Monitoring the microbiome for food safety and quality using deep shotgun sequencing

Running Title: Monitoring the microbiome for food safety and quality

Kristen L. Beck[1,8*+], Niina Haiminen[2,8+], David Chambliss[1,8], Stefan Edlund[1,8], Mark Kunitomi[1,8], B. Carol Huang[3,8], Nguyet Kong[3,8], Balasubramanian Ganesan[4,5,8], Robert Baker[4,8], Peter Markwell[4,8], Ban Kawas[1,8], Matthew Davis[1,8], Robert J. Prill[1,8], Harsha Krishnareddy[1,8], Ed Seabolt[1,8], Carl H. Marlowe[6,8], Sophie Pierre[7,8], André Quintanar[7,8], Laxmi Parida[2,8], Geraud Dubois[1,8], James Kaufman[1,8], and Bart C. Weimer[3,8*]

[*]Co-corresponding authors

[+]Contributed equally

Contact information: Kristen L. Beck, IBM Almaden Research Center, 650 Harry Road, San Jose CA, 95120 USA, klbeck@us.ibm.com, +1 408-927-1963

AUTHOR AFFILIATIONS:

[1]IBM Almaden Research Center, San Jose CA

[2]IBM T.J. Watson Research Center, Yorktown Heights, NY

[3]University of California Davis, School of Veterinary Medicine, 100K Pathogen Genome Project, Davis, CA 95616

[4]Mars Global Food Safety Center, Beijing, China

[5]Wisdom Health, A Division of Mars Petcare, Vancouver WA.

[6]Bio-Rad Laboratories, Hercules CA

[7]Bio-Rad, Food Science Division, MArnes-La-Coquette, France

[8]Consortium for Sequencing the Food Supply Chain, San Jose, CA

ABSTRACT:

In this work, we hypothesized that shifts in the food microbiome can be used as an indicator of unexpected contaminants or environmental changes. To test this hypothesis, we sequenced total RNA of 31 high protein powder (HPP) samples of poultry meal pet food ingredients. We developed a microbiome analysis pipeline employing a key eukaryotic matrix filtering step that improved microbe detection specificity to >99.96% during *in silico* validation. The pipeline identified 119 microbial genera per HPP sample on average with 65 genera present in all samples. The most abundant of these were *Bacteroides, Clostridium, Lactococcus, Aeromonas,* and *Citrobacter.* We also observed shifts in the microbial community corresponding to ingredient composition differences. When comparing culture-based results for *Salmonella* with total RNA sequencing, we found that *Salmonella* growth did not correlate with multiple sequence analyses. We conclude that microbiome sequencing is useful to characterize complex food microbial communities, while additional work is required for predicting specific species' viability from total RNA sequencing.

KEYWORDS:

microbiome, food safety, bioinformatics, shotgun sequencing, microbial ecology, pathogens

1. INTRODUCTION:

Sequencing the microbiome of food may reveal characteristics about the associated microbial content that culturing or targeted whole genome sequencing alone cannot. However, to meet the various needs of food safety and quality, next generation sequencing (NGS) and analysis techniques require additional development[1] with specific consideration for accuracy, speed, and applicability across the supply chain.[2] Microbial communities and their characteristics have been studied in relation to flavor and quality in fermented foods,[3–5] agricultural processes in grape[6] and apple fruit[7], and manufacturing processes and production batches in Cheddar cheese.[8] However, the advantage of using the microbiome specifically for food safety and quality has yet to be demonstrated.

Currently, food safety regulatory agencies including the Food and Drug Administration (FDA), Centers for Disease Control and Prevention (CDC), United States Department of Agriculture (USDA), and European Food Safety Authority (EFSA) are converging on the use of whole genome sequencing (WGS) for pathogen detection and outbreak investigation. Large scale WGS of food-associated bacteria was first initiated via the 100K Pathogen Genome Project[9] with the goal of expanding the diversity of bacterial reference genomes— a crucial need for foodborne illness outbreak investigation, traceability, and microbiome studies.[10,11] However, since WGS relies on culturing a microbial isolate prior to sequencing, there are inherent biases and limitations in its ability to describe the microorganisms and their interactions in a food sample. Such information would be very valuable for food safety and quality applications.

High throughput sequencing of total DNA and total RNA are promising approaches to characterize microbial niches in their native state without introducing bias due to culturing.[12–14] Additionally, total RNA sequencing has the potential to provide evidence of live and biologically active components of the sample.[14,15] It also provides accurate microbial naming, relative

microbial abundance, and better reproducibility than total DNA or amplicon sequencing.[14] Total RNA sequencing minimizes PCR amplification bias that occurs in single gene amplicon sequencing and overcomes the decreased detection sensitivity from using DNA sequencing in metagenomics.[14] Total RNA metatranscriptome sequencing, however, is yet to be examined in raw food ingredients as a method to provide robust characterization of the microbial communities and the interacting population dynamics.

From a single sequenced food microbiome, numerous dimensions of the sample can be characterized that may yield important indicators of safety and quality. Using total DNA or RNA, evidence for the eukaryotic food matrix can be examined. In Haiminen *et al*.,[16] we quantitatively demonstrated the utility of metagenome sequencing to authenticate the composition of complex food matrices. In addition, from total DNA or RNA, one can observe signatures from commensal microbes, pathogenic microbes, and genetic information for functional potential (from DNA) or biologically active function (from RNA).[14,15] Detecting active transcription from live microbes in food is very important to avoid spurious microbial observations that may instead be false positives due to quiescent DNA in the sample. Use of RNA in food analytics also offers the opportunity to examine expression of metabolic processes that are related to antibiotic resistance,[17,18] virulence factors, or replication genes, among others. Additionally, it has the potential to define viable microbes that are capable of replication in the food and even microorganisms that stop replicating but continue to produce metabolic activity that changes food quality and safety.[19–24]

Microorganisms are sensitive to changes in temperature, salinity, pH, oxygen content, and many other physicochemical factors that alter their ability to grow, persist, and cause disease. They exist in dynamic communities that change in response to environmental perturbation – just as the gut microbiome shifts in response to diet.[25–28] Shifts in microbiome composition or activity can be leveraged in the application of microbiome characterization to monitor the food supply chain. For

example, Noyes et al. followed the microbiome of cattle from the feed lot to the food packaging, concluding that the microbial community and antibiotic resistance characteristics change based on the processing stage.[17,18,29] We hypothesize that observable shifts in microbial communities of food can serve as an indicator of food quality and safety.

In this work, we examined 31 high protein powder samples (HPP; derived from poultry meal). HPP are commonly used raw materials in pet foods. They are subject to microbial growth prior to preparation and continued survival in powder form.[30] We subjected the HPP samples to deep total RNA sequencing with ~300 million reads per sample. In order to process the 31 samples collected over ~1.5 years from two suppliers at a single location, we defined and calibrated the appropriate methods– from sample preparation to bioinformatic analysis– needed to taxonomically identify the community members present and to detect key features of microbial growth. First, we removed the HPP's food matrix RNA content as eukaryotic background with an important bioinformatic filtering step designed specifically for food analysis. The remaining sequences were used for relative quantification of microbiome members and for identifying shifts based on food matrix content, production source, and *Salmonella* culturability. This work demonstrates that total RNA sequencing is a robust approach for monitoring the food microbiome for use in food safety and quality applications, while additional work is required for predicting pathogen viability.

2. RESULTS:

2.1 Evaluation of microbial identification capability in total RNA and DNA sequencing

Microbial identification in microbiomes often leverages shotgun DNA sequencing; however, total RNA sequencing can provide additional information about viable bacterial activity in a community via transcriptional activity. Since using total RNA to study food microbiomes is novel, each step of the analysis workflow (Figure 1) was carefully designed and scrutinized for accuracy.

For all analyses done in this study, we report relative abundance in reads per million (RPM) (Equation 1) as recommended by Gloor et al[31,32] and apply the conservative threshold of RPM > 0.1 to indicate presence as previously described by Langelier et al and Illot et al.[33,34] Numerically, this threshold translates to ~30 reads per genus per sample considering a sequencing depth of ~300 million reads per sample (Methods Section 4.4). First, we examined the effectiveness of RNA for taxonomic identification and relative quantification of microbes in the presence of food matrix reads. We observed that RNA sequencing results correlated (R^2 = 0.93) with the genus relative quantification provided by DNA sequencing (Supplementary Figure S1). RNA sequencing also detected more genera demonstrated by a higher α-diversity than the use of DNA (Supplementary Figure S2). Additionally, from the same starting material, total RNA sequencing resulted in 2.4-fold more reads classified to microbial genera compared to total DNA sequencing (after normalizing for sequencing depth). This increase is substantial as microbial reads are such a small fraction of the total sequenced reads. Considering these results, we further examined the microbial content from total RNA extracted from 31 high protein powder (HPP) samples (Supplementary Table 1) that resulted in an average of ~300 million paired end 150 bp sequencing reads per sample in this study.

2.2 Evaluation and application of *in silico* filtering of eukaryotic food matrix reads

Sequenced reads from the eukaryotic host or food matrix may lead to false positives for microbial identification in microbiome studies.[35] This may occur partly due to reads originating from low complexity regions of eukaryotic genomes, e.g. telomeric and centromeric repeats, being misclassified as spurious microbial hits.[36] In total DNA or RNA sequencing of clinical or animal or even plant microbiomes, eukaryotic content may often comprise > 90% of the total sequencing reads. This presents an important bioinformatic challenge that we addressed by filtering matrix

content using a custom-built reference database of 31 common food ingredient and contaminant genomes (Supplementary Table 2) using the *k*-mer classification tool Kraken.[37] This step allows for rapidly classifying all sequenced reads (~300 million reads for each of 31 samples) as matrix or non-matrix. The matrix filtering process yielded an estimate of the total percent matrix content for a sample. See our work in Haiminen et al.[38] on quantifying the eukaryotic food matrix components with further precision.

To validate the matrix filtering step, we constructed *in silico* mock food microbiomes with a high proportion of complex food matrix content and low microbial content (Supplementary Table 3). We then computed the true positive, false positive, and false negative rates of observed microbial genera and sequenced reads (Table 1). False positive viral, archaeal, and eukaryotic microbial genera (as well as bacteria) were observed without matrix filtering, although bacteria were the only microbes included in the simulated mixtures. Introducing a matrix filtering step to the pipeline improved read classification specificity to >99.96% (from 78–93% without filtering) in both simulated food mixtures, while maintaining zero false negatives. With this level of demonstrated accuracy, we used bioinformatic matrix filtering prior to further microbiome analysis.

2.3 High protein powder microbiome ecology

After filtering eukaryotic matrix sequences, we applied the remaining steps in the bioinformatic workflow (Figure 1) to examine the shift in the high protein powder (HPP) microbiome membership and to quantify the relative abundance of microbes at the genus level. Genus is the first informative taxonomic rank for food pathogen identification that can be considered accurate given current incompleteness of reference databases[11,39–42] and was therefore used in subsequent analyses. Overall, between 98 and 195 microbial genera (avg. 119) were

identified (RPM > 0.1) per HPP sample (Supplementary Table 4). When analyzing α-diversity i.e. the number of microbes detected per sample, inter-sample comparisons may become skewed unless a common number of reads is considered since deeper sequenced samples may contain more observed genera merely due to a greater sampling depth.[43,44] Thus, we utilized bioinformatic rarefaction i.e. subsampling analysis to showcase how microbial diversity was altered by sequencing depth. Examination of α-diversity across a range of *in silico* subsampled sequencing depths showed that the community diversity varied across samples (Figure 2A). One sample (MFMB-04) had 1.7 times more genera (195) than the average across other samples (avg. 116, range 98–143) and exhibited higher α-diversity than any other sample at each *in silico* sampled sequencing depth (Figure 2A). Rarefaction analysis further demonstrated that when considering fewer than ~67 million sequenced reads, the observable microbial population was not saturated (median elbow calculated as indicated in Satopää, et al.[45]). This observation suggests that deeper sequencing or more selective sequencing of the HPP microbiomes will reveal more microbial diversity.

Notably, between 2%–4% (approximately 5,000,000–14,000,000) of reads per sample remained unclassified as either eukaryotic matrix or microbe (Supplementary Figure S3). However, the unclassified reads exhibited a GC (guanine plus cytosine) distribution similar to reads classified as microbial (Supplementary Figure S4) indicating these reads may represent microbial content that is absent or sufficiently divergent from existing references.

We calculated β-diversity to study inter-sample microbiome differences and to identify any potential outliers among the sample collection. The Aitchison distances[46] of microbial relative abundances were calculated between samples (as recommended for compositional microbiome data[31,32]), and the samples were hierarchically clustered based on the resulting distances (Figure

2B). The two primary clades were mostly defined by supplier (except for MFMB-17). In Haiminen *et al*.,[38] we reported that three of the HPP samples contained unexpected eukaryotic species. We hypothesized that the presence of these contaminating matrix components (beef identifiable as *Bos taurus* and pork identifiable as *Sus scrofa*) would alter the microbiome as compared to chicken (identifiable as *Gallus gallus*) alone. Clustering HPP samples using their microbiome membership led to a distinctly different group of the matrix-contaminated samples, supporting this hypothesis (Figure 2B). These observations indicate that samples can be discriminated based on their microbiome content for originating source and supplier, which is necessary for source tracking potential hazards in food.

2.4 Comparative analysis of high protein powder microbiome membership and composition

We identified 65 genera present in all HPP samples (Figure 3A), whose combined abundance accounted for between 88-99% of the total abundances of detected genera per sample. *Bacteroides, Clostridium, Lactococcus, Aeromonas,* and *Citrobacter* were the five most abundant of these microbial genera. The identified microbial genera also included viruses, the most abundant of which was *Gyrovirus* (< 10 RPM per sample). *Gyrovirus* represents a genus of non-enveloped DNA viruses responsible for chicken anemia which is ubiquitous in poultry. While there were only 65 microbial genera identified in all 31 HPP samples, the α-diversity per sample was on average two-fold greater as previously indicated.

