Comparison of commercial DNA extraction kits for whole metagenome sequencing of human oral, vaginal, and rectal microbiome samples

Michelle L. Wright[1,2*], Jessica Podnar[3¶], Kayla D. Longoria[1], Tien Nguyen[4,5], Sungju Lim[1], Sarina Garcia[4], Dennis Wylie[3]

[1] School of Nursing, University of Texas at Austin, Austin, Texas, USA

[2] Department of Women's Health, Dell Medical School at The University of Texas at Austin, Austin, Texas, USA

[3] Center for Biomedical Research, University of Texas at Austin, Austin, Texas, USA

[4] College of Natural Sciences, University of Texas at Austin, Austin, Texas, USA

[5] University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA

[*] Corresponding author

Email: michelle.wright@utexas.edu

Author roles: **MLW:** conceptualization, data curation, investigation, methodology, formal analysis, project administration, resources, funding acquisition, writing – original draft preparation. **KDL, SL, TN, SG:** writing – review and editing **JP:** methodology, investigation, resources, writing – review and editing. **DW:** data curation, formal analysis, methodology, visualization, writing – original draft preparation

¶

Abstract

Introduction Advancements in DNA extraction and sequencing technologies have been fundamental in deciphering the significance of the microbiome related to human health and pathology. Whole metagenome shotgun sequencing (WMS) is gaining popularity in use compared to its predecessor (i.e., amplicon-based approaches). However, like amplicon-based approaches, WMS is subject to bias from DNA extraction methods that can compromise the integrity of sequencing and subsequent findings. The purpose of this study was to evaluate systematic differences among four commercially available DNA extraction kits frequently used for WMS analysis of the microbiome. **Methods** Oral, vaginal, and rectal swabs were collected in replicates of four by a healthcare provider from five participants and randomized to one of four DNA extraction kits. Two extraction blanks and three replicate mock community samples were also extracted using each extraction kit. WMS was completed with NovaSeq 6000 for all samples. Sequencing and microbial communities were analyzed using nonmetric multidimensional scaling and compositional bias analysis. **Results** Extraction kits differentially biased the percentage of reads attributed to microbial taxa across samples and body sites. The PowerSoil Pro kit performed best in approximating expected proportions of mock communities. While HostZero was biased against gram-negative bacteria, the kit outperformed other kits in extracting fungal DNA. In clinical samples, HostZERO yielded a smaller fraction of reads assigned to *Homo sapiens* across sites and had a higher fraction of reads assigned to bacterial taxa compared to other kits. However, HostZERO appears to bias representation of microbial communities and demonstrated the most dispersion by site, particularly for vaginal and rectal samples. **Conclusions** Systematic differences exist among four frequently referenced DNA extraction kits when used for WMS analysis of the human microbiome. Consideration of such differences in study design and data interpretation is imperative to safeguard the integrity of microbiome research and reproducibility of results.

Introduction

The diverse composition of the microbiome throughout the body plays an important role in human health. The advancement of sequencing technologies has made the assessment of microbiome composition possible through microbial DNA extraction and sequencing. Early studies of the human microbiome largely described bacterial communities using the 16S amplicon-based sequencing approach [1]. However, this amplicon-based approach does not allow for the discovery of fungi or viruses within a sample or identification of microbial genes for functional relevance such as antimicrobial resistance genes. Furthermore, results from amplicon based approaches can be biased by numerous factors including: varied microbial genetic diversity in common 16S regions, multiple 16S genome copy numbers in bacterial genomes, and choice of DNA extraction method [2]. To counter limitations of amplicon-based sequencing, there is increasing use of whole metagenome shotgun sequencing (WMS) to analyze human microbiome communities related to health outcomes.

WMS may also be impacted by the same limitations and biases as 16S amplicon-based approaches. DNA extraction is crucial to the success of microbiome sequencing and is known to introduce bias in 16S studies [3–6]. The wide variations of cell membrane structures and composition can pose significant challenges to efficiency and bias-free extraction of genomic DNA. Moreover, WMS studies investigating the human microbiome can produce a disproportionately large amount of sequencing reads attributed to the human host, instead of microbes, which limits the sensitivity of the WMS approach to detect low abundance microbial species [7]. Therefore, the selection of a reliable method for DNA extraction is of paramount importance to ensure a high DNA yield and a representative characterization of microbial communities in any given body site.

The purpose of this study was to evaluate if systematic differences occur when different commercially available DNA extraction kits are used for WMS analysis of the human microbiome.

Methods

Ethics approval

This study was approved and monitored by the University of Texas Institutional Review Board (protocol #2019-08-0027) and the Institutional Biosafety Committee (protocol #IBC-2019-00216). Informed consent was obtained from all participants before the collection of oral, vaginal, and rectal swab samples used for this study.

Sample collection

Participants were recruited from the Women's Health Integrated Practice Unit (IPU) at UT Health Austin. Individuals presenting to the IPU for pelvic floor physical therapy were approached and asked if they would be interested in participating in the study. If individuals expressed interest, written informed consent was obtained. No additional information was collected about the participants or their health history. Five participants provided a total of 12 samples at the sample collection visit. The healthcare provider collected oral, vaginal, and rectal samples in replicates of four using sterile flocked swabs (COPAN diagnostics, Murrieta CA), in the order listed. The samples were pre-labeled by site and order of collection (e.g., oral 1, oral 2, oral 3, oral 4, etc.). The provider collecting the samples changed gloves between each site (i.e., oral, vaginal, rectal). Oral swabs were collected by first swabbing the tongue, then the roof of the mouth, and lastly around the gum line (top and bottom) twice. Vaginal samples were obtained by inserting the swab into the vagina 3-4 inches avoiding contact with the outside skin and then rotating the swab 3-4 times. Rectal samples were obtained by inserting the swab 1 inch

84 into the rectum and then rotating the swab 3-4 times. The samples were immediately transported to the
85 laboratory and stored at -80°C until DNA extraction was completed.

86 DNA extraction

87 Patient samples were randomized across extraction kits using the website randomizer.org to
88 ensure the order of swab collection did not bias the results (**Table 1**). For example, for Participant 1 all
89 swabs that were collected first and labeled number one for each site (oral, vaginal, and rectal) were
90 extracted with the HostZERO kit, the second swab collected for each site was extracted with the
91 PowerSoil Pro kit, and so on. We used four commercially available extraction kits frequently referenced
92 in the metagenomic literature: Qiagen DNeasy PowerSoil Pro kit (catalog number 47014), HostZERO
93 Microbial DNA kit (catalog number D4310), PureLink Microbiome kit (catalog number A29790), and
94 Qiagen DNeasy Blood and Tissue kit (catalog number 69504).

