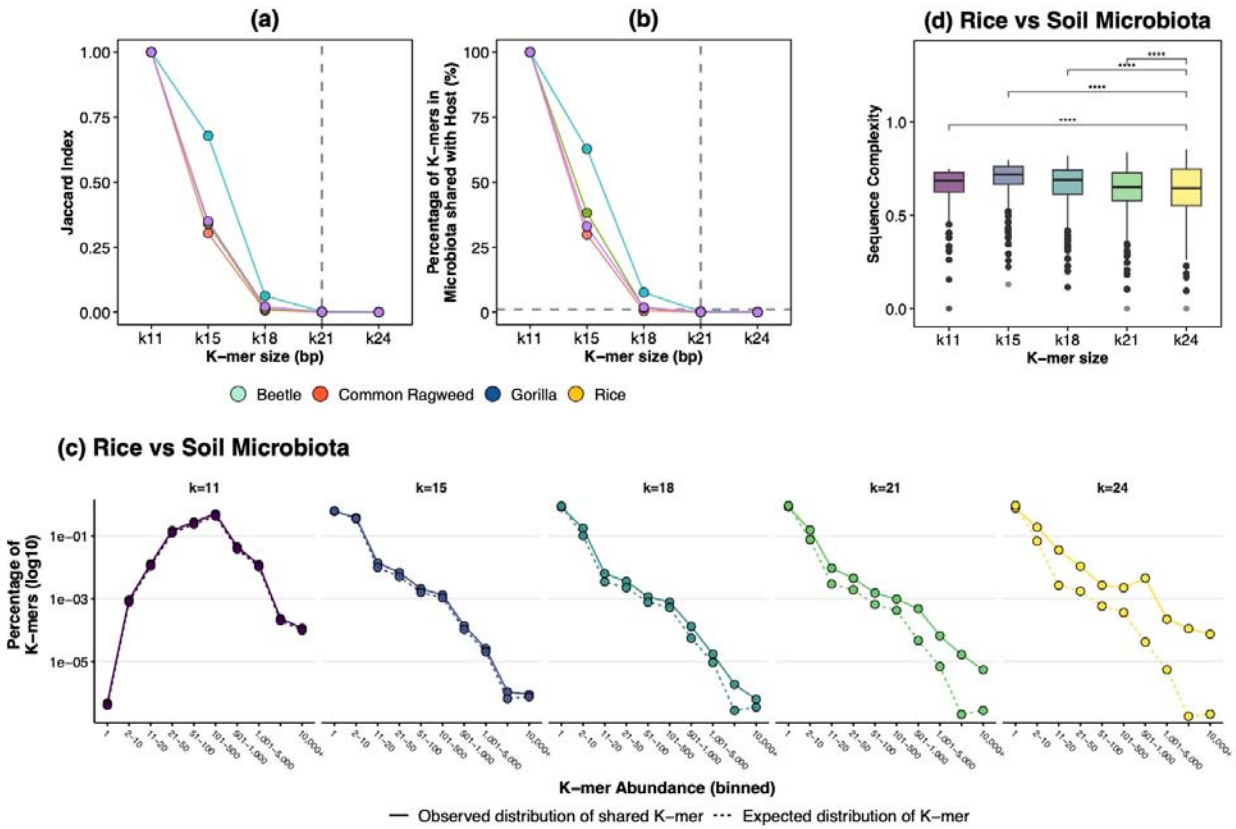


Figure 4. Shared k-mers among host species and representative MAGs of their respective associated microbiome (beetle vs soil microbiome, common ragweed vs soil microbiome, gorilla vs skin microbiome and rice vs soil microbiome). Jaccard index (a) and percentage of k-mers shared between host and associated microbiome (b) across k-mer sizes of 11, 15, 18, 21 and 24. (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 sizes.

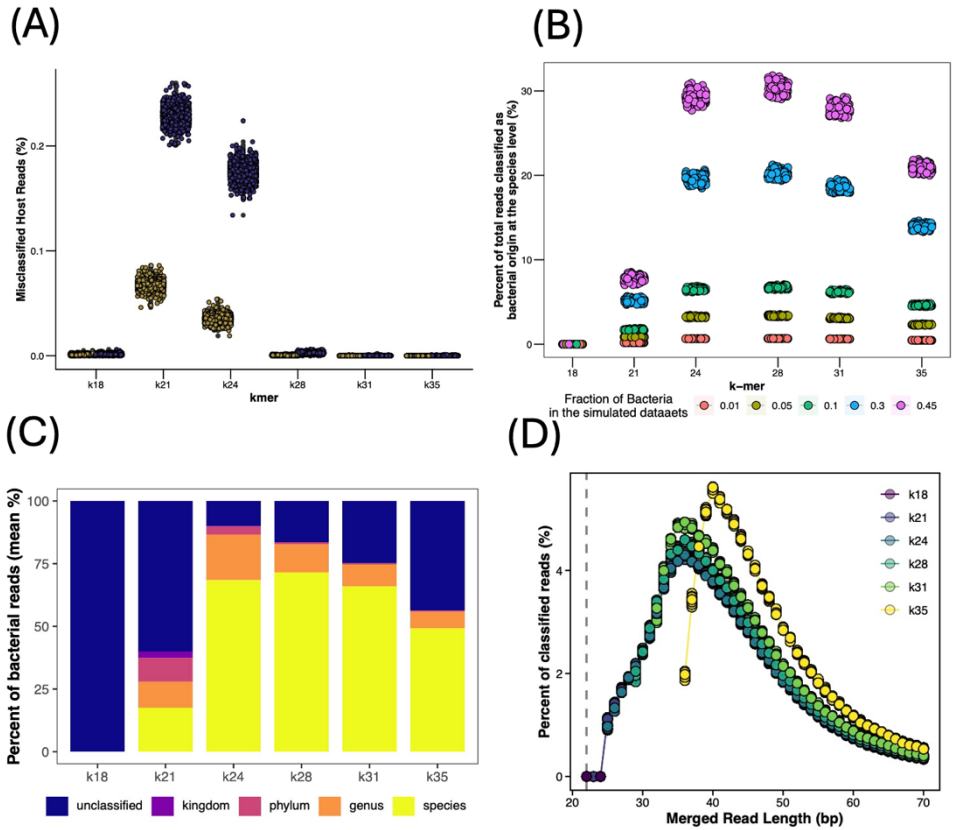
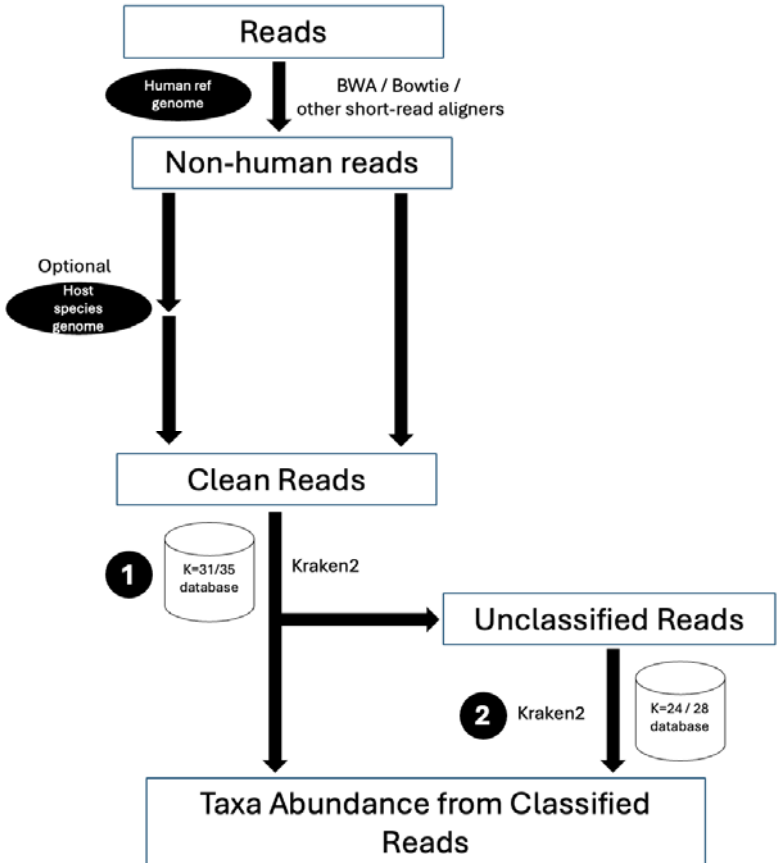