Beyond the collection of 65 microbes observed in all samples, there were an additional 164 microbes present in various HPP samples. Together, we identified a total of 229 genera among the 31 HPP samples tested (Figure 3B and 4, Supplementary Table 4). In order to identify genera that were most variable between samples, we computed the median absolute deviation (MAD)[47] using the normalized relative abundance of each microbe (Figure 5A). The abundance of *Bacteroides*

was the most variable among samples (median = 148.1 RPM, MAD = 30.6) and showed increased abundance in almost all samples from Supplier A compared to Supplier B (abundance for the 10 most variable genera shown in Figure 5B). *Clostridium* (median = 37.4 RPM, MAD = 24.2), *Lactococcus* (median = 36.8 RPM, MAD = 18.2), and *Lactobacillus* (median = 24.2, MAD = 7.2) were also highly variable and 3–4 fold more abundant in samples MFMB-04 and MFMB-20 compared to other samples (Figure 5B). *Pseudomonas* (median = 11.1 RPM, MAD = 12.2) was markedly more abundant in MFMB-83 than any other sample (Figure 5B). These genera highlight variability between microbiomes from a single food source and may provide insights into other dissimilarities in these samples.

2.5. Microbiome shifts in response to changes in food matrix composition

We tested the hypothesis that the microbiome composition will shift in response to changes in the food matrix and can be a unique signal to indicate contamination or adulteration. In 28 of the 31 HPP samples, >99% of the matrix reads were determined in our related work[38] to originate from poultry (*Gallus gallus*), which was the only ingredient expected based on ingredient specifications. However, three samples had higher pork and beef content compared to all other HPP samples: MFMB-04 (7.74% pork, 8.99% beef), MFMB-20 (0.53% pork, 1.00% beef), and MFMB-38 (0.92% pork, 0.29% beef) compared to the highest pork (0.01%) and beef (0.00%) content among the other 28 HPP samples (Supplementary Data by Haiminen *et al.*[38]). The microbiomes of these matrix contaminated samples also clustered into a separate sub-cluster (Figure 2B). This demonstrated that a shift in the food matrix composition was associated with an observable shift in the food microbiome.

MFMB-04 and MFMB-20 had the highest percentage of microbial reads compared to other samples (Supplementary Figure S3). They also exhibited an increase in *Lactococcus, Lactobacillus*, and *Streptococcus* relative abundances compared to other samples (Figure 5B), also reflected at respective higher taxonomic levels above genus (Supplementary Figure S5).

There were 53 genera identified uniquely in MFMB-04 and/or MFMB-20, but not present in any other sample. (MFMB-38 had a very low microbial load and contributed no uniquely identified genera above the abundance threshold.) MFMB-04 contained 44 unique genera (Figure 4) with the most abundant being *Macrococcus* (35.8 RPM), *Psychrobacter* (23.8 RPM), and *Brevibacterium* (18.1 RPM). Additionally, *Paenalcaligenes* was present only in MFMB-04 and MFMB-20 with an RPM of 6.4 and 0.3, respectively, compared to a median RPM of 0.004 among other samples. Notable differences in the matrix-contaminated samples' unique microbial community membership compared to other samples may provide microbial indicators associated with unanticipated pork or beef presence.

2.6. Genus level identification of foodborne microbes

We evaluated the ability of total RNA sequencing to identify genera of commonly known foodborne pathogens within the microbiome. We focused on fourteen pathogen-containing genera including *Aeromonas, Bacillus, Campylobacter, Clostridium, Corynebacterium, Cronobacter, Escherichia, Helicobacter, Listeria, Salmonella, Shigella, Staphylococcus, Vibrio*, and *Yersinia* that were found to be present in the HPP samples with varying relative abundances. Of these genera, *Aeromonas*, *Bacillus*, *Campylobacter*, *Clostridium*, *Corynebacterium*, *Escherichia*, *Salmonella*, and *Staphylococcus* were detected in every HPP with median abundance values between 0.58–48.31 RPM (Figure 6A). This indicated that a baseline fraction of reads can be attributed to foodborne microbes when using NGS. Of those genera appearing in all samples, there

261 was observed sample-to-sample variation in their abundance with some genera exhibiting longer
262 tails of high abundance, e.g. *Staphylococcus* and *Salmonella*, whereas others exhibit very low
263 abundance barely above the threshold of detection, e.g. *Bacillus* and *Yersinia* (Figure 6A). None
264 of the pathogen-containing genera were consistent with higher relative abundances due to
265 differences in food matrix composition. *Bacillus* and *Corynebacterium* exhibited slightly higher
266 relative abundances in sample MFMB-04 which contained 7.7% pork and 9.0% beef (Figure 6B).
267 Yet while MFMB-04 contained higher cumulative levels of these foodborne microbes, the next
268 highest sample was MFMB-93 which was not associated with altered matrix composition, and
269 both MFMB-04 and MFMB-93 contained higher levels of *Staphylococcus* (Figure 6B). Thus,
270 matrix composition alone did not explain variations of these pathogen-containing genera.

271 Interestingly, low to moderate levels of *Salmonella* were detected within all 31 HPP
272 microbiomes (Figure 6A). The presence of *Salmonella* in HPP is expected but the viability of
273 *Salmonella* is an important indicator of safety and quality. Thus, we further sought to delineate
274 *Salmonella* growth capability within these microbiomes by comparing culturability with multiple
275 established bioinformatic NGS methods for *Salmonella* relative abundances in the samples.

276 **2.7 Assessment of *Salmonella* culturability and total RNA sequencing**

277 Total RNA sequencing of food microbiomes has the potential to provide additional
278 sensitivity beyond standard culture-based food safety testing to confirm or reject the presence of
279 potentially pathogenic microbes. In all of the examined HPP samples, some portion of the
280 sequenced reads were classified as belonging to pathogen-containing genera (Figure 6); however,
281 the presence of RNA transcripts does not necessarily indicate current growth of the organism itself.
282 We further inspected one pathogen of interest, *Salmonella*, to determine the congruence between
283 sequencing-based and culturability results. Of the 31 samples examined with total RNA

284 sequencing, *Salmonella* culture testing was applied to 27 samples, of which four were culture-
285 positive. Surprisingly, *Salmonella* culture-positive samples were not among those with the highest
286 relative abundance of *Salmonella* from sequencing (Figure 7A). When ranking the samples by
287 decreasing *Salmonella* abundance, the culture-positive samples were not enriched for higher ranks
288 (p=0.86 from Wilcoxon rank sum test indicating that the distributions are not significantly
289 different, Table 2). To confirm that the microbiome analysis pipeline did not miss *Salmonella* reads
290 present, we completed two orthogonal analyses on the same data set used in the microbial
291 identification step. The reference genomes relevant to these additional analyses were publicly
292 available and closed high quality genomes available from the sources indicated below.
293 First, for a targeted analysis, we aligned the sequenced reads using a different tool, Bowtie
294 2,[48] to an augmented *Salmonella*-only reference database. This reference was comprised of the 264
295 *Salmonella* genomes extracted from NCBI RefSeq Complete (used in our previous microbial
296 identification step) as well as an additional 1,183 public *Salmonella* genomes which represent
297 global diversity within the genus.[49] The number of reads that aligned to the *Salmonella*-only
298 reference was on average 370-fold higher than identified as *Salmonella* by Kraken using the multi-
299 microbe NCBI RefSeq Complete. In this additional analysis, the culture-positive samples had
300 overall higher ranks compared to culture-negative samples (p=0.06, Table 2) indicating that
301 additional *Salmonella* genomic data in the reference significantly improved discriminatory
302 identification power. *Salmonella* culture-positive samples were still not the most abundant (Figure
303 7B), but with an enriched database, sequencing positioned all four culturable samples within the
304 top 10 ranking.
305 The second additional analysis examined alignment of the reads to a specific gene
306 required[50] for replication and protein production in actively dividing *Salmonella*— elongation
307 factor Tu (*ef-Tu*). This was done by aligning the reads to 4,846 gene sequences for *ef-Tu* extracted

for a larger corpus of *Salmonella* genomes from OMXWare.[51] The relative abundances of this transcript in culture-positive samples were still comparable to culture-negative samples (Figure 7C). Culture-positive samples did not exhibit higher ranks compared to culture-negative samples (p=0.56, Table 2), indicating that *ef-Tu* relative abundance alone was not sufficient to improve the lack of concordance in culturability vs sequencing. These two orthogonal analyses demonstrated that results from carefully developed culture-based testing and those from current high-throughput sequencing technologies, whether assessed at overall reads aligned or specific gene abundances, were not conclusively in agreement when detecting active *Salmonella* in food samples (Figure 7 and Table 2). However, the use of a reference database enriched in whole genome sequences of the specific organism of interested was found appropriate for food safety applications.

Since microbes compete for available resources within an environmental niche and therefore impact one another,[52] we investigated *Salmonella* culture results in conjunction with co-occurrence patterns of other microbes in the total RNA sequencing data (Figure 8). Point-biserial correlation coefficients (r_{pb}) were calculated between *Salmonella* culturability results (presence or absence which were available for 27 of the 31 samples) and microbiome relative abundance. We observed 31 genera that positively correlated and with *Salmonella* presence ($r_{pb} > 0.5$). *Erysipelothrix*, *Lactobacillus*, *Anaerococcus*, *Brachyspira*, and *Jeotgalibaca* exhibited the largest positive correlations. *Gyrovirus* was negatively correlated with *Salmonella* growth ($r_{pb} = -0.54$). In three of the four *Salmonella*-positive samples (MFMB-04, MFMB-20, and MFMB-38), food matrix contamination was also observed (Supplementary Data in Haiminen *et al.*[38]). The concurrency of *Salmonella* growth and matrix contamination was affirmed by the microbial co-occurrence (specifically *Erysipelothrix, Brachyspira,* and *Gyrovirus*). This highlights the complex dynamic and community co-dependency of food microbiomes, yet shows that multiple dimensions

of the data (microbiome composition, culture-based methods, and microbial load) will signal anomalies from typical samples when there is an issue in the supply chain.

3. DISCUSSION:

Accurate and appropriate tests for detecting potential hazards in the food supply chain are key to ensuring consumer safety and food quality. Monitoring and regular testing of raw ingredients can reveal fluctuations within the supply chain that may be an indicator of an ingredient's quality or of a potential hazard. Such quality is assessed by standardized tests for chemical and microbial composition to meet legal requirements and specifications from government agencies throughout the world. For raw materials or finished products to meet these bounds of safety and quality, their composition must usually have a low microbiological load (except in fermented foods) and be chemically identical in macro-components such as carbohydrate, protein, and fat. Methods in this space must avoid false negative results which could endanger consumers, while also minimizing false positives which could lead to unnecessary recalls and food loss.

Existing microbial detection technologies used in food safety today such as pulse field gel electrophoresis (PFGE) and whole genome sequencing (WGS) require microbial isolation. This provides biased outcomes as it removes microbes from their native environment where other biotic members also subsist, and selects microbes by culturability alone. Amplicon sequencing, while a low-cost alternative to metagenome or metatranscriptome sequencing for bacteria, also imparts PCR amplification bias and reduces detection sensitivity due to reliance on a single gene (16S ribosomal RNA).[14,53,54] We therefore investigated the utility of total RNA sequencing of food microbiomes and demonstrated that from this single test, we are able to yield several pertinent results about food safety and quality.

For this evaluation, we developed a pipeline to characterize the microbiome of typical food ingredient samples and to detect potentially hazardous outliers. Special considerations for food samples were made as computational pipelines for human or other microbiome analyses are not sufficient for applications in food safety without modification. In food, the eukaryotic matrix needs to be confirmed, may be mixed, and, as we and others have shown, affects the identification accuracy of microbes that are present.[35,36] By filtering food matrix sequence data properly, we avoid incorrect microbial identification and characterization of the microbiome[36] while also increasing the computational efficiency for downstream processing. The addition of this filtering step in the pipeline removed approximately 90% of false positive genera and provided results at 99.96% specificity when evaluating simulated mixtures of food matrix and microbes (Table 1).

Through the analysis of 31 high protein powder total RNA sequencing samples, we demonstrated the pipeline's ability to characterize food microbiomes and indicate outliers. In this sample collection, we identified a core catalog of 65 microbial genera found in all samples where *Bacteroides*, *Clostridium*, and *Lactococcus* were the most abundant (Supplementary Table 4). We also demonstrated that in these food microbiomes the overall diversity was 2-fold greater than the core microbe set. Fluctuations in the microbiome can indicate important differences between samples as observed here, as well as in the literature for grape berry[6] and apple fruit microbiomes (pertaining to organic versus conventional farming)[7] or indicate inherent variability between production batches or suppliers as observed here and during cheddar cheese manufacturing.[8] Specifically, we observed a shift in the microbial composition (Figure 2B) and the microbial load (Supplementary Figure S3) in high protein powder samples (derived from poultry meal) where unexpected pork and beef were observed. Matrix-contaminated samples were marked by increased relative abundances of specific microbes including *Lactococcus*, *Lactobacillus*, and *Streptococcus* (Figure 5B). This work shows that the microbiome shifts with observed food matrix contamination

from sources with similar macronutrient content and thus, the microbiome alone is a likely signal of compositional change in food.