95 **Table 1. Randomization of samples across extraction kit type**

	Participant				
Extraction Kit	**1**	**2**	**3**	**4**	**5**
PowerSoil Pro	2	1	3	1	4
HostZERO	1	4	4	3	2
PureLink	4	3	2	2	1
DNeasy	3	2	1	4	3

96

97 All samples using the same protocol and kit were extracted by MLW on the same day. Extraction
98 protocols were followed per manufacturer instructions, and each protocol was completed on separate
99 days to minimize any potential cross-contamination. All samples (i.e., oral, vaginal, rectal) that

corresponded with the order number were used for the same kit. Two extraction blanks of sterile DEPC-treated, DNAse, RNAse free water (Fisher BioReagents™, Pittsburgh, PA) and three replicate mock community samples (ZymoBIOMICS™ mock microbial community standards catalog number D3600, Irvine, CA) were also extracted using each extraction kit. The mock communities are composed of eight bacteria [3 Gram-negative (*Pseudomonas aeruginosa, Escherichia coli, and Salmonella enterica)* and 5 Gram-positive (*Lactobacillus fermentum, Enterococcus faecalis, Staphylococcus aureus, Listeria monocytogenes,* and *Bacillus subtilis*], with genomic DNA from each species contributing to 12% of the community composition, and two yeasts (*Saccharomyces cerevisiae* and *Cryptococcus neoformans*) with genomic DNA from each contributing 2% to the community composition. The concentration of DNA was determined by a Qubit Flex™ fluorometer (Invitrogen, Carlsbad, CA) (**S1 Table**). In total, 80 samples were extracted using four kits.

Library preparation and sequencing

The extracted DNA samples were prepared for Illumina next-generation sequencing using the NEBNext Ultra II FS DNA Prep kit from Illumina, part number E7805. The protocol was followed per the manufacturer's instructions and the samples were fragmented for 15 minutes. Before library preparation, a fragmentation test was performed to determine the optimal fragmentation time. Post library QC was completed first by running the samples on the Agilent BioAnalyzer High Sensitivity Assay and then performing qPCR using the Kapa SYBR green assay for Illumina. Sequencing was conducted on the NovaSeq 6000, a full SP PE250 Illumina run type generating 9.9 x 10^8 total reads.

Bioinformatics

All sequence data were stored with unique de-identified subject IDs. The metagenomic profiles of each sample were characterized using both Kraken2 [8] version 2.1.1 and Metaphlan 3.0 [9] version

122 3.0.7 standard processing pipelines, both run with default parameter values; the standard Kraken2
123 database was used for Kraken2 classifications. These two pipelines were both applied to determine if,
124 and how, analysis results might vary by algorithm; Kraken applies a *k*-mer based taxonomic
125 classification approach, while Metaphlan bases its classifications on a catalog of clade-specific marker
126 genes.

127 Quality control

128 After initial bioinformatic processing, stacked bar plots of microbiome composition were
129 assessed by site and kit (**Fig. 3**). Upon visual inspection, it was clear that one of the oral and vaginal
130 swabs extracted using the HostZERO kit was incorrect: note, samples labeled "52V" and "52O" are
131 shown in **Fig. 3** in their presumed correct positions; "52V" specifically appears to have an abundance of
132 *H. parainfluenzae*, *R. mucilaginosa*, and *S. mitis*, all of which it shares in common with the other oral
133 swab samples processed with the HostZero kit. Taxa identified from the "52O" sample are primarily
134 contaminants that were observed in multiple samples. It is likely these samples were swapped either at
135 time of collection or during the extraction procedure. Therefore, during analysis, the samples were
136 switched back to their presumed correct positions. We conducted our analysis with the sample excluded,
137 in the likely swapped positions, and with the samples switched to their presumed correct position.
138 Analysis results were consistent regardless of inclusion or exclusion of the samples. We chose to present
139 the results with the samples in their presumed correct positions.

140 Statistical analysis

141 We measured alpha diversity using the Shannon and Chao1 indexes and observed species
142 richness, using rarefied samples. Each alpha diversity index was compared between extraction kit types
143 using the ANOVA test if the index was normally distributed, or the Kruskal-Wallis test if the index is

144 non-normal. For comparing between test kits using paired within individual samples, beta diversity was
145 measured using the Bray-Curtis distances and exploratory analysis with principal coordinate analysis
146 (PCoA) plots or Non-metric Multi-dimensional Scaling (NMDS) was completed. To obtain a global
147 overview of the sequencing results, nonmetric multidimensional scaling (NMDS) analysis was
148 performed. We applied NMDS as implemented by the metaMDS function of the R package vegan [10]
149 to the square-root-transformed and Wisconsin-standardized [11] matrix of species-level abundance
150 estimates resulting from Metaphlan. To identify which extraction kit results best approximate the known
151 composition of the mock communities, we used compositional bias analysis using the R package
152 metacal [12]. We scored the ability of each kit to reproduce the pattern of relative abundances according
153 to the geometric standard deviation—defined as the exponential of the standard deviation of log-
154 transformed values, analogous to the geometric mean—of the bias values across all taxa. A metacal bias
155 value of 1 for a given taxa indicates that there is no bias in relative abundance, while values above or
156 below 1 indicate that the relative abundance of the taxa is over- or under-estimated, respectively. The
157 geometric mean of metacal bias values across all taxa is always unity since metacal aims to assess bias
158 in relative, not absolute, abundance. A geometric standard deviation of 1 would indicate that the
159 proportions of all taxa are in perfect accord with the input amounts, while higher values of the geometric
160 standard deviation of bias values indicate bias values further away from the ideal value of 1. All
161 statistical analyses were completed in R.

162 Results and discussion

163 Contamination

164 We identified known laboratory contaminants in our extraction blanks and mock community
165 samples. Other groups have previously identified microbial communities that were associated with

166 specific extraction kits or laboratory reagents often referred to as a "kitome" [13–16]. The most
167 prevalent contaminant we identified was *Sphingomonas melonis,* which is a gram-negative bacterium
168 from the genus *Sphingomonas* and is commonly found in soil, water, and plants [17]. *S. melonis* was
169 detected in all extraction blanks, which contained sterile water. The contaminant was not present in all
170 samples, although was identified in two mock communities and 11 clinical samples (four oral, six rectal,
171 and one vaginal sample). *Burkholderia contaminans* were only detected in blanks and mock community
172 samples that were extracted using the PureLink kit. *B. contaminans* were not detected in any of the
173 clinical samples using any kit. Lastly, *Lactobacillus iners* and *Mobiluncus curtisii* were identified in two
174 blank samples. The contaminants were in separate blank samples with very low abundance (0.07, 0.09
175 respectively). However, both samples containing these contaminants were extracted using the PureLink
176 kit. Both bacteria are present in our clinical samples, and cross-contamination from clinical samples is
177 possible.