Figure 5. The analyses of the simulated rice datasets. **(A)** The percentage of host reads misclassified as microbial taxa. The simulated datasets were processed with (gold) and without (purple) host DNA content removal step. **(B)** The number of classified bacterial reads in the simulated rice aDNA datasets after the removal of host DNA content (rice). 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.

(A)



(B)

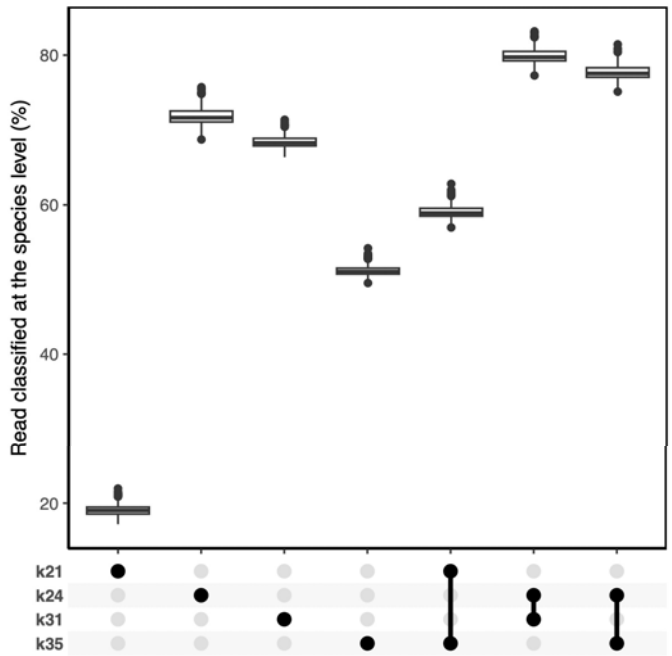


Figure 6 **(A)** Proposed 2-step analysis workflow for the taxonomic annotation of metagenomic reads from historical samples. **(B)** The percentage of microbial reads identified by one k-mer or 2-kmer search strategies in the simulated rice datasets. The k-mer value(s) used for the annotation was highlighted with a circle in black in the bottom layer of the plot.

Dataset	Sample Size (N) [§]	With Host DNA Removal			Without Host removal			OTU Removed by Host DNA removal [#]	OTU# with >5% change in read count between workflows	Mean read count	
		Before filtering [¶]	After filtering	% filtered	Before filtering	After filtering	% filtered			for OTU with >5% change	for OTU with <5% change
Contemporary Rice	330	9,164	7,184	21.6%	9,164	7,188	21.6%	0 (0.0%)	6 (0.1%)	68.7	2,464.9
Herbarium Rice	5	8,197	798	90.3%	8,279	1,034	87.5%	82 (1.0%)	357 (34.5%)	57.2	194.8
Herbarium Ragweed	5	8,706	5,862	32.7%	8,773	6,956	20.7%	67 (0.8%)	3,443 (49.5%)	165.3	818.1
Museum Beetle	5	5,647	285	95.0%	6,575	330	95.0%	928 (16.4%)	25 (7.6%)	22.8	242.1
Museum Gorilla	2	10,358	578	94.4%	10,544	662	93.7%	186 (1.8%)	68 (10.3%)	5,461.2	4,639.7

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

[¶] Filtering cutoff: read count = 50, prevalence = 10% 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 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.

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."

Byrd, A. L., Y. Belkaid and J. A. Segre (2018). "The human skin microbiome." Nat Rev Microbiol **16**(3): 143-155.

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.

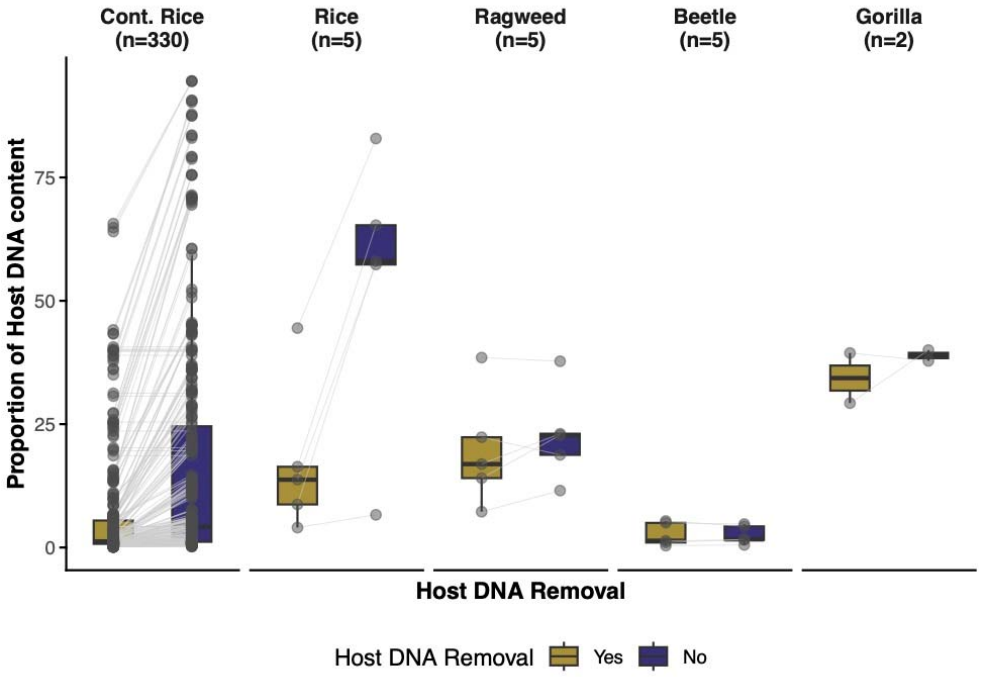
Gancz, A. S., A. G. Farrer, M. P. Nixon, S. Wright, L. Arriola, C. Adler, E. R. Davenport, N. Gully, A. Cooper, K. Britton, K. Dobney, J. D. Silverman and L. S. Weyrich (2023). "Ancient dental calculus reveals oral microbiome shifts associated with lifestyle and disease in Great Britain." Nat Microbiol **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., M. Herrmann, M. J. Paulsen, C. Rödelsperger, A. Dréau, W. Röseler, R. J. Sommer and H. A. Burbano (2020). "Museum phylogenomics of extinct Oryctes beetles from the Mascarene Islands." bioRxiv: 2020.2002.2019.954339.
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.
Morgulis, A., E. M. Gertz, A. A. Schaffer and R. Agarwala (2006). "A fast and symmetric DUST implementation to mask low-complexity DNA sequences." J Comput Biol **13**(5): 1028-1040.