Beyond shifts in the microbiome, we focused on a set of well-defined foodborne-pathogen containing genera and explored their relative abundances observed from total RNA sequencing. Of these genera, *Aeromonas*, *Bacillus*, *Campylobacter*, *Clostridium*, *Corynebacterium*, *Escherichia*, *Salmonella*, and *Staphylococcus* were detected in every HPP sample. This highlights that when using NGS there may be an observable baseline of sequences assigned to potentially pathogenic microbes. For this ingredient type, this result lends a range of normalcy of relative abundance generated by NGS. Further work is needed to establish a definitive and quantitative range of typical variation in samples of a particular food source and the degree of anomaly for a new sample or genus abundance. However, preliminary studies of this nature can inform the development of guidelines when working with increasingly sensitive shotgun metagenomic or metatranscriptomic analysis.

Furthermore, sequenced DNA or RNA alone does not imply microbial viability. Therefore, we investigated the relatedness of culture-based tests and total RNA sequencing for the pathogenic bacterium *Salmonella* in the high protein powder samples. As has been reported for human gut[55] and deep sea[56] microbiomes, we also did not dretect a correlation between *Salmonella* read abundance and culturability (Figure 7 and Table 2). Sequence reads matching *Salmonella* references were observed for all samples (both culture-positive and culture-negative) as determined by multiple analysis techniques: microbiome classification, alignment to *Salmonella* genomes, and targeted growth gene analysis. When ranking the high protein powder samples based on *Salmonella* abundance from whole genome alignments, the culture-positive samples were enriched for higher ranks ($p = 0.06$). However, the culture-positive samples were still intermixed in ranking with culture-negative samples. This indicated that there was no clear minimum

threshold of sequence data as evidence for culturability and that this analysis alone is not predictive of pathogen growth. One possible reason for this is that the culture-positive variant of *Salmonella* is missing from existing reference data sets. Potentially, *Salmonella* attained a nonculturable state wherein it was detected by sequencing techniques yet remained nonculturable from the HPP sources. Successful isolation of total RNA and DNA and gene expression analysis from experimentally known nonculturable bacteria has been demonstrated by Ganesan *et al*. in multiple studies in other genera.[19,22] Physiological state should thus be taken under consideration when benchmarking sequencing technologies in comparison with culture-based methods. Thus, total RNA sequencing of food samples may identify shifts that standard food testing does not, but the incongruity between sequencing read data and culture-based results highlights the need to perform more benchmarking in food microbiome analysis for pathogen detection.

The characterization of HPP food microbiomes leveraged current accepted public reference databases, yet it is known that these databases are still inadequate.[1,2,11,57,58] Furthermore, when considering congruence between *Salmonella* culturability and NGS read mapping techniques, the genetic breadth and depth of multi-genome reference sequences is essential. For example, focusing on *ef-Tu,* a known marker gene for *Salmonella* growth, was not sufficient to mirror viability of *in vitro* culture tests. This highlights the limitations of single gene approaches for identification. When the sequenced reads were examined in the context of an augmented reference collection of *Salmonella* genomes, we observed improved ranking and read mapping rate for culture-positive samples (yet we did not achieve complete concordance). This improvement underlined the increased analytical robustness yielded from a multi-genome reference. We also recognize that the read mapping rate may be exaggerated as reads from non-*Salmonella* genomes could map to *Salmonella* in the absence of any other reference genomes. Overall for robust analysis and applicability to food safety and quality, microbial references must be expanded to include more

genetically diverse representatives of pathogenic and spoilage organisms. Description of food microbiomes will only improve as additional public sequence data is collected and leveraged.

In our sample collection, 2-4% (effectively 5 to 14 million) of reads remain unclassified. The GC content distribution of unclassified reads matched microbial GC content distribution (Supplementary Figure S4) suggesting that these reads may have been derived from microbes missing from the current reference database that have not yet been isolated or sequenced. By sequencing the microbiome, we sampled environmental niches in their native state in a culture-independent manner and therefore collected data from diverse and potentially never-before seen microbes. Tracking unclassified reads will also be essential for monitoring food microbiomes. The inability to provide a name from existing references does not eliminate the possibility that the sequence is from an unwanted microbe or indicates a hazard. In addition to tracking known microbes, quantitative or qualitative shifts in the unclassified sequences might be used to detect when a sample is different from its peers.

We demonstrated the potential utility of analyzing food microbiomes for food safety using raw ingredients. This study resulted in the detection of shifts in the microbiome composition corresponding to unexpected matrix contaminants. This signifies that the microbiome is likely an important and effective hazard indicator in the food supply chain. While we have used total RNA sequencing for detection of microbiome membership, the technology has future applicability for detection of antimicrobial resistance, virulence, and biological function for multiple food sources, and for other sample types. Notably, while this pipeline was developed for food monitoring, with applicable modifications and identification of material-specific indicators, it can be applied to other microbiomes including human and environmental.

4. METHODS:

450

451 **4.1 Sample Collection, Preparation, and Sequencing**

452 High protein powder (HPP, 2.5 kg) samples were each collected from a train car in Reno, NV,
453 USA between April 2015 and February 2016 in four batches from two suppliers and shipped to
454 the Weimer lab at the University of California, Davis (Davis, CA). Each HPP sample was
455 composed of five sub-samples from random locations within the train car prior to shipment.
456 Sample preparation, total RNA extraction and integrity confirmation, cDNA construction, and
457 library construction for these samples was previously described by Haiminen et al.[38]

458 Sequencing was performed by BGI@UC Davis (Sacramento, CA) using Illumina HiSeq
459 4000 (San Diego, CA) with 150 paired end chemistry for each sample except the following: HiSeq
460 3000 with 150 paired end chemistry was used for MFMB-04 and MFMB-17. All total RNA
461 sequencing data are available via the 100K Pathogen Genome Project BioProject (PRJNA186441)
462 at NCBI (Supplementary Table 1).

463 For evaluation of total RNA sequencing for microbial classification in paired processing
464 steps, total RNA and total DNA were extracted from the same sample and denoted as MFMB-03
465 and MFMB-08, respectively. Total RNA was extracted and sequenced as described above. Total
466 DNA was extracted and sequenced as described previously.[10,59–64] The Illumina HiSeq 2000 with
467 100 paired end chemistry was used for MFMB-03 and MFMB-08.

468

469 **4.2 Sequence Data Quality Control**

470 Illumina Universal adapters were removed and reads were trimmed using Trim Galore[65]
471 with a minimum read length parameter 50 bp. The resulting reads were filtered using Kraken[37], as
472 described below in Section 4.3, with a custom database built from the PhiX genome (NCBI
473 Reference Sequence: NC_001422.1). Removal of PhiX content is suggested as it is a common

474 contaminant in Illumina sequencing data.[66] Trimmed non-PhiX reads were used in subsequent
475 matrix filtering and microbial identification steps.
476
477 **4.3 Matrix Filtering Process and Validation**
478 Kraken[37] with a *k*-mer size of 31 bp (optimal size described in the Kraken reference
479 publication) was used to identify and remove reads that matched a pre-determined list of 31
480 common food matrix and potential contaminant eukaryotic genomes (Supplementary Table 2).
481 These food matrix organisms were chosen based on preliminary eukaryotic read alignment
482 experiments of the HPP samples as well as high-volume food components in the supply chain. Due
483 to the large size of eukaryotic genomes in the custom Kraken[37] database, a random *k*-mer reduction
484 was applied to reduce the size of the database by 58% using kraken-build with option --max-db-
485 size, in order to fit the database in 188 GB for in-memory processing. A conservative Kraken score
486 threshold of 0.1 was applied to avoid filtering microbial reads. The matrix filtering database
487 includes low complexity and repeat regions of eukaryotic genomes to capture all possible matrix
488 reads. This filtering database with the score threshold was also used in the matrix filtering *in silico*
489 testing as described below.
490 Matrix filtering was validated by constructing synthetic paired end reads (150 bp) using
491 DWGSIM[67] with mutations from reference sequences using the following parameters: base error
492 rate (e) = 0.005, outer distance between the two ends of a read pair (d) = 500, rate of mutations (r)
493 = 0.001, fraction of indels (R) = 0.15, probability an indel is extended (X) = 0.3. Reference
494 sequences are detailed in Supplementary Table 3. We constructed two *in silico* mixtures of
495 sequencing reads by randomly sampling reads from eukaryotic reference genomes. Simulated
496 Food Mixture 1 was comprised of nine species with the following number of reads per genome:
497 2M cattle, 2M salmon, 1M goat, 1M lamb, 1M tilapia (transcriptome), 962K chicken

(transcriptome), 10K duck, 1K horse, and 1K rat totaling 7.974M matrix reads. Simulated Food Mixture 2 contained 5M soybean, 4M rice, 3M potato, 2M corn, 200K rat, and 10K drain fly reads, totaling 14.210M matrix reads. Both simulated food mixtures included 1,000 microbial sequence reads generated from 15 different microbial species for a total of 15K sequence reads (Supplementary Table 3).

4.4 Microbial Identification

Remaining reads after quality control and matrix filtering were classified using Kraken[37] against a microbial database with a *k*-mer size of 31 bp to determine the microbial composition within each sample. NCBI RefSeq Complete[68] genomes were obtained for bacterial, archaeal, viral, and eukaryotic microorganisms (~7,800 genomes retrieved April 2017). Low complexity regions of the genomes were masked using Dustmasker[69] with default parameters. A threshold of 0.05 was applied to the Kraken score in an effort to maximize the F-score of the result (as demonstrated in Kraken's operating manual.[70] Taxa-specific sequence reads were used to calculate a relative abundance in reads per million (RPM; Equation 1) where R_T represents the reads classified per microbial entity (e.g. the genus *Salmonella*) and R_Q represents the number of sequenced reads remaining after quality control (trimming and PhiX removal) for an individual sample, including any reads classified as eukaryotic:

$$RPM = \frac{R_T}{R_Q} \times 1{,}000{,}000 \qquad Equation\ 1$$

This value provides a relative abundance of the microbial entity of interest and was used in comparisons of taxa among samples. Genera with a conservative threshold of RPM > 0.1 were

521 defined as present, as previously applied by others in the contexts of human infectious disease and
522 gut microbiome studies.[33,34] Pearson correlation of resulting microbial genus counts was
523 computed.
524
525 **4.5 Community Ecology Analysis**
526 Rarefaction analysis at multiple subsampled read depths R_D was performed by multiplying
527 the microbial genus read counts with R_D/R_Q and rounding the results down to the nearest integer
528 to represent observed read counts. Here R_Q is the total number of reads in the sample after quality
529 control (including microbial, matrix, and unclassified reads). Resulting α-diversity at read depth
530 R_D was computed as the number of genera with resulting RPM > 0.1 and plotted at five million
531 read intervals: R_D = 5M, 10M, 15M, …, R_Q. If, due to random sampling and rounding effects, the
532 computed α-diversity was lower than the diversity computed at any previous depth, the previous
533 higher α-diversity was used for plotting. The median elbow was calculated as previously
534 described[45] using the R package kneed.[45]
535 In compositional data analysis,[31] non-zero values are required when computing β-diversity
536 based on Aitchison distance.[46] Therefore, reads counts assigned to each genus were pseudo-
537 counted by adding one in advance of computation of RPM (Eq. 1) prior to calculating the Aitchison
538 distance for the microbial table. β-diversity was calculated using the R package robCompositions[71]
539 and hierarchical clustering was performed using base R function hclust using the "ward.D2"
540 method as recommended for compositional data analysis.[31]
541
542 **4.6 Unclassified Read Analysis**

543 The GC percent distributions of matrix (from matrix filtering), microbial, and remaining
544 unclassified reads per sample were computed using FastQC[72] and collated across samples with
545 MultiQC.[73]
546
547 **4.7 Analysis of *Salmonella* Culturability**
548 Growth of *Salmonella* was determined using a real-time quantitative PCR method for the
549 confirmation of *Salmonella* isolates for presumptive generic identification of foodborne
550 *Salmonella*. Testing was performed fully in concordance with the Bacteriological Analytical
551 Manual (BAM) for *Salmonella*[74,75] for this approach that is also AOAC-approved. All samples
552 with positive results for *Salmonella* were classified as containing actively growing *Salmonella*. To
553 compare culture results with those from total RNA sequencing, *Salmonella* RPM values were
554 parsed from the genus-level microbe table (described in Section 4.4).
555 Two additional approaches were employed to examine *Salmonella* read mapping with a
556 more sensitive tool and broader reference databases. Quality controlled matrix-filtered reads were
557 aligned using Bowtie2[48] with very-sensitive-local-mode to 1. an expanded collection of whole
558 *Salmonella* genomes and 2. to a curated growth gene reference for elongation factor Tu (*ef-Tu*).
559 For results from both complete genome and *ef-Tu* gene alignments, the relative abundance (RPM)
560 was computed as shown in Equation 1.
561 For whole genome alignments, a reference was constructed from 1,183 recently published
562 *Salmonella* genomes[49] in addition to the 264 *Salmonella* genomes extracted from the
563 aforementioned NCBI RefSeq Complete collection (see Methods Section 4.4).
564 To construct a curated growth gene (*ef-Tu*) reference, gene sequences annotated in
565 *Salmonella* genomes as "elongation factor Tu", "EF-Tu" or "eftu" (case insensitive) were retrieved
566 from OMXWare[51] using its Python package. This query yielded 4,846 unique gene sequences from

a total of 36,242 *Salmonella* genomes which were assembled or retrieved from the NCBI Sequence Read Archive or RefSeq Complete Sequences as previously described.[51] The retrieved *ef-Tu* gene sequences were subsequently used to build a custom Bowtie2[48] reference. Read alignment was completed with very-sensitive-local-mode.

The read counts for each sample were ranked and Wilcoxon rank sum test was computed between the rank vectors of 4 *Salmonella*-positive and 23 *Salmonella*-negative samples. The 4 samples with unknown *Salmonella* status were excluded from the rankings.