178 Mock community samples

179 The expected composition of the mock samples should be equal for eight bacterial species and
180 two fungal species, with each representing 12% and 2% of genomic reads respectively. We also
181 observed generally consistent results of microbial classification when using Kraken and Metaphlan to
182 assign taxonomy. However, if investigators are interested in analyzing fungi, only the Metaplan database
183 contains fungal sequences and should be used. All but one of the bacterial species—*Bacillus subtilis*—
184 were observed in the data (**Fig. 1**). However, *Bacillus intestinalis* was observed in all mock samples and
185 all kits, which were not supposed to be present in the mock community samples. Therefore, we
186 suspected that *B. subtilis* was misclassified as *B. intestinalis* and included *B. subtilis* in the results in
187 Figure 1. We then investigated the sequences included in the Metaphlan taxonomy database and
188 discovered there were only two sequences covering 2.7 kb for *B. subtilis* and 86 sequences covering 79

189 kb included for *B. intestinalis* (by contrast, the Kraken database contains 307 sequences totaling 754 Mb
190 for *B. subtilis* and only one sequence of 4 Mb for *B. intestinalis*). Sequences for the two species in the
191 16S region are 99% similar and may explain the misclassification, despite the WMS approach generally
192 providing higher accuracy in microbial species classification [18,19]. This suggests that even when using
193 WMS, there may be some species and strain level misclassification of microbial taxa that are closely
194 related [20].

195 **Fig. 1 Percentage of mock community microbial reads extracted by kit.** The composition of the mock samples
196 from WMS analysis by the four DNA extraction kits. Eight bacterial species and two fungal species were
197 expected to represent 12% and 2% of genomic reads respectively.

198 Consistent naming of taxa has become more complicated because the discovery of novel genera
199 and species has grown exponentially due to increased use of metagenomic approaches in research. The
200 International Code of Nomenclature of Prokaryotes delineates rules for the naming of Archaea and
201 Bacteria. Validly published, continuously updated taxa names can be found on the List of Prokaryotic
202 names with Standing in Nomenclature site (https://www.bacterio.net) [20,21]. The lack of integration of
203 updated taxonomy into databases used in bioinformatics pipelines results in different groups reporting
204 the same sequence as different species or levels of taxonomy, making it difficult to compare results
205 across studies. Since most mock community samples with known percentages of bacteria or genomic
206 DNA contain few species, investigators should periodically review changes in taxa nomenclature in
207 addition to including analysis of mock communities with their clinical samples as a ground truth. When
208 changes in the nomenclature of taxa commonly reported in the literature occur, it is important to note the
209 previous name to allow for accurate comparison with previous studies.

210 Next, we determined that extraction kits differentially biased the percentage of reads attributed to
211 microbial taxa. For example, *S. aureus* and *L. monocytogenes* were detected at the highest abundance

212 using the HostZERO and PureLink kits, while *L. fermentum* was overrepresented when using the
213 DNeasy method (**Fig. 2**). *S. cerevisiae* was detected at the highest level using the HostZERO kit. The
214 other yeast species included in the mock community, *C. neoformans*, was only detected in samples
215 extracted with the HostZERO kit. Both fungi included in the mock community were underrepresented in
216 all mock community samples, with a median relative abundance for *C. neoformans* of 0.02% compared
217 to 1.05% for *S. cerevisiae*, instead of the expected 2%. We also observed overrepresentation of *S. aureus*
218 and *L. monocytogenes,* at the expense of representation of all gram-negative bacteria (*P. aeruginosa, E.*
219 *coli, S. enterica)*. None of the gram-negative bacteria present in the mock communities were identified
220 in the sequencing results of samples extracted using the HostZERO kit when analyzed using Metaphlan
221 (**Fig. 1 & 2**). Interestingly, a small percentage of reads were assigned to the gram-negative bacteria
222 when using Kraken to assign taxonomy for samples extracted with the HostZERO kit. Given that the
223 gram negative bacteria that were present in higher abundance were correctly annotated when using
224 Metaphlan and Kraken, it is likely because Metaphlan is very conservative relative to Kraken when
225 assigning taxonomy to low abundance microbes [22]. Overall, Kraken has been shown to be a fast and
226 accurate method. However, others have reported that features that allow for speed may result in more
227 false positives when using Kraken, which can be remedied by filtering out low abundance reads prior to
228 assigning taxonomy [23].

229 **Fig 2. Metacal bias parameter estimates of mock communities.** 1.0 indicates no relative bias for a particular
230 species; values less than/greater than 1 indicate a bias toward lower/higher abundance values for the kit in
231 question. The bias estimates are indicated on a log scale.

232 General differences between kit protocols also likely contribute to the observed differences in
233 mock community results. For example, HostZERO was designed to remove >90% of human reads from
234 microbiome samples by degrading eukaryote DNA before the DNA purification step. Despite being

designed to recover >85% of microbial DNA from samples, we found that the protocol was biased against gram-negative bacteria yet performed better for extracting DNA of fungi from mock communities than all other kits tested. However, compared to other kits, the relative abundance of bacteria within samples extracted using the HostZERO kit was distinctly different from the other kits. The PureLink and PowerSoil Pro kits were both formulated to increase yields from difficult to lyse samples and had extraction results that were the most similar. Results for samples extracted using the PowerSoil Pro kit also best approximated the expected proportions of the mock communities, as quantified by compositional bias analysis using the R package metacal [12]; see **Table 2**. The geometric standard deviations of metacal bias value estimates suggest that extraction with the HostZERO kit yields the highest geometric standard deviations in both Kraken and Metaphlan-based analyses, owing to the considerably overestimated abundances of *S. aureus*, *E. faecalis*, and *S. cerevisiae* contrasting with the notably underestimated abundances of *S. enterica*, *E. coli*, *L. fermentum*, and *P. aeruginosa*.

Table 2. Dispersion of metacal bias estimates, as measured by geometric standard deviation (geo sd)

Kit	Kraken geo sd	Metaphlan geo sd
DNeasy	3.09	2.60
HostZERO	11.2	7.62
PowerSoil	3.22	1.81
PureLink	3.37	2.60

Clinical samples

The HostZERO kit yielded smaller fractions of reads assigned to *Homo sapiens* across all three sites sampled (oral, rectal, and vaginal) than did each of the other kits. When using HostZERO for extraction, in oral swabs, an average of 16.4% of reads in oral swabs were mapped to *H. sapien* as

253 compared to 85.4%, 73.6%, and 67.7% for DNeasy, PowerSoil, and PureLink, respectively; in rectal
254 swabs, the HostZERO *H. sapien* mapping rate was 42.4% versus 60.5%, 66.9%, and 58.7% for the other
255 kits; finally, in vaginal swabs, the rate for HostZERO was 74.5% versus 87.5%, 87.9%, and 83.3% (**Fig.**
256 **3**). The HostZERO-prepared samples had correspondingly higher fractions of reads assigned to bacterial
257 taxa than did the samples prepared using other kits. For investigators interested in minimizing the
258 number of human reads in their sequencing run, HostZERO could be the preferable kit. However,
259 minimizing host reads also appears to bias the representation of microbial communities, as samples
260 extracted using HostZERO look more distinctly different than differences between other extraction kits
261 (**Fig. 4**). These differences are consistent with what we observed in the mock community samples.

