1 **Negligible effect of host DNA on metagenomic analysis enables microbial ecology**

2 **investigation in historical samples**

3

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

5

6 1 – Royal Botanic Gardens, Kew, Richmond, Surrey, United Kingdom

7 * Corresponding author

Abstract

Microbiome composition and function are shaped by its environment with major shifts induced by the intensified anthropogenic pressures over the past centuries. This time interval extends beyond experimental and longitudinal studies typically designed to investigate microbiome change. The extensive collection of historical samples available in museums and herbaria worldwide is an untapped resource for studying host-microbiome interactions over a prolonged time and in a global scale. Currently, their potential remains unfulfilled owing to incompatibilities with analytical pipelines and limited understanding of optimal classification parameters. Host DNA removal was deemed necessary for accurate taxonomic assignment of metagenomic reads, however, for many host species in historical samples it is impractical due to the absence of reference genomes. Here, we demonstrated that host DNA content has no significant impact on key microbial ecological metrics such as alpha- and beta-diversity. In addition, sequencing reads from historical samples are often highly fragmented due to degradation processes. Through k-mer analysis of genomics sequences from host species and their associated microbiomes, we observed that the shortest length of metagenomics read that can produce meaningful results for taxonomic profiling in historical samples is 21 bp. Overall, this study establishes a solid foundation for integrating extensive natural history collections into host-associated microbiome research, offering important insights into the anthropogenic impacts on microbiomes.

Introduction

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, soil microbiome contributes to plant growth through biogeochemical cycling of nutrients and organic matter (Jansson and Hofmockel 2020, Banerjee and van der Heijden 2023), endophytic microbiome contributes to plant survival under drought (Naylor and Coleman-Derr 2017, Trivedi, Leach et al. 2020) and 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 study) 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, Morais, Winkler et al. 2024), however, cannot be attained with these samples. Historical samples including 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 microbiome using metagenomic approach (Warinner, Herbig et al. 2017, Weyrich, Duchene et al. 2017). Examples of fundamental changes in the last two centuries include

52 raise of antimicrobial resistance in microbes in response to introduction of antibiotics and decrease
53 in soil nitrogen fixing activity in response to widespread application of synthetic fertilizers
54 (Naamala, Jaiswal et al. 2016, Aeron, Khare et al. 2020). Historical samples provide a global and
55 extended temporal perspective that conventional studies simply cannot achieve (Burbano and
56 Gutaker 2023).
57
58 High abundance of host DNA and low endogenous microbial biomass (Eisenhofer, Minich
59 et al. 2019) make metagenomic analysis with biological samples challenging. Additionally, nucleic
60 acid material extracted from historical samples is prone to damage and environmental
61 contamination. The impact of host DNA content on taxonomic classification of microbial
62 community has recently been investigated (Pereira-Marques et al. 2019) and the removal of host
63 DNA prior to data analysis has been seen as a necessary step. Several commercial DNA extraction
64 kits are designed to perform differential lysis and digestion of host cells to deplete their DNA,
65 however these kits are only effective for human or animal host cells but not for plant nor host cells
66 with complex cell wall structure. Alternatively, host removal can be performed *in silico* with the
67 help of reference genomes from the host species or closely related species. The breadth of microbe-
68 host-environment interactions, however, is best studied with diverse species including non-model
69 organisms (e.g. diverse angiosperm plants (Zuntini, Carruthers et al. 2024)), where representative
70 reference genomes are often absent.
71
72 Several critical questions regarding microbial analyses in historical samples remain
73 unexplored. First, is host DNA removal necessary when analyzing historical samples, especially
74 those from non-model organisms? While host DNA content might slightly reduce the sensitivity

of detecting rare taxa (Pereira-Marques et al. 2019), it is still unclear if ecological metrics such as alpha and beta diversity are affected. Second, microbiome profiling in historical samples usually relies on non-alignment-based tools like Kraken2 (Lu, Rincon et al. 2022), where the sensitivity and specificity of the classification are determined by k-mer size or seed length. Nucleic acid material from historical samples are often degraded (e.g. cytosine deamination) and highly fragmented (<50bp) (Gutaker and Burbano 2017). Short reads are discarded if they are annotated with a database built with a k-mer size larger than their length (e.g. k=35). Therefore, inclusion of these short reads in downstream analyses requires adjustments of k-mer size. It however not clear what are the shortest read length and optimal k-mer size that can produce reliable results (limited matching of sequence by chance). In this study, we investigated the effect of host DNA removal on the profiling of host associated microbiome and to explore the shortest read length and appropriate k-mer size for microbial profiling in historical samples.

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 soil and skin microbiomes. The soil MAGs catalogue (SMAG catalogue) contains 21,077 MAGs reconstructed from 3,304 soil

111 metagenomes and resolved at the species level, were obtained under the accession PRJNA983538
112 (Ma, Lu et al. 2023). The skin microbial genome collection (SMGC) comprises 251 isolate
113 genomes and 7,535 MAGs reconstructed from 4,802 skin metagenomes were retrieved from the
114 EBI FTP server
115 (http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/skin_microbiome/01_allprok/)
116 (Saheb Kashaf, Proctor et al. 2022). The datasets included in this study were summarized in
117 Supplementary Table 1.
118
119
120 **Sample collection and preparation of rice metagenomics**
121
122 In this study, we generated data for rice to compare the impact of host DNA removal in
123 contemporary and historical samples. We identified five herbarium specimens of rice at the
124 herbarium collection of Smithsonian Institution, collected between 1906 and 1914 in the
125 Philippines. A small fragment of leaf (approximately 20mm x 5mm) was removed and processed
126 in the semi-clean room facilities (without air filtering) at the New York University Centre for
127 Genomics and Systems Biology. Standard precautions for handling historical and ancient DNA
128 were taken (disposable PPE, use of dedicated safety cabinets, frequent cleaning regimes and UV-
129 sterilization). Sampled tissue was ground 'wet' with PTB lysis buffer in 37°C. Subsequently, DNA
130 isolation and genomic library preparation protocols established for historical plant tissues by
131 Latorre et al. (2020) were followed. Libraries were amplified with 10 polymerase chain reaction
132 cycles, quantified using RT-qPCR, normalized and pooled to equimolar concentrations. Libraries

133 were then sent to Macrogen for sequencing on a single lane of S4 flow cell with NovaSeq6000
134 instrument and demultiplexed.
135
136
137 **Preprocessing of shotgun metagenomic data**
138
139 For reproducibility, metagenomic analysis of the raw sequencing data was carried out with
140 the nf-core/mag pipeline, an open-source and community curated best-practice bioinformatics
141 pipeline for preprocessing, assembly, binning and annotation of metagenomes based on the
142 Nextflow nf-core framework (Ewels, Peltzer et al. 2020, Krakau, Straub et al. 2022). This pipeline
143 streamlines high throughput metagenomic data analysis on large-scale computing facilities. To
144 assess the effects of host genome removal on non-model organisms, we utilized the quality control
145 and taxonomic classification modules of the pipeline. Quality control of the raw metagenomic
146 sequencing reads in FASTQ format was performed using FastQC (v.0.12.0) (Andrews 2015).
147 Adapters and low-quality reads were trimmed, and overlapping paired-end sequences were merged
148 with an all-in-one fastp preprocessing tool on default settings (Chen, Zhou et al. 2018). PhiX
149 derived reads were discarded by mapping the quality filtered reads to the PhiX genome using
150 Bowtie2 (v.2.3.4) with the option --sensitive (Langmead and Salzberg 2012).
151
152
153 **Removal of host derived sequences**
154