Point-biserial correlation coefficients (r_{pb}) were calculated between *Salmonella* growth indicated by culture results (+1 and -1 for presence and absence, respectively) and observed relative abundance from total RNA sequencing results using the R package ltm.[76] The point-biserial correlation is a special case of the Pearson correlation that is better suited for a binary variable e.g. when *Salmonella* is reported as present or absent (a sample's *Salmonella* status).

Data Availability:

All high protein powder (HPP) poultry meal sequences are available through the 100K Pathogen Genome Project (PRJNA186441) in the NCBI BioProject (see Supplementary Table 1 for a complete list of accession numbers).

Code Availability:

The pipeline and microbial or matrix references were constructed from publicly available tools and reference sequences as described in the Methods section. Automated usability of this pipeline is available through membership in the Consortium for Sequencing the Food Supply Chain.

Acknowledgements:

We'd like to acknowledge the IBM Research OMXWare team for their data management support and availability for the retrieval and processing of microbial genomes. This research project was financially supported by the Consortium for Sequencing the Food Supply Chain. Funding for the total RNA sequencing of high protein powder factory ingredients was provided by Mars, Incorporated to B.C.W. with specific interest in metagenomics of the food microbiome.

Contributions:

KLB and NH conceived of the experimental design, developed the approach, completed and oversaw the experiments, performed analyses, and wrote the paper; DC, SE, MK, BK, MD, RP, HK, ES developed the approach, analyzed data, and revised the manuscript; BCH completed nucleic acid extraction method development and sequencing library construction, and contributed to data analysis and writing; NK coordinated sample collection and processing, nucleic acid extraction and contributed to writing; RB and PM conceived of the experimental design, developed the approach, and reviewed the paper; BG contributed to the experimental design, developed the approach, and wrote the paper; GD, CHM, SP, AQ participated to the conception of the experimental design and to the review of the manuscript; LP conceived of the experiment, contributed to the data analysis, and wrote the paper; JHK conceived of the experiment, developed the approach, and wrote the paper; BCW conceived of the experimental design, developed the approach, oversaw the experiments, performed analyses, and wrote the paper

Competing Interests:

The authors were employed by private or academic organizations as described in the author affiliations at the time this work was completed. IBM Corporation, Mars Incorporated, and Bio-

615 Rad Laboratories are members of the Consortium for Sequencing the Food Supply Chain. The
616 authors declare no other competing interests

617

618 **Supplementary information is available at npj Science of Food's website**

619

620 **REFERENCES:**

621 1. Kovac, J., Bakker, H. den, Carroll, L. M. & Wiedmann, M. Precision food safety: A
622 systems approach to food safety facilitated by genomics tools. *TrAC Trends Anal. Chem.*
623 (2017). doi:10.1016/j.trac.2017.06.001
624 2. Weimer, B. C. *et al.* Defining the food microbiome for authentication, safety, and process
625 management. *IBM J. Res. Dev.* **60**, 1 (2016).
626 3. Walsh, A. M. *et al.* Microbial Succession and Flavor Production in the Fermented Dairy
627 Beverage Kefir. *mSystems* **1**, (2016).
628 4. Walsh, A. M. *et al.* Species classifier choice is a key consideration when analysing low-
629 complexity food microbiome data. *Microbiome* **6**, 50 (2018).
630 5. Duru, I. C. *et al.* Metagenomic and metatranscriptomic analysis of the microbial
631 community in Swiss-type Maasdam cheese during ripening. *Int. J. Food Microbiol.* **281**,
632 10–22 (2018).
633 6. Martins, G. *et al.* Grape berry bacterial microbiota: Impact of the ripening process and the
634 farming system. *Int. J. Food Microbiol.* **158**, 93–100 (2012).
635 7. Abdelfattah, A., Wisniewski, M., Droby, S. & Schena, L. Spatial and compositional
636 variation in the fungal communities of organic and conventionally grown apple fruit at the
637 consumer point-of-purchase. *Hortic. Res.* **3**, 16047 (2016).
638 8. Williams, A. G., Choi, S.-C. & Banks, J. M. Variability of the species and strain
639 phenotype composition of the non-starter lactic acid bacterial population of cheddar
640 cheese manufactured in a commercial creamery. *Food Res. Int.* **35**, 483–493 (2002).
641 9. Weimer, B. C. 100K Pathogen Genome Project. *Genome Announc.* **5**, e00594-17 (2017).
642 10. Emond-Rheault, J.-G. *et al.* A Syst-OMICS Approach to Ensuring Food Safety and
643 Reducing the Economic Burden of Salmonellosis. *Front. Microbiol.* **8**, 996 (2017).
644 11. Kaufman, J. H. *et al.* Insular microbiogeography. (2017).
645 12. Bashiardes, S., Zilberman-Schapira, G. & Elinav, E. Use of Metatranscriptomics in
646 Microbiome Research. *Bioinform. Biol. Insights* **10**, 19–25 (2016).
647 13. McGrath, K. C. *et al.* Isolation and analysis of mRNA from environmental microbial
648 communities. *J. Microbiol. Methods* **75**, 172–176 (2008).
649 14. Cottier, F. *et al.* Advantages of meta-total RNA sequencing (MeTRS) over shotgun
650 metagenomics and amplicon-based sequencing in the profiling of complex microbial
651 communities. *npj Biofilms Microbiomes* **4**, 2 (2018).
652 15. Macklaim, J. M. *et al.* Comparative meta-RNA-seq of the vaginal microbiota and
653 differential expression by Lactobacillus iners in health and dysbiosis. *Microbiome* **1**, 12
654 (2013).
655 16. Haiminen, N. *et al.* Food authentication from shotgun sequencing reads with an
656 application on high protein powders. *npj Sci. Food* **3**, (2019).

17. Lakin, S. M. *et al.* MEGARes: an antimicrobial resistance database for high throughput sequencing.
18. Noyes, N. R. *et al.* Resistome diversity in cattle and the environment decreases during beef production. *Elife* **5**, e13195 (2016).
19. Ganesan, B., Dobrowolski, P. & Weimer, B. C. Identification of the Leucine-to-2-Methylbutyric Acid Catabolic Pathway of Lactococcus lactis. *Appl. Environ. Microbiol.* **72**, 4264–4273 (2006).
20. Ganesan, B., Seefeldt, K., Koka, R. C., Dias, B. & Weimer, B. C. Monocarboxylic acid production by lactococci and lactobacilli. *Int. Dairy J.* **14**, 237–246 (2004).
21. Ganesan, B., Seefeldt, K. & Weimer, B. C. Fatty Acid Production from Amino Acids and -Keto Acids by Brevibacterium linens BL2. *Appl. Environ. Microbiol.* **70**, 6385–6393 (2004).
22. Ganesan, B., Stuart, M. R. & Weimer, B. C. Carbohydrate Starvation Causes a Metabolically Active but Nonculturable State in Lactococcus lactis. *Appl. Environ. Microbiol.* **73**, 2498–2512 (2007).
23. Ganesan, B. *et al.* Probiotic bacteria survive in Cheddar cheese and modify populations of other lactic acid bacteria. *J. Appl. Microbiol.* **116**, 1642–1656 (2014).
24. Ganesan, B. & Weimer, B. C. *Cheese : chemistry, physics, and microbiology.* (Elsevier, 2004).
25. Sheflin, A. M., Melby, C. L., Carbonero, F. & Weir, T. L. Linking dietary patterns with gut microbial composition and function. *Gut Microbes* **8**, (2017).
26. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**, e00031-18 (2018).
27. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The Impact of the Gut Microbiota on Human Health: An Integrative View. *Cell* **148**, 1258–1270 (2012).
28. Richards, J. L., Yap, Y. A., McLeod, K. H., Mackay, C. R. & Mariño, E. Dietary metabolites and the gut microbiota: an alternative approach to control inflammatory and autoimmune diseases. *Clin Trans Immunol* **5**, e82 (2016).
29. Yang, X. *et al.* Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Pathogens within the Microbiome of the Beef Production Chain. *Appl Env. Microbiol* **82**, 2433–2443 (2016).
30. Hofacre, C. L. *et al.* Characterization of antibiotic-resistant bacteria in rendered animal products. *Avian Dis.* **45**, 953–61
31. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8**, 2224 (2017).
32. Gloor, G. B. & Reid, G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* **62**, 692–703 (2016).
33. Langelier, C. *et al.* Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E12353–E12362 (2018).
34. Ilott, N. E. *et al.* Defining the microbial transcriptional response to colitis through integrated host and microbiome profiling. *ISME J.* **10**, 2389–2404 (2016).
35. Ripp, F. *et al.* All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing. *BMC Genomics* **15**, 639 (2014).
36. Lee, A. Y., Lee, C. S. & Gelder, R. N. Van. Scalable metagenomics alignment research tool (SMART): a scalable, rapid, and complete search heuristic for the classification of metagenomic sequences from complex sequence populations. *BMC Bioinformatics* **17**, 292 (2016).

37. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
38. Haiminen, N. *et al.* Food authentication from shotgun sequencing reads with an application on high protein powders. *npj Sci. Food* in press (2019).
39. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
40. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060 (2009).
41. Kyrpides, N. C. *et al.* Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains. *PLoS Biol.* **12**, e1001920 (2014).
42. Kyrpides, N. C., Eloe-Fadrosh, E. A. & Ivanova, N. N. Microbiome Data Science: Understanding Our Microbial Planet. *Trends Microbiol.* **24**, 425–427 (2016).
43. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457 (2017).
44. Nayfach, S. & Pollard, K. S. Toward Accurate and Quantitative Comparative Metagenomics. *Cell* **166**, 1103–1116 (2016).
45. Satopää, V., Albrecht, J., Irwin, D. & Raghavan, B. *Finding a 'Kneedle' in a Haystack: Detecting Knee Points in System Behavior*.
46. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. & Pawlowsky-Glahn, V. Logratio Analysis and Compositional Distance. *Math. Geol.* **32**, 271–275 (2000).
47. Di Palma, M. A. & Gallo, M. A co-median approach to detect compositional outliers. *J. Appl. Stat.* **43**, 2348–2362 (2016).
48. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–9 (2012).
49. Kong, N. *et al.* Draft Genome Sequences of 1,183 Salmonella Strains from the 100K Pathogen Genome Project. *Genome Announc.* **5**, (2017).
50. Tubulekas, I. & Hughes, D. A Single Amino Acid Substitution in Elongation Factor Tu Disrupts Interaction between the Ternary Complex and the Ribosome. *J. Bacteriol.* 240–250 (1993).
51. Seabolt, E. E. *et al.* OMXWare, A Cloud-Based Platform for Studying Microbial Life at Scale. *arXiv* **1911.02095**, (2019).
52. Zelezniak, A. *et al.* Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl. Acad. Sci.* **112**, 6449–6454 (2015).
53. Jones, M. B. *et al.* Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci U S A* (2015). doi:10.1073/pnas.1519288112
54. Pollock, J., Glendinning, L., Wisedchanwet, T. & Watson, M. The madness of microbiome: Attempting to find consensus 'best practice' for 16S microbiome studies. *Appl. Environ. Microbiol.* AEM.02627-17 (2018). doi:10.1128/AEM.02627-17
55. Browne, H. P. *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546 (2016).
56. Eilers, H., Pernthaler, J., Glöckner, F. O. & Amann, R. Culturability and In situ abundance of pelagic bacteria from the North Sea. *Appl. Environ. Microbiol.* **66**, 3044–51 (2000).
57. Hinchliff, C. E. *et al.* Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci U S A* **112**, 12764–12769 (2015).
58. Knight, R. *et al.* Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).
59. Weis, A. M. *et al.* Genomic Comparison of Campylobacter spp. and Their Potential for Zoonotic Transmission between Birds, Primates, and Livestock. *Appl. Environ. Microbiol.*

82, 7165 LP – 7175 (2016).

60. Miller, B. *et al.* A novel, single-tube enzymatic fragmentation and library construction method enables fast turnaround times and improved data quality for microbial whole-genome sequencing. *Kapa Biosyst. Appl. Note* 1–8 (2015). doi:10.13140/RG.2.1.4534.3440
61. Lüdeke, C. H. M., Kong, N., Weimer, B. C., Fischer, M. & Jones, J. L. Complete genome sequences of a clinical isolate and an environmental isolate of Vibrio parahaemolyticus. *Genome Announc.* **3**, e00216-15 (2015).
62. Jeannotte, R. *et al.* High-Throughput Analysis of Foodborne Bacterial Genomic DNA Using Agilent 2200 TapeStation and Genomic DNA ScreenTape System. *Agil. Appl. Note* 1–8 (2015). doi:doi:10.6084/m9.figshare.1372504
63. Arabyan, N. *et al.* Salmonella Degrades the Host Glycocalyx Leading to Altered Infection and Glycan Remodeling. *Sci. Rep.* **6**, 1–11 (2016).
64. Kong, N. *et al.* Draft Genome Sequences of 1,183 *Salmonella* Strains from the 100K Pathogen Genome Project. *Genome Announc.* **5**, e00518-17 (2017).
65. Krueger, F. TrimGalore. (2018).
66. Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand. Genomic Sci.* **10**, 18 (2015).
67. Homer, N. DWGSIM. (2011).
68. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
69. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).
70. Wood, D. Kraken's operating manual.
71. Templ, M., Hron, K. & Filzmoser, P. robCompositions: An R-package for Robust Statistical Analysis of Compositional Data. in *Compositional Data Analysis* 341–355 (John Wiley & Sons, Ltd, 2011). doi:10.1002/9781119976462.ch25
72. Andrews, S. FastQC.
73. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
74. Andrews, W. H., Wang, H., Jacobson, A. & Hammack, T. Bacteriological Analytical Manual (BAM) Chapter 5: Salmonella. in *Bacteriological Analytical Manual* (U.S. Food and Drug Administration, 2018).
75. Grim, C. J. *et al.* High-Resolution Microbiome Profiling for Detection and Tracking of Salmonella enterica. *Front. Microbiol.* **8**, 1587 (2017).
76. Rizopoulos, D. **ltm** : An *R* Package for Latent Variable Modeling and Item Response Theory Analyses. *J. Stat. Softw.* **17**, 1–25 (2006).

