Assessing and removing the effect of unwanted technical variations in microbiome data

Muhamad Fachrul[1,2], Guillaume Méric[1,3], Michael Inouye[1,2,4,5,6,7,8], Sünje Johanna Pamp[9,10], Agus Salim[11,12,13,14*]

[1] Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, VIC 3004, Australia.

[2] Department of Clinical Pathology, University of Melbourne, Parkville, VIC 3010, Australia.

[3] Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, VIC 3004, Australia.

[4] Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[5] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[6] British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

[7] Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK

[8] The Alan Turing Institute, London, UK

[9] National Food Institute, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

[10] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

[11] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC 3010, Australia.

[12] School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC 3010, Australia.

[13] Department of Population Health, Baker Heart and Diabetes Institute, Melbourne, VIC 3004, Australia.

[14] Department Mathematics and Statistics, La Trobe University, Bundoora, VIC 3086, Australia.

*Correspondence: salim.a@unimelb.edu.au

Keywords: unwanted variations, microbiome, bioinformatics, batch effects.

Abstract

Varying technologies and experimental approaches used in microbiome studies often lead to irreproducible results due to unwanted technical variations. Such variations, often unaccounted for and of unknown source, may interfere with true biological signals, resulting in misleading biological conclusions. In this work, we aim to characterize the major sources of technical variations in microbiome data and demonstrate how a state-of-the art approach can minimize their impact on downstream analyses. We analyzed 184 pig faecal metagenomes encompassing 21 specific combinations of deliberately introduced factors of technical and biological variations. We identify several known experimental factors, specifically storage conditions and freeze-thaw cycles, as a likely major source of unwanted variation in metagenomes. We also observed that these unwanted technical variations do not affect taxa uniformly, with freezing samples affecting taxa of class *Bacteroidia* the most, for example. Additionally, we benchmarked the performance of a novel batch correcting tool used in this study, RUV-III-NB (https://github.com/limfuxing/ruvIIInb/), to other popular batch correction methods, including ComBat, ComBat-seq, RUVg, and RUVs. While RUV-III-NB performed consistently robustly across our sensitivity and specificity metrics, most other methods did not remove unwanted variations optimally, with RUVg even overcorrecting and removing some of the true biological signals from the samples. Our analyses suggests that a careful consideration of possible technical confounders is critical in the experimental design of microbiome studies to ensure accurate biological reading of microbial taxa of interest, and that the inclusion of technical replicates is necessary to efficiently remove unwanted variations computationally.

Introduction

The technological advancement of sequencing technologies has made microbiome studies more accessible and meaningful. From amplifying short 16S rRNA hypervariable regions to taking advantage of long-read sequencing, the breadth of data options has enabled the field to flourish in the past couple of decades, allowing a better understanding of the role of microbiomes in numerous ecological, environmental and clinical contexts[1]. For example, the dynamics of the gut microbiome is now known to be influenced by environment and diet, and perturbations (or "dysbiosis") have been linked to chronic conditions such as cardiometabolic diseases and type 2 diabetes (T2D)[2–4]. As a result, the human microbiome is now considered an important biological basis for potential therapeutic targets[5,6]. However, the multiplication of microbiome studies using a very wide range of technological methods and experimental designs comes with a cost on reproducibility. Any valuable application from microbiome studies has indeed been largely hindered by the lack of reproducibility due to the presence of unwanted technical variations.

Microbiome studies can differ considerably when it comes to the experimental approaches; and each step of the workflow, including variations introduced by the experimenter, has the potential to introduce artificial results due to unwanted technical variations[7,8]. For example, under-sampling might occur due to the lack of consideration during initial collection process, resulting in zero reads detected for certain microbiota that actually exist in the environment due to their underrepresentation in the samples[9]. Additionally, specific stool collection kits have been shown to impact microbial abundance of faecal samples differently[10]. DNA extraction, library preparation kits, storage condition, storage time, and choice of sequencing platforms have also been found to introduce artificial variations in microbial abundance[7,8,11,12]. As an example, outgrowth of certain microorganisms has been reported between sample preservation methods, as some methods are unable to prevent the growth of facultative anaerobic organisms[12]. Contamination of external microbial taxa can also contribute to unwanted variations and can happen at any stage of the experiment. For instance, recent studies have found that some library preparation kits used before sequencing could introduce specific microbial taxa coming from reagents in the kit[13]. This issue is observed to be even more critical in low-biomass samples, even leading to debated interpretations on the existence of a microbiome in environments that might not have any, such as the human placenta or the meconium[13–16]. Left unaddressed, such unwanted variations can considerably confound true biological signals and result in misleading conclusions.

Various computational methods have been developed to correct for batch effects within experiments. Popular methods such as ComBat[17] and RUV[18], originally developed for microarray

and RNA-seq transcriptomics datasets, have also been considered for microbiome studies[7,19] but do not particularly take into account the specific characteristics of microbiome data such as compositionality and zero-inflation. Indeed, contrary to features in transcriptomics datasets, the presence of each taxa is not independent from the rest of the taxa in the microbiome, and raw abundance information acquired after sequencing is not directly representative of the actual abundance in the environment due to the limit of the sequencing depth that each platform has[20], a problem that is not commonly addressed in current studies[21]. Additionally, microbiome data are typically zero-inflated and very sparse, as many features are present in only a very few samples. Log-transformation of raw abundances, a process allowing for more robust statistical analyses to be performed, prerequires to substitute zero-values with a constant arbitrary number, also known as “pseudocount”. Despite being part of the compositional data analysis (CoDA) standards, there is an argument against adding pseudocount then using log-transformation in analyzing highly sparse datasets in the form of counts per sample, as it changes the ratio of taxa abundance substantially, diminishes variance from less abundant taxonomic groups and artificially exaggerates the differences between zero and non-zero values[22,23].

In this study, we aim to characterize the major sources of technical variations in microbiome data and demonstrate how a state-of-the art approach can minimize their impact on downstream analyses. Using a dataset containing 184 faecal microbiome samples from pigs comprising up to 21 unique combinations of technical variations, we use RUV-III-NB (https://github.com/limfuxing/ruvIIInb/), a novel robust batch correction tool which utilizes Negative Binomial (NB) distribution to estimate and adjust for unwanted variations without the need to add pseudocounts, to identify parts of the experimental workflow that introduce critical unwanted variations that affect observed microbial abundances. Then, we compare the performance of RUV-III-NB to other popular tools including ComBat, ComBat-Seq, RUVg, and RUVs.

Results

For the main analyses of assessing the contributions of unwanted variations towards microbial abundance, we utilized the novel RUV-III-NB, which is able to estimate and adjust sequencing count data for unwanted technical variations based on the negative control taxa and sample replicate information in the dataset. For benchmarking purposes, we refer to methods utilizing batch information for control measure (ComBat and ComBat-Seq) as ComBat-based methods, and methods utilizing control features (RUVg, RUVs, and RUV-III-NB) as RUV-based methods.

Unwanted variations correlate with known sources of technical variations, and mostly affect highly abundant taxa.

We acquired 7 unwanted factors (referred to as W1 to W7) from using RUV-III-NB on the spiked pig metagenomes, representing detected sources of unwanted variations within the dataset. The primary unwanted source of variation (called W1) correlated with log library size (r=0.648), which suggests that it is capturing the variability caused by sequencing depths between samples. It also correlated with log geometric mean (r=0.964), highlighting the ability of RUV-III-NB to identify the optimal form of normalization. In this case, it supports the method of utilizing log geometric mean as scaling factor, such as with the centered log-ratio (CLR) transformation, as a more effective approach than total sum scaling (TSS) normalization at providing the most accurate abundance estimate. A number of known experimental factors were also found to correlate well with other unwanted variations: we found storage conditions and freeze-thaw cycles correlating with W2 (r=0.770) and W3 (r=0.782), respectively **(Figure 1B)**. Differences in the library preparation kit used were also found to weakly correlate with W5 (r=-0.522) and W6 (r=0.500).

Microbial taxa belonging to the *Bacteroidia* class were found to be the most affected by W2, which correlated to storage conditions of the samples **(Figure 1C)**. Members of the *Bacilli* and *Clostridia* classes were also among the most affected across W3–W7. Incidentally, *Bacteroidia, Bacilli,* and *Clostridia* were the most abundant classes in the dataset, with 2,153 species belonging to *Bacteroidia*, totaling over 656 million reads (62.8% of total). To estimate the extent to which each of the detected unwanted factor affected our samples, we calculated the proportion of read counts belonging to the 100 most affected taxa, defined by the taxa with the highest pseudo-R^2 for each unwanted factor (**Figure 1D**). We found that storage conditions (referring here whether samples were kept frozen or at room temperature before processing), represented by unwanted factor W2, affected up to 25.32% of a sample's total abundance — a notable amount considering only 100 out of 8,453 taxa were taken into account **(Figure 1D)**. This suggests that after library size, storage conditions are the primary source of unwanted variation in microbiome studies. Unwanted

142 factors W3 and W4 contributed to 4.72% and 4.29% of overall sample abundance, respectively,
143 whereas W5–W7 each contributed to under 2%.