155 We evaluated the effect of host DNA content on metagenomic analyses of contemporary
156 and historical samples by comparing the microbial community metrics between the same dataset
157 obtained by nf-core/mag pipeline with (i) host DNA removal or (ii) without this step (Figure 1).
158 For the host removal step (i), filtered reads were aligned to host genome using Bowtie2 with default
159 settings. Reads that did not align to the host genome were then subject to taxonomic classification.
160 For the workflow without host removal step, filtered reads were taxonomically annotated
161 immediately after quality check and adaptor removal.
162
163
164 **Taxonomic classification**
165
166 Taxonomic classification of DNA reads processed with and without host removal step, was
167 performed using Kraken2 (v2.0.9) (Wood, Lu et al. 2019) with PlusPFP database (version
168 2024.06.05), which contains archaea, bacteria, viral, plasmid, human, UniVec_Core along with
169 protozoa, fungi and plants. Kraken2 classifies individual metagenomic reads by mapping all k-
170 mers to the lowest common ancestor (LCA) of all genomes present in the database. Following
171 Kraken2 analysis, Bracken (Bayesian Re-estimation of Abundance with Kraken) was used to
172 estimate the taxon abundances at the species level (Lu, Breitwieser et al. 2017). The taxonomic
173 lineage information for each operational taxonomic unit (OTU) was parsed and converted to TSV
174 format using the kraken-biom software (Dabdoub 2016). To normalize sequencing depth, samples
175 were rarefied (down-sampling) to the minimum value of bacterial read sums in their respective
176 dataset (contemporary rice=510,147; rice herbarium=4,409; museum ragweed=69,042; museum

177 beetle=625; museum gorilla skin=1,632,030) OTUs were further filtered with a minimum read
178 count of 50 and prevalence of 10%.
179
180
181 **K-mer analysis of host and resident genomes**
182
183 To evaluate the similarity between host genome DNA and resident microbiome DNA, k-
184 mer analysis of genomic sequences was performed with JellyFish (Marcais and Kingsford 2011).
185 For soil microbiome, a single representative OTU was shortlisted for each of 107 pangenomes in
186 the SMAG catalogue based on their genome completeness (the number of expected single-copy
187 marker genes typically detected in a microbial genome assembly) and contamination (the
188 percentage of redundant single-copy marker genes) using CheckM (Parks, Imelfort et al. 2015).
189 The 107 pangenomes were generated from 2,200 species-level genomic bins with >10 high quality
190 MAGs by clustering conspecific protein sequences at 90% amino acid identity (Ma, Lu et al. 2023).
191 For the representative microbes with a reference genome available on NCBI (sequence accession
192 with a GCF_* or GCA_* prefix), we retrieved them from the NCBI GenBank using "download"
193 command of the NCBI datasets tool. For skin microbiome, 837 genomic sequences (622
194 prokaryotic species derived from 7,535 MAGs and 251 isolate genomes) representing core skin
195 microbes in the SMGC were included (Saheb Kashaf, Proctor et al. 2022). Counting of k-mer
196 shared between host and microbial genome sequences was carried out with JellyFish "count"
197 command in 5 different k-mer size (--mer-len=11, 15, 18, 21 and 24), with Bloom filter disabled
198 and both strands considered (--canonical). K-mer substrings were then exported to FASTA format
199 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.

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 was estimated using Chao1 and Shannon indices. 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. 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. Part of the analyses in this study was conducted on CropDiversity-HPC (Percival-Alwyn 2024).

222 **Results**

223

224 **Effect of host DNA removal on taxonomic assignments**

225

226 The host DNA removal is generally considered as necessary for the purpose of accurate
227 microbial profiling. To evaluate the effects of host DNA removal on taxonomic classification,
228 metagenomic datasets generated from contemporary and historical samples associated with 4
229 different host species (rice, ragweed, extinct rhinoceros beetle and eastern gorilla) were included
230 in this study (Supplementary Table 1). Each dataset was processed in two separate workflows in
231 parallel: with host DNA removal step and without it (Figure 1). During host removal step, the
232 reference genome for exact host species was used whenever possible. The reference genome for
233 rice and common ragweed were publicly available whereas genomic sequences for extinct
234 rhinoceros beetle and eastern gorilla were unavailable. Instead, we used the reference genomes of
235 two species closely related to them, scarab beetle (*Oryctes borbonicus*) and western lowland
236 gorilla (*Gorilla gorilla gorilla*) (Supplementary Table 1). To evaluate the impact of host DNA
237 removal on metagenomics analyses, we considered the number of reads, OTUs obtained and
238 microbial community diversity between the two workflows.

239

240 Overall, the samples included in this study covered a wide spectrum of host DNA content
241 (reads mapped to non-prokaryotic species), spanning from less than 5% in contemporary rice and
242 museum beetle to over 50% in some of gorilla and herbarium rice samples (Supplementary Table
243 2). As expected, host DNA content was markedly reduced when the samples were processed with
244 the host DNA removal. The percentage of endogenous non-bacterial DNA content was decreased,

245 especially in herbarium rice, museum gorilla and contemporary rice datasets (Supplementary
246 Figure 1). On average, host derived reads (either from plant or metazoa) had as much as 25-fold
247 difference between the workflows (Supplementary Table 3). Meanwhile, the number of reads
248 assigned to bacterial taxa was consistent (~1.0-fold difference between workflows), regardless of
249 host DNA removal (Figure 2 and Supplementary Table 3).
250
251 The taxonomic assignment was performed using Kraken2 with the prebuilt PlusPFP
252 database (Figure 1). Except rice, genome sequences for other host species were not available in
253 the database. The most closely related species in the database for common ragweed and eastern
254 gorilla were common sunflower (*Helianthus annus*) and human (*homo sapiens*), respectively
255 (Supplementary Table 4). Detected OTUs were then filtered based on their abundance and
256 prevalence. Before filtering, the numbers of OTUs were between 8,200-10,600, regardless of host
257 DNA removal (Table 1). Most of them were rare taxa with a relative abundance of 0.01% or below
258 and removed in the subsequent filtering step (Supplementary Figure 2). Of 330 contemporary rice
259 samples (3 replicates for each rice variety), same number of OTUs (9,164) were detected in both
260 workflows, and 21.6% of the OTUs were removed in subsequent filtering step with a filtering
261 criterion applied (a minimum read count of 50 and prevalence of 10%). Only 4 OTUs (0.1%) failed
262 the filtering step when the host removal step was applied, resulting in 7,184 OTUs. The number
263 of OTUs with more than 5% change in read count (abundance proxy) was less than 0.1%. Host
264 DNA removal exerted a negligible effect on the contemporary rice samples. For the historical
265 samples, a significant portion of detected OTUs (>80%) were filtered out, except in herbarium
266 ragweed samples. Museum beetle had the lowest number of filtered OTUs (330) while 6,956 OTUs
267 were retained in herbarium ragweed samples after filtering. The host DNA removal step led to