801
802
803 **FIGURE and TABLE LEGENDS: (corresponding to their order at end of merged document)**

804

805 **Figure 1:** Pipeline description of bioinformatic steps applied to high protein powder
806 metatranscriptome samples. Black arrows indicate data flow and blue boxes describe outputs
807 from the pipeline.

808

809 **Table 1:** Accuracy of microbial identification using *in silico* constructed Simulated Food
810 Mixtures with expected food matrix and microbial sequences.

811

812 **Figure 2A**: Alpha diversity (number of genera) for all (n = 31) high protein powder
813 metatranscriptomes is compared to total number of sequenced reads for a range of *in silico*
814 subsampled sequencing depths. The dashed vertical line indicates the median elbow (at approx.
815 67 million reads).

816

817 **Figure 2B:** Hierarchical clustering of Aitchison distance values of poultry meal samples based
818 on microbial composition. Samples were received from Supplier A (blue and red) and Supplier B
819 (green). Matrix-contaminated samples are additionally marked in red.

820

821 **Figure 3A:** Phylogram of the 65 microbial genera present in all samples with RPM > 0.1

822

823 **Figure 3B:** Phylogram of all microbes observed in *any* sample. Log of the median RPM value
824 across samples is indicated. Grey indicating a median RPM value of 0.

825

826

827 **Figure 4:** Heatmap (log_{10}-scale) of high protein powder microbial composition and relative
828 abundance (RPM) where absence (RPM < 0.1) is indicated in grey. Genera are ordered by
829 summed abundance across samples. Samples were received from Supplier A (blue) and Supplier
830 B (green). Red stars indicate matrix-contaminated samples (from Supplier A).

831

832 **Figure 5A:** All identified microbial general are plotted with median value and median absolute
833 deviation (MAD) of RPM abundance. Genera with MAD > 5 are labeled with the genus name.

834

835 **Figure 5B:** Heatmap (log_{10}-scale) of ten microbial genera with the largest median absolute
836 deviation (MAD) across samples. Genera are ordered by decreasing MAD from top to bottom.
837 Samples were received from Supplier A (blue) and Supplier B (green). Red stars indicate matrix
838 contaminated samples (from Supplier A).

839

840 **Figure 6A:** Relative abundance of microbes with high relevance to food safety and quality from
841 high protein powder total RNA sequenced microbiomes. Width of violin plot indicates density of
842 samples with relative abundance at that value. Observation threshold of RPM = 0.1 is indicated
843 with the horizontal black line.

844

845 **Figure 6B:** Foodborne microbe relative abundances are shown across samples of high protein
846 powder total RNA sequenced samples.

847

848 **Figure 7:** *Salmonella* culturability status and high-throughput sequencing read abundance
849 (RPM) from *k*-mer classification to NCBI Microbial RefSeq Complete **(A)**, from alignments to

1,447 *Salmonella* genomes **(B)**, and from alignments to 4,846 EF-Tu gene sequences **(C)**. *Salmonella* presence (red) indicates culture-positive result, absence (green) indicates culture-negative result, and no record (black) indicates samples for which no culture test was completed.

Table 2: The ranks for *Salmonella*-positive samples and the associated p-values from Wilcoxon rank sum test are shown for high-throughput sequencing read abundance (RPM) for multiple analyses: *k*-mer classification to NCBI Microbial RefSeq Complete (left), alignments to 1,447 *Salmonella* genomes (middle), and alignments to 4,846 *ef-Tu* gene sequences (right). The corresponding *Salmonella* relative abundances are shown in Figure 7A–C.

Figure 8: *Salmonella* status correlations with genus relative abundances. Only those genera with absolute value of the correlation coefficient > 0.5 are shown. Positive and negative correlations are indicated in grey and blue, respectively.

SUPPLEMENTAL INFORMATION:

Supplemental Figures (pdf): Supplemental Figures S1–S5

Supplemental Table 1 (.xlsx) - Sample Descriptions

Supplemental Table 2 (.xlsx) - Matrix Filtering Genomes

Supplemental Table 3 (.xlsx) - Simulated Food Mixtures

Supplemental Table 4 (.xlsx) - Microbial Genera

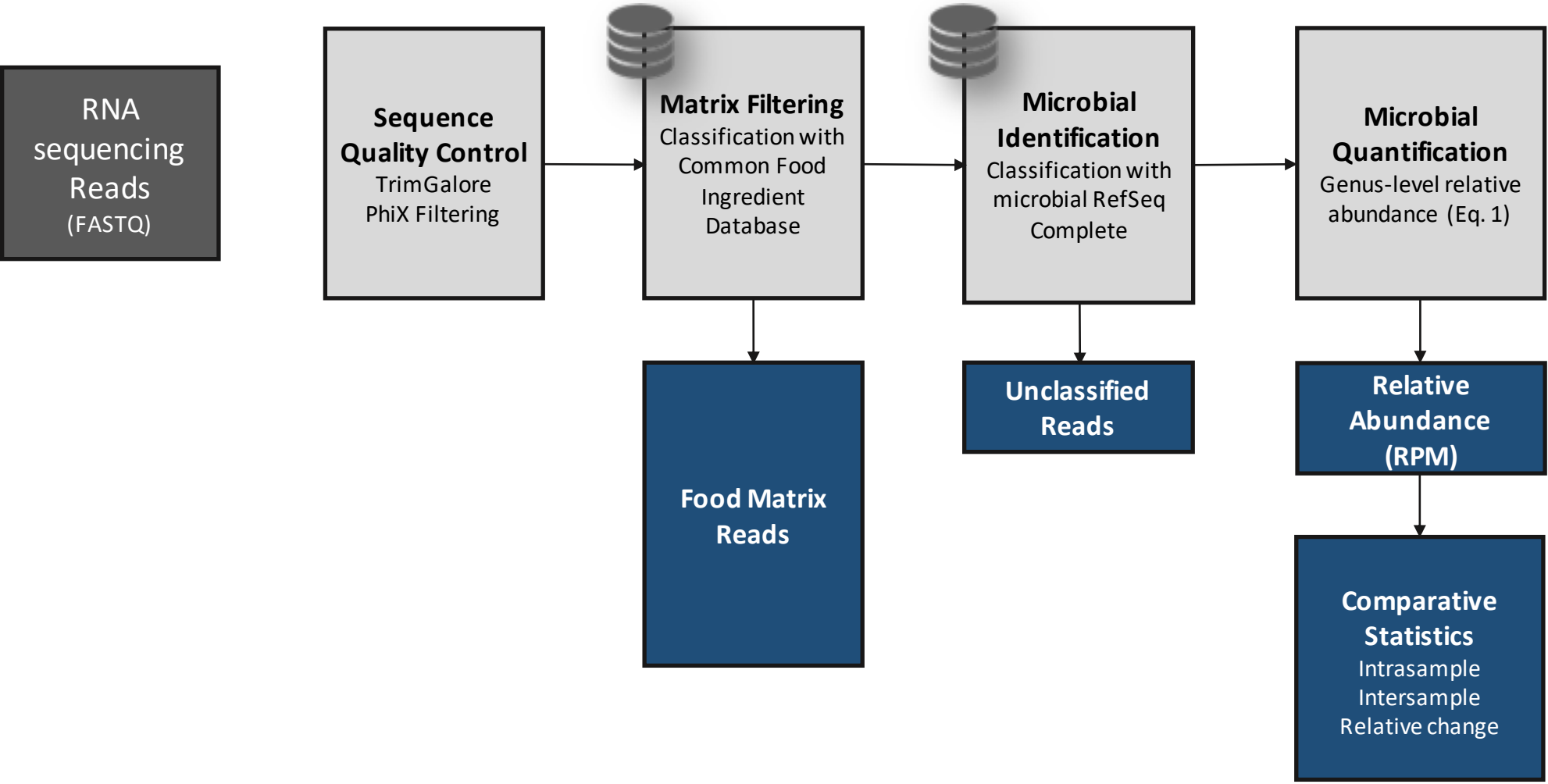
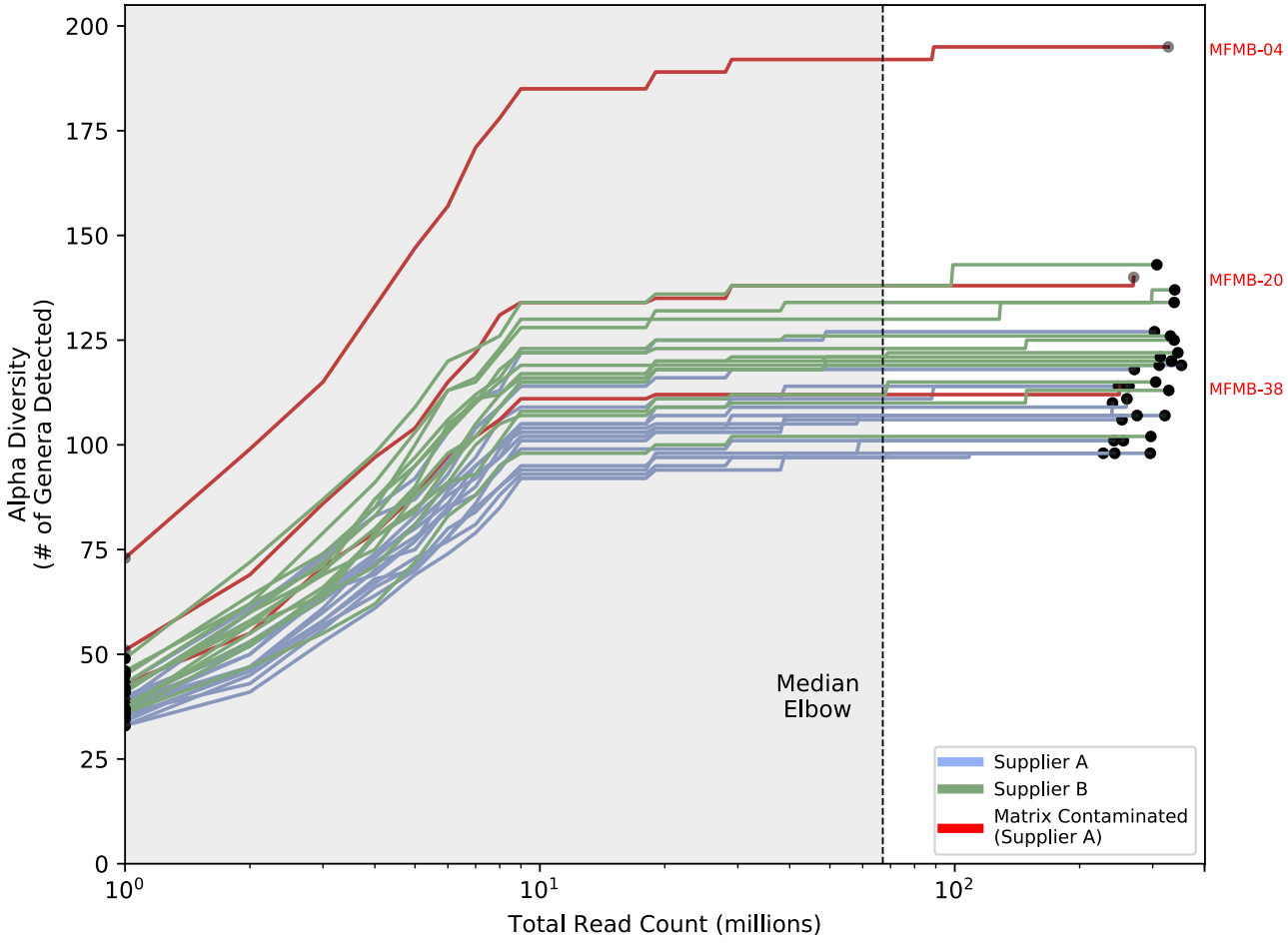
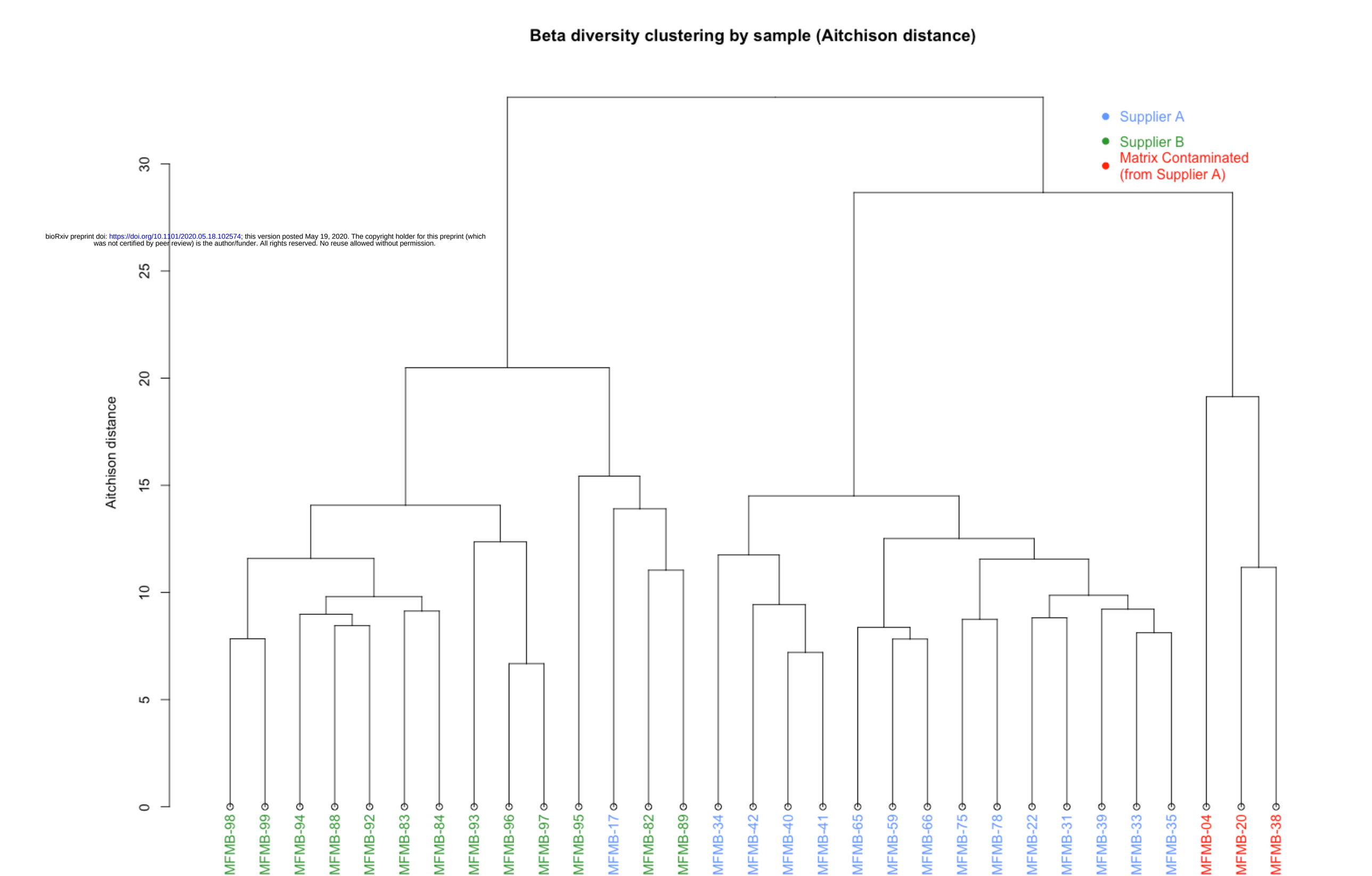
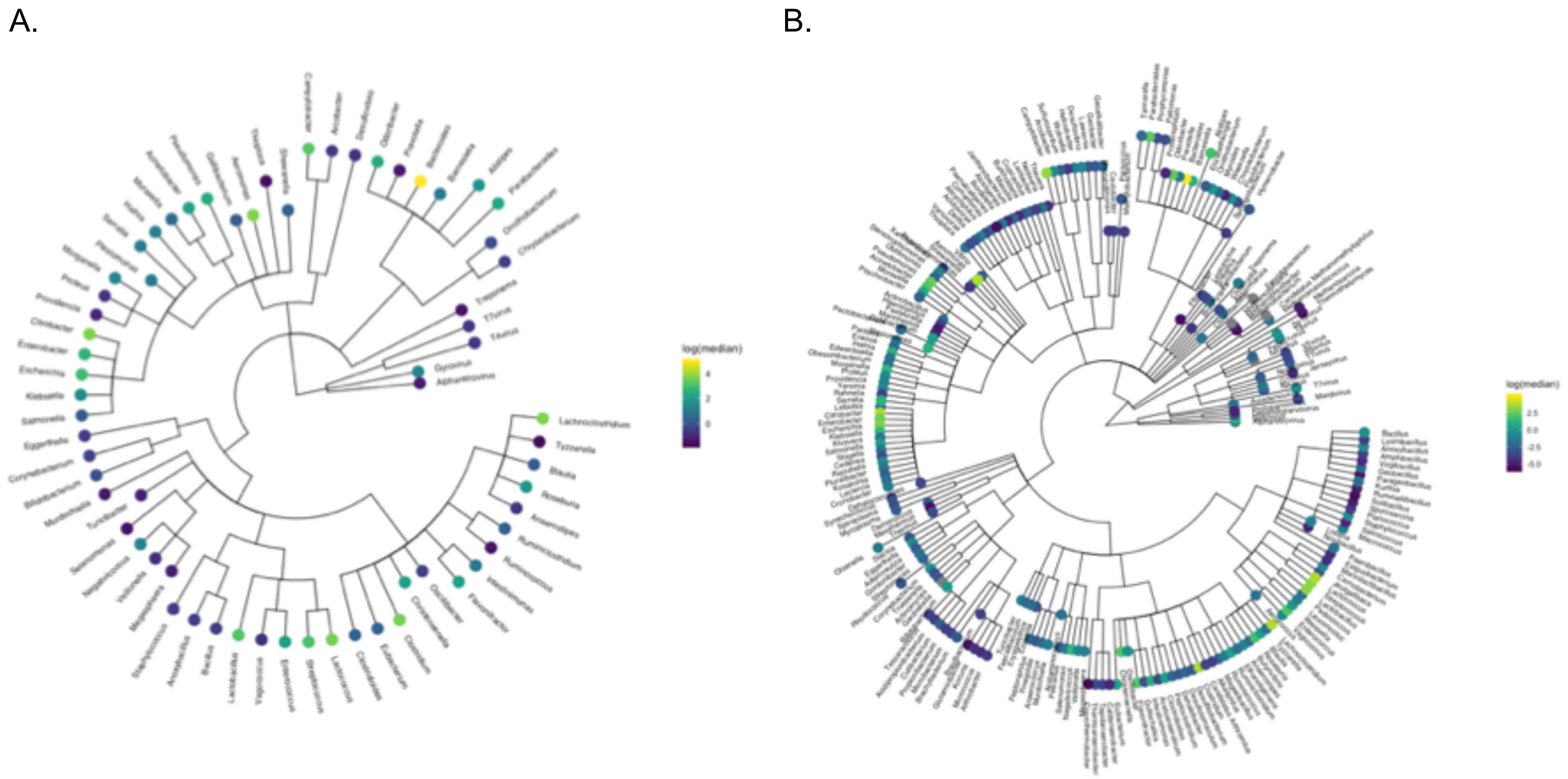