144 Technical variations do not affect different taxa uniformly.

145 The previous observation suggests that not every taxon is affected similarly by variations in
146 technical procedures, and that some taxa are biologically more sensitive or resistant to varying
147 experimental conditions than others. To further this, we analyzed abundances from several
148 microbial classes in samples from the same source (P1) but subjected to different technical
149 variations (**Figure 1A).** We found the highly abundant bacterial classes to still be differentially
150 abundant between batch groups, including storage conditions, freeze-thaw cycles, and library
151 preparation kits. Taxa from classes *Bacteroidia*, *Bacilli*, *Clostridia*, and *Alphaproteobacteria* were
152 found to significantly differ in abundance when compared between storage conditions and freeze-
153 thaw cycles, with taxa from *Alphaproteobacteria* also differed between samples generated from
154 different library preparation kits **(Figure S1)**. Moreover, the differential abundance patterns, as in
155 the state of the taxa in question being more abundant in one condition than the other, were not
156 shared between bacterial classes in the same conditions: *Bacteroidia* was found to be more
157 abundant in unfrozen samples, whereas *Bacilli*, *Clostridia*, and *Alphaproteobacteria* were less
158 abundant. The pattern also differed for these classes in samples generated with varying freeze-
159 thawing cycles and library preparation kits, with *Bacilli*, *Clostridia* sharing a similar abundance
160 pattern across the different cycles. This suggests that the effect of technical variations is not
161 uniform across different taxa.

162 Most RUV-based methods remove unwanted variations better than ComBat-based
163 methods

164 We then plotted relative log expression (RLE) plots before and after correction using ComBat,
165 ComBat-seq, RUVg, RUVs, and RUV-III-NB, to assess how much overall unwanted variations
166 were removed from each approach, by calculating a metric based on the variance of the medians
167 and interquartile ranges. Unlike with factors of unwanted variations or principal components (PCs)
168 where we could assess how each of them correlates with specific technical variations, RLE plots
169 only assess the whole/overall unwanted variations that exist within a dataset; hence, the term
170 overall unwanted variations here is used when referring to results from RLE analysis. The RLE
171 plots and principal component analysis (PCA) plots before and after RUV-III-NB correction
172 (**Figure 2A, Figure 2B)** visually represent the removal of overall unwanted variations. Based on
173 the RLE metric, better corrections were seen for RUV-based methods than ComBat-based
174 methods, with the exception of RUVs using spike-in taxa only **(Figure 2C)**. RUVg had the best

RLE metric in average ($\bar{x}$ *rle*=8.64), followed by RUV-III-NB ($\bar{x}$ *rle*=6.95). RUVs removed unwanted variations comparably to RUV-III-NB when using both combination and only empirical sets of control features yet did not remove much overall unwanted variations when using only spike-in taxa (*rle*=0.49), scoring even lower than uncorrected data (log2 *rle*=2.93, CLR *rle*=3.98). Both tested ComBat methods, using log2 and CLR transformations, had RLE metrics higher than uncorrected data, but still scored lower on average ($\bar{x}$ *rle*=4.58) than the lowest-scoring RUVg (*rle*=5.15) and RUV-III-NB (*rle*=4.98) runs. ComBat-Seq did not seem to remove the entirety of the unwanted variations as the resulting RLE metric was similar to uncorrected data with log2 transformation (*rle*=2.94). This analysis suggests that RUV-based methods generally perform better in removing overall unwanted variations in microbiome data than ComBat-based methods.

To get a visual representation of overall variation in datasets and the effects of correcting for batch effects, we performed principal component analysis (PCA) on the microbiome composition matrix data and calculated the silhouette statistics (ss) for clustering by a technical factor of main batch separation (in this case storage conditions), using the top four principal components (PCs) before and after correction (**Figure 2B**). Specific clustering of samples based on storage conditions and the library preparation kit used could be seen in the PCA plots before any batch effect correction, specifically captured by PC3 and PC4, respectively, but this was no longer apparent after correction with RUV-III-NB using both spike-in and empirical taxa **(Figure 2B)**. With the exception of RUVs using only empirical control taxa *(ss*=0.51*)*, all approaches had lower silhouette scores compared to uncorrected data (uncorrected *ss*=0.488; ComBat *ss*=0.11–0.26; ComBat-Seq *ss*=0.11; RUVg *ss*=0.14–0.17; RUV-III-NB *ss*=0.09–0.16; RUVs *ss*=0.08-0.51), suggesting successful removal of storage condition effects **(Figure S2)**. The effect of using different normalization methods did not seem to alter the silhouette scores in uncorrected data, unlike when using ComBat approaches. Overall, correction using RUV-III-NB and ComBat-Seq yielded the lowest average scores within this dataset, indicating their robustness in removing unwanted variations from the major technical variation source.

To identify taxa most affected by different unwanted sources of variation, we performed differential abundance analysis using edgeR between the different storage conditions, in which only spiked P1 samples were used. After FDR correction and fold change filtering (FDR<0.05, |log2(FC)| >1), samples ran using ComBat-Seq had the lowest number of differentially abundant taxa after correction, followed by RUV-III-NB and RUVs with combination control taxa **(Table S1)**. Yet, correction with RUV-III-NB resulted in the highest estimates of the proportion of null (pi0; non-differentially expressed) features, with correction using combination and solely empirical taxa

208 reaching 0.964 and 0.950, respectively — higher than ComBat-Seq (pi0=0.905) **(Figure 2D)**.
209 Correction using RUV-III-NB also resulted in the highest average pi0 compared to the other RUV
210 methods, followed by RUVs which performed comparably when using combination control
211 features (pi0=0.955). RUV methods with solely spike-in control taxa all performed comparably,
212 with higher results than uncorrected data but lower than ComBat-Seq. Taken together, these
213 observations suggest that RUV-III-NB performed the most consistently robust compared to the
214 other methods in removing unwanted variations.

215 Technical replicates and choice of control features are crucial for RUV methods

216 To assess how well correction methods preserve individual information, we calculated silhouette
217 statistics of individual separation after removal of unwanted variations, a metric describing how
218 well samples from two conditions (here, pigs P1 and P2) separate after clustering, based on how
219 close each point in one cluster is to points in the neighboring clusters; in this case, high value
220 indicates precise clustering and separation of samples based on the pig individuals and, whereas
221 low value suggests otherwise. RUVs using empirical control taxa had the highest silhouette score
222 (ss=0.797), followed by RUVs and RUV-III-NB with combination control taxa (ss=0.790 and 0.761,
223 respectively) **(Figure 3A)**. However, RUV-III-NB in average had a higher silhouette score than
224 RUVs, as correction with solely spike-in control taxa using the latter yielded a silhouette score
225 lower than uncorrected data (RUVs $\bar{x}$ *ss*=0.64, RUV-III-NB $\bar{x}$ *ss*=0.75). ComBat-based methods
226 performed comparably to uncorrected data, with correction using ComBat using log2 having
227 higher silhouette score. Interestingly, RUVg yielded the lowest score overall, with all three control
228 feature options scoring just below zero. This is visualized in the PCA plots before and after
229 correction, in which separation of P1 and P2 samples are explained by PC1 and PC2 even before
230 any correction. The discrimination between individual metagenomic profiles is retained after RUV-
231 III-NB correction, but not after using RUVg **(Figure 3B)**. This analysis suggests that sample
232 replicate information is necessary when correcting unwanted variations using RUV-based
233 methods to avoid over-correction. That being said, despite both methods utilizing sample replicate
234 information, RUV-III-NB still prevailed over RUVs when taking into consideration performance
235 consistency over the different sensitivity and specificity metrics.

236 Discussion

237 The impact of technical variations in microbiome analysis is an important topic that has been
238 explored numerous times, and past studies have reported on the best practices at every stage of
239 sample processing[24]. However, our study is the first to quantify the contribution of different
240 technical factors towards unwanted variations in a shotgun metagenomic sequencing dataset,

241 and how to best remove them. To do so, we used RUV-III-NB, a novel tool accessible on
242 (https://github.com/limfuxing/ruvIIInb/), which we also compare to other available published
243 methods.