268 0.0%-16.4% decrease in the number of OTUs, of which almost all were in low-abundance with a
269 read count of 100 or less, suggesting that these OTUs were rare or misclassified taxa (Table 1 and
270 Supplementary Figure 2). For the retained OTUs, most (50.5% of OTUs in herbarium ragweed,
271 92.8% in museum beetle) had less than 5% difference in the number of assigned reads between
272 workflows. Those with highly variable read counts (>5% difference) between the workflows had
273 a mean read count much lower than those OTUs with a consistent read count. Altogether, low
274 abundance or misclassified OTUs were more prone to the host DNA removal step (Table 1).
275
276 Distinct taxonomic profiles were observed across different host-associated microbiota
277 (Figure 3a and Supplementary Figure 3). Genera with relative abundance lower than 2% were
278 dominant in the samples of rice and ragweed, while taxa from the genus *Curtobacterium* and
279 *Acinetobacter* were overrepresented in herbarium and contemporary rice samples, respectively.
280 Overall, relative abundance of taxa at phylum and genus level showed no consistent difference
281 between the workflows with and without host DNA removal. For example, the exclusion of the
282 host DNA removal step increased share of metagenomic assignments to Actinomycetota in rice
283 and gorilla, but decreased in ragweed and beetles.
284
285
286 **Effects of host DNA removal on microbial diversity**
287
288 We sought to determine if the host DNA removal step would alter the community diversity
289 and richness metrics of the host-associated microbiome by measuring alpha and beta diversity.
290 Alpha diversity was calculated using several metrics to account for different components of

291 community diversity. Chao1 diversity metrics included an estimation of undetected species in the
292 sample, and Shannon diversity index, which accounts for species presence/absence and evenness.
293 For the contemporary rice samples, Chao1 diversity metrics were not significantly different
294 (p=0.97) between the workflows with and without host DNA removal step, with average values of
295 1,667.7 ±1,200.1 and 1,660.2 ±1,201.7, respectively. A similar pattern was obtained for Shannon
296 diversity index (with host removal=2.45 ±1.29 and without=2.46 ±1.29), where no significant
297 differences (p=0.99) was observed (Figure 3a). The overall microbial diversity and community
298 structure in historical, museum samples were consistent between workflows and non-significantly
299 different, despite of a higher percentage of OTUs discarded after host DNA removal (Figure 3b-
300 e).
301
302 To better understand the differences in the community diversity of the taxonomic profiles
303 between the two workflows with and without host DNA removal, we assessed the differences in
304 microbial community structure. Principal coordinate analysis (PCoA) plot of beta diversity for rice
305 phyllosphere metagenome using Bray-Curtis dissimilarity index showed that host genome removal
306 has minimally influenced the microbial community structure. Analysis of similarities (ANOSIM)
307 showed no significant difference in microbial community structure between the two bioinformatic
308 treatments (R=-0.003228, p=1), even in historical samples (Figure 3). These results suggest that
309 the community metrics were not sensitive to the host DNA removal procedure.
310
311
312 **K-mer analysis and optimal k-mer size for historical samples**
313

314 To evaluate the impacts of host DNA on downstream microbiome analysis, we
315 decomposed the host genome and its associated microbiome into constituent k-mers, followed by
316 assessing the proportion of overlap between these constituent k-mers (shared k-mers) using
317 Jaccard index. This approach provided an overview of the extent to which sequence fragments
318 from the host genome could be mistakenly assigned to its associated microbiota purely by chance.
319 The exact reference genomes were obtained for rice and ragweed while the genomes of closely
320 related species were used for beetle and western gorilla. For host-associated microbiomes, we
321 focused on skin and soil microbiota whose comprehensive catalogues of pan-genomes of
322 metagenome-assembled genome (MAG) bins and microbial reference genomes (microbial
323 isolates) were available, 837 in skin and 107 in soil respectively. These pan-genomes were
324 reconstructed from a large number (3,000+) of metagenomes, suggesting that they robustly
325 represent the core microbes in their respective biome. For each microbiome, we counted distinct
326 k-mers with two methods: (a) collective combination of all pan-genome MAGs, and (b)
327 individually with each pan-genome MAG. The number of distinct k-mers depends on genome size
328 and k-mer size (Supplementary Table 5). Rice genome is smallest (391Mb) among the host
329 genomes analyzed in this study. The number of distinct k-mers was 10-fold different between k=11
330 and k=24 (2,097,152 vs 282,788,287). In comparison, west gorilla genome (3.1Gb) produced
331 1000-fold more distinct k-mers at k=11 compared to k=24 (2,097,081 vs 2,324,433,502). For skin
332 and soil, the number of distinct k-mers depends on the richness of microbial community (the total
333 genomic size of MAGs in this study). Roughly 1.51 billions of distinct k-mers were obtained from
334 skin microbiome (2.7Gb) at k-mer size of 24, compared to 0.37 billion in soil microbiome
335 (388Mb). We observed a similar pattern for k-mer numbers for individual pan-genome MAGs
336 (Supplementary Table 3).

The number of shared k-mers between microbiome and their respective host organisms (e.g. gorilla vs 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 their 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 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 4). 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 associated microbiome were significantly lower with increasing k-mer size (Figure 4d), indicating that low complexity sequences could more easily be misclassified

360 **Computational costs of host genome removal**

361

362 We then investigated the influence of host genome removal on the computational resources
363 for each of the workflow processes. To do this, we focused on the currently available best-practice
364 pipeline for performing each step of a typical metagenomic dataset analysis. In the box plots, we
365 estimated the amount of memory in gigabytes (GB), CPU usage, and the time duration consumed
366 by individual step in the workflow. Quality control check of the reads was performed twice using
367 FASTQC before filtering and after filtering the raw sequencing reads using FASTP. Host DNA
368 was removed using bowtie2, an ultrafast and memory-efficient tool for aligning sequencing reads
369 to long reference sequences (Langmead and Salzberg 2012). The average amount of memory used
370 by all processes in the pipeline with host removal step enabled was 32.5 ± 56.9 GB, while it was
371 reduced to 27.5 ± 52.9 GB when the removal step was disabled (Supplementary Figure 6). Among
372 the steps in the pipeline, Kraken2 consumed the highest amount of system memory, with an
373 average of 146 GB, regardless of host removal step enabled or not. All other steps had an average
374 memory usage of 10 GB. The total runtime for the workflows with and without host removal was
375 8.5 ± 8.2 and 8.8 ± 7.4 minutes, respectively. The pipeline spent the longest CPU time on host
376 removal step (bowties2). Besides, there was no significant difference between the two pipelines
377 on CPU usage.

