Microbiome: Pooled analysis bacteria

Indoor dust bacterial and fungal microbiota composition and allergic diseases: a scoping review

Javier Mancilla Galindo, MSc student
Supervisors: Inge M. Wouters and Alex Bossers
Examiner: Lidwien A.M. Smit

2024-06-18

Table of contents

Session

```
R version 4.2.2 Patched (2022-11-10 r83330)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.6 LTS

Matrix products: default

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] report_0.5.8        htmltools_0.5.5     gt_0.10.1
 [4] pals_1.7            RColorBrewer_1.1-3  phyloseq_1.42.0
 [7] vegan_2.6-4         lattice_0.22-5      permute_0.9-7
[10] lubridate_1.9.2     forcats_1.0.0       stringr_1.5.0
[13] dplyr_1.1.2         purrr_1.0.1         readr_2.1.4
[16] tidyr_1.3.0         tibble_3.2.1        ggplot2_3.4.2
[19] tidyverse_2.0.0     BiocManager_1.30.22
```

Description

A selection of studies included in the scoping review that had available raw sequence data and a comparable dust collection method were identified to retrieve sequence data to undergo processing in the same pipeline, as a proof of concept that pooled analyses could offer insight into the composition of the indoor dust microbiome and to identify potential needs and difficulties for the performance of such pooled microbiome analyses. The type of indoor environment was restricted to dwellings as this was the most common type of indoor environment. A total of **3 studies**(Amin et al., 2022; Fakunle et al., 2023; Vestergaard et al., 2018) following the electrostatic dust collector (EDC) method and **2 studies**(Adams et al., 2020; Hickman et al., 2022) using the petri dish method were selected due to the known comparability of these sample collection methods.(Adams et al., 2015) Additionally, previously unpublished data from a study sampling Dutch households with the EDC method was also included. Out of the 6 studies, 5 studies reported bacterial indoor dust sequences, whereas 4 made fungal ITS sequences publicly available.

This report includes the processing and analysis of **bacterial 16S** sequences.

Nigerian(Fakunle et al., 2023) and Dutch (unpublished) households

Dutch and Nigerian indoor dust samples follow similar methods of collection through EDC and were processed and analyzed together in the laboratory. These samples have been sequenced for the 16S rRNA combined V5-V6 hypervariable region for bacteria, and the ITS1 region between the 18S and 5.8S rRNA subunits for fungi. After processing, these were compared against the SILVA (bacteria) and UNITE datasets to determine the microbial composition of the samples.

The Dutch samples are from a cross-sectional study aiming to assess environmental factors as determinants of the indoor house dust microbial composition. The study population consisted of households located in Utrecht and surrounding areas of whose inhabitants were invited between April and May 2017 to participate in a survey in which diverse household characteristics, socioeconomic data, and health status variables were collected, as well as airborne dust samples from the household. Participants were instructed to place electrostatic dust collectors at a height of at least 1.25 meters above floor level in the main living room. Out of 600 homes selected to represent different housing characteristics, to which invitations were sent, seventy-nine ultimately provided the answered questionnaires and household dust samples. Only one adult per household provided responses to the questionnaire including their health status and was responsible for collecting and delivering the samples.

The Nigerian samples are from a case-control study of the relationship between indoor airborne dust microbiome and childhood lower respiratory tract infections.(Fakunle et al., 2023)

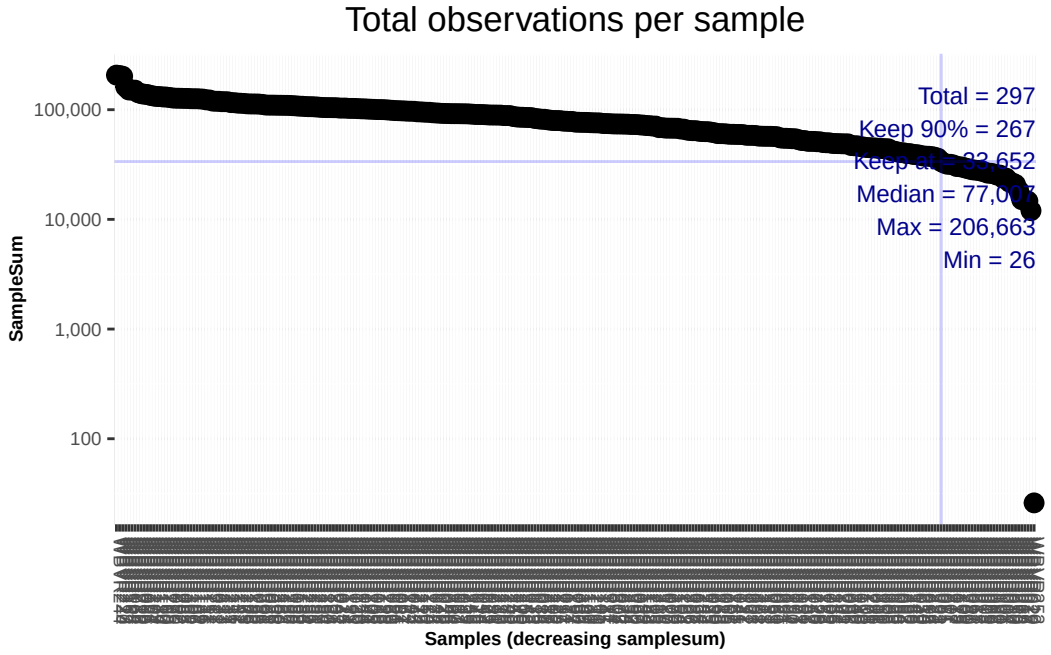
Phyloseq object containing both NL and NG samples + controls:

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 140881 taxa and 318 samples ]
sample_data() Sample Data:       [ 318 samples by 55 sample variables ]
tax_table()   Taxonomy Table:    [ 140881 taxa by 7 taxonomic ranks ]
```

Only residential households were kept, since this was similarly done with retrieved samples from studies captured in the literature review to reduce the number of samples to process. This was done under the assumption that quality control has already been done for all published and unpublished studies.

Phyloseq object containing both NL and NG samples, after removing controls:

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 140881 taxa and 297 samples ]
sample_data() Sample Data:       [ 297 samples by 55 sample variables ]
tax_table()   Taxonomy Table:    [ 140881 taxa by 7 taxonomic ranks ]
```

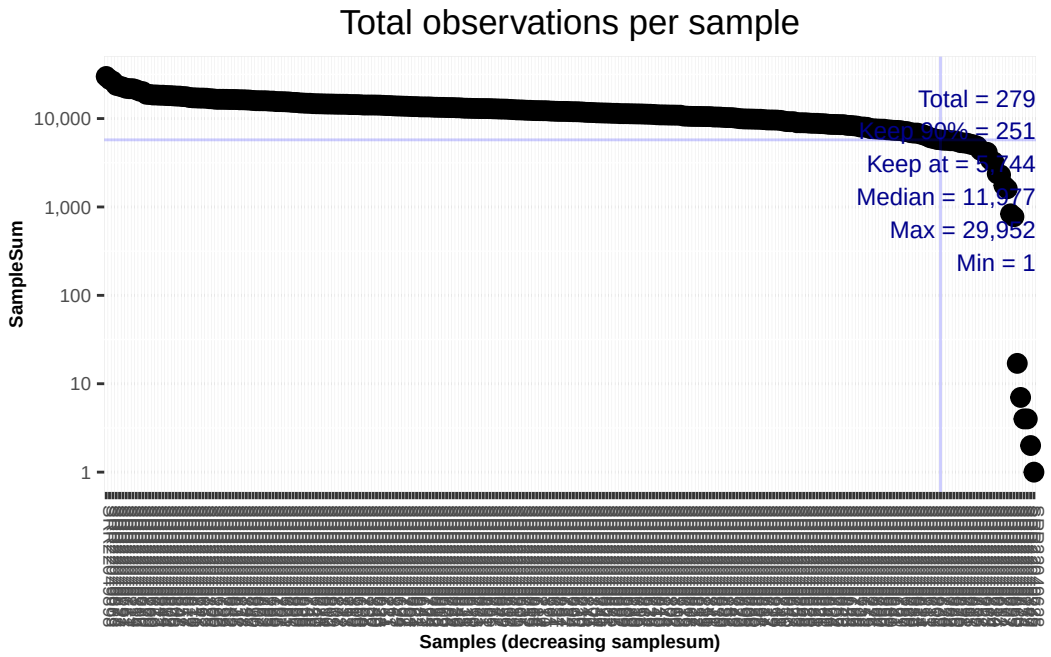


Since there is limited metadata from other studies that would allow comparison by additional environmental determinants or health outcome, this analysis will be limited to a comparison by the country of origin of the samples.

Finnish households (Hickman et al., 2022)

PRJNA892469

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 9402 taxa and 279 samples ]
sample_data() Sample Data:       [ 279 samples by 1 sample variables ]
tax_table()   Taxonomy Table:    [ 9402 taxa by 8 taxonomic ranks ]
```

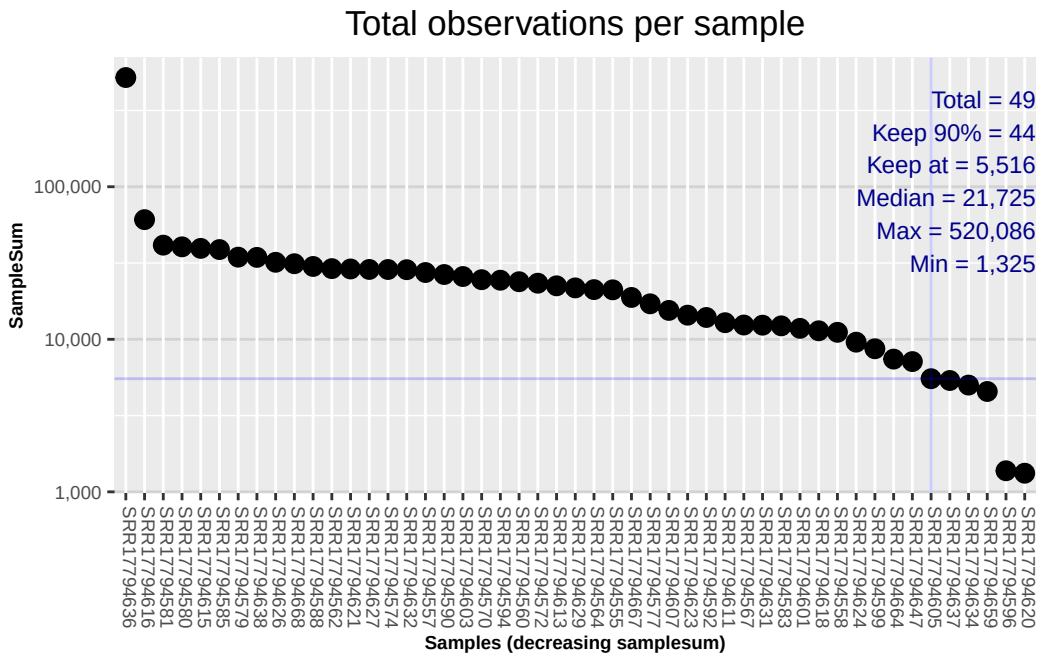


Danish households

PRJNA801418 corresponds to cow farmers' homes(Amin et al., 2022), and **PRJNA417363 / SRP124427**, to pig farmers' homes and suburban dwellings.(Vestergaard et al., 2018)

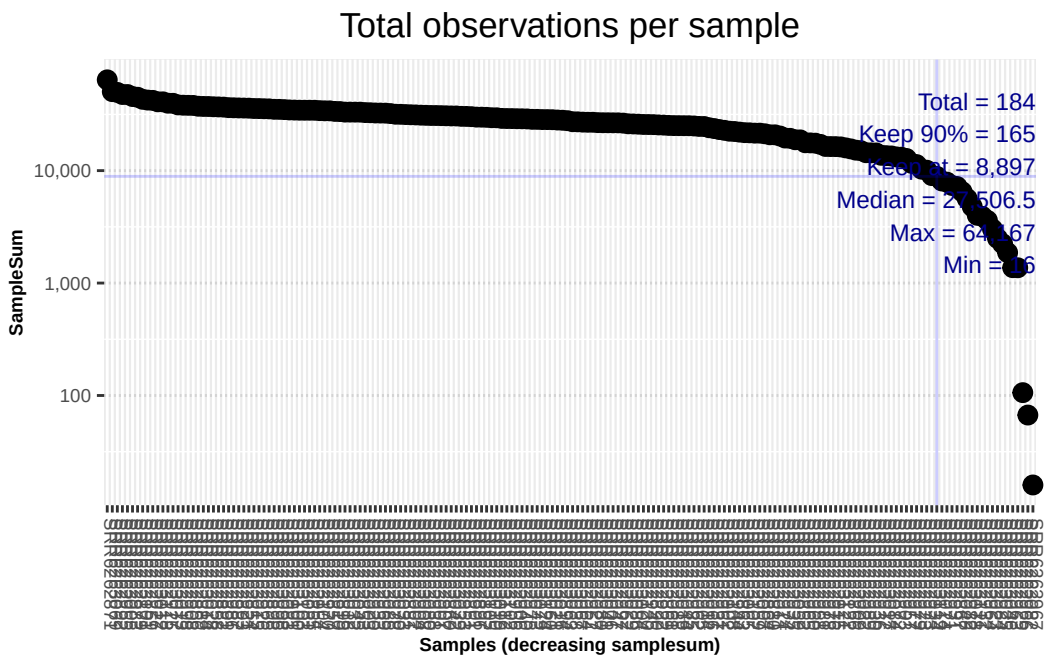
(Amin et al., 2022)

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 15604 taxa and 49 samples ]
sample_data() Sample Data:       [ 49 samples by 1 sample variables ]
tax_table()   Taxonomy Table:    [ 15604 taxa by 8 taxonomic ranks ]
```