244 Inconsistent sample storage can be a large source of unwanted variations in 245 microbiome studies

246 The variety of storage conditions – which encompass factors such as temperature, time, and
247 storing method used – have been previously found to introduce considerable effects on microbial
248 abundance[11,12,25,26]. The impact of different storage conditions is also not limited to metagenomic
249 data, as Hickl et al. also reported significant differences in microbial protein identification between
250 samples stored using flash freezing and RNAlater[27]. Our study does not aim to replicate and
251 validate the findings of such studies, but to show how, if left unaddressed, the differences in
252 sample storage conditions could introduce significant changes in microbial abundance of
253 abundant taxa which are potentially important for our understanding of particular conditions or
254 diseases. The three most-affected bacterial classes in our study, namely *Bacteroidia*, *Clostridia*
255 and *Bacilli*, are among the most dominant microbial classes in the mammalian gut[28–30], therefore
256 it is not surprising to see them strongly affected by technical variations. Both *Bacteroidia* and
257 *Clostridia* classes in particular have been found to be significantly impacted by different storage
258 methods in both metagenomic and metaproteomic data[27]. In clinical cases, the dysbiosis of
259 members of the classes *Bacteroidia* and *Clostridia* have been linked to a wide variety of
260 conditions, including T2D, Crohn's disease, lupus, HIV infection, and major depressive
261 disorders[30–35]. Hence, extra precautions are necessary in making sure the experimental design
262 minimizes sources of unwanted variations as possible.

263 Metagenomic sequencing depth and data normalization also contribute to 264 unwanted variations in microbiome studies

265 Aside from storage conditions, RUV-III-NB identified that library size and geometric mean
266 correlated with unwanted factor W1 in our study, which is unsurprising. In microbial
267 metagenomics, sequencing depth typically impacts the robust classification of reads into genes
268 or microbial taxa, and optimal depth to use depends on each study objectives[36,37]. Our dataset
269 included a range of sequencing depths, which explains why library size was captured as a source
270 of variation by RUV-III-NB. Furthermore, although normalization and log transformation alone are
271 not enough to remove unwanted variations, normalization by library size, also known as TSS
272 normalization, remains widely used in the field mostly due to conceptual simplicity[38]. However,

273 this approach does not account for heteroskedasticity and is very sensitive to pseudocount[39]. With
274 unwanted factor W1 correlating much stronger with geometric mean than library size, our work
275 suggests that normalizing by geometric mean (e.g., CLR transformation) rather than by library
276 size (e.g. TSS) potentially limits unwanted variations in observed microbial abundances.

277 **RUV-III-NB is a robust and consistent microbiome batch correction method**

278 Although the observation of best practice methods and preventive measures are always
279 recommended when it comes to avoiding unwanted technical variations in any experimental
280 setting, there are a few existing computational approaches to rectify batch effect variation after a
281 dataset has been processed and sequenced. Here, we show that, despite deliberately introduced
282 complications in our gut metagenomics experimental workflow, the effective removal of unwanted
283 variations was still possible during downstream analysis, with different methods varying in
284 efficiency. Methods such as ComBat, RUVg, and RUVs require normalization of the data prior to
285 the actual batch removal step, but RUV-III-NB is able to take into account the most optimal
286 normalization method and identify unwanted variations in one step without any prior pseudocount
287 addition.

288 Additionally, ComBat-based methods rely on known batch information and might not be suitable
289 in the realistic situation of multiple batch effects conflating with true biological signals.
290 Nevertheless, when batch information is clearly defined, ComBat-based methods remove
291 unwanted variations efficiently from our samples, but also impact the recovery of true biological
292 signals separating individuals. In our work, ComBat-based methods were also unable to identify
293 and adjust for unknown unwanted variations, which are expected to be common and unavoidable
294 in reality. RUV methods avoid this hurdle by adjusting samples based on the variations of the
295 control features, which are assumed to be in constant amount across technical variations.

296 One potential issue with RUV methods is that, unlike with RNA-seq data where housekeeping
297 genes have been established[40,41] and are routinely used to normalize expression, there is no
298 obvious parallel in microbiome features, as taxa composition varies considerably across datasets.
299 The choice of control features is especially crucial for the methods to work as intended; therefore,
300 the presence of spike-ins is highly beneficial when utilizing RUV methods in microbiome studies.
301 In our case, the mock community consisted of 6 bacterial and 2 eukaryotic microorganisms. Over
302 3,000 genes have been established to be consistently expressed in human cells and are used to
303 normalize expression in RNA-seq studies, which is considerably larger than our spike-ins. We
304 therefore supplemented the spike-in taxa with a set of empirical control taxa. In this study,
305 correction with only the spike-in control taxa removed unwanted variations to an extent, though

using the additional empirical control taxa significantly improved it. Additionally, we observed that using only empirical taxa without spike-ins was still effective in removing unwanted variations, indicating that it is possible to effectively use RUV methods without spike-in taxa, which is more convenient when preparing samples.

When comparing between RUV methods, we showed how inclusion of sample replicate information is important, as demonstrated by RUVg's tendency to overcorrect and lose individual separation in the process. Both RUVs and RUV-III-NB utilize sample replicate information, and are therefore able to avoid overcorrection, yet the former performed less consistent across the performance tests. RUV-III-NB has an advantage over RUVs due to its direct modelling of mean-variance relationship of count data – a feature not shared by RUVs.

In this study, we show how technical variations – especially storage conditions – may introduce unwanted variations in microbiome data, affecting the observed abundances of important and dominant microbial taxa. Left unaddressed, this might be problematic for studies focusing on such taxa, therefore careful considerations are crucial in designing the experimental workflow of microbiome studies. Minimizing the possibility of introducing unwanted variations by limiting the presence of batches and utilizing consistent storage conditions and equipment are highly suggested. We finally show that for existing datasets, post-processing corrective measures can still be performed *in silico* to remove unwanted variations stemming from variations in experimental techniques, and we suggest the use of *RUV-III-NB* as a consistent and robust method.

327 # Materials and Methods

328 ## Sample processing, library preparation, and sequencing conditions

329 The fecal samples investigated in this study originated from two individual pigs (P1 and P2)
330 collected right after defecation from two different conventional pig farms in Denmark as previously
331 described[42,43]. Each individual pig faecal sample was homogenized with a sterile wooden spatula
332 and then separated into two large aliquots. One of the aliquots was spiked with a freshly prepared
333 mock community consisting of 6 bacterial and 2 eukaryotic microorganisms, namely
334 *Propionibacterium freudenreichii* DSM 20271, *Bacteroides fragilis* NCTC 9343, *Staphylococcus*
335 *aureus* NCTC 8325*, Fusobacterium nucleatum* ATCC 25586*, Escherichia coli* ATCC 25922*,*
336 *Salmonella* Typhimurium str. ATCC 14028S*, Cryptosporidium parvum* IOWA II isolate, and
337 *Saccharomyces cerevisiae* S288C*.* The spiked fecal samples were homogenized using a sterile
338 wooden spatula, and small aliquots for each sample storage condition were prepared in
339 Eppendorf tubes for both, the spiked and unspiked aliquots. DNA was isolated immediately from
340 the aliquots for the initial time point (storage for 0h) using a modified QIAamp Fast DNA Stool
341 Mini Kit (Qiagen) protocol with an initial bead beating step (MoBio garnet beads)[44]. The remaining
342 aliquots were stored at different storage conditions comprising different temperature and time
343 combinations **(Figure 1A)**. Samples were stored at 22°C, 5°C, -20°C, and -80°C for several hours
344 (days) (16h (0.67 days), 40h (1.6 days), 64h (2.6 days), 88h (3.6 days)) and for the temperatures
345 -20°C and -80°C also for months and up to one year (4m, 8m, 12m). Aliquots stored for 40h and
346 88h, as well as a subset of 64h, also underwent 2–4 freeze-thaw cycles. All fecal samples
347 underwent the same DNA isolation method (see above)[43,44] prior to library preparation.

348 Three different library preparation kits (the PCR-free NEXTflex and KAPA, as well as Nextera)
349 and two sequencing platforms (HiSeq 4000 and NextSeq 500) were used in the study, similarly
350 as previously described[42,43]. All samples were sequenced paired-end with a read length of 150
351 bp. Two to three technical replicates were performed for each treatment combination.

352 Together, a total of 184 different samples spread across 60 different replicate groups were
353 acquired, encompassing 21 specific combinations of storage conditions, spiking status, library
354 preparation kit used, as well as thawing cycles. The raw reads are deposited at the European
355 Nucleotide Archive (ENA) (Project acc.: PRJEB31650).

356 ## Quality control and taxonomic classification

357 Quality control of sequencing files was done following DTU Food's in-house pipeline,
358 FoodQCpipeline (https://bitbucket.org/RolfKaas/foodqcpipeline), in which BBMap's bbduk2

359 (v38.71, https://jgi.doe.gov/data-and-tools/bbtools/) was used to trim reads with a length of at least
360 50bp, Phred score of at least 20, and also remove a custom list of Illumina adapters. FastQC
361 (v0.11.8) was also applied to the files before and after trimming to assess the quality of the
362 reads[45].