RNA sequencing Reads
(FASTQ)
Sequence Quality Control
TrimGalore
PhiX Filtering
Matrix Filtering
Classification with Common Food Ingredient Database
Microbial Identification
Classification with microbial RefSeq Complete
Microbial Quantification
Genus-level relative abundance (Eq. 1)
Food Matrix Reads
Unclassified Reads
Relative Abundance (RPM)
Comparative Statistics
Intrasample
Intersample
Relative change

Table 1: Microbial Identification Accuracy from Simulated Food Microbiome Mixtures

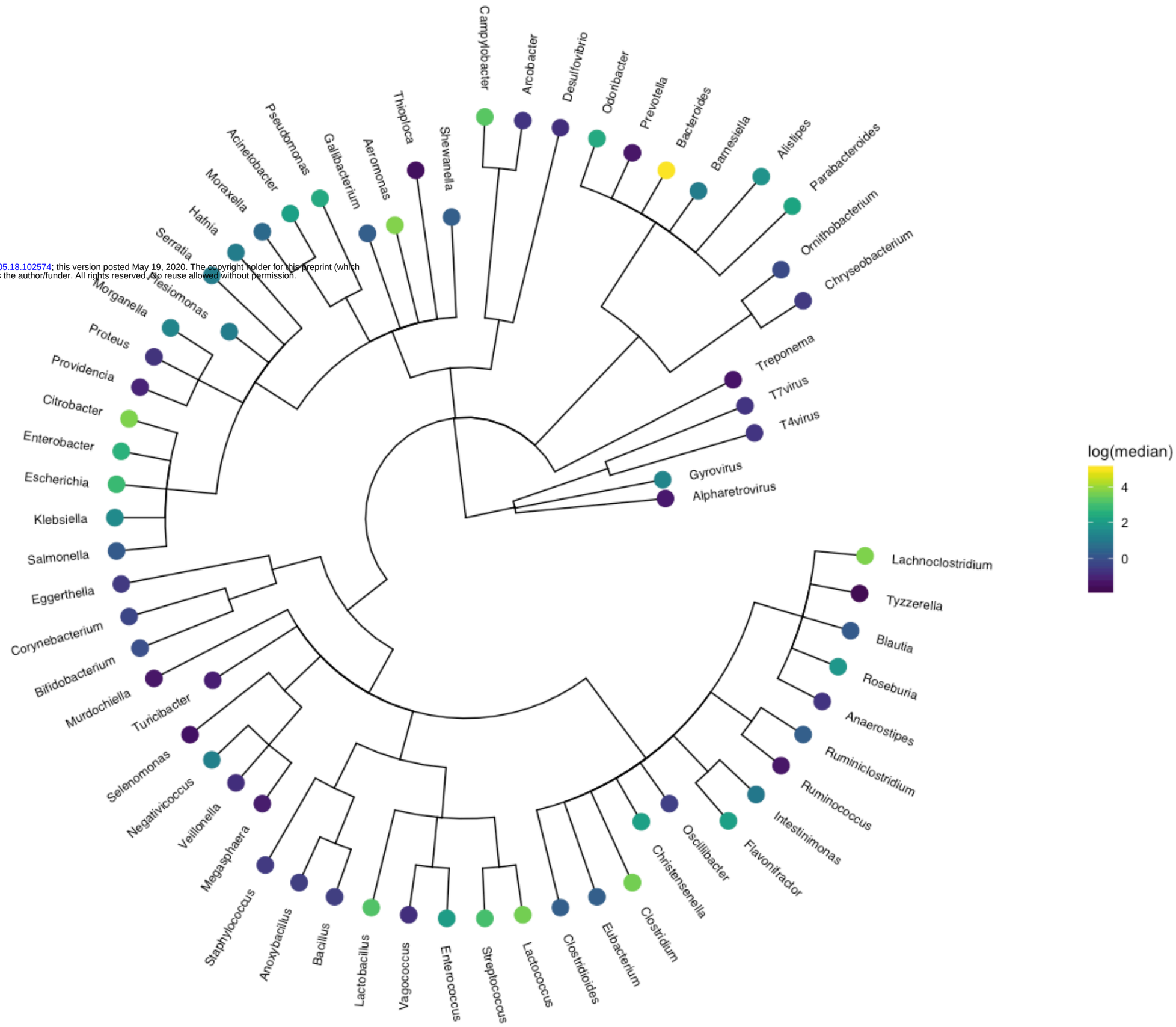
	Simulated Mixture 1				Simulated Mixture 2			
	With Matrix Filtering		No Matrix Filtering		With Matrix Filtering		No Matrix Filtering	
	# GENERA	**GENUS READS**	# GENERA	GENUS READS	**# GENERA**	**GENUS READS**	# GENERA	GENUS READS
Bacteria in Simulated Mixture (Expected Content)	**14**	**15,000**	**14**	**15,000**	**14**	**15,000**	**14**	**15,000**
Observed Microbial Content								
Bacteria	18	13,517	34	13,700	15	13,551	33	13,999
Viruses	0	0	9	563	0	0	4	328
Archaea	0	0	1	1	0	0	1	3
Eukaryota	0	0	4	104	0	0	4	799
Total Observed	**18**	**13,517**	**48**	**14,368**	**15**	**13,551**	**42**	**15,129**
True Positives (as a % of total observed)	14 (78%)	13,511 (99.96%)	14 (29%)	13,571 (94.45%)	14 (93%)	13,548 (99.98%)	14 (33%)	13,623 (90.05%)
False Positives (as a % of total observed)	4 (22%)	6 (0.04%)	34 (71%)	797 (5.55%)	1 (7%)	3 (0.02%)	28 (67%)	1,506 (9.95%)
False Positives Removed with Matrix Filtering (as a % of false positives without filtering)	**30 (88.2%)**	**791 (99.2%)**			**27 (96.4%)**	**1,503 (99.8%)**		


Alpha Diversity
(# of Genera Detected)
200
175
150
125
100
75
50
25
0
10^0
10^1
10^2
Total Read Count (millions)
Median
Elbow
MFMB-04
MFMB-20
MFMB-38
Supplier A
Supplier B
Matrix Contaminated
(Supplier A)