378

379

380 **Discussion**

381

382 Historical samples, such as those from museum or herbarium collections, offer valuable
383 insights into how the diversity and community structure of host-associated microbiomes are
384 shaped by environmental changes over time. Profiling host-associated microbiome in these
385 samples presents challenges due to the DNA from host species or other non-target organisms, such
386 as humans or arthropods. These extraneous sequences have been suggested to interfere with
387 downstream analyses, potentially leading to false identification of taxa or misassembly of sequence
388 contigs (Schmieder and Edwards 2011). The removal of non-target sequences prior to downstream
389 analyses is considered to be an established procedure. However, many historical samples come
390 from non-model or extinct species, for which reference genomes are usually unavailable.
391 Therefore, it raises the question to what extent the microbiome analyses are affected when host
392 DNA is not removed. To this end, we profiled the host-associated microbiome in contemporary
393 and historical samples, and compared the microbial diversity and community structure between
394 the workflows with host DNA removal and without it. Our findings suggests that host DNA has a
395 negligible effect on microbial diversity metrics, regardless of their content in the samples. In
396 modern rice, historical rice and ragweed, the host DNA content varied from less than 5% to nearly
397 90%. Our findings are not conditional on the presence of host genome in the database (the pre-
398 built kraken2 database) for taxonomic assignment of sequencing reads. Notably, the rice genome
399 is included in the database, whereas human, the closet relatives of eastern gorilla which shares a
400 high degree of genomic similarity, is the only primate species available in the database (Scally,
401 Dutheil et al. 2012). Moreover, plant species and arthropods are underrepresented in many
402 biological databases, including but not limited to the prebuilt database used in this study. The

403 database contains sequences from the common sunflower, which is the closest possible species
404 available to common ragweed within the Asteroideae subfamily (Supplementary Table 1). The
405 abundance of taxa called differentially between our two tested approaches is generally low (with
406 the exception of the gorilla metagenome, Table 1). In contrast to host DNA removal step, OTU
407 filtering step has a greater impact on downstream microbiome analysis, reducing the number of
408 falsely identified taxa (Table 1), which are typically in low abundance and prevalence (Cao, Sun
409 et al. 2020). It is important therefore to consider the abundance filtering criteria thoroughly,
410 especially when excluding the host DNA removal step. Other technical factors, such as sample
411 size and sequencing depth, can also influence the identification of differential taxa (Nearing,
412 Douglas et al. 2022). However, we did not examine these factors in this study, as they are often
413 beyond the control of investigators when working with historical samples.
414
415 DNA isolated from historical samples is often degraded and highly fragmented, which
416 complicates bioinformatic analysis (Burbano and Gutaker 2023). This is likely reflected in our
417 findings that OTU abundances are more stable in contemporary rice samples, compared to
418 historical rice samples, regardless of whether host DNA removal step has been performed (Figure
419 1). The typical workflow for profiling community composition of host-associated microbiome in
420 historical samples relies on k-mer-based taxonomic classification tools such as Kraken2 and
421 Centrifuge (Kim, Song et al. 2016). The sensitivity and specificity of these tools can be adjusted
422 through k-mer size or seed length. By default, prebuilt kraken2 databases were compiled with a k-
423 mer size of 35 bp (Arizmendi Cardenas, Neuenschwander et al. 2022, Lu, Rincon et al. 2022). This
424 setting is ideal for general use as it leverages between specificity and sensitivity but not optimal
425 for degraded and heavily fragmented DNA sequences such as those from historical samples.

426 Different values of k-mer size have been proposed for various scenarios and types of datasets. The
427 databases used for filtering exogenous DNA in historical samples were built with a k-mer size of
428 29 (Ravishankar, Perez et al. 2024). For metagenomic datasets, k-mer size of 21 was shown to be
429 sufficient to obtain accurate results in most cases (Ondov, Treangen et al. 2016). However, these
430 values may not be optimal for profiling host-associated microbial DNA in historical samples. We
431 conducted k-mer analyses on host genome and its associated microbiome to yield insights into the
432 optimal k-mer size and supporting evidence to validate the negligible effects of host DNA on
433 microbiome analysis in historical samples.
434
435 Microorganisms in the rhizosphere play vital roles in nutrient cycling and other
436 biogeographical processes in soil (Philippot, Chenu et al. 2024). Similarly, skin-colonizing
437 microbes help maintain cutaneous homeostasis and serve as an additional barrier against the
438 invasion of exogenous pathogens (Byrd, Belkaid et al. 2018). The composition of both skin and
439 soil microbial communities is influenced by environmental perturbations, therefore historical
440 samples are useful in gaining insights into the microbial community dynamics and factors shaping
441 them over long-time scales. We focused on core microbes in the soil and skin microbiome, which
442 were reconstructed from an extensive collection of soil and skin metagenomes. The constituent k-
443 mers shared between host genome and associated microbiome was considered as an indicator of
444 taxonomic misclassification due to host DNA contamination. As expected, the number of k-mers
445 shared between host genome and core microbes in their respective associated microbiome depends
446 on the size of k-mer and their genomes. When k-mer size is small (e.g. k=11), all constituent k-
447 mers were shared simply by chance alone. Nonetheless, with a larger k-mer size (e.g. k=18 or
448 above), the overlap in k-mer space was significantly reduced (Jaccard index < 0.1), suggesting that

449 the misclassification of host DNA into microbial taxa become unlikely at a large k-mer size. Based
450 on our observation, k-mer size of 18-21 is sufficient for profiling host-associated microbiome in
451 historical samples from non-model species with a low level of misclassification (0.1% or less).
452 This finding also explains why host DNA has a negligible impact on microbial diversity metrices
453 when taxonomic classification was performed using databases compiled with a large k-mer, such
454 as k=35 in PlusPFP database. Furthermore, we observed that at k=24, 5.9% of the k-mers shared
455 between host genome and associated microbiome (0.01% or less of total k-mers) were low-
456 complexity k-mers with high occurrence (≥10) (Supplementary Table 5-6). It strongly suggests
457 that a sequence complexity filter, such as DUST algorithm (Morgulis, Gertz et al. 2006), should
458 be applied prior to taxonomic classification to ameliorate the impact of low-complexity sequences.
459
460 In this study, we demonstrated that host DNA has an negligible impact on microbial
461 analyses of host-associated microbiome, even in samples with high content of host DNA. As the
462 host removal step is often computational demanding, omitting it may enhance workflow efficiency
463 in some cases. Our findings also suggest that meaningful microbial analyses can be conducted on
464 historical samples where the genomic sequences of host species are unavailable. Furthermore, we
465 showed that microbial reads as short as 21 bp are distinguishable from those from host species and
466 informative results can still be obtained. To annotate the short reads of this length, a k-mer size of
467 21 could be used to build the database, to increase the sensitivity of annotation at a cost of decrease
468 in specificity. Additionally, a sequence complexity filter should be applied to minimize the
469 misclassification of low-complexity host DNA sequences as microbial reads, particularly when
470 host DNA removal is not feasible. Finally, this manuscript does not consider the problem of post-
471 mortem microbial colonization that is common in historical samples. Future work should identify

472 microbial contaminants through constructing a post-mortem reference database or constructing
473 relevant metagenomic classifiers, likely unique to each kingdom. This work marks an important
474 milestone in enabling metagenomic investigation of historical samples, contributing to our
475 understanding of how anthropogenic changes have shaped host and associated microbiome.
476
477
478
479
480
481

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.