363 Taxonomic classification was done with Kraken2[46], followed by taxa abundance re-estimation
364 using bracken[47] at species level with read length of 150 and minimum taxa threshold of 1. A
365 custom genomic index was used for the classification based on GTDB release 89[48,49], which
366 includes 23,458 bacterial and 1,248 archaeal species. After merging individual reports, taxa
367 filtering was done based on CPM>4 in at least 15% of all samples. A total of 8,453 taxa were
368 included for downstream analysis.

369 **RUV-III-NB batch correction method**

370 The novel RUV-III-NB (https://github.com/limfuxing/ruvIIInb/) method extends the RUV-III method
371 that was previously developed for array-based gene expression data[50]. Instead of using linear
372 model, it uses a Negative Binomial-based generalized linear model (GLM) with log link function
373 to model the effect of wanted biological signals and unwanted variations on the sequencing count
374 at taxon-level. The method uses negative control taxa (i.e., taxa whose variations are assumed
375 solely due to unwanted variations) to estimate the latent unwanted factors, followed by the
376 utilization of replicate information to estimate the taxon-specific effect of these latent unwanted
377 factors on the variations of the sequencing count. Once these are estimated, RUV-III-NB uses
378 percentile-invariant adjusted count to output the sequencing count matrix that has been adjusted
379 or corrected for the unwanted variations. The percentile-invariant adjusted count was calculated
380 based on a modification version of the randomized quantile residual method[51].

381 **Sensitivity and specificity of batch correction methods**

382 We compare the effectiveness of the following methods for removing unwanted variation:
383 ComBat[52], ComBat-Seq[53], RUVg[18], RUVs[18], and RUV-III-NB (see RUV-III-NB subsection).
384 ComBat and ComBat-Seq both utilize known batch variable to remove batch effects from a
385 dataset, though the former allows normalized and/or log-transformed matrix as input, whereas the
386 latter requires raw integer count matrix as input[52,53]. RUVg, RUVs, and RUV-III-NB all utilize
387 control features in estimating unwanted factors. Both RUVg and RUVs assume an underlying
388 Normal model and take normalized and/or log-transformed matrix as input, while RUV-III-NB only
389 accepts integer count matrix; though both RUVs and RUV-III-NB require sample replicate
390 information[18]. Out of all the methods, only ComBat-Seq and RUV-III-NB deal directly with the

391 mean-variance relationship in the count data by using NB distribution. We tested two different log-
392 transformation approaches for uncorrected data and prior to ComBat correction: normalization by
393 library size / total sum scaling (TSS) followed by log2 transformation, as well as Aitchison's
394 centered log-ratio (CLR) transformation[54–56]. Prior to RUVg and RUVs, CLR transformation was
395 also performed. For RUV methods, we set k – which represents the number of unwanted
396 variations – to 7 as it is the highest possible number for the smallest set of control features. For
397 RUV-III-NB, we also set the parameters *lambda.a*=0.01 and *lambda.b*=5.

398 In addition to the 8 spike-in taxa, we curated a set of empirical control taxa – which is defined as
399 a set of microbial taxa that are present and empirically concordant in abundance in both groups
400 that we would like to compare against – to be used as control features in RUV methods. To identify
401 the empirical control taxa, the following procedure was used:

402 - Differential abundance analysis was done between spiked and unspiked samples using
403 edgeR[57,58] for each unique replicate group of each pig. In average, only <1% of the taxa
404 were significantly differentially abundant in all the replicate groups.
405 - We then took the intersection of the least significant taxa from each replicate group within
406 the same pig, specifically the bottom 86% of overall taxa based on FDR<0.05, as control
407 taxa for each pig.
408 - From the control taxa of both pigs, we then took the intersection and identified a total of
409 471 overlapping taxa. This set is what we refer to as the empirical control taxa from this
410 point on.

411 For each RUV method, performance was assessed while using either solely spike-in taxa (8 taxa),
412 solely empirical taxa (471 taxa), or the combination of both as control features (479 taxa).

413 We generated relative log expression (RLE)[59] and principal component analysis (PCA) plots
414 before and after correction to assess how well the different methods perform. RLE plots visualize
415 the presence of unwanted variations by calculating the deviations of from the median of each
416 feature, in this case microbial taxa, whereas PCA plots visualize clustering of certain batches
417 through unsupervised dimensionality reduction[59,60]. For RUV-III-NB, the log of percentile-invariant
418 adjusted count was used as the adjusted data matrix for the visualizations. To compare the RLE
419 plots between methods, we calculated a metric capturing the variance of medians of samples
420 from the same individual as well as the variance of the RLE's interquartile ranges (IQR):

421 - We subtracted each sample's median with the average median of all samples from the
422 same pig source, then find the variance of all the medians.

- After the addition of the variance of medians and the variance of all the IQRs from the RLE, we then performed log transformation to the total.
- We took the negation of the transformed value; hence, higher metric implies more effective removal of unwanted variations.

For PCA plots, we calculated the silhouette statistics using the 'cluster' package [61] (version 2.1.0) to assess two different comparisons:

- Preservation of individual source information, in which we calculated the score of separation between P1 and P2 samples and higher scores signify better preservation.
- Batch dispersion, in which we calculated the scores to see clustering of batch groups within each pig group, with lower scores signifying better removal.

We also performed an additional specificity assessment for the methods through differential abundance analysis using edgeR[57,58] on samples from the same source (P1) between groups of the main batch – in this case, between frozen and unfrozen storage conditions. Since ComBat only produces a transformed adjusted data matrix and not the estimated factors of unwanted variations, and therefore lacking the factors of unwanted variations necessary for covariates in differential abundance analysis, it is excluded from this analysis. Output from ComBat-Seq was directly used as input for the analysis. For RUV methods, untransformed integer counts were used as input along with factors of unwanted variations (W) as covariates in the linear model. Out of all the differentially abundant microbial taxa, we set the significance threshold to those with p-value < 0.05 after FDR correction and with absolute log-foldchange of over 1. We also performed q-value estimation of the results, which measures the proportion of true null p-values (pi0).

Quantify relative contribution from unwanted factors

We used RUV-III-NB to quantify unwanted factors in the data and analyze their relative contribution towards microbial taxa abundance. Hence, for our main analysis only spiked samples were used as the approach requires control features as input, which in our case included the 8 spike-in taxa and an additional set of 471 empirically constant taxa explained in the previous section. We set the same number of k as stated in the previous section for consistency. We then used a negative binomial generalized linear model (*glm.nb*) and model the regression of each microbial taxa as the response variable and the progressive accumulation of factors of unwanted variations W as the predictor. We then calculated the Pseudo-R^2 from each model using the PseudoR2 function from the 'DescTools' R package (version 0.99.38)[62] with Veall-Zimmermann correction[63], as it is among the closest approximations to ordinary least square R^2. Excluding the first factor, the contribution of each individual unwanted factor (W) towards each microbial taxon

456 is represented by the difference in pseudo-R^2 between the cumulative n^{th} and $n\text{-}1^{th}$ models (i.e.
457 contribution of W3 was calculated by subtracting the pseudo-R^2 accounting for unwanted factors
458 W1+W2+W3 with the value accounting for unwanted factors W1+W2). To see the effect of known
459 technical variations on microbial abundance, we then took the top 100 affected microbial taxa by
460 each of the individual factor of unwanted variations, ranked based on their pseudo-R^2 and
461 grouped based on taxonomic class. Since factor of unwanted variation W1 correlated strongly
462 with library size and geometric mean instead of any known technical variation, we only included
463 factors W2 – W7 in this step. From the top 100 affected microbial taxa in each factor of unwanted
464 variation, we calculated the proportion of their abundance in an average sample. For each of the
465 most dominantly affected microbial classes, we also performed Wilcoxon signed rank test
466 between storage conditions and Kruskal-Wallis test to compare the different freeze-thawing
467 cycles and library preparation kits on log2-transformed, TSS-normalized data.