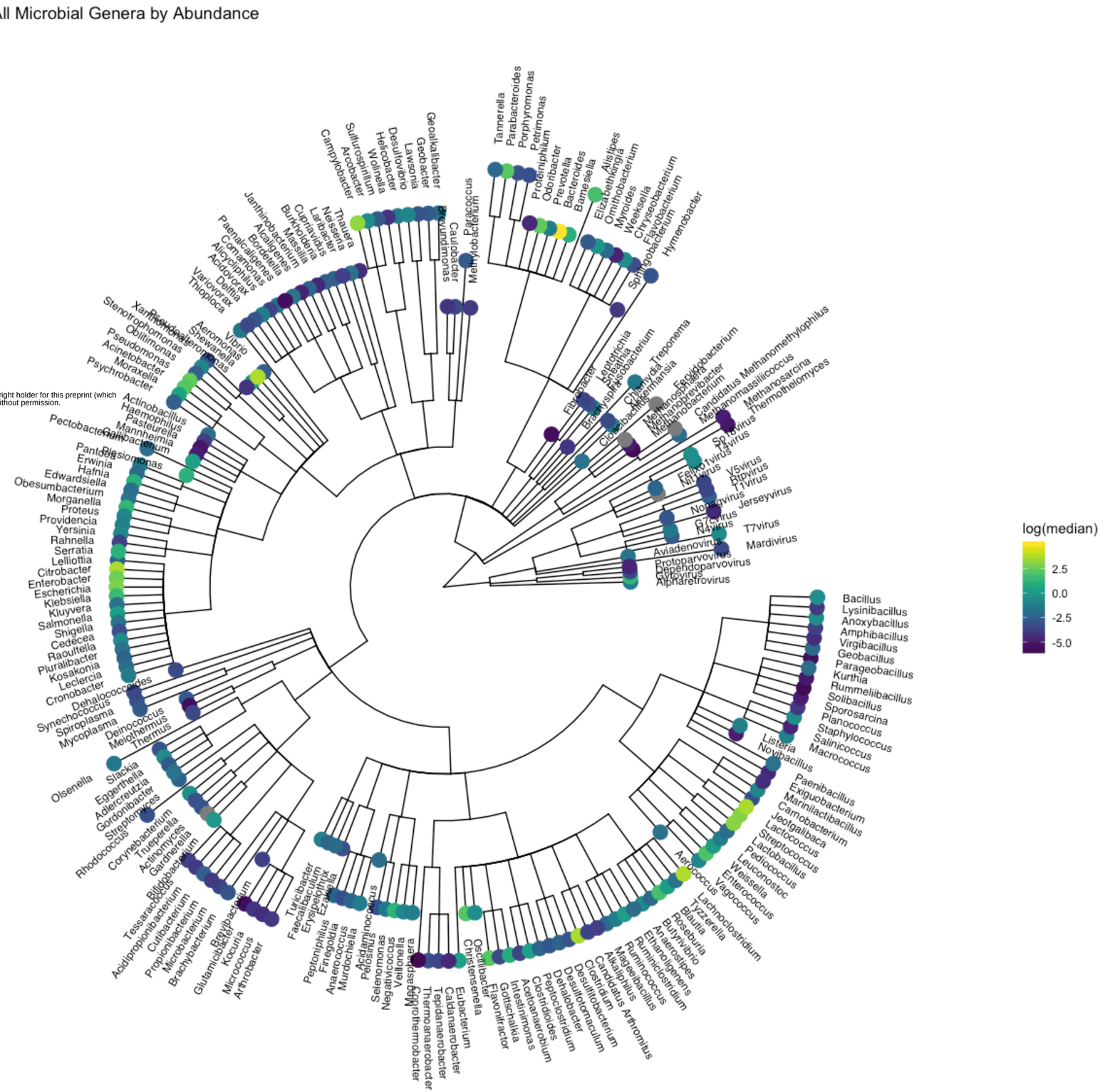


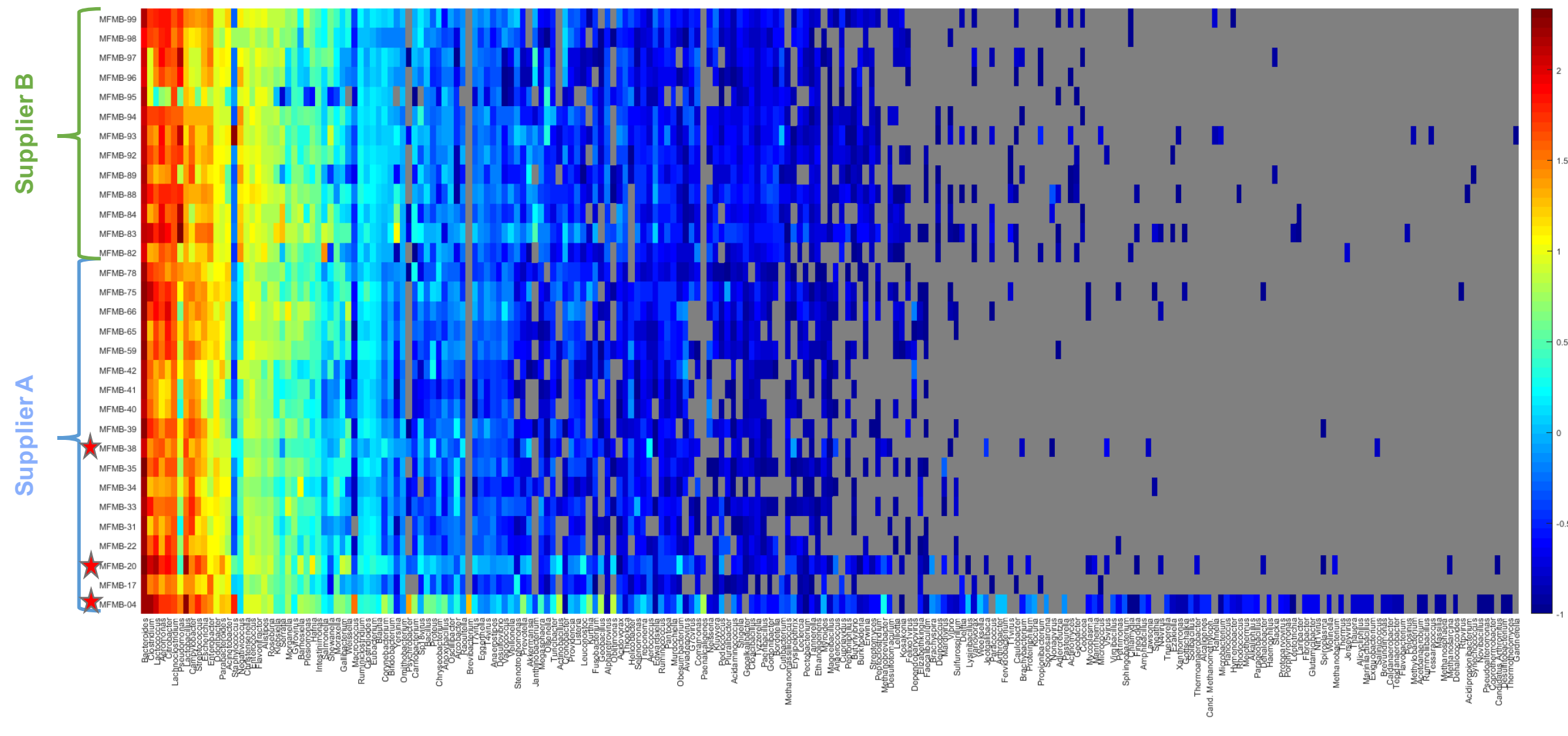

A.
B.
log(median)
4
2
0
2.5
0.0
-2.5
-5.0
Lachnoclostridium
Tyzzerella
Blautia
Roseburia
Anaerostipes
Ruminiclostridium
Ruminococcus
Intestinimonas
Flavonifractor
Oscillibacter
Clostridium
Eubacterium
Lactococcus
Streptococcus
Enterococcus
Vagococcus
Lactobacillus
Bacillus
Staphylococcus
Turicibacter
Bifidobacterium
Corynebacterium
Eggerthella
Salmonella
Klebsiella
Escherichia
Enterobacter
Citrobacter
Providencia
Proteus
Morganella
Campylobacter
Arcobacter
Gyrovirus
Alphatorquevirus

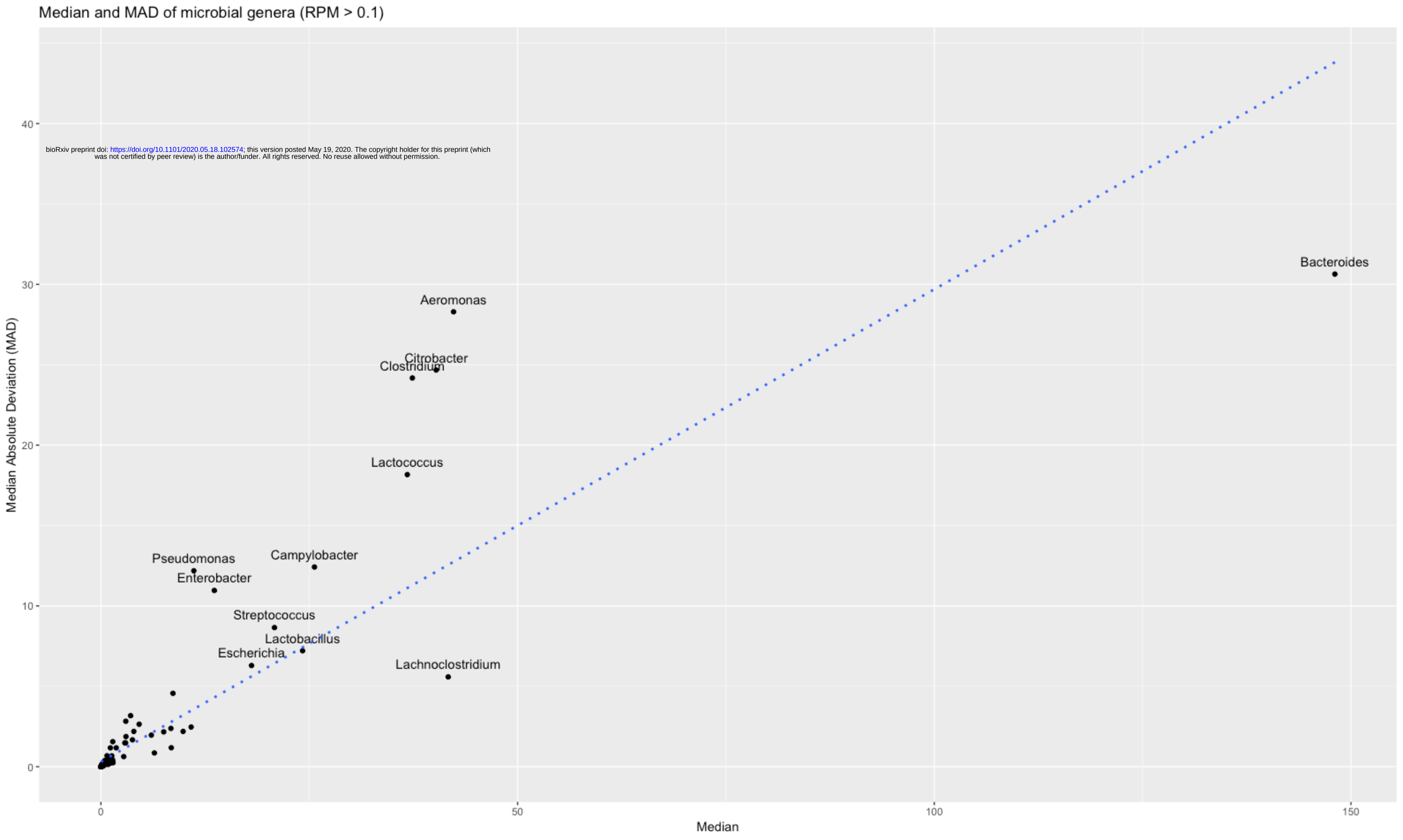
Phylogram of Consensus Microbial Genera by Abundance