(Vestergaard et al., 2018)

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 25737 taxa and 184 samples ]
sample_data() Sample Data:       [ 184 samples by 1 sample variables ]
tax_table()   Taxonomy Table:    [ 25737 taxa by 8 taxonomic ranks ]
```



Merge phyloseq objects

Check if rank names are equal

```
[1] "Kingdom" "Phylum"  "Class"   "Order"   "Family"  "Genus"   "Species"
```

```
[1] "Kingdom"      "Phylum"       "Class"        "Order"        "Family"
[6] "Genus"        "Species"      "GenusSpecies"
```

```
[1] "Kingdom"      "Phylum"       "Class"        "Order"        "Family"
[6] "Genus"        "Species"      "GenusSpecies"
```

```
[1] "Kingdom"      "Phylum"       "Class"        "Order"        "Family"
[6] "Genus"        "Species"      "GenusSpecies"
```

The GenusSpecies rank is missing in the NL and NG data. I will remove it and check if they are now equal:

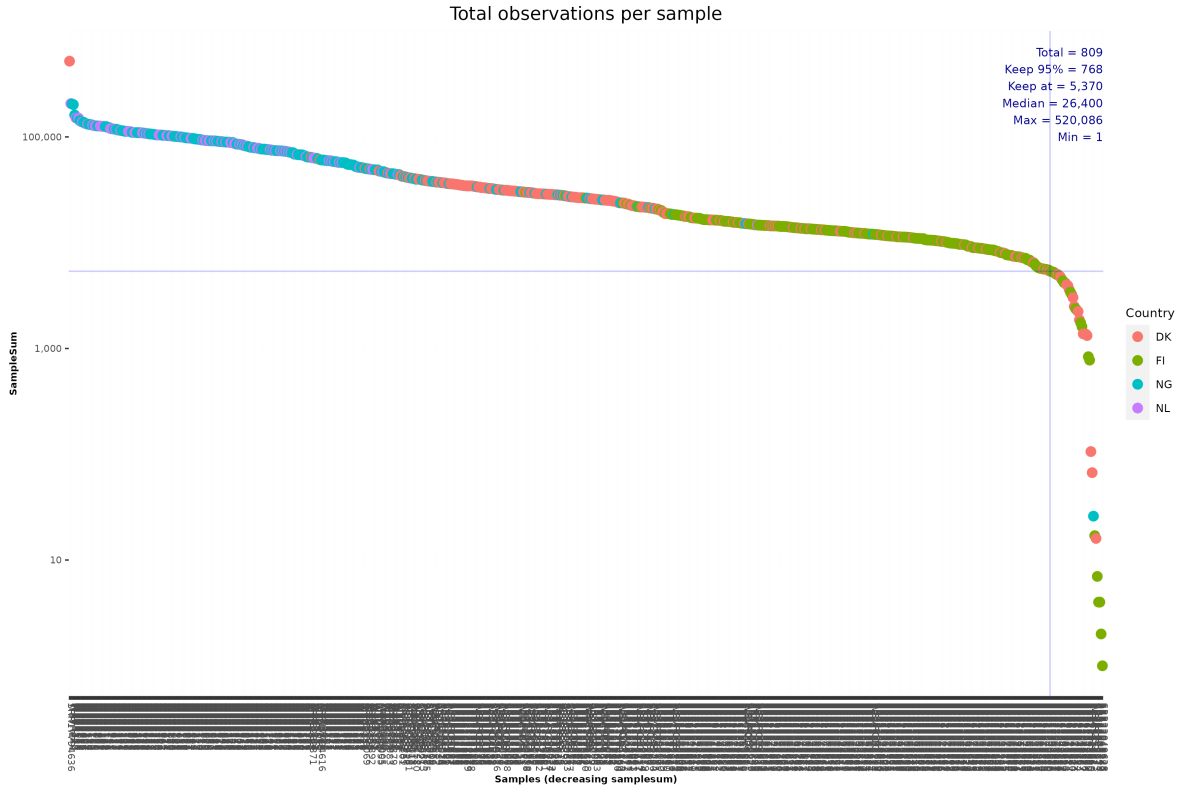
```
[1] TRUE
```

Merge into a single phyloseq object

Save unrarefied data:

```
saveRDS(
  bacter,
  paste0(psfolder,"/bacter_combined.rds")
)
```

Sample sums plot:



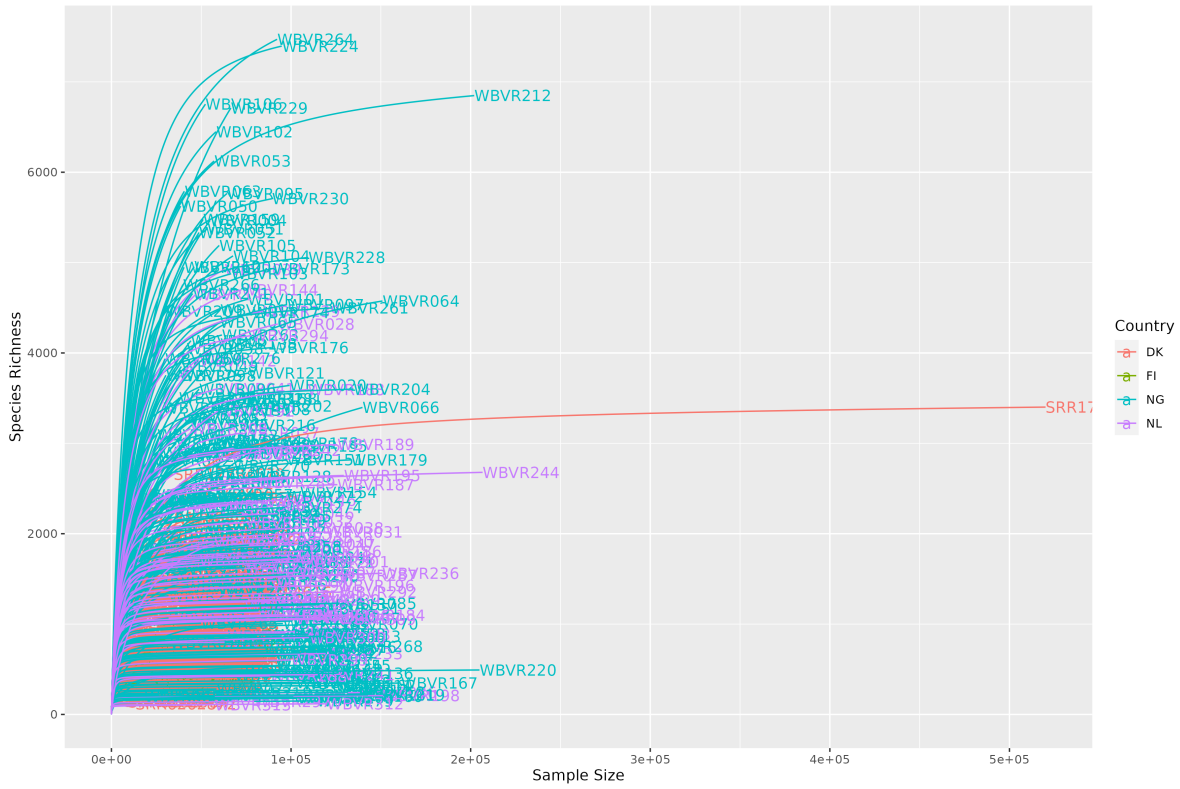
I will keep 95% of samples.

Save `bacter_keep`:

```
saveRDS(
  bacter,
  paste0(psfolder,"/bacter_keep.rds")
)
```

The construction of rarefaction curves was done with the following script:

```
source("scripts/rarefaction_plot_bacter.R")
```



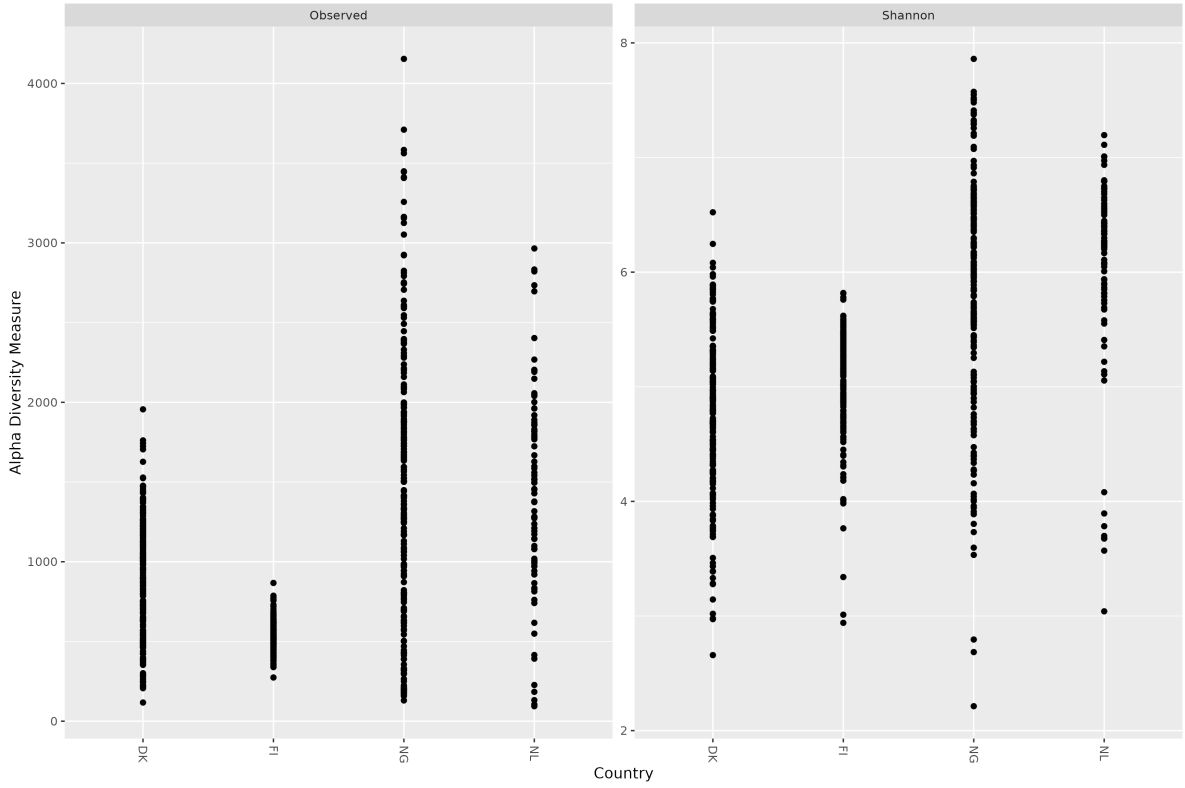
Rarefying above 10,000 leads to very large losses of samples. Thus, I will rarefy at 10,000 which leads to loosing close to ~10% (n=80) of samples.