468 ## References

469 1. Morgan, X. C. & Huttenhower, C. Chapter 12: Human Microbiome Analysis. *PLoS*
470 *Comput. Biol.* **8**, e1002808 (2012).
471 2. Tilg, H. & Moschen, A. R. Microbiota and diabetes: An evolving relationship. *Gut* **63**,
472 1513–1521 (2014).
473 3. Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* **569**, 641–648
474 (2019).
475 4. Wang, J. *et al.* A metagenome-wide association study of gut microbiota in type 2
476 diabetes. *Nature* **490**, 55–60 (2012).
477 5. Wong, A. C. & Levy, M. New Approaches to Microbiome-Based Therapies. *mSystems* **4**,
478 (2019).
479 6. Lemon, K. P., Armitage, G. C., Relman, D. A. & Fischbach, M. A. Microbiota-targeted
480 therapies: An ecological perspective. *Science Translational Medicine* (2012).
481 doi:10.1126/scitranslmed.3004183
482 7. Wang, Y. & LêCao, K.-A. Managing batch effects in microbiome data. *Brief. Bioinform.*
483 (2019). doi:10.1093/bib/bbz105
484 8. Nearing, J. T., Comeau, A. M. & Langille, M. G. I. Identifying biases and their potential
485 solutions in human microbiome studies. *Microbiome* **9**, 113 (2021).
486 9. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is
487 inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).
488 10. Byrd, D. A. *et al.* Comparison of Methods To Collect Fecal Samples for Microbiome
489 Studies Using Whole-Genome Shotgun Metagenomic Sequencing. *mSphere* **5**, (2020).
490 11. Ma, J. *et al.* Variations of Gut Microbiome Profile Under Different Storage Conditions and
491 Preservation Periods: A Multi-Dimensional Evaluation. *Front. Microbiol.* **11**, 972 (2020).
492 12. Pribyl, A. L. *et al.* Critical evaluation of faecal microbiome preservation using
493 metagenomic analysis. *ISME Commun.* 1–10 (2021). doi:10.1038/s43705-021-00014-2
494 13. de Goffau, M. C. *et al.* Human placenta has no microbiome but can contain potential
495 pathogens. *Nature* **572**, 329–334 (2019).
496 14. Drago, L. *et al.* Microbiota network and mathematic microbe mutualism in colostrum and
497 mature milk collected in two different geographic areas: Italy versus Burundi. *ISME J.* **11**,
498 875–884 (2017).
499 15. Zuo, T. *et al.* Gut mucosal virome alterations in ulcerative colitis. *Gut* **68**, 1169–1179
500 (2019).
501 16. Kennedy, K. M. *et al.* Fetal gut colonization: meconium does not have a detectable

502 microbiota before birth. *bioRxiv* 2021.02.17.431710 (2021).
503 doi:10.1101/2021.02.17.431710
504 17. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package
505 for removing batch effects and other unwanted variation in high-throughput experiments.
506 *Bioinformatics* **28**, 882 (2012).
507 18. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor
508 analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
509 19. Gibbons, S. M., Duvallet, C. & Alm, E. J. Correcting for batch effects in case-control
510 microbiome studies. *PLoS Comput. Biol.* **14**, e1006102 (2018).
511 20. Rivera-Pinto, J. *et al.* Balances: a New Perspective for Microbiome Analysis. *mSystems*
512 **3**, e00053-18 (2018).
513 21. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome
514 datasets are compositional: And this is not optional. *Front. Microbiol.* **8**, 1–6 (2017).
515 22. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and
516 dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome*
517 *Biol.* **20**, 295 (2019).
518 23. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-
519 seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
520 24. Wu, W. K. *et al.* Optimization of fecal sample processing for microbiome study — The
521 journey from bathroom to bench. *Journal of the Formosan Medical Association* **118**, 545–
522 555 (2019).
523 25. Shaw, A. G. *et al.* Latitude in sample handling and storage for infant faecal microbiota
524 studies: The elephant in the room? *Microbiome* **4**, 40 (2016).
525 26. Choo, J. M., Leong, L. E. X. & Rogers, G. B. Sample storage conditions significantly
526 influence faecal microbiome profiles. *Sci. Rep.* **5**, 1–10 (2015).
527 27. Hickl, O. *et al.* Sample preservation and storage significantly impact taxonomic and
528 functional profiles in metaproteomics studies of the human gut microbiome.
529 *Microorganisms* **7**, (2019).
530 28. King, C. H. *et al.* Baseline human gut microbiota profile in healthy people and standard
531 reporting template. *PLoS One* **14**, e0206484 (2019).
532 29. Barelli, C. *et al.* The Gut Microbiota Communities of Wild Arboreal and Ground-Feeding
533 Tropical Primates Are Affected Differently by Habitat Disturbance. *mSystems* **5**, (2020).
534 30. Jiao, N. *et al.* Gut microbiome may contribute to insulin resistance and systemic
535 inflammation in obese rodents: A meta-analysis. *Physiol. Genomics* **50**, 244–254 (2018).

536 31. Vujkovic-Cvijin, I. *et al.* Dysbiosis of the gut microbiota is associated with HIV disease
537 progression and tryptophan catabolism. *Sci. Transl. Med.* **5**, (2013).
538 32. Capuco, A. *et al.* Current Perspectives on Gut Microbiome Dysbiosis and Depression.
539 *Advances in Therapy* **37**, 1328–1346 (2020).
540 33. Kaakoush, N. O. *et al.* Microbial dysbiosis in pediatric patients with Crohn's disease. *J.*
541 *Clin. Microbiol.* **50**, 3258–3266 (2012).
542 34. Cheung, S. G. *et al.* Systematic review of gut microbiota and major depression. *Frontiers*
543 *in Psychiatry* **10**, 34 (2019).
544 35. Hevia, A. *et al.* Intestinal dysbiosis associated with systemic lupus erythematosus. *MBio*
545 **5**, 1548–1562 (2014).
546 36. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology*
547 **17**, (2016).
548 37. Zaheer, R. *et al.* Impact of sequencing depth on the characterization of the microbiome
549 and resistome. *Sci. Rep.* **8**, 5890 (2018).
550 38. Paulson, J. N., Bravo, H. C. & Pop, M. Reply to: A fair comparison. *Nat. Methods* **11**,
551 359–360 (2014).
552 39. McKnight, D. T. *et al.* Methods for normalizing microbiome data: An ecological
553 perspective. *Methods Ecol. Evol.* **10**, 389–400 (2019).
554 40. Dheda, K. *et al.* Validation of housekeeping genes for normalizing RNA expression in
555 real-time PCR. *Biotechniques* **37**, 112–119 (2004).
556 41. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends in*
557 *Genetics* **29**, 569–574 (2013).
558 42. Poulsen, C., Pamp, S., Ekstrøm, C. & Aarestrup, F. Library preparation and sequencing
559 platform introduce bias in metagenomics characterisation of microbial communities.
560 *bioRxiv* 592154 (2019). doi:10.1101/592154
561 43. Poulsen, C. S., Kaas, R. S., Aarestrup, F. M. & Pamp, S. J. Standard Sample Storage
562 Conditions Impact on Inferred Microbiome Composition and Antimicrobial Resistance
563 Patterns. (2021).
564 44. Knudsen, B. E. *et al.* Impact of Sample Type and DNA Isolation Procedure on Genomic
565 Inference of Microbiome Composition. *mSystems* **1**, 2021 (2016).
566 45. Andrews, S. Babraham Bioinformatics - FastQC A Quality Control tool for High
567 Throughput Sequence Data. Available at:
568 https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 19th March
569 2020)

46. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
47. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **2017**, e104 (2017).
48. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
49. Méric, G., Wick, R. R., Watts, S. C., Holt, K. E. & Inouye, M. Correcting index databases improves metagenomic studies. *bioRxiv* 712166 (2019). doi:10.1101/712166
50. Molania, R., Gagnon-Bartsch, J. A., Dobrovic, A. & Speed, T. P. A new normalization for Nanostring nCounter gene expression data. *Nucleic Acids Res.* **47**, 6073–6083 (2019).
51. Dunn, P. K. & Smyth, G. K. Randomized Quantile Residuals. *J. Comput. Graph. Stat.* **5**, 236 (1996).
52. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
53. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinforma.* **2**, (2020).
54. Calle, M. L. Statistical Analysis of Metagenomics Data. *Genomics Inform.* **17**, e6 (2019).
55. Aitchison, J. *The Statistical Analysis of Compositional Data*. (Chapman and Hall, 1986).
56. Quinn, T. P., Erb, I., Richardson, M. F. & Crowley, T. M. Understanding sequencing data as compositions: An outlook and review. *Bioinformatics* **34**, 2870–2878 (2018).
57. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
58. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gks042
59. Gandolfo, L. C. & Speed, T. P. RLE plots: Visualizing unwanted variation in high dimensional data. *PLoS One* **13**, e0191629 (2018).
60. Ringnér, M. What is principal component analysis? *Nature Biotechnology* **26**, 303–304 (2008).
61. Maechler, M. *et al.* Finding Groups in Data: Cluster Analysis Extended Rousseeuw et al. *Cran* (2019).
62. Signorell, A. DescTools: Tools for descriptive statistics. *R package version 0.99.38.* (2020).

63. Veall, M. R. & Zimmermann, K. F. Pseudo-R2 measures for some common limited dependent variable models. *J. Econ. Surv.* **10**, 241–259 (1996).

Authors' contributions

AS, MI, and MF conceived the original project. AS and GM supervised the experiments. AS constructed the tool highlighted in the study. SJP provided the microbiome dataset used in the study. MF performed the experiments, analyzed the data, prepared the figures, and wrote the bulk of the manuscript. GM, MI, AS, and SJP wrote and revised the manuscript. All authors read and approved the manuscript.