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).
Saheb Kashaf, S., D. M. Proctor, C. Deming, P. Saary, M. Holzer, N. C. S. Program, M. E. Taylor, H. H. Kong, J. A. Segre, A. Almeida and R. D. Finn (2022). "Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions." Nat Microbiol **7**(1): 169-179.
Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, S. McCarthy, S. H. Montgomery, P. C. Schwalie, Y. A. Tang, M. C. Ward, Y. Xue, B. Yngvadottir, C. Alkan, L. N. Andersen, Q. Ayub, E. V. Ball, K. Beal, B. J. Bradley, Y. Chen, C. M. Clee, S. Fitzgerald, T. A. Graves, Y. Gu, P. Heath, A. Heger, E. Karakoc, A. Kolb-Kokocinski, G. K. Laird, G. Lunter, S. Meader, M. Mort, J. C. Mullikin, K. Munch, T. D. O'Connor, A. D. Phillips, J. Prado-Martinez, A. S. Rogers, S. Sajjadian, D. Schmidt, K. Shaw, J. T. Simpson, P. D. Stenson, D. J. Turner, L. Vigilant, A. J. Vilella, W. Whitener, B. Zhu, D. N. Cooper, P. de Jong, E. T. Dermitzakis, E. E. Eichler, P. Flicek, N. Goldman, N. I. Mundy, Z. Ning, D. T. Odom, C. P. Ponting, M. A. Quail, O. A. Ryder, S. M. Searle, W. C. Warren, R. K. Wilson, M. H. Schierup, J. Rogers, C. Tyler-Smith and R. Durbin

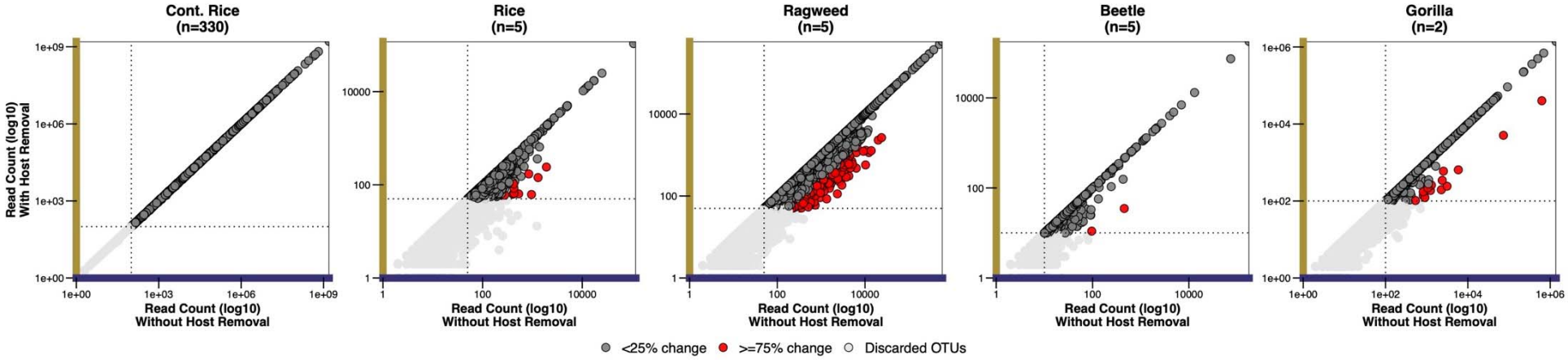
(2012). "Insights into hominid evolution from the gorilla genome sequence." Nature **483**(7388): 169-175.
Schmieder, R. and R. Edwards (2011). "Fast identification and removal of sequence contamination from genomic and metagenomic datasets." PLoS One **6**(3): e17288.
Su, P., W. A. Wicaksono, C. Li, 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.
van der Valk, T., D. Diez-Del-Molino, T. Marques-Bonet, K. Guschanski and L. Dalen (2019). "Historical Genomes Reveal the Genomic Consequences of Recent Population Decline in Eastern Gorillas." Curr Biol **29**(1): 165-170 e166.
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.