492 **Data Accessibility and Benefit-sharing**

493

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

495 PRJEB87646.

496

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

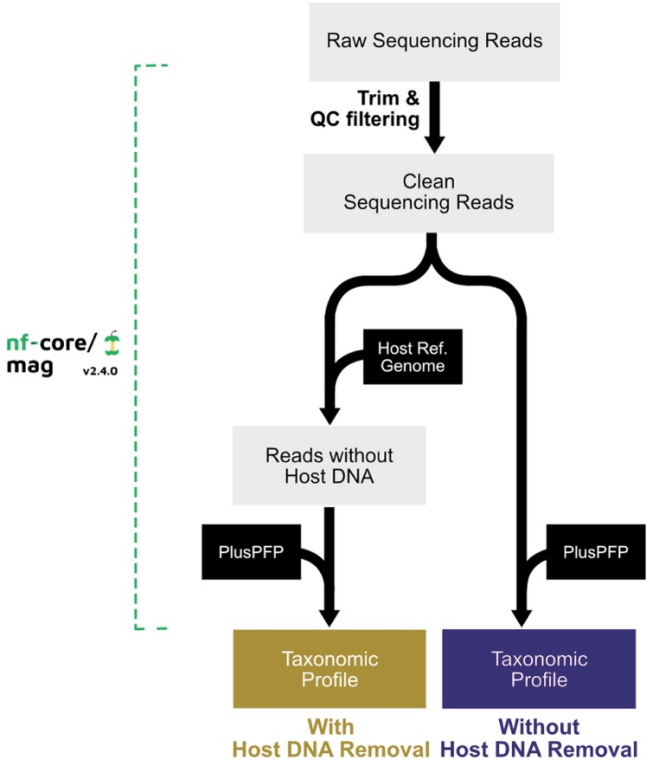


Figure 1. Bioinformatic workflow. The Nextflow metagenomic analysis workflow (nf-core/mag) was used in this study. Two separate workflows (gold and purple) were performed in parallel with identical steps except the host removal step prior to taxonomic classification.

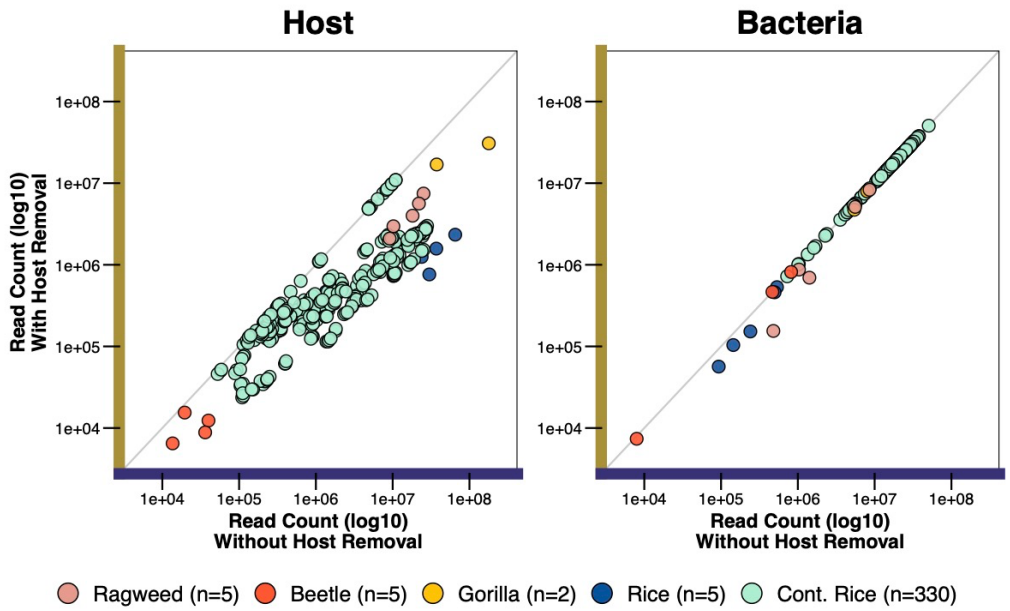


Figure 2. Proportion of sequencing reads (log10) assigned to host (plant or metazoa) and microbial taxa with (y-axis; gold) and without (x-axis; purple) host DNA removal. Each circle represents a sample and is color-coded with datasets.

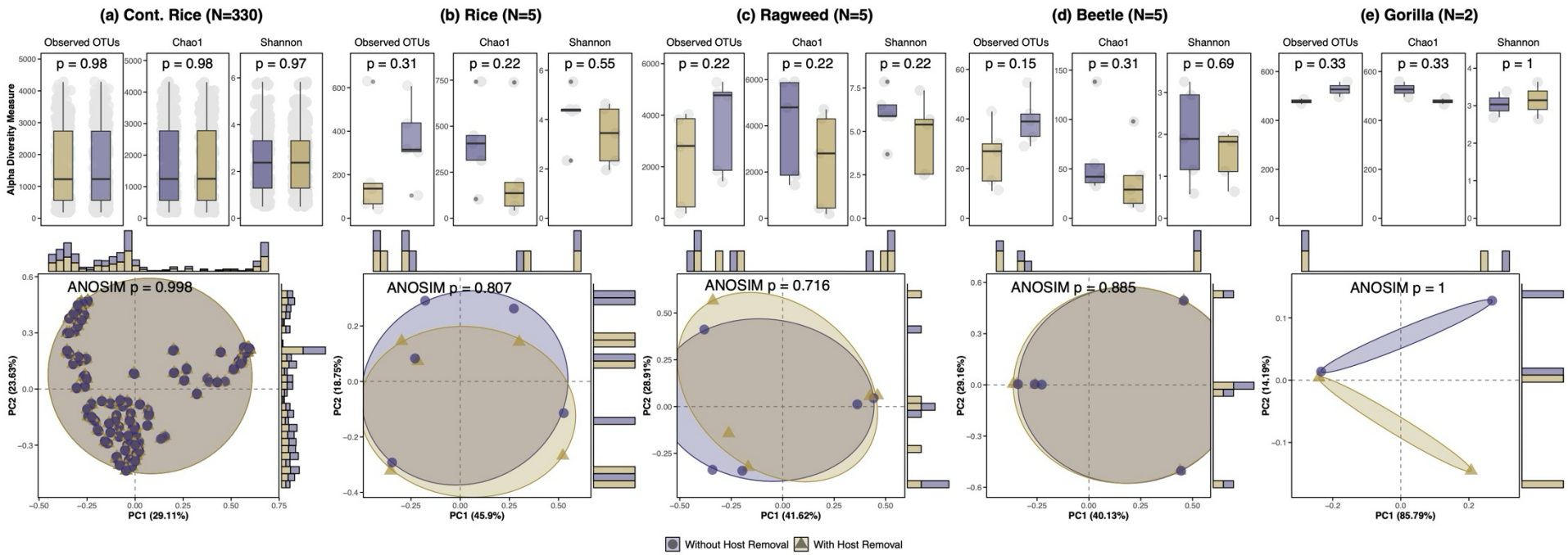


Figure 3. Alpha and beta diversity analysis of microbial community in the samples of contemporary rice (a), herbarium rice (b), herbarium ragweed (c), museum beetle (d) and museum gorilla (e). Microbial diversity analysis using Chao1 and Shannon diversity indices (top panels) and microbial community 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.