262 **Fig 3**. **Fraction of reads assigned to H. sapiens versus microbial reads**. Host reads are highly abundant among
263 samples extracted with all kits. HostZERO does reduce the fraction of host reads (light grey), especially for the
264 oral swabs better than other extraction kits. Some of the reads which are not mapped to any of the indicated taxa
265 (including "other bacterial species") remain unidentified by Metaphlan but may be identified as bacterial by
266 kraken2; this is particularly true for the mock specimens.

267 **Fig. 4 Top microbial species extracted for each clinical site by kit type.** Five columns in each box represent
268 the five samples from the volunteers.

269 However, as illustrated in **Figure 5**, when we assess the samples using NMDS the samples
270 separate primarily by site, with the vaginal swabs exhibiting a more substantial dispersion than the other
271 sites. Samples characterized as outliers (based on root-mean-square (RMS) displacement, a.k.a.
272 Euclidean distance, from the corresponding within-kit-and-site NMDS coordinate centroid position
273 exceeding a value of 2) are labeled in a "<kit>: <site>" format on the NMDS plot of **Fig.5**: all four such
274 samples were processed using the HostZERO kit, and three out of four were vaginal swab samples. Two

of these outlier samples, including the lone non-vaginal (rectal) swab outlier, co-locate with the blank specimens.

Fig. 5. Sample Analysis Using Non-metric Multi-dimensional Scaling (NMDS). Samples cluster most tightly by type or site with HostZERO samples being most dispersed, particularly for vaginal samples.

Table 3. Average root-mean-square (RMS) displacements from within-group centroids for two-dimensional NMDS coordinates by kit and site.

Kit	Analysis	Blank	Mock	Oral swab	Rectal swab	Vaginal swab
DNeasy	Kraken	0.335	0.038	0.316	0.363	0.976
HostZERO	Kraken	0.65	0.382	0.113	0.729	1.429
PowerSoil	Kraken	0.044	0.035	0.174	0.532	0.875
PureLink	Kraken	0.233	0.062	0.306	0.352	1.01
DNeasy	Metaphlan	0.059	0.003	0.56	0.622	1.103
HostZERO	Metaphlan	0	0.613	0.492	1.203	1.907
PowerSoil	Metaphlan	0	0.031	0.401	0.526	0.913
PureLink	Metaphlan	0.439	0.176	0.465	0.46	1.039

The average RMS displacements of the samples from their corresponding group centroids within each kit/site combination are displayed in **Table 3**. All blank samples processed with the HostZERO and PowerSoil kits are assigned identical NMDS coordinates, leading to the 0 dispersion values shown in the first column of the table. For the mock samples, as well as both the rectal- and vaginal-site swabs, the HostZERO-processed sample sets have notably higher average RMS dispersions than any of the other kits. For the oral swabs, however, there is a much narrower range of dispersion values between kits, with the DNeasy-processed samples more dispersed than the HostZERO-processed samples. Overall, the

290 communities represented by the HostZERO kit were most different from the other kits for rectal and
291 vaginal sites and DNeasy community representation was most different for the oral microbiome
292 samples. For investigators comparing microbiome communities across studies or conducting meta-
293 analyses, the extraction method used may be an important covariate to include if there is variability
294 across studies.

295 Differences in the fractions assigned to fungal taxa or viruses appeared to be both smaller and
296 less consistent across kits; this may be partially attributable to the lower overall abundance of these
297 species, leading to reduced precision in the estimated abundance. For example, we did not detect any
298 fungal taxa in our clinical samples. Although our mock communities did not include virus as a ground
299 truth, we did identify some viral taxa in oral and rectal clinical samples for all extraction kits.
300 Surprisingly, we also detected virus, predominantly *Staphylococcus* bacteriophage Andhra, in mock
301 communities that did not contain virus using all kits except PowerSoil (**Fig. 6**). Additional studies to
302 ground truth the mycobiome, virome and microbiome concurrently may be informative for investigators
303 interested in studying the interaction across microbial communities.

304 **Figure 6. Total vial abundance for clinical samples by kit.** Total viral (relative) abundance levels above are
305 taken from Kraken2 analysis as viral sequences are not present in the Metaphlan database.

306 Conclusion

307 Similar to reports of bias introduced by extraction kits when conducting amplicon-based
308 sequencing, we observe bias in relative abundance of microbial taxa for WMS by extraction method. We
309 observed very low levels of contamination that varied by kit and significant differences in relative
310 abundance in mock community extractions by method. Although HostZERO was the only kit to extract
311 both fungal species from the mock communities, it was at the expense of gram-negative bacteria.

312 HostZERO was also superior in reducing the number of human reads present in a sample, although this
313 may contribute to greater deviation from the expected yields when using this extraction kit. Given the
314 biased introduced by extraction method, it is important for investigators conducting meta-analyses to
315 include this as a covariate in their analyses when different methods are used across studies. Investigators
316 should weight the benefits and biases introduced by different methods to minimize influences to their
317 analysis.

318 Acknowledgments

319 DNA library prep and sequencing were performed by the Genomic Sequencing and Analysis Facility at
320 UT Austin, Center for Biomedical Research Support. RRID#: SCR_021713.

321 References

322 1. Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Brief
323 Bioinform. 2021;22: 178–193. doi:10.1093/bib/bbz155
324 2. Xu W, Chen T, Pei Y, Guo H, Li Z, Yang Y, et al. Characterization of Shallow Whole-Metagenome
325 Shotgun Sequencing as a High-Accuracy and Low-Cost Method by Complicated Mock
326 Microbiomes. Front Microbiol. 2021;12. Available:
327 https://www.frontiersin.org/articles/10.3389/fmicb.2021.678319
328 3. Fiedorová K, Radvanský M, Němcová E, Grombiříková H, Bosák J, Černochová M, et al. The
329 Impact of DNA Extraction Methods on Stool Bacterial and Fungal Microbiota Community
330 Recovery. Front Microbiol. 2019;10: 821. doi:10.3389/fmicb.2019.00821
331 4. Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Priemé A, Aarestrup FM, et al. Impact of
332 Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition
333 Improved DNA isolation method for metagenomics and 16S rRNA gene profiling + new

334 microbiome findings on human feces, pig feces, and sewage Downloaded from. 2016 [cited 11 May
335 2020]. doi:10.6084/m9.figshare.3475406
336 5. Panek M, Čipčić Paljetak H, Barešić A, Perić M, Matijašić M, Lojkić I, et al. Methodology
337 challenges in studying human gut microbiota-Effects of collection, storage, DNA extraction and
338 next generation sequencing technologies. Sci Rep. 2018;8: 5143. doi:10.1038/s41598-018-23296-4
339 6. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, et al. The truth about
340 metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. 2015;15.
341 doi:10.1186/s12866-015-0351-6
342 7. Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, Van Doorn LJ, et al. Impact of
343 host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for
344 microbiome analysis. Front Microbiol. 2019;10: 1277. doi:10.3389/FMICB.2019.01277/BIBTEX
345 8. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol.
346 2019;20: 1–13. doi:10.1186/S13059-019-1891-0/FIGURES/2
347 9. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating
348 taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3.
349 eLife. 2021;10. doi:10.7554/ELIFE.65088
350 10. Oksanen J. Vegan: ecological diversity. R Package Version 2.4-4. 2017.
351 11. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol
352 Monogr. 1957;27: 326–349. doi:10.2307/1942268
353 12. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing
354 experiments. eLife. 2019;8: e46923. doi:10.7554/eLife.46923