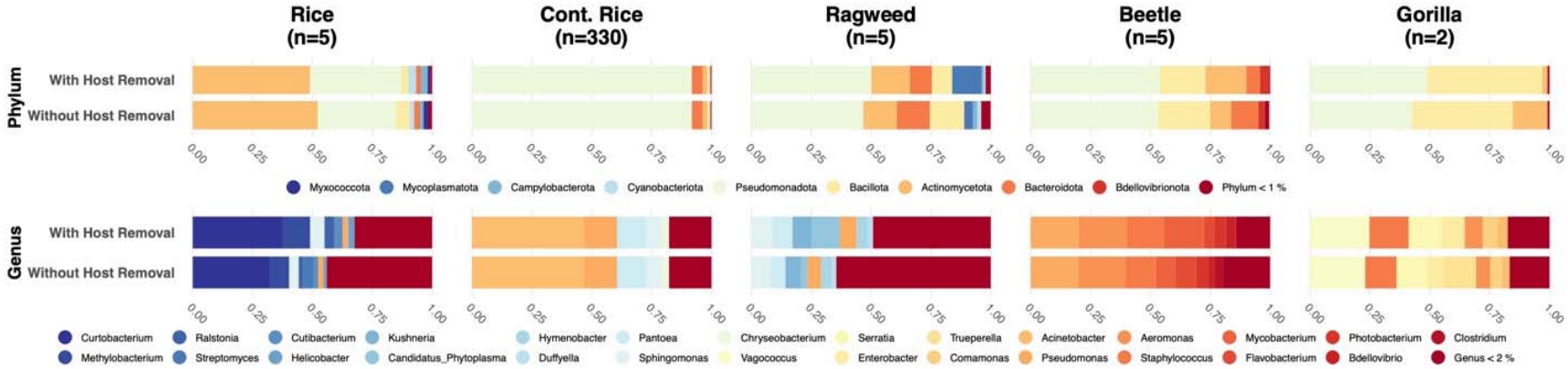
Supplementary Table and Figures



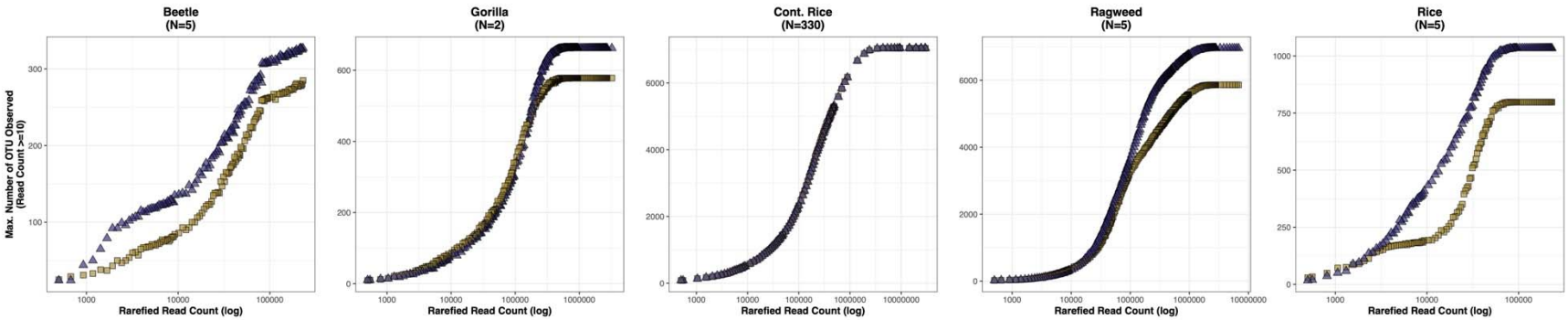
Supplementary Figure 1. Proportion of host DNA content in each sample (point in grey) across datasets between the workflow with (gold) and without (purple) host DNA removal.



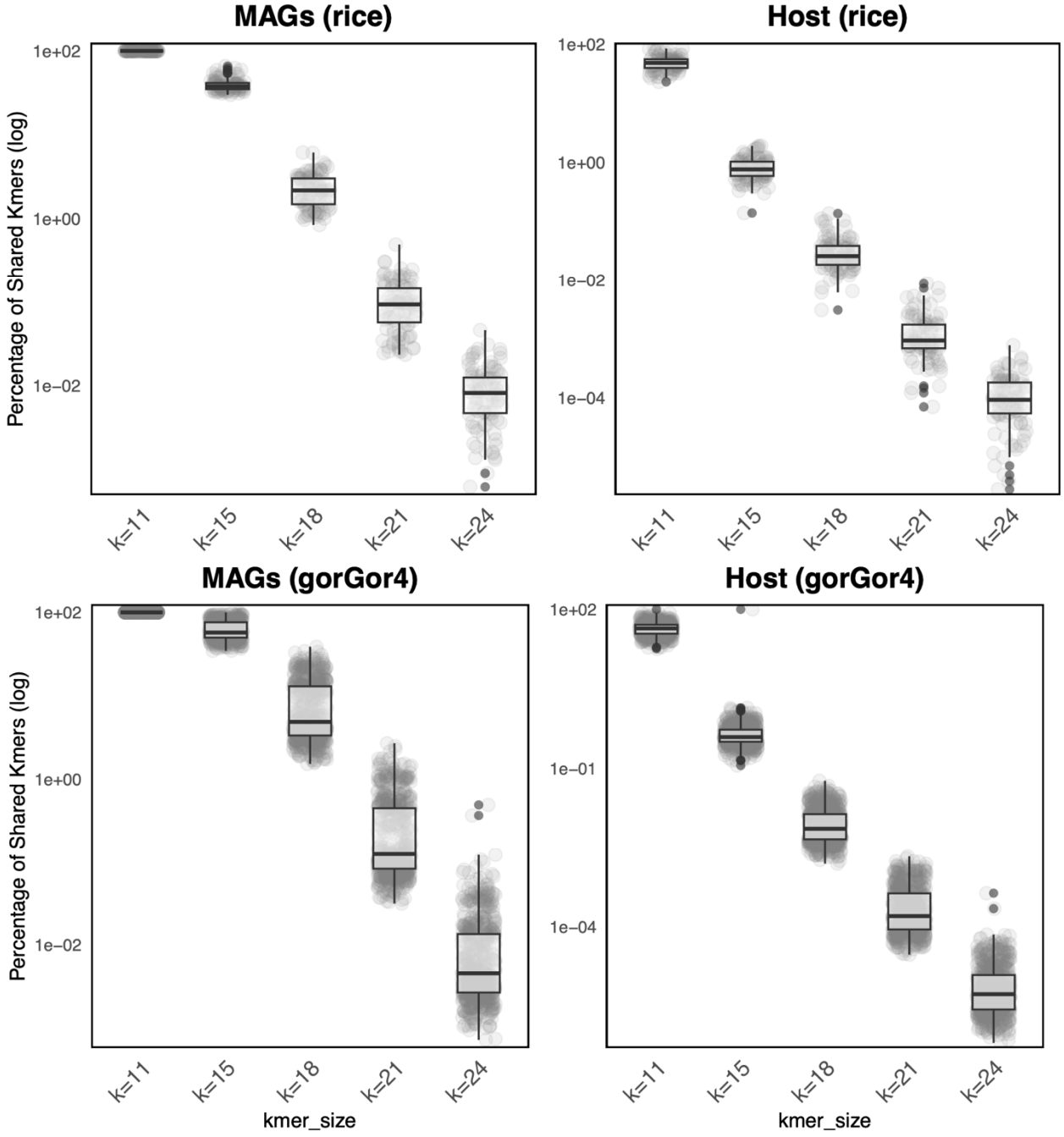
Supplementary Figure 2. The distribution of bacterial OTUs in assigned read count (log10) between two workflows with (y-axis; gold) and without (x-axis; purple) host DNA removal. Each circle represents a bacterial OTU. Those with more than 75% change in read count between two workflows are highlighted in red.



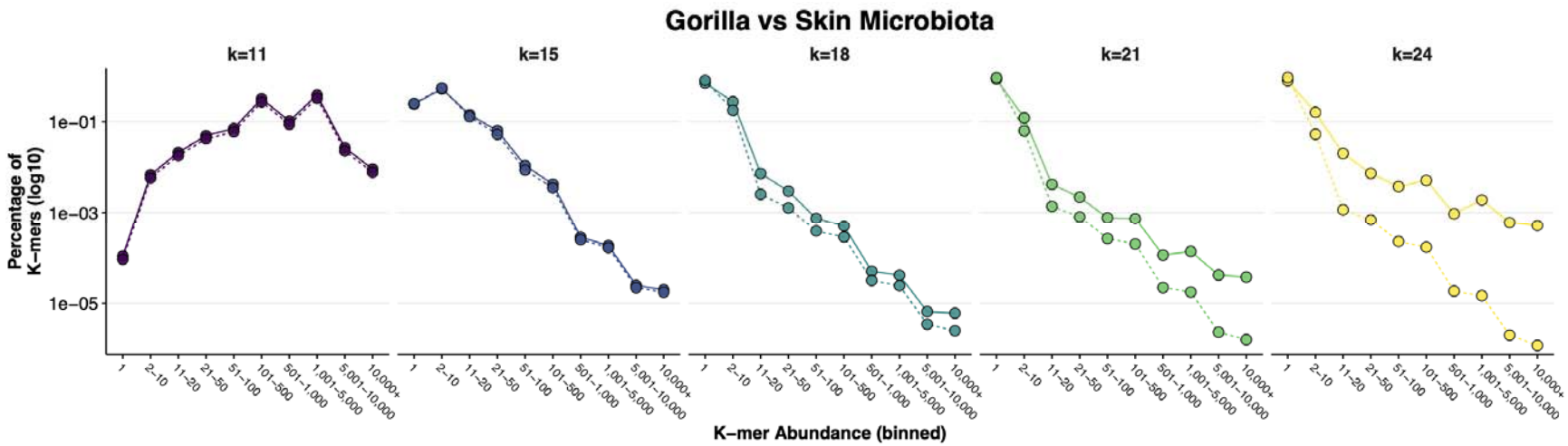
Supplementary Figure 3. Stacked-bar charts showing the relative abundance of top bacterial taxa at different taxonomic ranks between the two workflows (before: without host DNA removal; after: with host DNA removal) at the Phylum (up) and Genus (bottom) level.