Acknowledgements

MI was supported by the Munz Chair of Cardiovascular Prediction and Prevention. MF was supported by a Melbourne Research Scholarship from The University of Melbourne jointly funded by the Baker Heart and Diabetes Institute. This work was supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. This study was also supported by the Victorian Government's Operational Infrastructure Support (OIS) program.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed in this manuscript are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Data Availability

The raw sequence data was deposited at the European Nucleotide Archive (ENA) under accession number PRJEB31650.

Figures

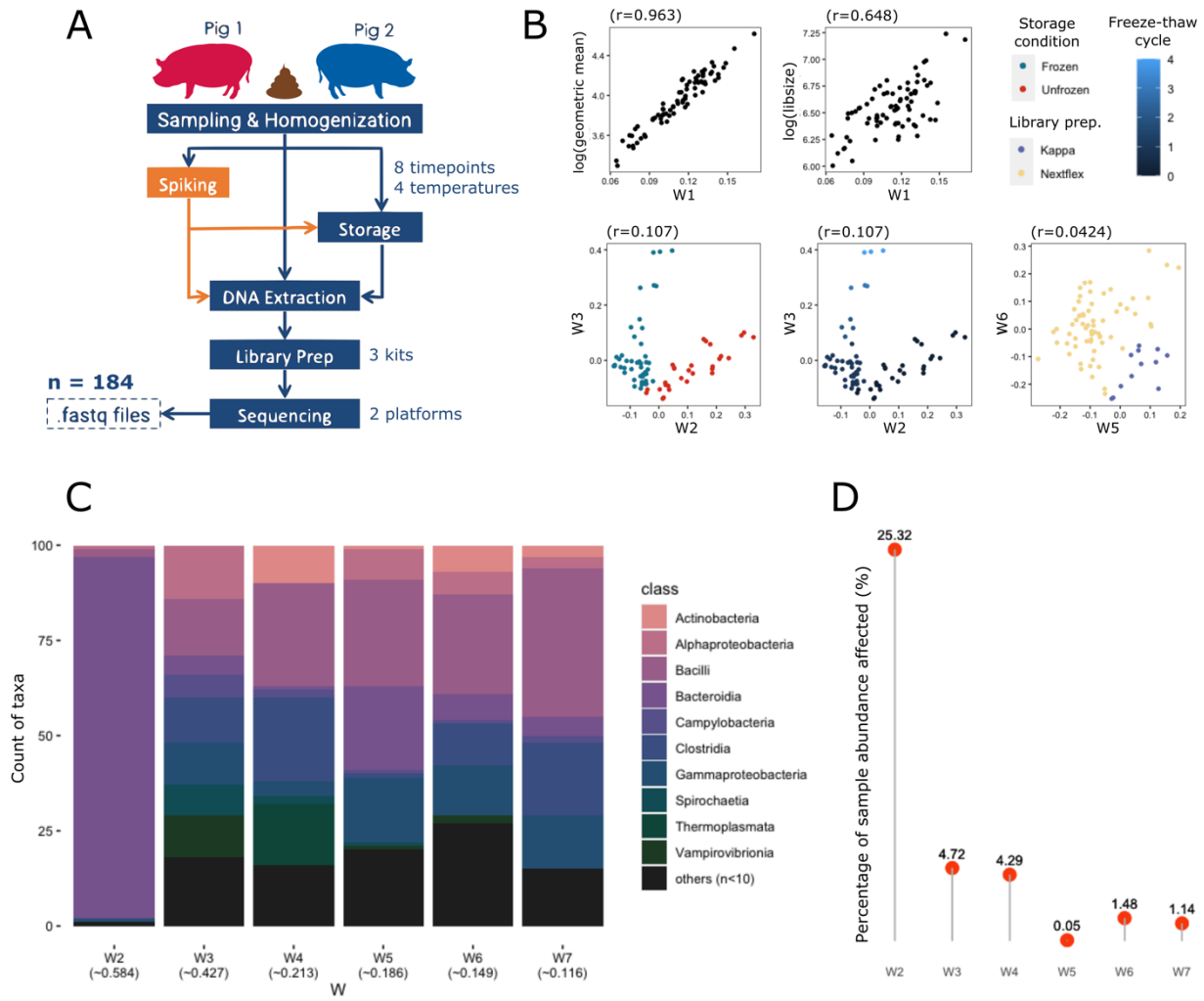

Figure 1: (A) Experimental workflow of the dataset, showing the sources of possible unwanted variations, some of which correlated with the calculated unwanted factors; **(B)** unwanted factor 1 (W1) correlated with log library size and log geometric mean, W2 correlated with storage conditions, W3 correlated with freeze-thaw cycles, and both W4 & W5 correlated with library preparation kits. **(C)** The top 100 microbial taxa most affected by each W colored by Class, ranked based on Veall-Zimmermann pseudo R-squared (mean listed under every W). Classes with less than two microbial taxa were merged as “others”. **(D)** The proportion of a sample’s total abundance affected by the top 100 microbial taxa for each W.