355 13. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of
356 extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass
357 samples. Gut Pathog. 2016;8: 24. doi:10.1186/s13099-016-0103-7
358 14. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory
359 contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12: 87.
360 doi:10.1186/s12915-014-0087-z
361 15. Sheik CS, Reese BK, Twing KI, Sylvan JB, Grim SL, Schrenk MO, et al. Identification and
362 Removal of Contaminant Sequences From Ribosomal Gene Databases: Lessons From the Census of
363 Deep Life. Front Microbiol. 2018;9. Available:
364 https://www.frontiersin.org/articles/10.3389/fmicb.2018.00840
365 16. Tavares M, Kozak M, Balola A, Coutinho CP, Godinho CP, Hassan AA, et al. Adaptation and
366 Survival of Burkholderia cepacia and B. contaminans During Long-Term Incubation in Saline
367 Solutions Containing Benzalkonium Chloride. Front Bioeng Biotechnol. 2020;8. Available:
368 https://www.frontiersin.org/articles/10.3389/fbioe.2020.00630
369 17. Buonaurio R, Stravato VM, Kosako Y, Fujiwara N, Naka T, Kobayashi K, et al. Sphingomonas
370 melonis sp. nov., a novel pathogen that causes brown spots on yellow Spanish melon fruits. Int J
371 Syst Evol Microbiol. 2002;52: 2081–2087. doi:10.1099/00207713-52-6-2081
372 18. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages
373 of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun.
374 2016;469: 967–77. doi:10.1016/j.bbrc.2015.12.083
375 19. Tetz V, Tetz G. Draft Genome Sequence of a Strain of Bacillus intestinalis sp. nov., a New Member
376 of Sporobiota Isolated from the Small Intestine of a Single Patient with Intestinal Cancer. Genome
377 Announc. 2017;5. doi:10.1128/GENOMEA.00489-17

20. Parte AC, Carbasse JS, Meier-Kolthoff JP, Reimer LC, Göker M. List of prokaryotic names with standing in nomenclature (LPSN) moves to the DSMZ. Int J Syst Evol Microbiol. 2020;70: 5607–5612. doi:10.1099/IJSEM.0.004332/CITE/REFWORKS
21. Pallen MJ, Telatin A, Oren A. The Next Million Names for Archaea and Bacteria. Trends Microbiol. 2021;29: 289–298. doi:10.1016/J.TIM.2020.10.009/ATTACHMENT/DE7C894D-2875-4045-9580-611062A6241B/MMC6.XLSX
22. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell. 2019;178: 779–794. doi:10.1016/j.cell.2019.07.010
23. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 2017;18: 182. doi:10.1186/s13059-017-1299-7

Supporting information

S1 Table – Concentrations of extracted DNA.

Data and code availability – Sequence data in the process of being uploaded to SRA and code is available at https://github.com/mlwright97/WMSextraction.

Financial disclosure statement

MLW is supported by the National Institute of Nursing Research (K01NR017903). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research is also supported by the St. David's Center for Health Promotion and Disease Prevention Research and research start-up funding from the University of Texas at Austin. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

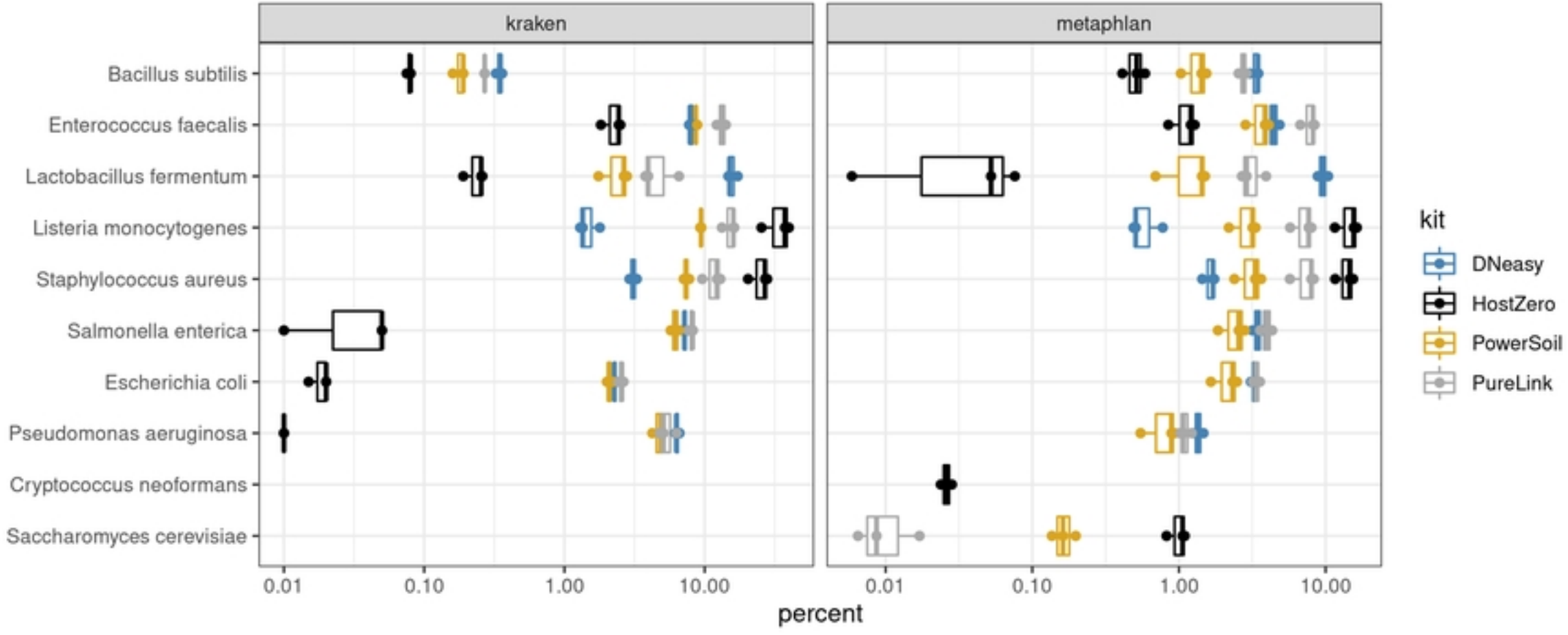
kraken
metaphlan
Bacillus subtilis
Enterococcus faecalis
Lactobacillus fermentum
Listeria monocytogenes
Staphylococcus aureus
Salmonella enterica
Escherichia coli
Pseudomonas aeruginosa
Cryptococcus neoformans
Saccharomyces cerevisiae
0.01
0.10
1.00
10.00
percent
kit
DNeasy
HostZero
PowerSoil
PureLink