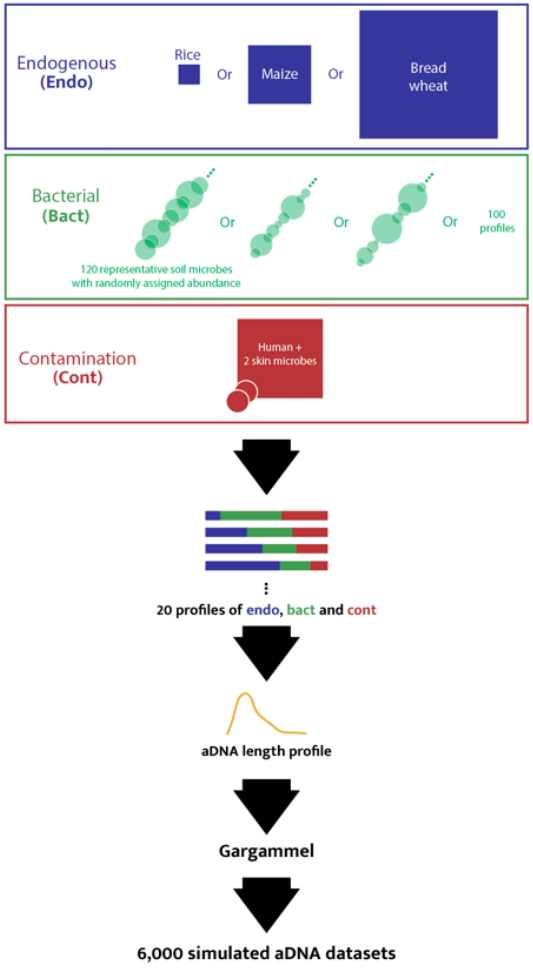
Supplementary Figure 4. The maximum number of detected OTUs for each dataset in rarefaction analyses between two workflows with (square in gold color) and without (triangle in purple color) host DNA removal.



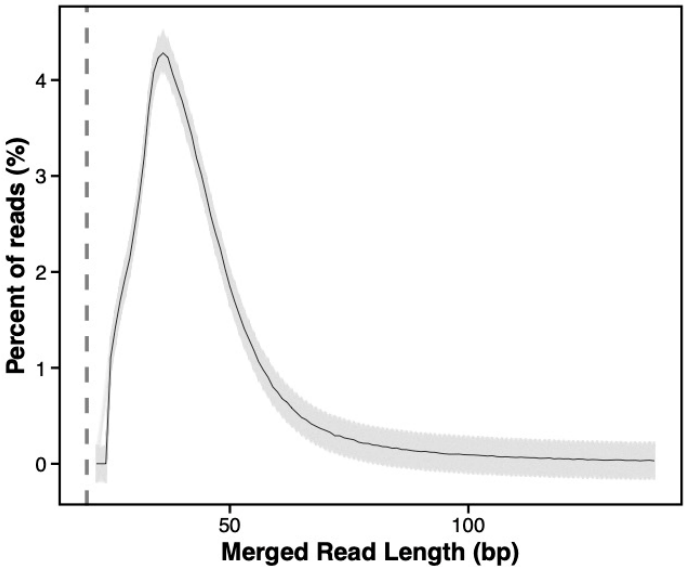
Supplementary Figure 5. Log-transformed percentage of shared k-mers across k-mer size (k=11, 15, 18, 21 and 24) in soil microbiota (a), rice (b), skin microbiota (c) and gorilla (d).



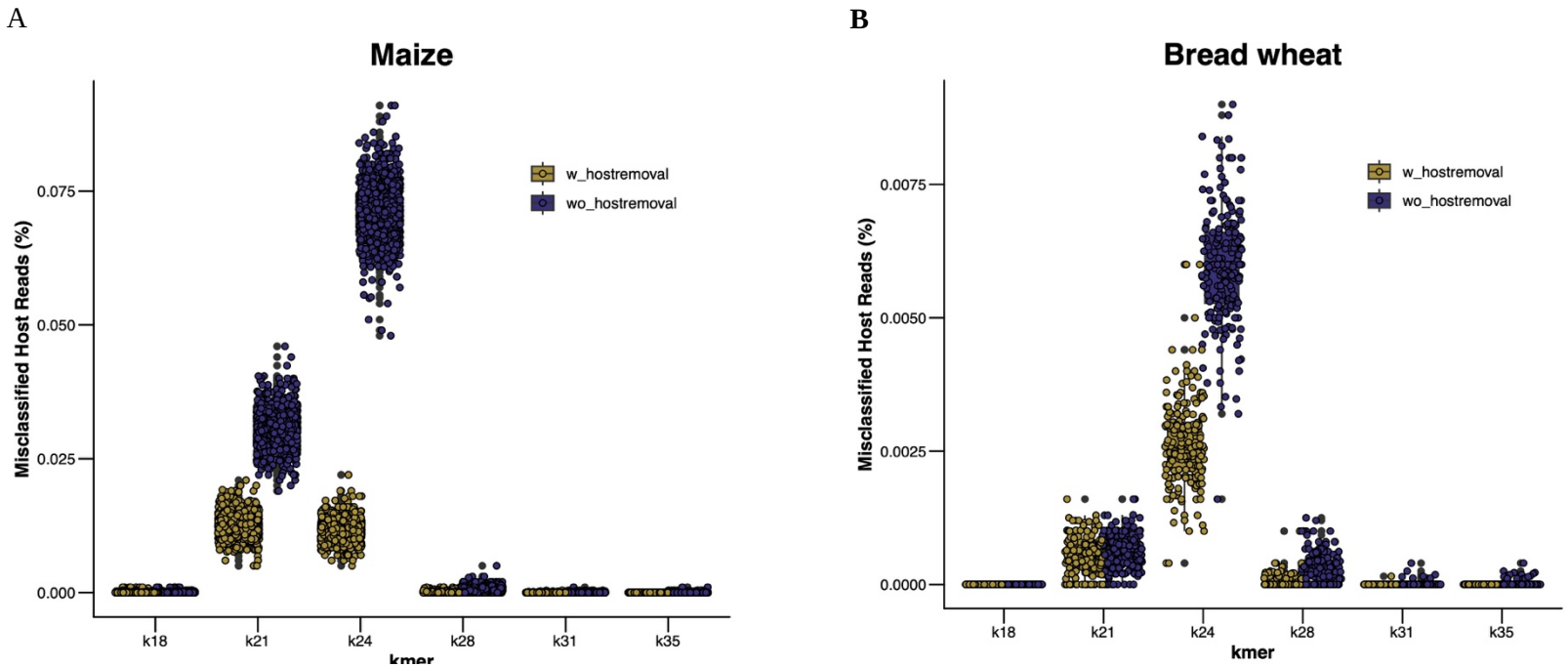
Supplementary Figure 6. The expected distribution of k-mers in host (dashed line) and those shared between gorilla and representative MAGs from skin microbiome (solid line). K-mers were grouped into intervals (x-axis) based on their occurrence in host genome (1, 2-10, 11-20, etc)



Supplementary Figure 7. The workflow for generating simulated aDNA datasets using 3 plant species and representative soil microbes.