```
bacter_rar <- rarefy_even_depth(
  bacter,
  sample.size=10000,
  rngseed=seed,
  replace=FALSE
  )

bacter_rar
```

```
80 samples removedbecause they contained fewer reads than sample.size. 53231OTUs
were removed because they are no longer present in any sample after random
subsampling
```

The number of samples is now 688.

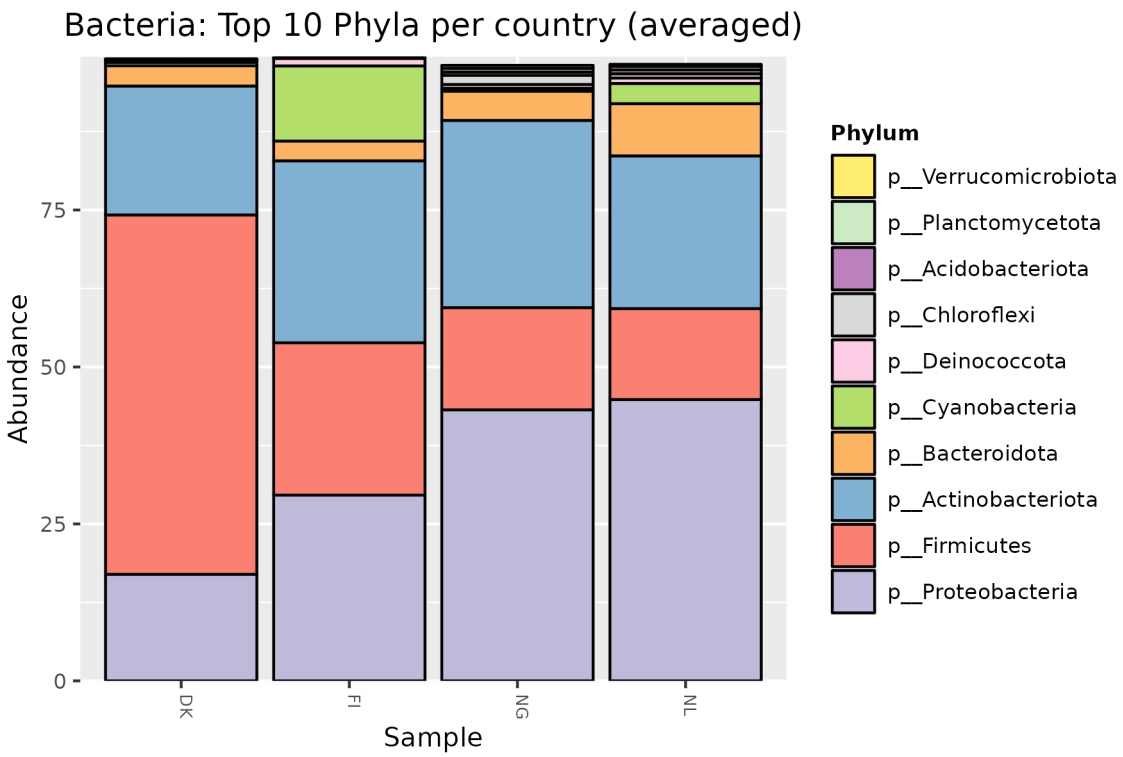
Alpha diversity

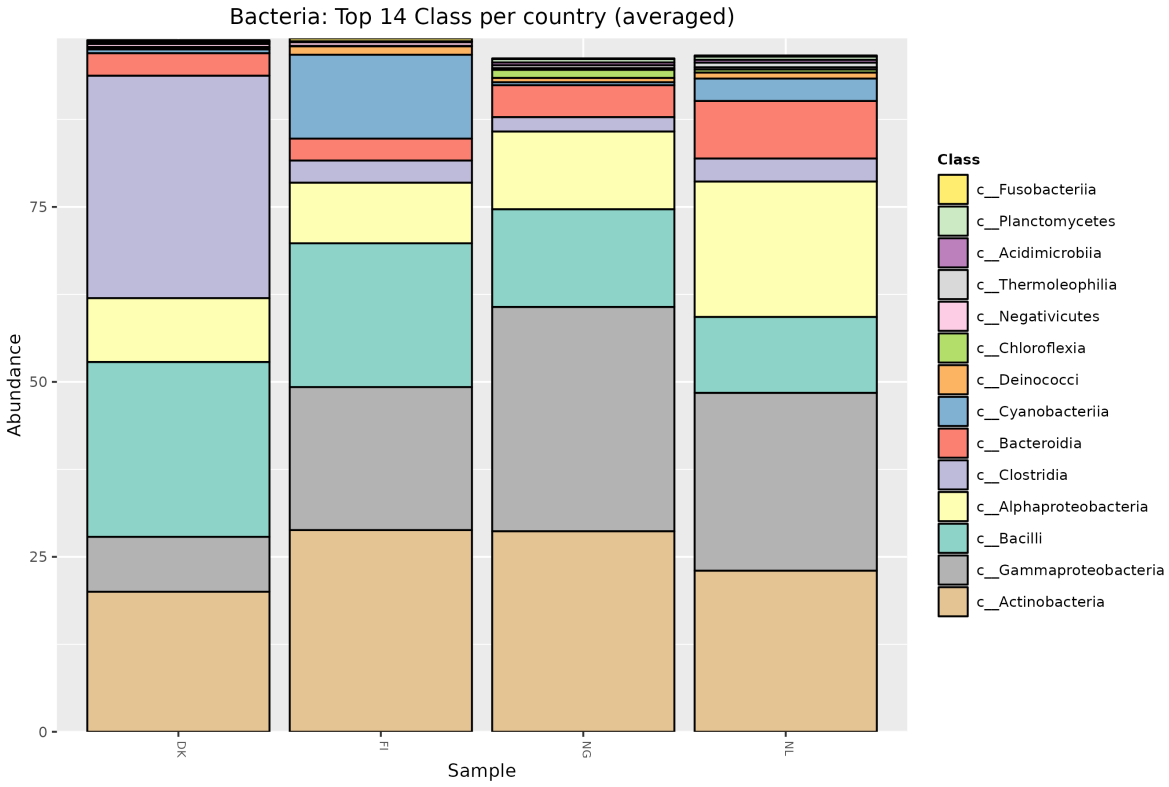


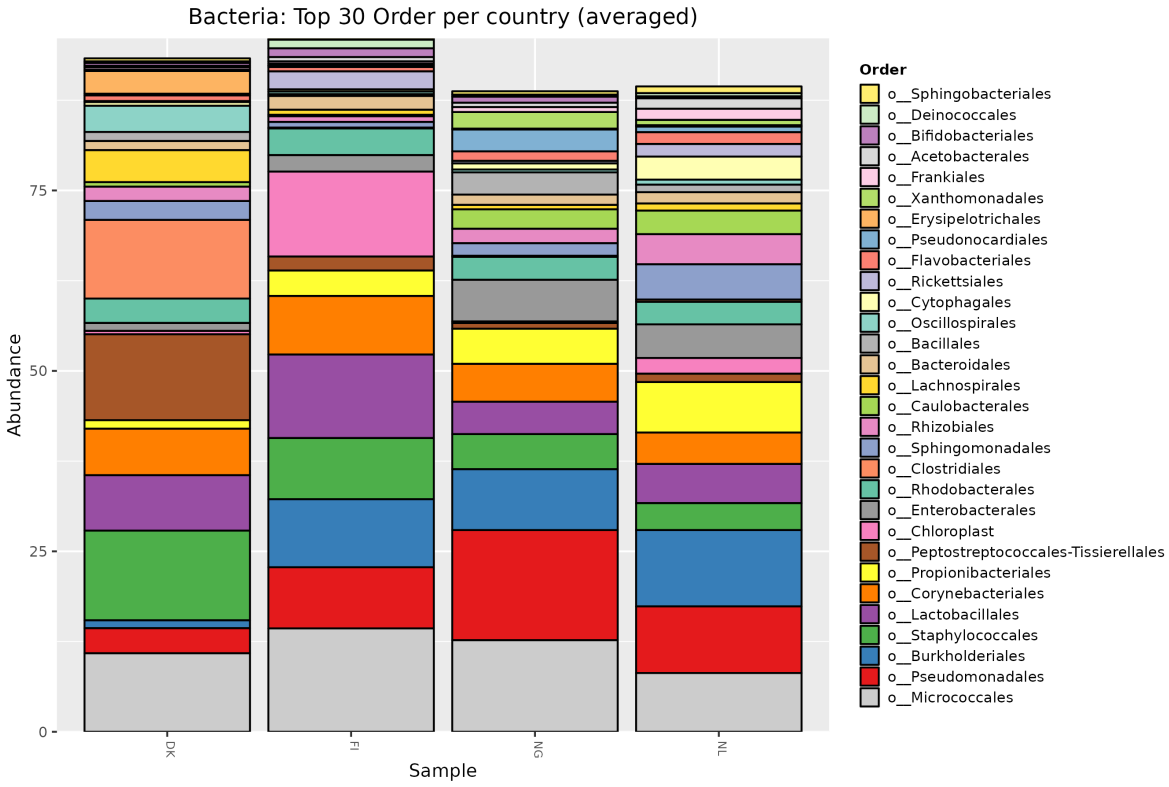
Relative abundance

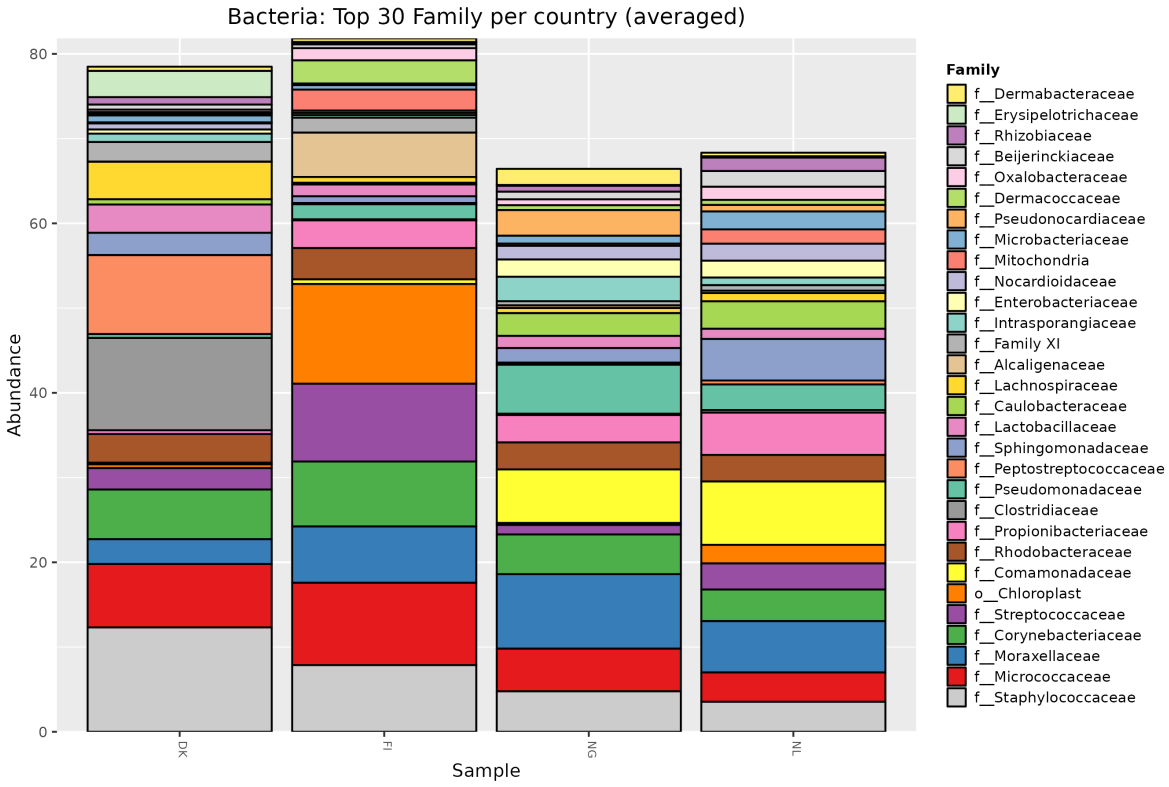
The figures were generated with the following script:

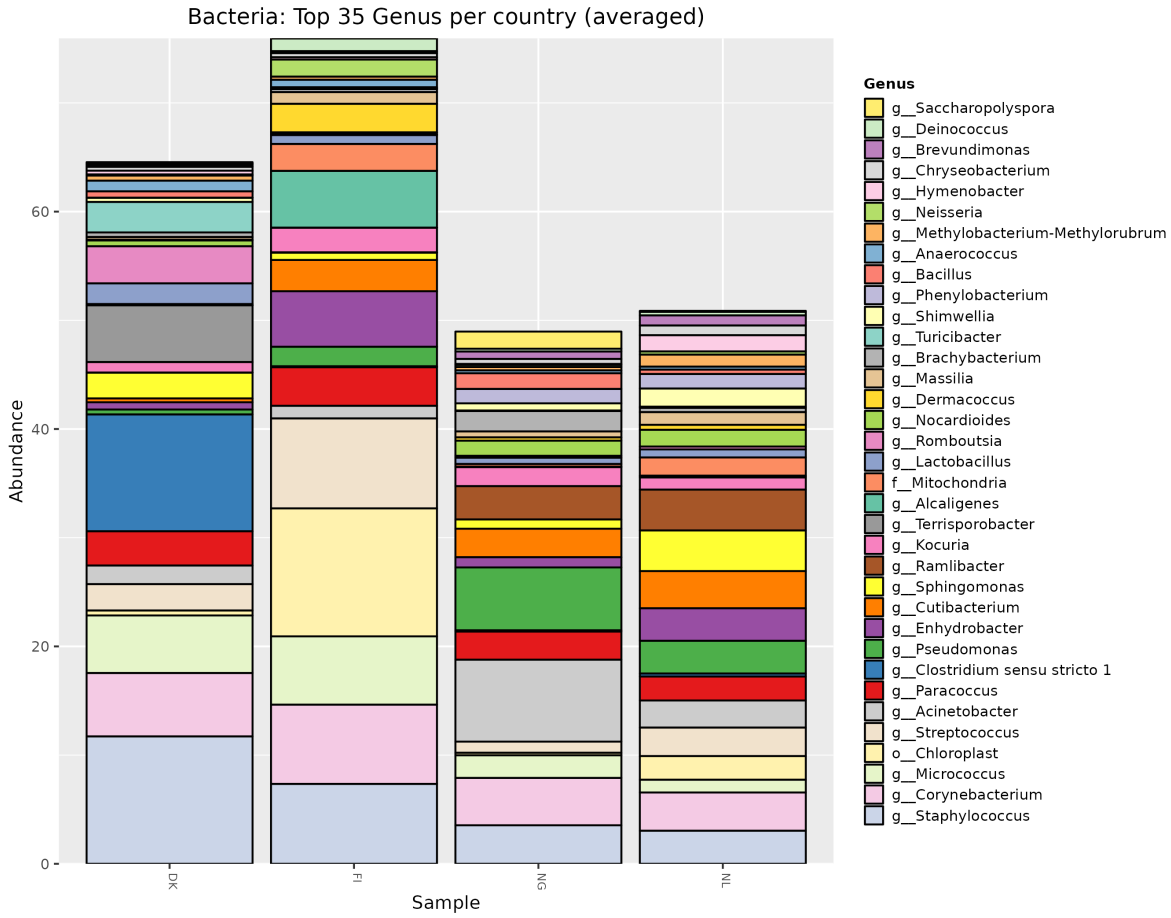
```
source("scripts/relative_abundance_plots_bacter.R")
```




Bacteria: Top 14 Class per country (averaged)
Abundance
0
25