Figure 1

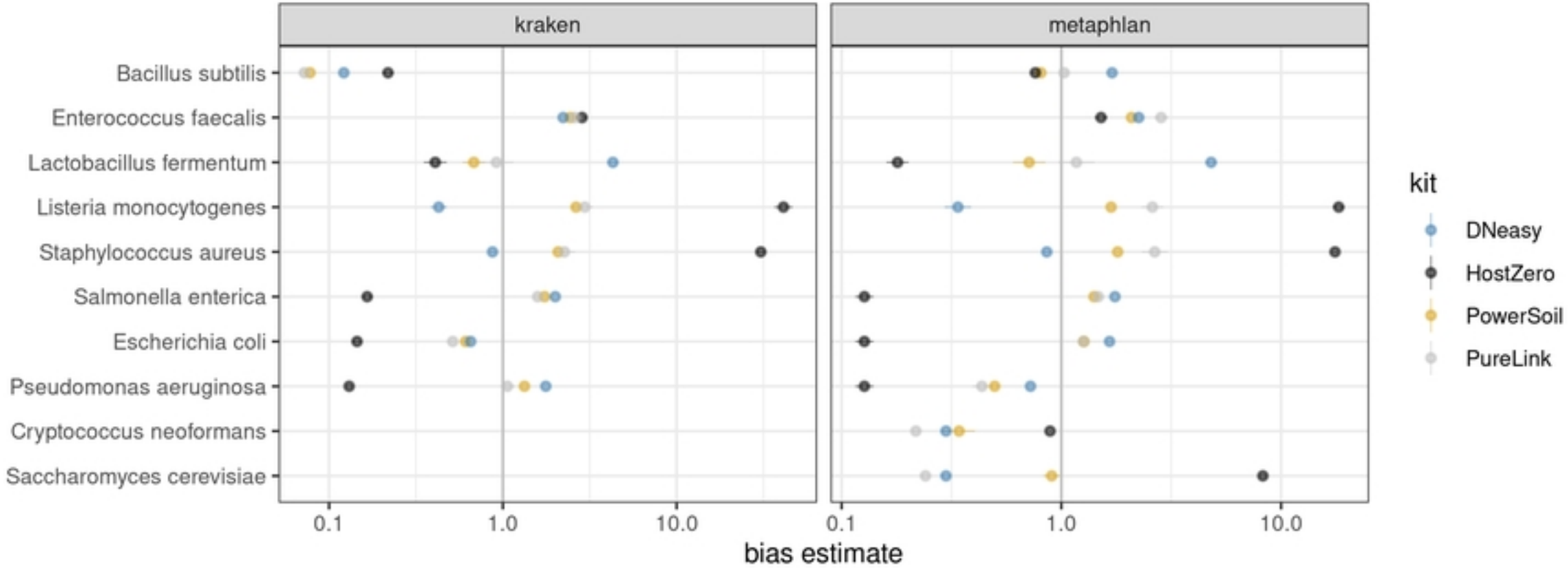
kraken
metaphlan
Bacillus subtilis
Enterococcus faecalis
Lactobacillus fermentum
Listeria monocytogenes
Staphylococcus aureus
Salmonella enterica
Escherichia coli
Pseudomonas aeruginosa
Cryptococcus neoformans
Saccharomyces cerevisiae
0.1
1.0
10.0
bias estimate
kit
DNeasy
HostZero
PowerSoil
PureLink

Figure 2

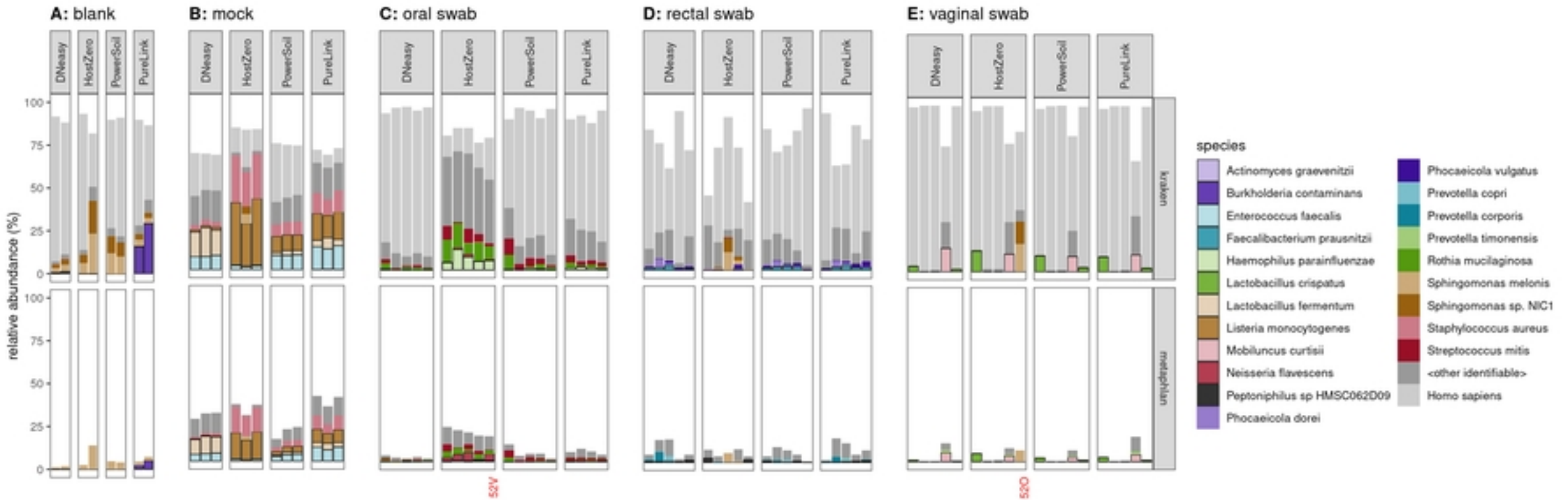
A: blank
B: mock
C: oral swab
D: rectal swab
E: vaginal swab
DNeasy
HostZero
PowerSoil
PureLink
relative abundance (%)
100
75
50
25
0
kraken
metaphlan
S2V
S2O
species
Actinomyces graevenitzii
Burkholderia contaminans
Enterococcus faecalis
Faecalibacterium prausnitzii
Haemophilus parainfluenzae
Lactobacillus crispatus
Lactobacillus fermentum
Listeria monocytogenes
Mobiluncus curtisii
Neisseria flavescens
Peptoniphilus sp HMSC062D09
Phocaeicola dorei
Phocaeicola vulgatus
Prevotella copri
Prevotella corporis
Prevotella timonensis
Rothia mucilaginosa
Sphingomonas melonis
Sphingomonas sp. NIC1
Staphylococcus aureus
Streptococcus mitis
<other identifiable>
Homo sapiens