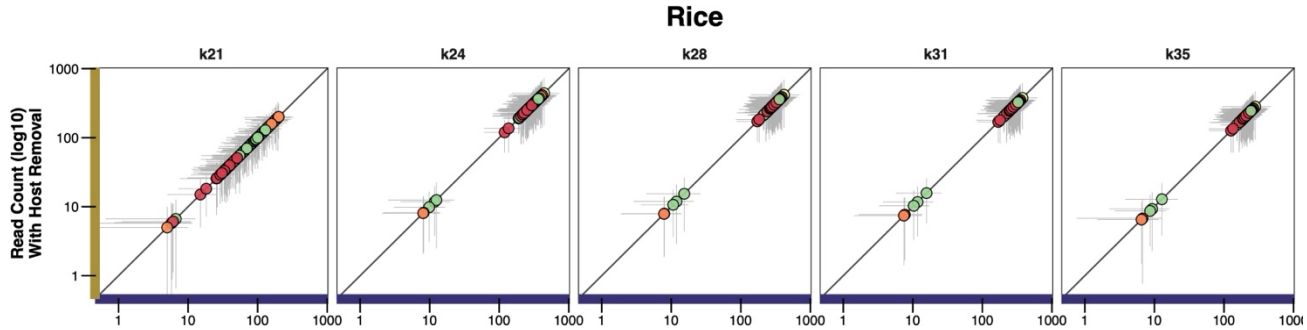


Supplementary Figure 8. The read length distribution of 1-million merged reads summarized from 2,000 simulated aDNA rice datasets.

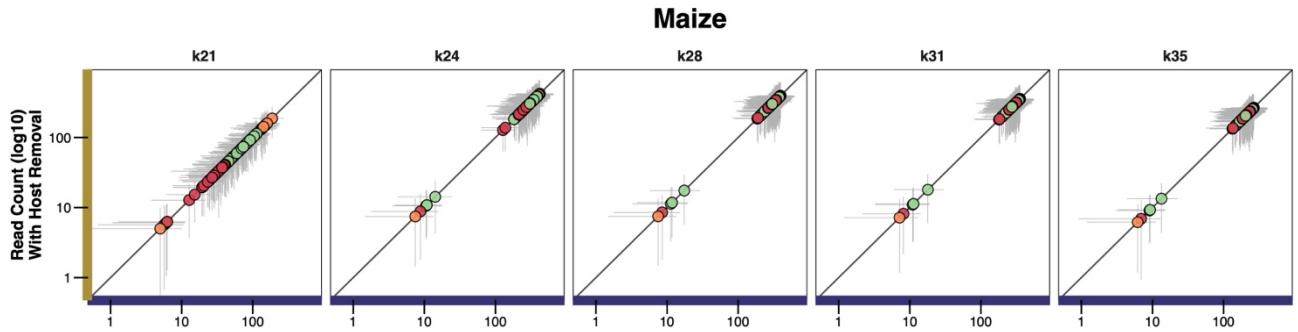


Supplementary Figure 9. The percentage of host reads misclassified as microbial taxa in the simulated **(A)** maize and **(B)** bread wheat datasets. Merged reads from 2,000 simulated datasets were annotated with databases created with different k-mer sizes (18, 21, 24, 28, 31 and 35). The simulated datasets were processed with (gold) and without (purple) host DNA content removal step.

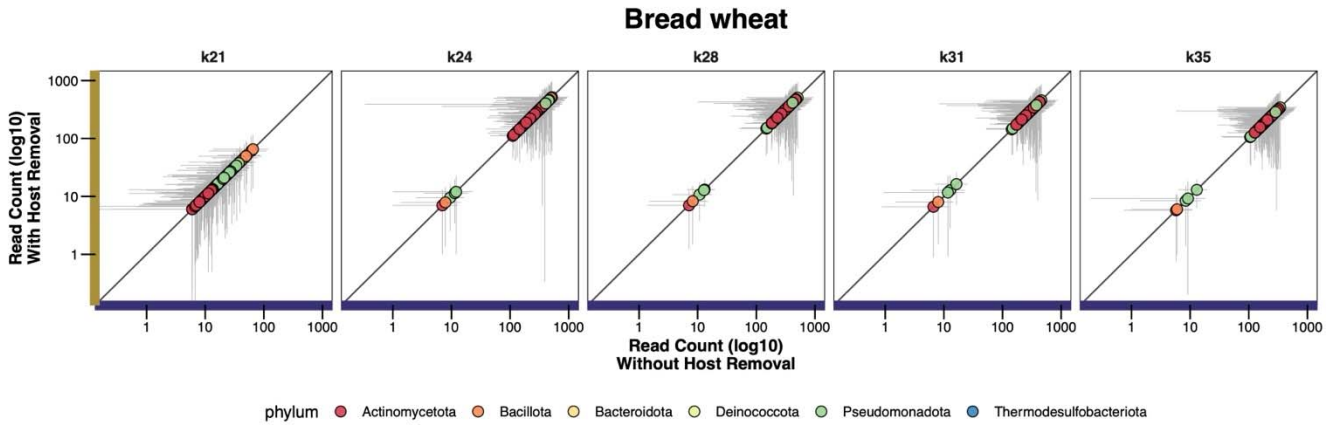
(A)


Rice
k21
k24
k28
k31
k35
Read Count (log10)
With Host Removal
1000
100
10
1

(B)

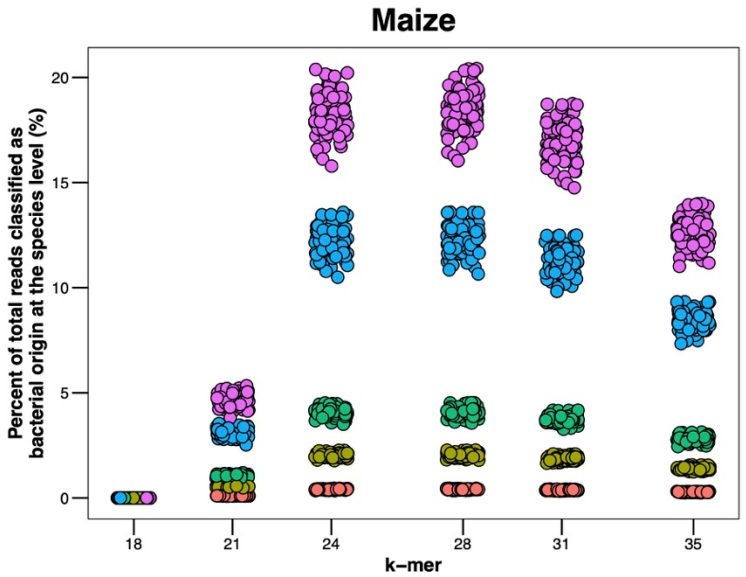

Maize
k21
k24
k28
k31
k35
Read Count (log10)
With Host Removal
100
10
1

(C)