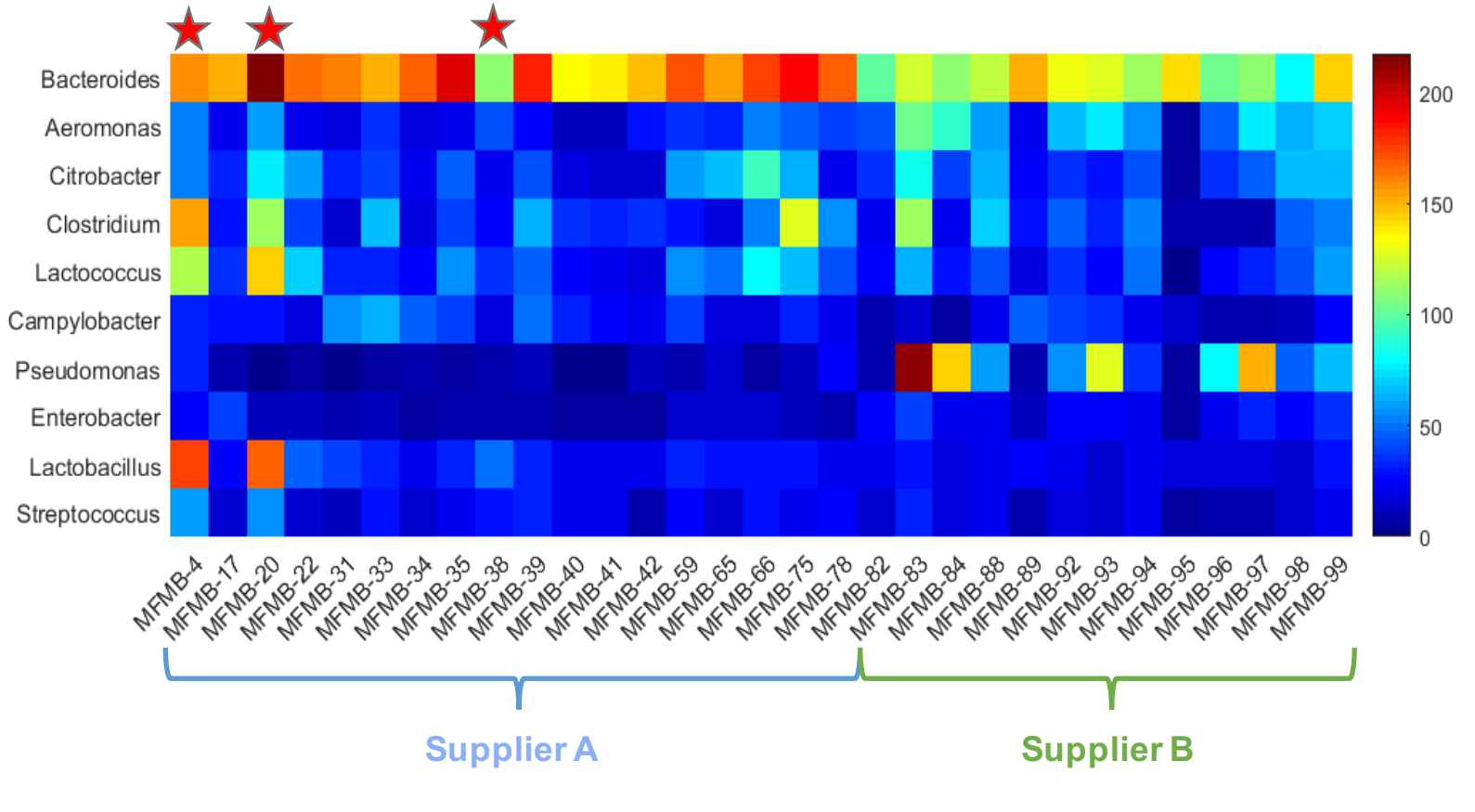


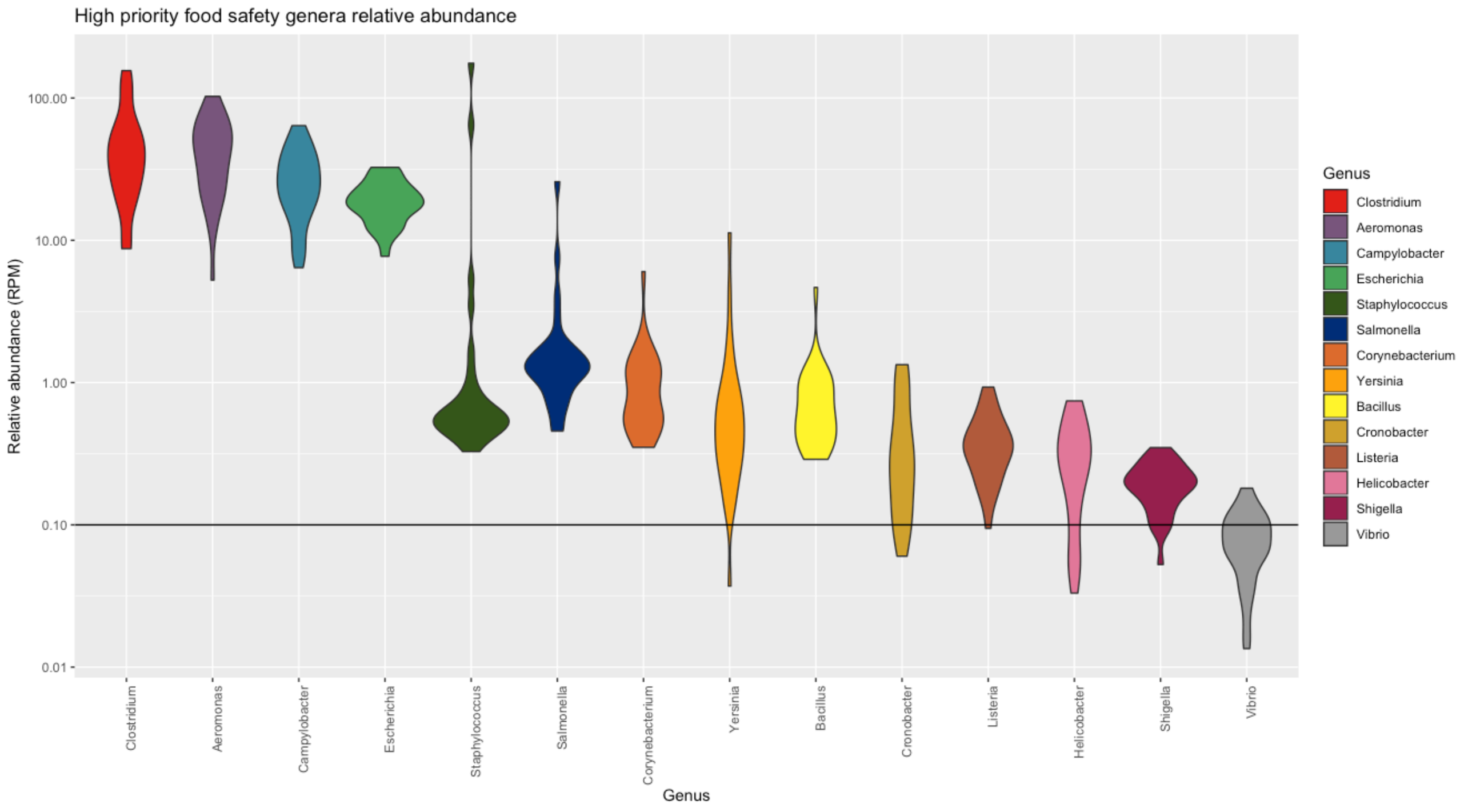
Phylogram of All Microbial Genera by Abundance

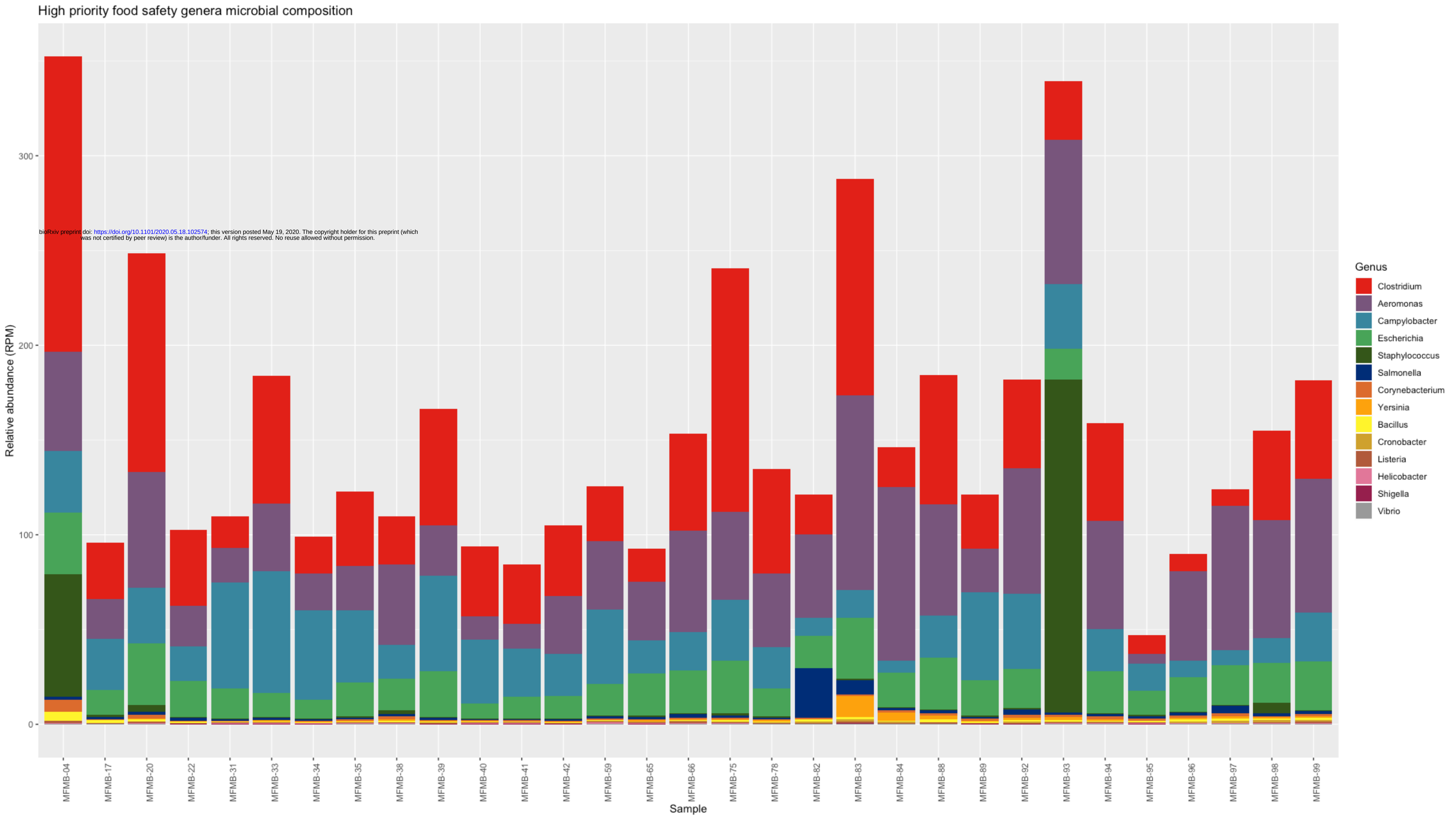


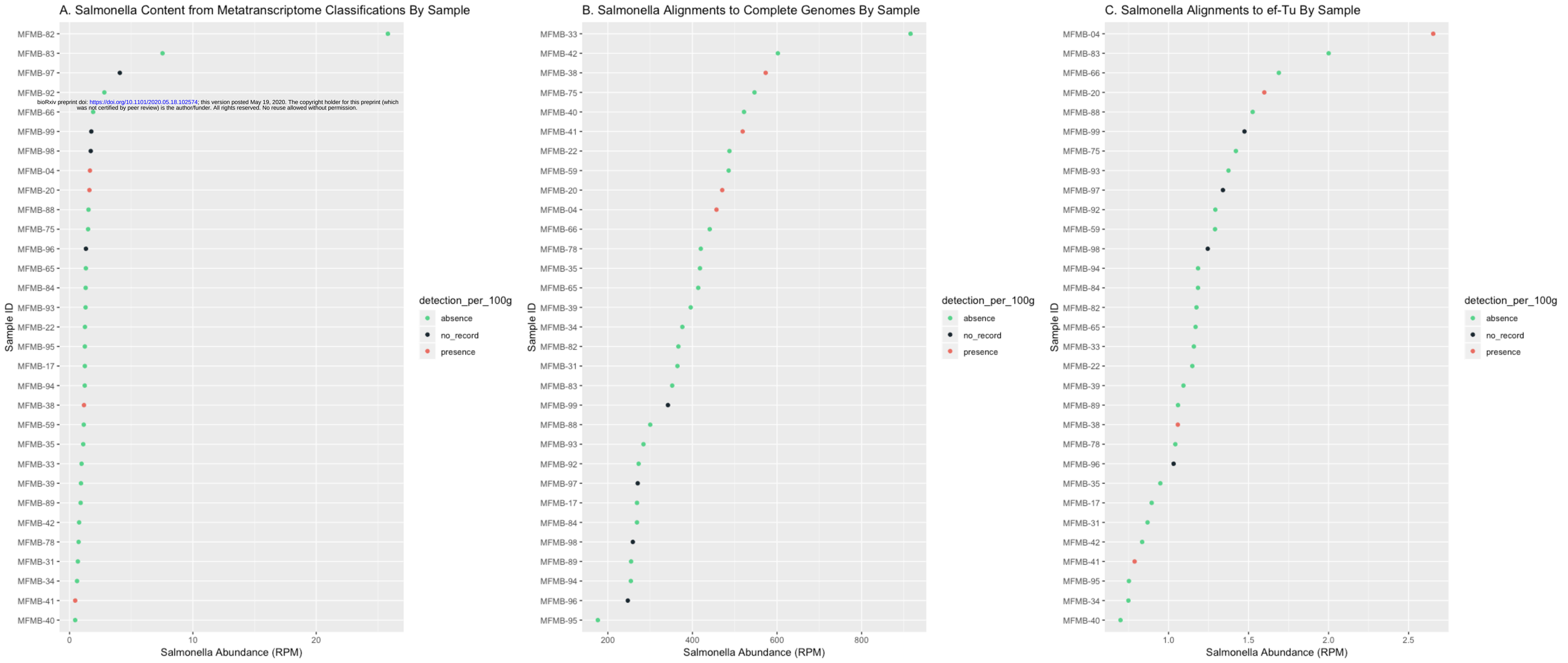

Supplier B
Supplier A
MFMB-99
MFMB-98
MFMB-97
MFMB-96
MFMB-95
MFMB-94
MFMB-93
MFMB-92
MFMB-89
MFMB-88
MFMB-84
MFMB-83
MFMB-82
MFMB-78
MFMB-75
MFMB-66
MFMB-65
MFMB-59
MFMB-42
MFMB-41
MFMB-40
MFMB-39
MFMB-38
MFMB-35
MFMB-34
MFMB-33
MFMB-31
MFMB-22
MFMB-20
MFMB-17
MFMB-04
2
1.5
1
0.5
0
-0.5
-1




Bacteroides
Aeromonas
Citrobacter
Clostridium
Lactococcus
Campylobacter
Pseudomonas
Enterobacter
Lactobacillus
Streptococcus
200
150
100
50
0
MFMB-4
MFMB-17
MFMB-20
MFMB-22
MFMB-31
MFMB-33
MFMB-34
MFMB-35
MFMB-38
MFMB-39
MFMB-40
MFMB-41
MFMB-42
MFMB-59
MFMB-65
MFMB-66
MFMB-75
MFMB-78
MFMB-82
MFMB-83
MFMB-84
MFMB-88
MFMB-89
MFMB-92
MFMB-93
MFMB-94
MFMB-95
MFMB-96
MFMB-97
MFMB-98
MFMB-99
Supplier A
Supplier B


High priority food safety genera relative abundance
Relative abundance (RPM)
100.00
10.00
1.00
0.10
0.01
Clostridium
Aeromonas
Campylobacter
Escherichia
Staphylococcus
Salmonella
Corynebacterium
Yersinia
Bacillus
Cronobacter
Listeria
Helicobacter
Shigella
Vibrio
Genus
Genus
Clostridium
Aeromonas
Campylobacter
Escherichia
Staphylococcus
Salmonella
Corynebacterium
Yersinia
Bacillus
Cronobacter
Listeria
Helicobacter
Shigella
Vibrio





Salmonella-positive sample	k-mer Classification	Whole Genome Alignment	*ef-Tu* Alignment
MFMB-04	8th	10th	1st
MFMB-20	9th	9th	4th
MFMB-38	20th	3rd	21st
MFMB-41	30th	6th	28th
Rank sum test p-value	p=0.86	p=0.06	p=0.56