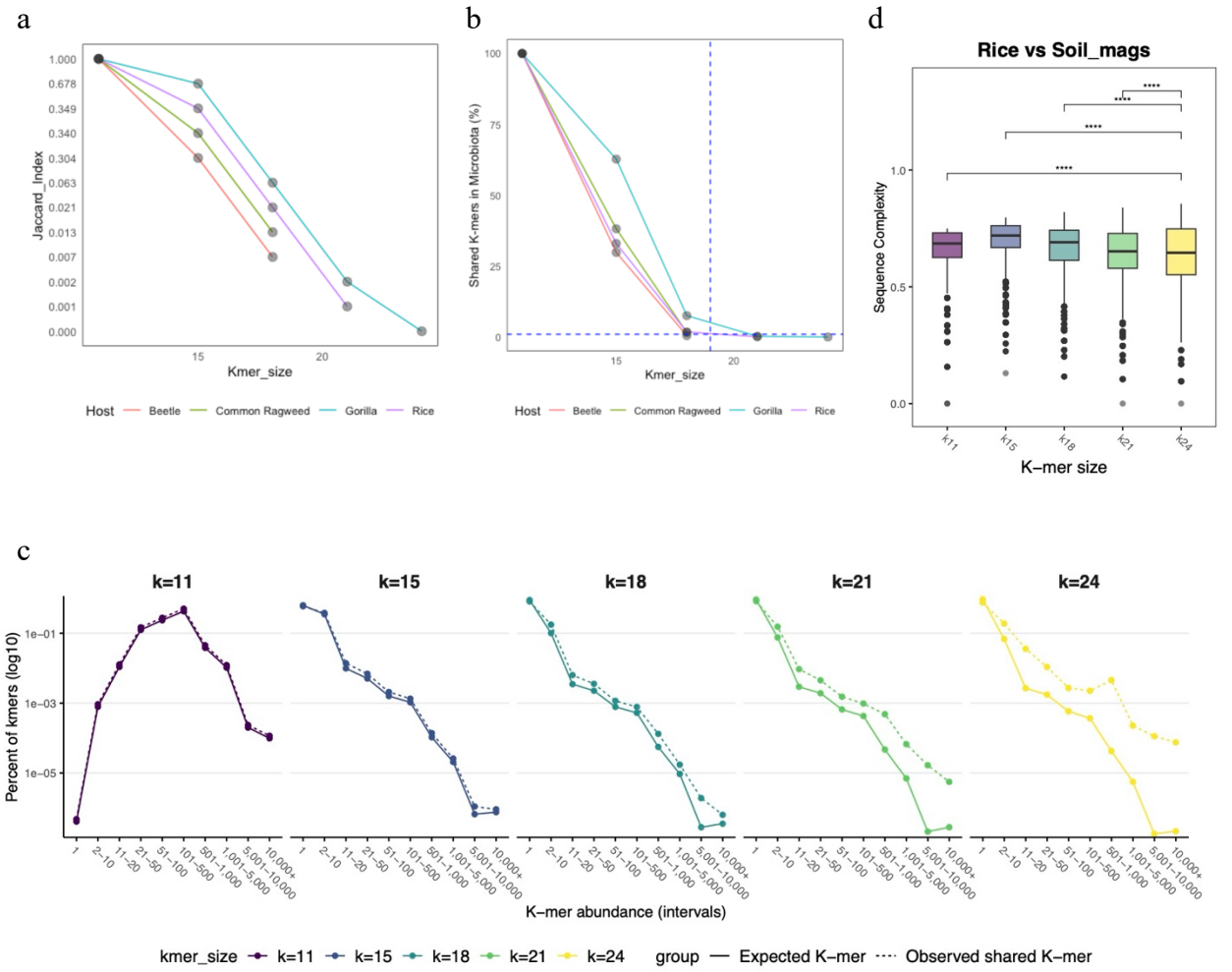


Figure 4. Shared k-mers among host species and their respective associated microbiome (beetle-soil microbiome, common ragweed-soil microbiome, gorilla-skin microbiome and rice-soil microbiome). Jaccard index (a) and percentage of k-mers in associated microbiome shared with host species (b) across k-mer sizes of 11, 15, 18, 21 and 24. (c) The distribution of k-mers in rice (solid line) and those shared between rice and soil microbiome (dashed 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 soil microbiome across k-mer sizes.

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			OTU with >5% change	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.

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

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.

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.

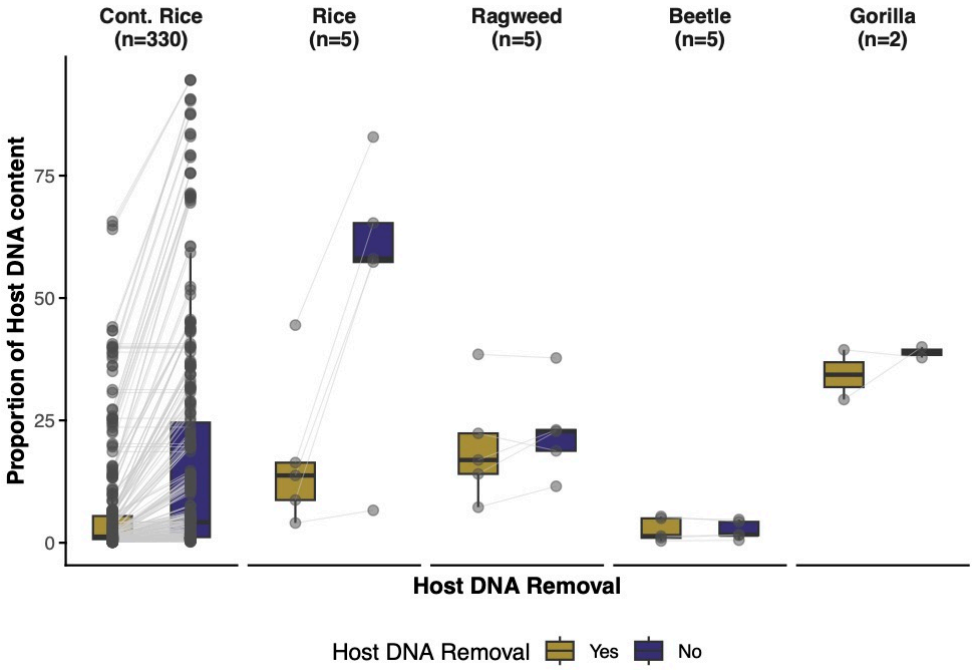
Gould, A. L., V. Zhang, L. Lamberti, E. W. Jones, B. Obadia, N. Korasidis, A. Gavryushkin, J. M. Carlson, N. Beerenwinkel and W. B. Ludington (2018). "Microbiome interactions shape host fitness." Proc Natl Acad Sci U S A **115**(51): E11951-E11960.

Gutaker, R. M. and H. A. Burbano (2017). "Reinforcing plant evolutionary genomics using ancient DNA." Curr Opin Plant Biol **36**: 38-45.
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).