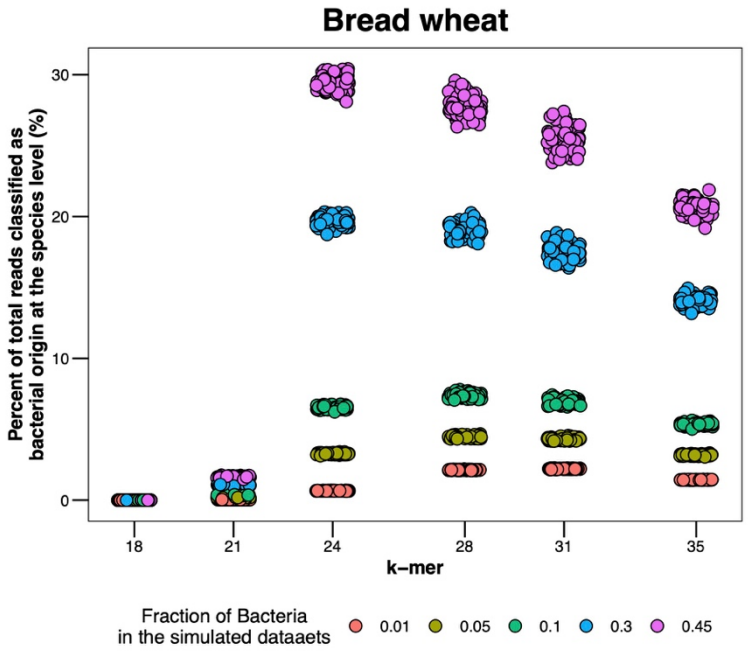

Bread wheat
k21
k24
k28
k31
k35
Read Count (log10)
With Host Removal
1000
100
10
1
Read Count (log10)
Without Host Removal
phylum
Actinomycetota
Bacillota
Bacteroidota
Deinococcota
Pseudomonadota
Thermodesulfobacteriota

Supplementary Figure 10. The distribution of soil microbes between two workflows with (y-axis; log-transformed; gold) and without (x-axis; log-transformed; purple) host DNA removal in **(A)** simulated rice datasets, **(B)** simulated maize datasets, and **(C)** simulated wheat datasets. Merged reads from each dataset was annotated with databases built with different k-mer values (k=21 to 35). K18 was not included as there was no read assigned to the species level. Each circle represented one of 107 soil microbes summarized from 100 simulated aDNA datasets and was color-coded by their phylum. The vertical and horizontal grey lines behind each circle indicated the first quantile of the read counts in the simulated datasets.

(A)

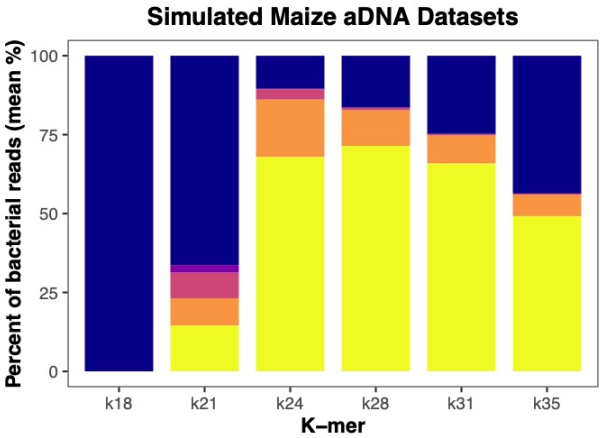


(B)

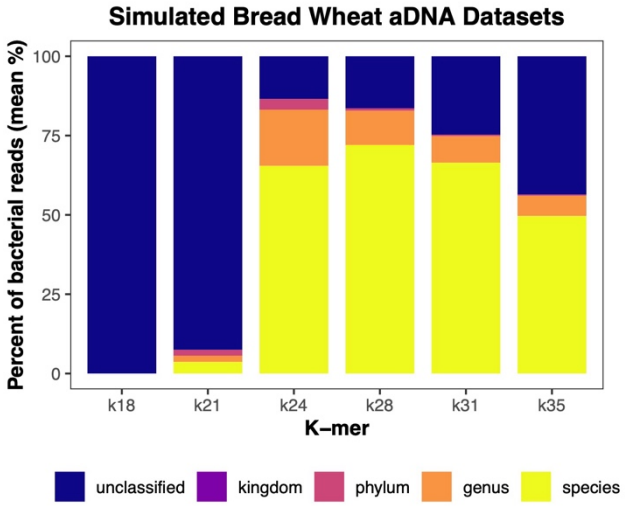


Supplementary Figure 11. The number of classified reads in simulated (A) maize and (B) bread wheat aDNA datasets

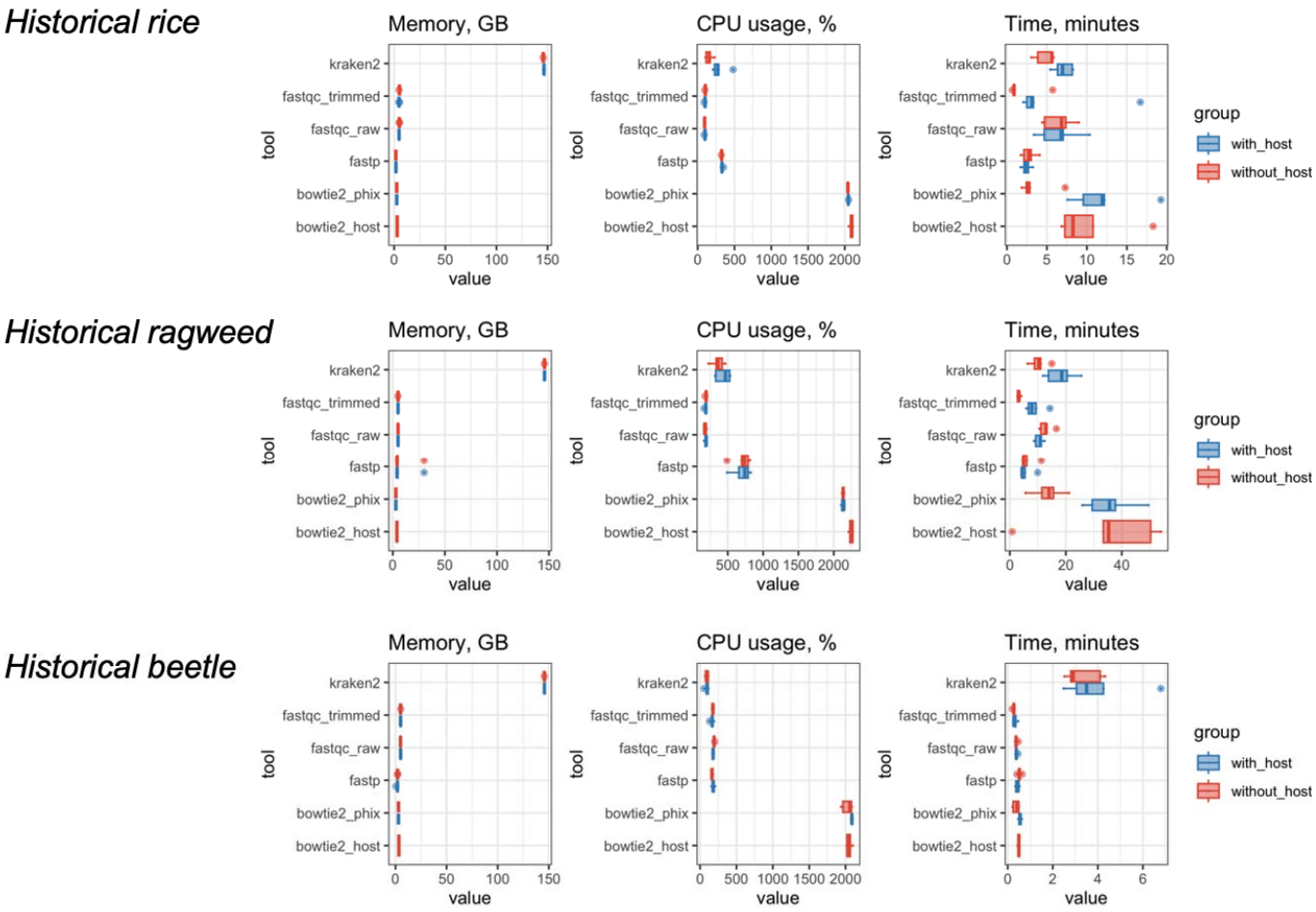
(A)