Figure 3

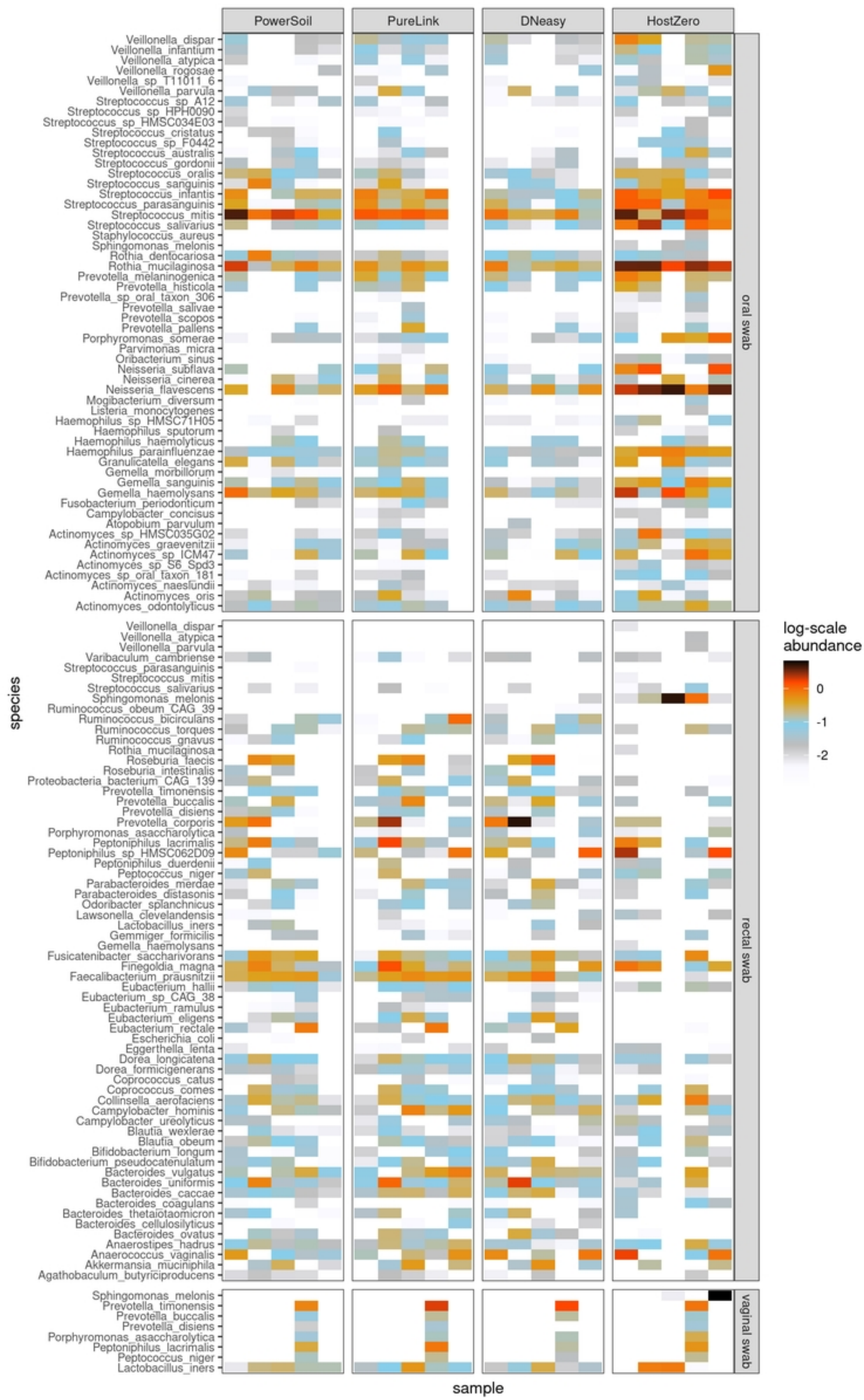
PowerSoil
PureLink
DNeasy
HostZero
oral swab
rectal swab
vaginal swab
species
sample
log-scale abundance
0
-1
-2
Veillonella_dispar
Veillonella_infantium
Veillonella_atypica
Veillonella_rogosae
Veillonella_sp_T11011_6
Veillonella_parvula
Streptococcus_sp_A12
Streptococcus_sp_HPH0090
Streptococcus_sp_HMSC034E03
Streptococcus_cristatus
Streptococcus_sp_F0442
Streptococcus_australis
Streptococcus_gordonii
Streptococcus_oralis
Streptococcus_sanguinis
Streptococcus_infantis
Streptococcus_parasanguinis
Streptococcus_mitis
Streptococcus_salivarius
Staphylococcus_aureus
Sphingomonas_melonis
Rothia_dentocariosa
Rothia_mucilaginosa
Prevotella_melaninogenica
Prevotella_histicola
Prevotella_sp_oral_taxon_306
Prevotella_salivae
Prevotella_scopos
Prevotella_pallens
Porphyromonas_somerae
Parvimonas_micra
Oribacterium_sinus
Neisseria_subflava
Neisseria_cinerea
Neisseria_flavescens
Mogibacterium_diversum
Listeria_monocytogenes
Haemophilus_sp_HMSC71H05
Haemophilus_sputorum
Haemophilus_haemolyticus
Haemophilus_parainfluenzae
Granulicatella_elegans
Gemella_morbillorum
Gemella_sanguinis
Gemella_haemolysans
Fusobacterium_periodonticum
Campylobacter_concisus
Atopobium_parvulum
Actinomyces_sp_HMSC035G02
Actinomyces_graevenitzii
Actinomyces_sp_ICM47
Actinomyces_sp_S6_Spd3
Actinomyces_sp_oral_taxon_181
Actinomyces_naeslundii
Actinomyces_oris
Actinomyces_odontolyticus
Veillonella_dispar
Veillonella_atypica
Veillonella_parvula
Varibaculum_cambriense
Streptococcus_parasanguinis
Streptococcus_mitis
Streptococcus_salivarius
Sphingomonas_melonis
Ruminococcus_obeum_CAG_39
Ruminococcus_bicirculans
Ruminococcus_torques
Ruminococcus_gnavus
Rothia_mucilaginosa
Roseburia_faecis
Roseburia_intestinalis
Proteobacteria_bacterium_CAG_139
Prevotella_timonensis
Prevotella_buccalis
Prevotella_disiens
Prevotella_corporis
Porphyromonas_asaccharolytica
Peptoniphilus_lacrimalis
Peptoniphilus_sp_HMSC062D09
Peptoniphilus_duerdenii
Peptococcus_niger
Parabacteroides_merdae
Parabacteroides_distasonis
Odoribacter_splanchnicus
Lawsonella_clevelandensis
Lactobacillus_iners
Gemmiger_formicilis
Gemella_haemolysans
Fusicatenibacter_saccharivorans
Finegoldia_magna
Faecalibacterium_prausnitzii
Eubacterium_hallii
Eubacterium_sp_CAG_38
Eubacterium_ramulus
Eubacterium_eligens
Eubacterium_rectale
Escherichia_coli
Eggerthella_lenta
Dorea_longicatena
Dorea_formicigenerans
Coprococcus_catus
Coprococcus_comes
Collinsella_aerofaciens
Campylobacter_hominis
Campylobacter_ureolyticus
Blautia_wexlerae
Blautia_obeum
Bifidobacterium_longum
Bifidobacterium_pseudocatenulatum
Bacteroides_vulgatus
Bacteroides_uniformis
Bacteroides_caccae
Bacteroides_coagulans
Bacteroides_thetaiotaomicron
Bacteroides_cellulosilyticus
Bacteroides_ovatus
Anaerostipes_hadrus
Anaerococcus_vaginalis
Akkermansia_muciniphila
Agathobaculum_butyriciproducens
Sphingomonas_melonis
Prevotella_timonensis
Prevotella_buccalis
Prevotella_disiens
Porphyromonas_asaccharolytica
Peptoniphilus_lacrimalis
Peptococcus_niger
Lactobacillus_iners