50
75
DK
FI
NG
NL
Sample
Class
c__Fusobacteriia
c__Planctomycetes
c__Acidimicrobiia
c__Thermoleophilia
c__Negativicutes
c__Chloroflexia
c__Deinococci
c__Cyanobacteriia
c__Bacteroidia
c__Clostridia
c__Alphaproteobacteria
c__Bacilli
c__Gammaproteobacteria
c__Actinobacteria


Bacteria: Top 30 Order per country (averaged)
Abundance
0
25
50
75
DK
FI
NG
NL
Sample
Order
o__Sphingobacteriales
o__Deinococcales
o__Bifidobacteriales
o__Acetobacterales
o__Frankiales
o__Xanthomonadales
o__Erysipelotrichales
o__Pseudonocardiales
o__Flavobacteriales
o__Rickettsiales
o__Cytophagales
o__Oscillospirales
o__Bacillales
o__Bacteroidales
o__Lachnospirales
o__Caulobacterales
o__Rhizobiales
o__Sphingomonadales
o__Clostridiales
o__Rhodobacterales
o__Enterobacterales
o__Chloroplast
o__Peptostreptococcales-Tissierellales
o__Propionibacteriales
o__Corynebacteriales
o__Lactobacillales
o__Staphylococcales
o__Burkholderiales
o__Pseudomonadales
o__Micrococcales


Bacteria: Top 30 Family per country (averaged)
Abundance
80
60
40
20
0
DK
FI
NG
NL
Sample
Family
f__Dermabacteraceae
f__Erysipelotrichaceae
f__Rhizobiaceae
f__Beijerinckiaceae
f__Oxalobacteraceae
f__Dermacoccaceae
f__Pseudonocardiaceae
f__Microbacteriaceae
f__Mitochondria
f__Nocardioidaceae
f__Enterobacteriaceae
f__Intrasporangiaceae
f__Family XI
f__Alcaligenaceae
f__Lachnospiraceae
f__Caulobacteraceae
f__Lactobacillaceae
f__Sphingomonadaceae
f__Peptostreptococcaceae
f__Pseudomonadaceae
f__Clostridiaceae
f__Propionibacteriaceae
f__Rhodobacteraceae
f__Comamonadaceae
