Illumina MiSeq read quality assessment and taxonomic profiling was performed on a high performance compute cluster using a custom Nextflow pipeline, available at (<https://github.com/h3abionet/16S-rDNA-dada2-pipeline>). Quality was assessed with FastQC [Andrews] and MultiQC [Ewels et al], primers were removed by trimming the first 17 and 21 bp from the start of the 251bp forward and reverse reads, respectively (--trimFor 17 and --trimRev 21). Default settings were used for the remainder of the pipeline, which uses the DADA2 [Calahan et al.] method to group reads into amplicon sequence variants (ASVs). The main difference compared to the older operational taxonomic unit (OTU) clustering methods is that the DADA2 method detects exact ASVs, which unlike OTUs consist of a single unique sequence as opposed to a cluster of closely related (most commonly 97% identical) sequences. This is made possible by DADA2’s error correction capabilities, which relies on sequence quality information to build a machine learning error model, alternating info on sample composition and error rates until convergence. This allows assignment of all relevant reads to an error-corrected sequence. Taxonomic assignment was performed against the RefSeq-RDP 16S database (v3 May 2018) [Alishum et al.].

Downstream statistical analyses were performed in R, using the packages phyloseq [McMurdie et al.] for alpha and beta diversity analyses, MetagenomeSeq [Paulson et al.] for differential abundance testing, vegan [Oksanen et al] for principal coordinates analysis (PCoA) and NMF [Gajoux et al] for annotated heat maps. For PCoA and heatmaps raw reads were standardized so that all samples had equal total read counts. An abundance filter was applied to remove ASVs with less than 10 counts in less than 10% of samples or that made up less than 0.1% of the total read count for a given sample, leaving 599 of the original 2952 ASVs. For the heatmap samples were clustered using complete linkage clustering of the Bray-Curtis dissimilarity matrix. Per-sample genus- and phylum-level barplots were constructed using the bar.plots() function available in the public Github gist <https://gist.github.com/kviljoen/97d36c689c5c9b9c39939c7a100720b9>, excluding low-abundance taxa for ease of interpretation.

Differences in microbial compositions between fermented (amaas) vs. unfermented samples (milks) were assessed using the MetagenomeSeq MRfulltable() function, applied to raw reads merged at the lowest available taxonomic level; a custom filter was applied to identify high quality, significant features, as implemented in the super.fitZig.kv() function, which can be found in the aforementioned Github gist. Taxa were deemed significantly different (in terms of abundance and/or absence/presence) between fermented vs. unfermented samples if they exhibited a fold change (beta coefficient) of ≥ 1.5 and had an adjusted p-value of ≤ 0.05 and if at least one of the two groups compared had ≥ 60% of samples with the given ASV/taxon, or, if the result of Fisher’s exact test was significant (after multiple-testing correction by the Benjamini-Hochberg method [Benjamini et al.]). ASVs were first merged at the lowest available taxonomic level (e.g. for ASVs with *Lactobacillus* as the lowest available taxonomic annotation counts were summed, while ASVs with additional species-level annotation, e.g. *Lactobacillus acidophilus*, were summed at the species level instead). This taxonomic merging was performed using the tax\_glom.kv() function available in the aforementioned Github gist.

Reference:

FastQC: Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

MultiQC: Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, Volume 32, Issue 19, 1 October 2016, Pages 3047–3048, <https://doi.org/10.1093/bioinformatics/btw354>

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, *13*(7), 581

Ali Alishum. DADA2-formatted 16S rRNA gene sequences for both bacteria & archaea (Version 3). <https://zenodo.org/record/3266798#.XhxpjpMzZTY>

McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217. <https://doi.org/10.1371/journal.pone.0061217>.

Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat Methods 10:1200 –1202. <https://doi.org/10.1038/nmeth.2658>

Oksanen AJ, Blanchet FG, Kindt R, Legendre P, Minchin PR, Hara RBO, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2016. Package “vegan.” R Foundation for Statistical Computing, Vienna, Austria

Gaujoux R. 2014. Generating heatmaps for nonnegative matrix factorization. R Foundation for Statistical Computing, Vienna, Austria.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a

practical and powerful approach to multiple testing. J R Stat Soc 57:

289–300.