Figure 4

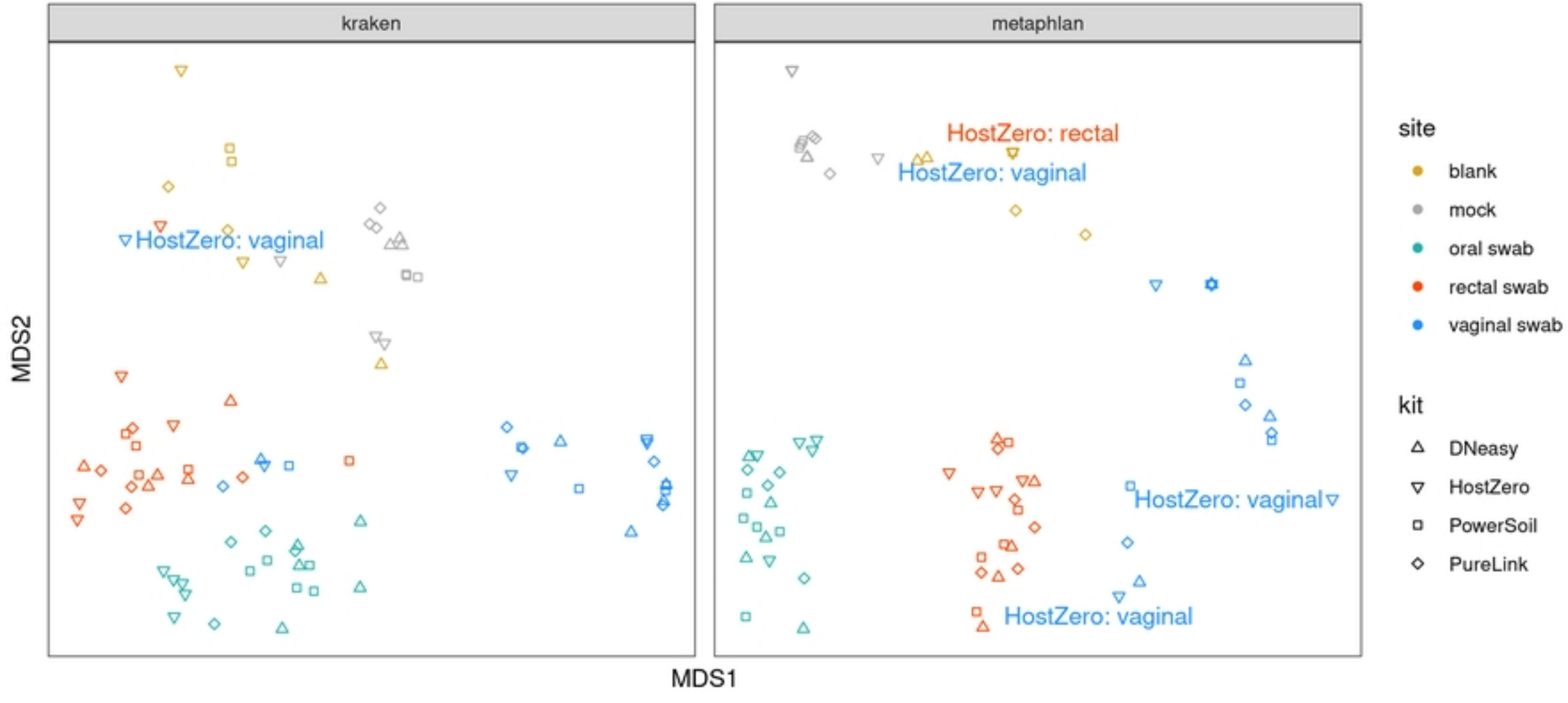
kraken
metaphlan
HostZero: vaginal
HostZero: rectal
HostZero: vaginal
HostZero: vaginal
HostZero: vaginal
MDS2
MDS1
site
blank
mock
oral swab
rectal swab
vaginal swab
kit
DNeasy
HostZero
PowerSoil
PureLink

Figure 5

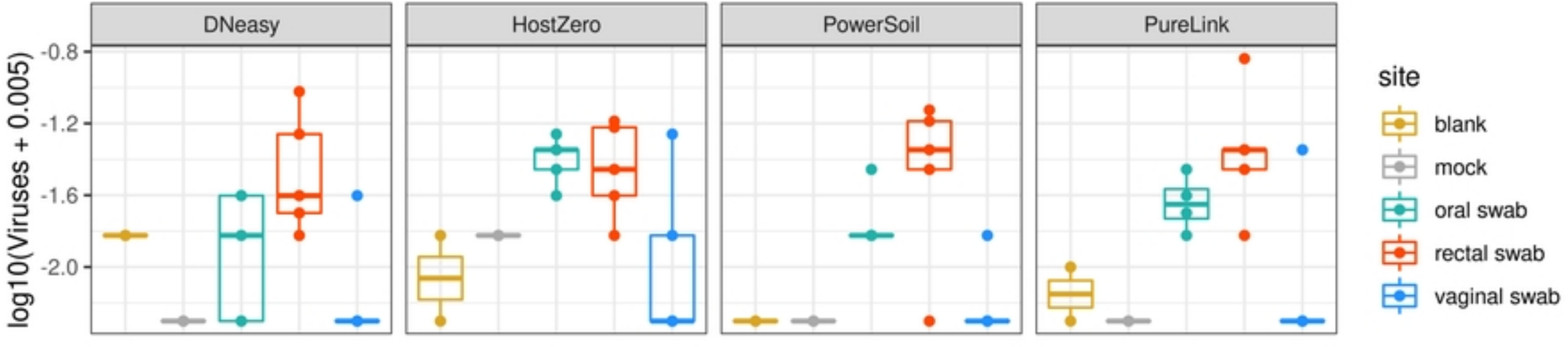
DNeasy
HostZero
PowerSoil
PureLink
log10(Viruses + 0.005)
-0.8
-1.2
-1.6
-2.0
site
blank
mock
oral swab
rectal swab
vaginal swab

Figure 6