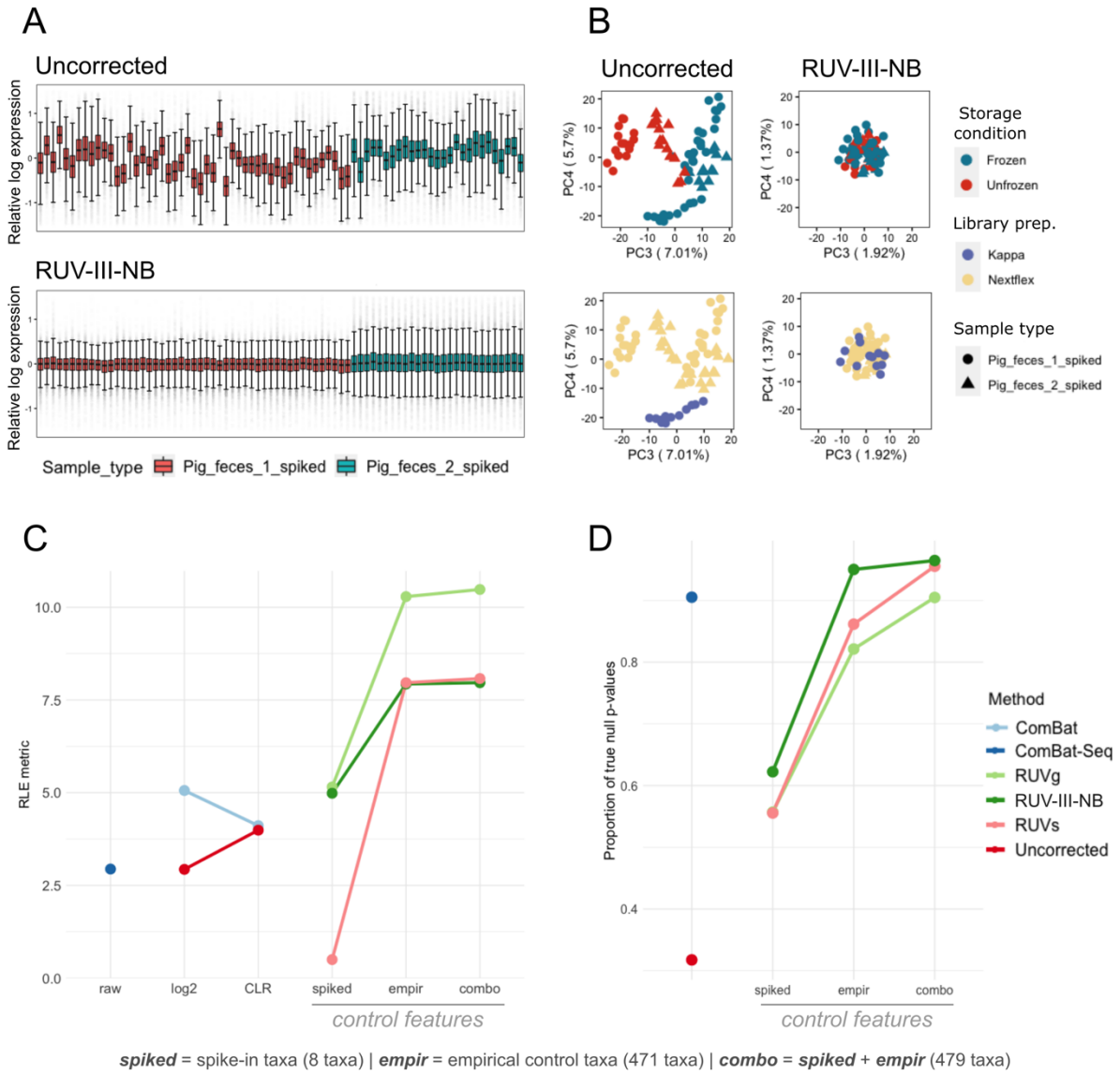

Figure 2: (A) RLE plots of all spiked samples, before correction (top) and after RUV-III-NB correction with combination control taxa (bottom). In RLE plots, successful removal of unwanted variations is represented by the medians of sample from the same individual being as close to zero as possible and as linear to each other as possible, as well as similar interquartile ranges (IQR) between the samples visualized by the size of the boxplot. **(B)** PCA plots of the same set of samples, before (left) and after RUV-III-NB correction with combination control taxa (right). Clustering based on storage conditions and library preparation kits could be explained prior to correction by PC3 and PC4, respectively – which is no longer the case after correction. The lack of visible clustering suggests successful removal of unwanted variations. **(C)** Comparison of correction method performances based on relative log expression (RLE metric), where higher number indicates better removal of overall unwanted variations. With the exception of RUVs with solely spike-in taxa as control features, RUV methods in average performed better compared than ComBat-based methods. **(D)** Proportion of true null p-values (pi0) of uncorrected and corrected data using different correction methods after differential abundance analysis. Since storage conditions were found to be the main batch variable in this dataset, comparison was done between frozen and unfrozen samples of samples from the same source (pig 1). Since edgeR requires integer counts as input, ComBat was omitted from this comparison. RUV-III-NB resulted in the highest pi0 overall when using the combination set of control taxa, and still performed better than ComBat-seq when using solely empirical taxa as control features.

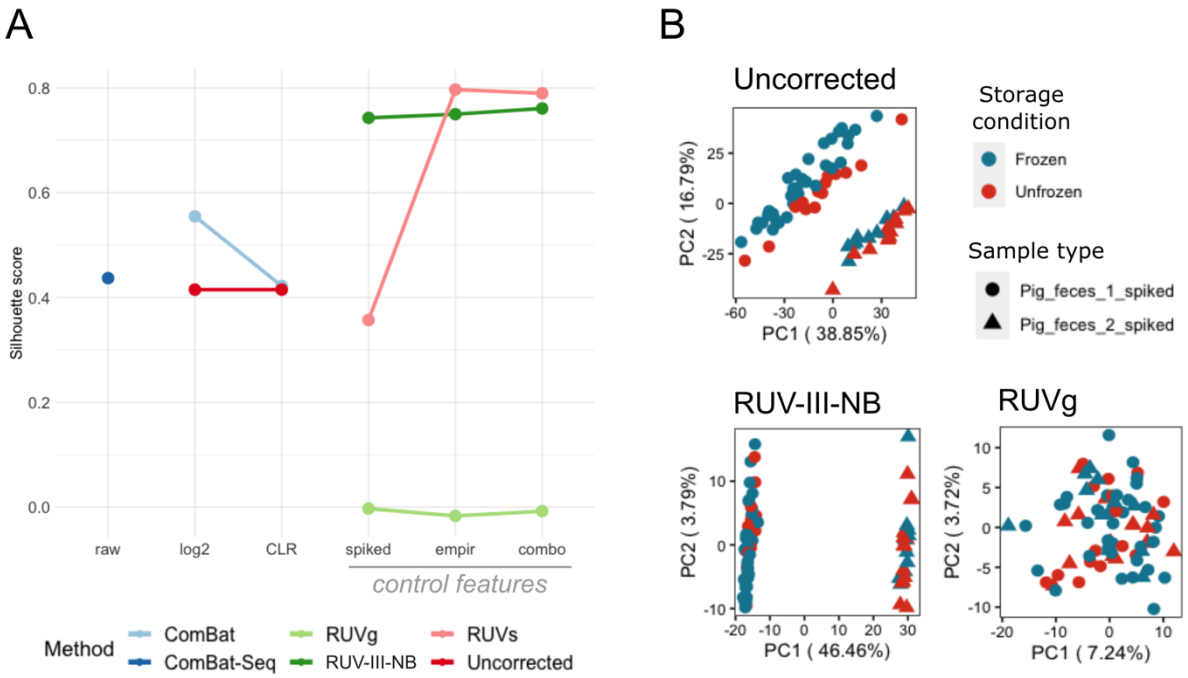

Figure 3: (A) Comparison of correction method performances in preserving individual biological information through silhouette scores, which are calculated based on the main principal components (PCs) of spiked samples and explain how well samples separate between defined groups (higher = better). RUVs using solely empirical control features had the highest silhouette score in separating pig 1 and pig 2 samples, yet correction using RUV-III-NB had the highest average silhouette score overall – indicating its consistency despite the different set of control features. RUVg performed really poorly overall, with silhouette scores using all the different control features placing significantly lower than even uncorrected data. **(B)** Based on the PCA plots of all spiked samples, separation of pig 1 and pig 2 samples can be explained by the main PCs in uncorrected data (top) and is even further defined after RUV-III-NB correction, clearly explained by PC1 (bottom-left). Separation of the two pig sources is lost after RUVg correction, as the two defined clusters are no longer visible.

Supplementary Materials

Supplementary Table

Table S1: Results of differential abundance analysis of spiked pig 1 samples between storage conditions (frozen vs unfrozen) using edgeR. Since all originated from the same individual and were spiked, little-to-no taxa should be detected as differentially abundant. Out of 8,355 taxa, our analysis with ComBat-Seq detected the lowest number of differentially abundant taxa (2) in the dataset, yet RUV-III-NB with combination control taxa resulted in the highest proportion of true null (pi0), indicating higher specificity.

Method	Variation	Number of differentially abundant taxa			Proportion of true null
		FDR<0.05, |log2(FC)| >1	p<0.01	p<0.05	
Uncorrected	-	273	3441	4418	0.317
ComBat-Seq	-	2	60	425	0.905
RUVg	spike-in control	414	1585	2451	0.556
RUVg	empirical control	67	446	1054	0.821
RUVg	combination control	42	320	848	0.904
RUVs	spike-in control	410	1602	2454	0.555
RUVs	empirical control	35	263	733	0.861
RUVs	combination control	31	215	637	0.955
RUV-III-NB	spike-in control	363	1538	2520	0.622
RUV-III-NB	empirical control	7	258	793	0.950
RUV-III-NB	combination control	14	214	708	0.964