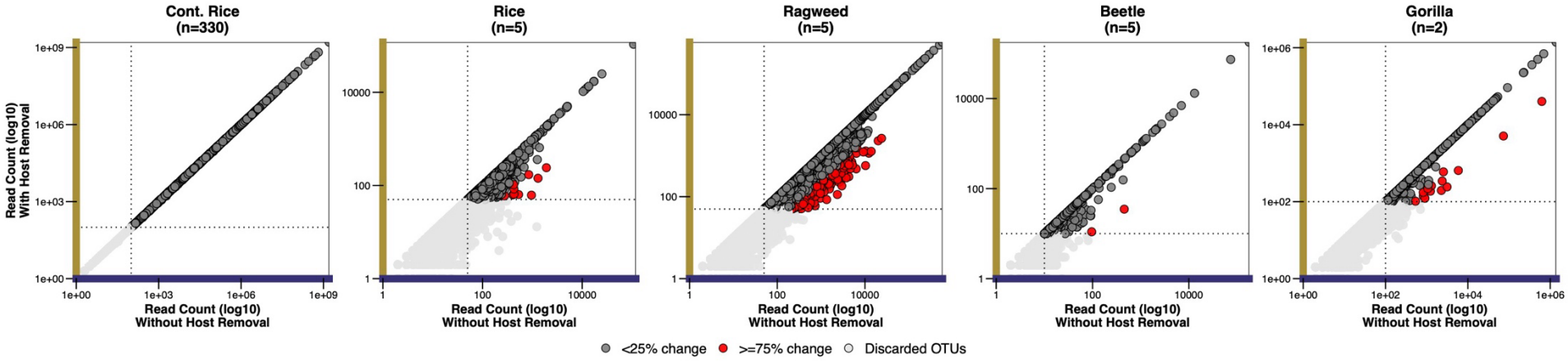
"UKCropDiversity-HPC: A collaborative high-performance computing resource approach for sustainable agriculture and biodiversity conservation." Plants, People, Planet.
Petersen, C., I. K. Hamerich, K. L. Adair, H. Griem-Krey, M. Torres Oliva, M. P. Hoeppner, B. J. M. Bohannan and H. Schulenburg (2023). "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).
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.

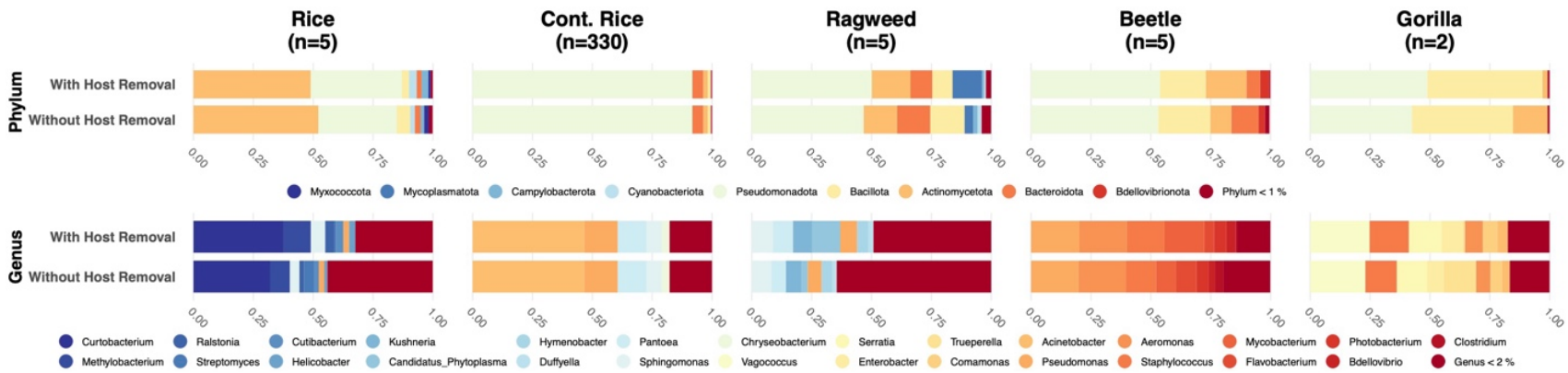
Supplementary Table and Figures



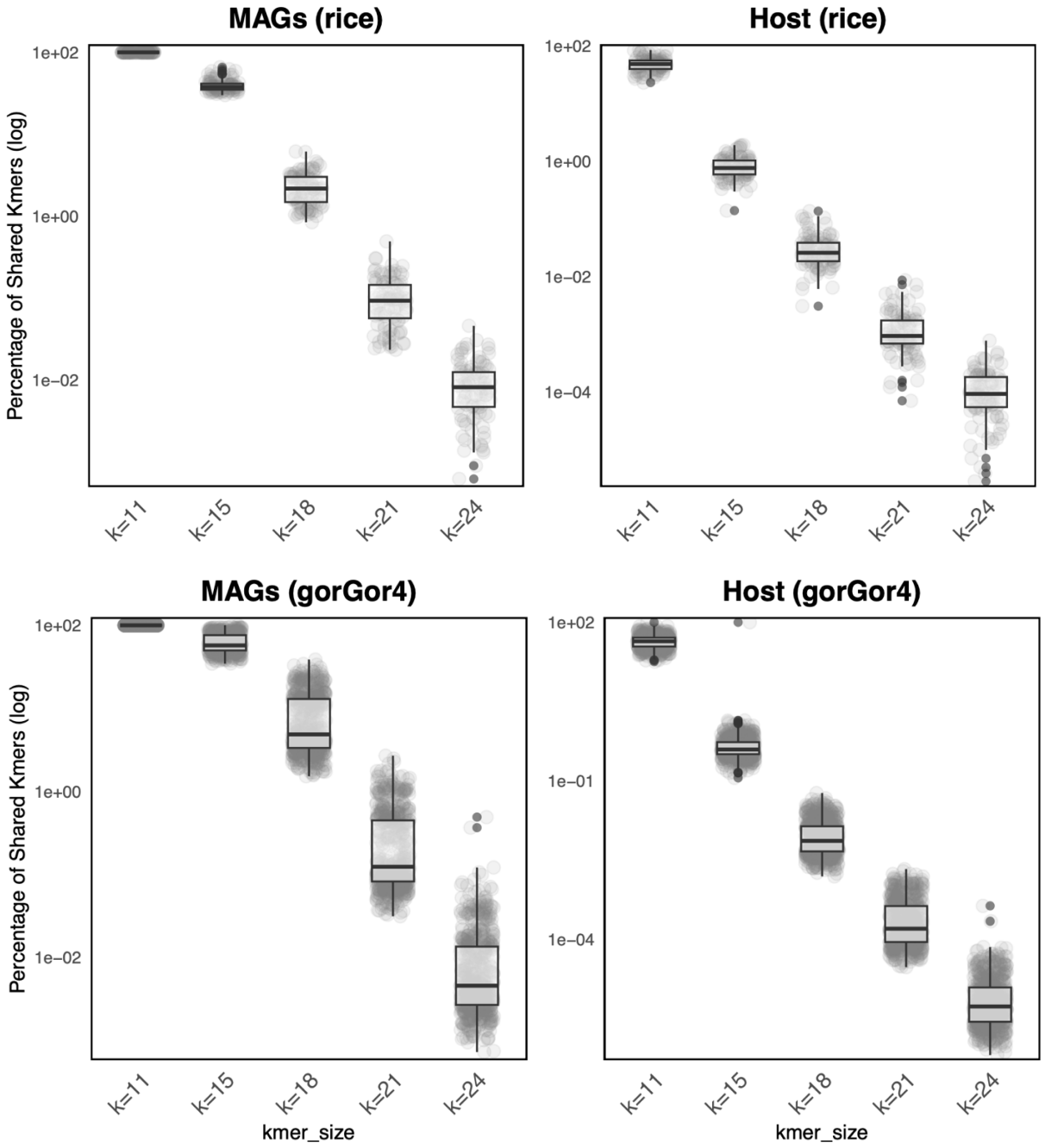
Supplementary Figure 1. Proportion of host DNA content in each sample (point in grey) across datasets between the workflow with (gold) and without (purple) host DNA removal.