(B)



Supplementary Figure 11. The percentage of classified reads assigned at lowest taxonomic ranks in the simulated (A) maize and (B) bread wheat aDNA datasets.



Supplementary Figure 13. Computational resources utilized in different steps.

Dataset	Sample Size (N)	SRA Accession	Reference for Dataset	Host Species	Host Genome	NCBI Accession for host genome sequence
Contemporary Rice	330	PRJEB45634	Su, Wicaksono et al. 2022	Rice (*Oryza sativa*)	Oryza sativa AGIS1.0 + chloroplast	GCA_034140825.1, MZ424305.1
Herbarium Rice	5	PRJEB87646	This study	Rice (*Oryza sativa*)	Oryza sativa AGIS1.0 + chloroplast	GCA_034140825.1, MZ424305.1
Herbarium Ragweed	5	PRJEB48563	Bieker, Battlay et al. 2022)	Common ragweed (*Ambrosia artemisiifolia*)	Common ragweed + chloroplast	GCA_024762085.1, MG019037.1
Museum Beetle	5	PRJEB34604	Latorre, Herrmann et al. 2020	Extinct species of rhinoceros beetle (genus Oryctes)	Scarab beetle (*Oryctes borbonicus*)	GCA_001443705.1
Museum Gorilla	2	PRJEB29503	van der Valk, Diez-Del-Molino et al. 2019	Estern gorilla (*Gorilla beringei*)	Western lowland gorilla (*Gorilla gorilla gorilla*)	GCF_000151905.1

Supplementary Table 1. Datasets and responding host genome used in this study.

A separate Excel file

Supplementary Table 2. Classified reads across different samples between two workflows

Dataset	Host (Plant or Metazoa)		Bacterial	
	Mean	SD	Mean	SD
Herbarium rice	25.45	14.44	1.34	0.29
Contemporary rice	4.97	3.89	1.00	0.00
Museum Gorilla	4.01	2.54	1.08	0.10
Herbarium Ragweed	5.26	2.06	1.69	0.89
Museum Beetle	1.74	0.59	1.84	1.13

Supplementary Table 3. Fold difference in read count (without host removal / with host removal)

Host Species	Available in PlusPFP	Closet available species in PlusPFP
Rice (*Oryza sativa*)	Yes	-
Common ragweed (*Ambrosia artemisiifolia*)	No	Common sunflower (*Helianthus annus*)
Rhinoceros beetle (*Genus Oryctes*)	No	Nil
Eastern gorilla (*Gorilla beringei*)	No	Human (*Homo sapiens*)

Supplementary Table 4. Availability of host genome sequence in Kraken2 database PlusPFP.

Host	Microbiome	K-mer size	Host Genome Size (bp)	Total Microbiome Genome Size (bp)	K-mer Count			Jaccard Index	Percentage of Shared K-mers in Microbiome
					Host	Microbiome	Shared		
Beetle	Soil MAGs	11	429,075,705	388,583,972	2,097,147	2,097,152	2,097,147	1.00	100.0%
Beetle	Soil MAGs	15	429,075,705	388,583,972	173,564,277	179,983,403	53,685,368	0.30	29.8%
Beetle	Soil MAGs	18	429,075,705	388,583,972	161,246,728	349,332,584	1,727,965	0.01	0.5%
Beetle	Soil MAGs	21	429,075,705	388,583,972	364,212,684	374,246,117	108,090	0.00	0.0%
Beetle	Soil MAGs	24	429,075,705	388,583,972	373,256,704	377,032,845	4,732	0.00	0.0%
Gorilla	Skin MAGs	11	3,114,439,439	267,669,037	2,097,081	2,097,152	2,097,081	1.00	100.0%
Gorilla	Skin MAGs	15	3,114,439,439	267,669,037	333,353,609	391,936,415	245,898,104	0.68	62.7%
Gorilla	Skin MAGs	18	3,114,439,439	267,669,037	1,813,644,458	1,294,388,208	97,615,278	0.06	7.5%
Gorilla	Skin MAGs	21	3,114,439,439	267,669,037	2,225,178,794	1,482,534,256	4,005,268	0.00	0.3%
Gorilla	Skin MAGs	24	3,114,439,439	267,669,037	2,324,433,502	1,516,563,307	148,042	0.00	0.0%
Common Ragweed	Soil MAGs	11	1,274,100,190	388,583,972	2,097,152	2,097,152	2,097,152	1.00	100.0%
Common Ragweed	Soil MAGs	15	1,274,100,190	388,583,972	224,100,937	179,983,403	68,719,202	0.34	38.2%
Common Ragweed	Soil MAGs	18	1,274,100,190	388,583,972	530,844,797	349,332,584	5,506,202	0.01	1.6%
Common Ragweed	Soil MAGs	21	1,274,100,190	388,583,972	613,360,411	374,246,117	192,334	0.00	0.1%
Common Ragweed	Soil MAGs	24	1,274,100,190	388,583,972	653,026,928	377,032,845	12,666	0.00	0.0%
Rice	Soil MAGs	11	391,303,510	388,583,972	2,097,152	2,097,152	2,097,152	1.00	100.0%
Rice	Soil MAGs	15	391,303,510	388,583,972	160,205,633	179,983,403	59,369,404	0.35	33.0%
Rice	Soil MAGs	18	391,303,510	388,583,972	254,985,333	349,332,584	6,318,722	0.02	1.8%
Rice	Soil MAGs	21	391,303,510	388,583,972	273,093,572	374,246,117	361,396	0.00	0.1%
Rice	Soil MAGs	24	391,303,510	388,583,972	282,788,287	377,032,845	26,773	0.00	0.0%

Supplementary Table 5. Distinct k-mer in host genome and microbial pan-genomes and shared k-mer between host and microbiota

A separate Excel file

Supplementary Table 6. Reference sequences for soil microbes included in the simulation study