Supplementary Figures

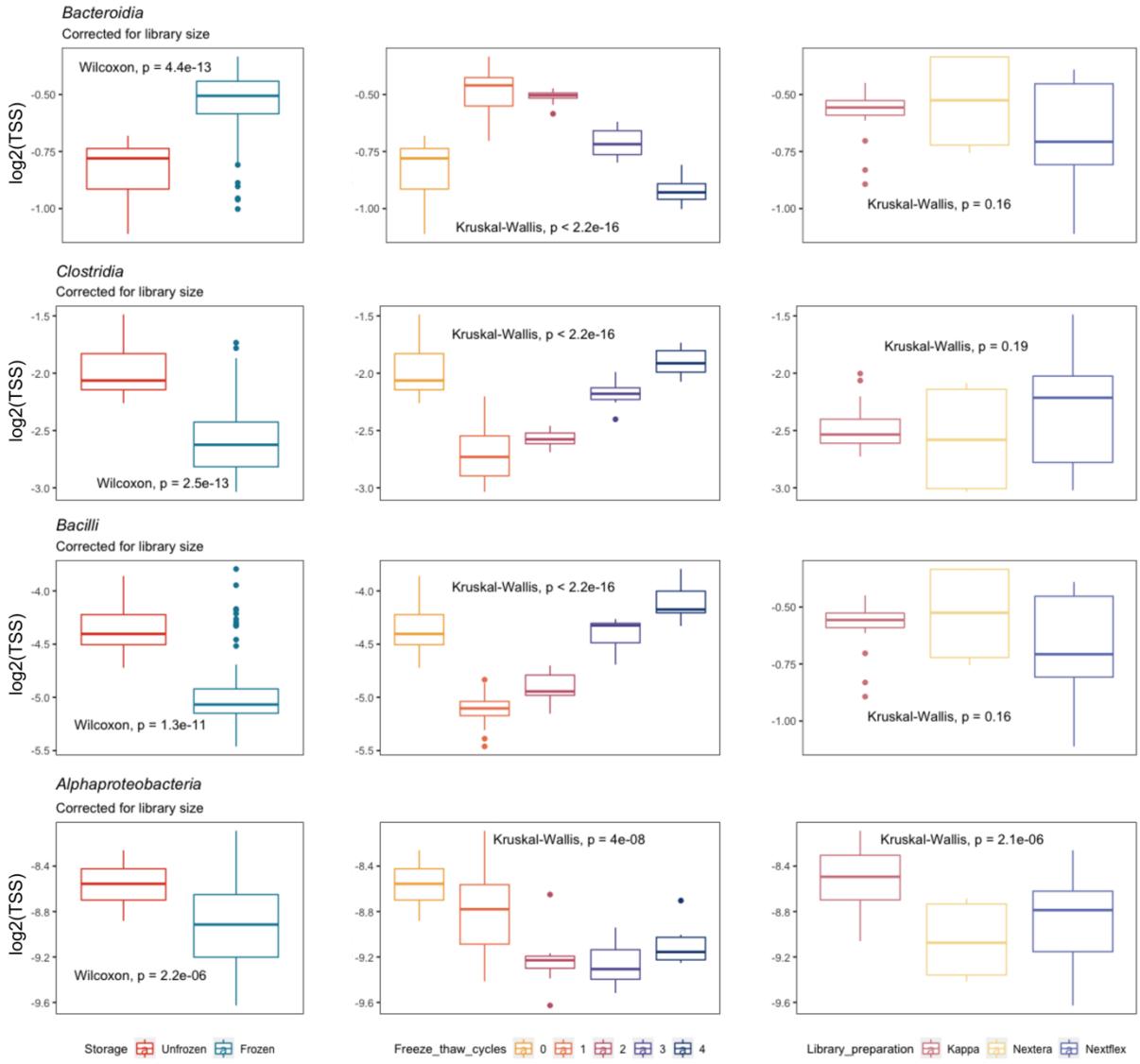

Figure S1: Differential abundance between variables of different batches (storage conditions, freeze-thaw cycles, and library preparation kits) in four of the most dominantly affected bacterial classes by unwanted variations: Bacteroidia, Clostridia, Bacilli, and Alphaproteobacteria. All classes were found to be differentially abundant between storage conditions and freeze-thaw cycles, with Alphaproteobacteria also being differentially abundant between library preparation kits.

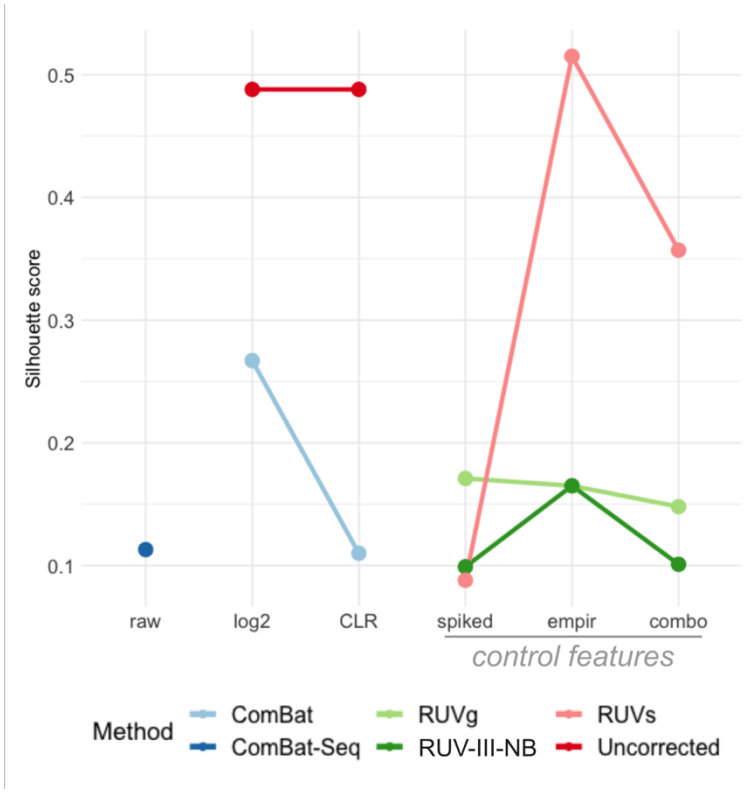

Figure S2: Comparison of correction method performances in removing unwanted variations from the main source (storage conditions). We calculated silhouette statistics based on the main principal components (PCs) of spiked samples, in which lower scores indicate poor clustering of batch variables, and therefore better removal of unwanted variations. With the exception of RUVs using only empirical control taxa (ss=0.51), all approaches had lower silhouette scores compared to uncorrected data (uncorrected ss=0.488; ComBat ss=0.11–0.26; ComBat-Seq ss=0.11; RUVg ss=0.14–0.17; RUV-III-NB ss=0.09–0.16; RUVs ss=0.08-0.51).

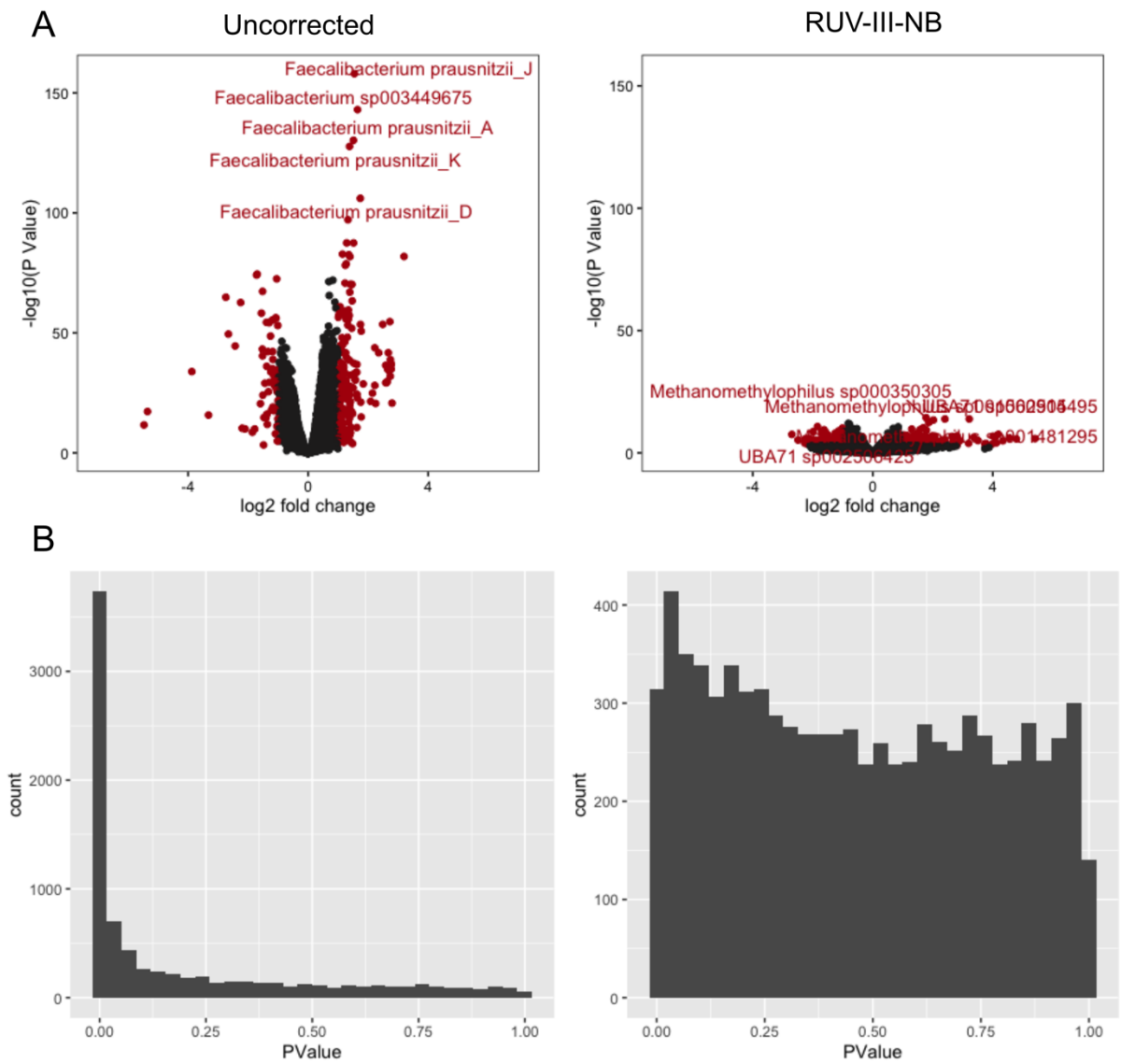

Figure S3: Results of differential abundance analysis of spiked pig 1 samples between storage conditions (frozen vs unfrozen), before (left) and after (right) correction using RUV-III-NB (with combination control features). Since the samples come from the same biological source, there should be little-to-no differences between the storage conditions. **(A)** volcano plot before correction shows over 200 significantly differentially abundant taxa (FDR<0.05, |log2(FC)|>1), with only 14 differentially abundant taxa after RUV-III-NB correction. **(B)** The distribution of p-values before correction is anti-conservative, noted by the peak on the left due to all the differentially abundant taxa. A more uniform p-value distribution can be seen after RUV-III-NB correction, suggesting minimum number of differentially abundant taxa.