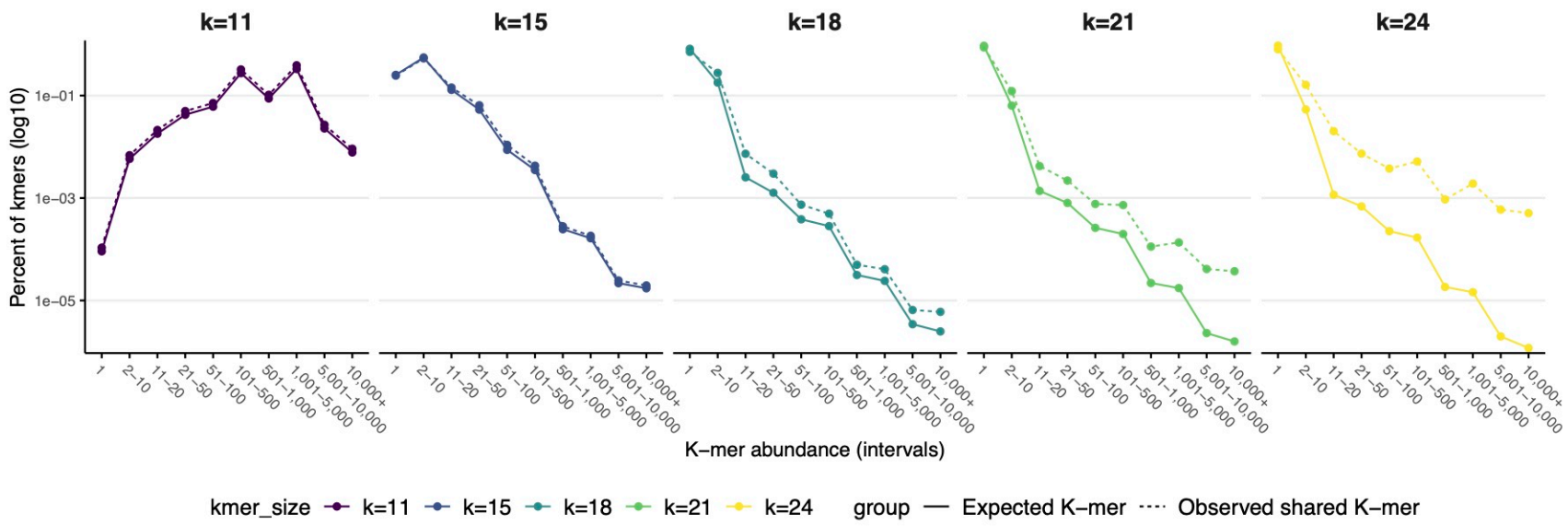
Supplementary Figure 2. Distribution of bacterial OTUs in assigned read count (log10) between two workflows with (y-axis; gold) and without (x-axis; purple) host DNA removal. Each circle represents a bacterial OTU. Those with more than 75% change in read count between two workflows are highlighted in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



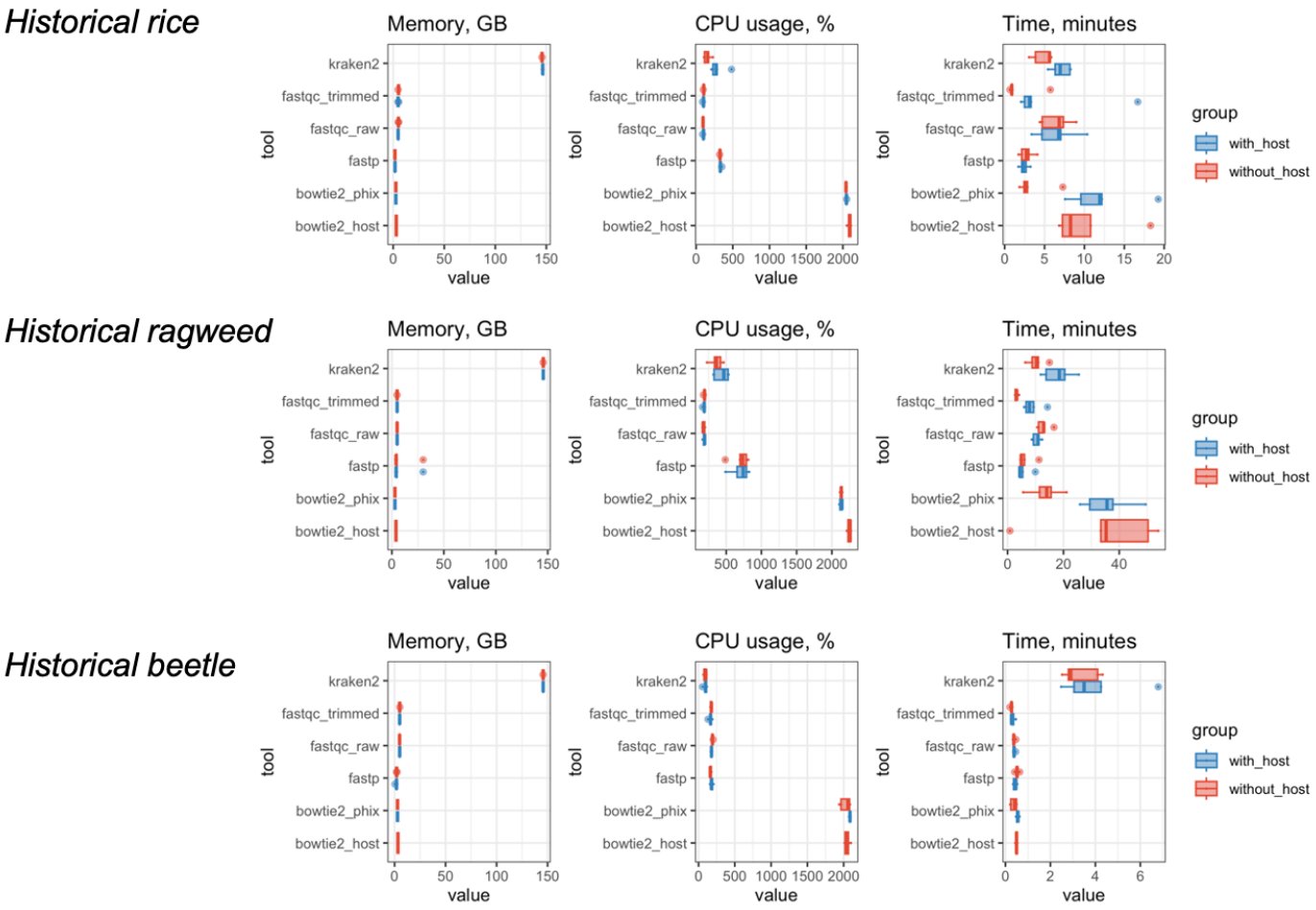
Supplementary Figure 3. Stacked-bar charts showing the relative abundance of top bacterial taxa at different taxonomic ranks between the two workflows (before: without host DNA removal; after: with host DNA removal) at the Phylum (up) and Genus (bottom) level.



Supplementary Figure 4. 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 5. The distribution of k-mers in host (solid line) and those shared between gorilla and representative MAGs from skin microbiome (dashed line). K-mers were grouped into intervals (x-axis) based on their occurrence in host genome (1, 2-10, 11-20, etc)



Supplementary Figure 6. 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.

Separate Excel file

Supplementary Table 2. Assigned read count across individual sample

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