o__Chloroplast
f__Streptococcaceae
f__Corynebacteriaceae
f__Moraxellaceae
f__Micrococcaceae
f__Staphylococcaceae

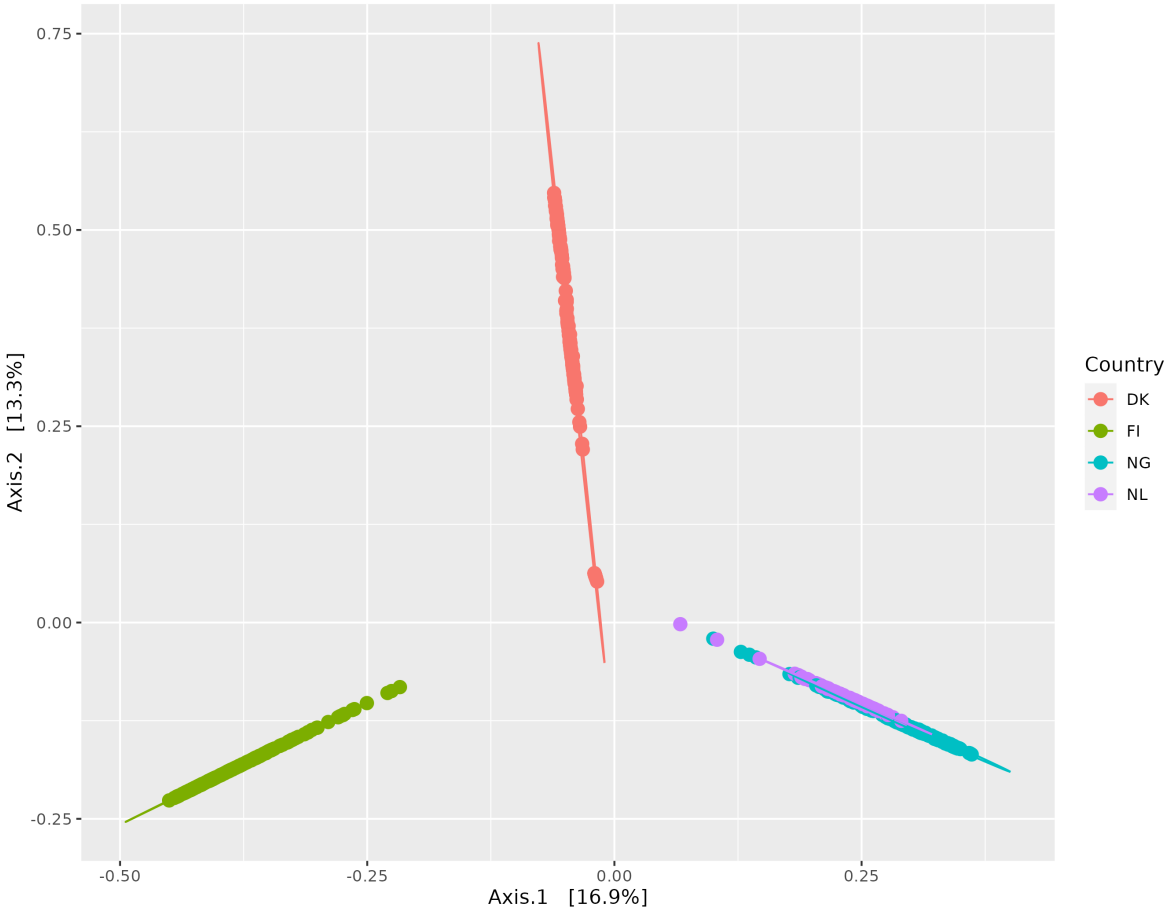

Bacteria: Top 35 Genus per country (averaged)
Abundance
0
20
40
60
DK
FI
NG
NL
Sample
Genus
g__Saccharopolyspora
g__Deinococcus
g__Brevundimonas
g__Chryseobacterium
g__Hymenobacter
g__Neisseria
g__Methylobacterium-Methylorubrum
g__Anaerococcus
g__Bacillus
g__Phenylobacterium
g__Shimwellia
g__Turicibacter
g__Brachybacterium
g__Massilia
g__Dermacoccus
g__Nocardioides
g__Romboutsia
g__Lactobacillus
f__Mitochondria
g__Alcaligenes
g__Terrisporobacter
g__Kocuria
g__Ramlibacter
g__Sphingomonas
g__Cutibacterium
g__Enhydrobacter
g__Pseudomonas
g__Clostridium sensu stricto 1
g__Paracoccus
g__Acinetobacter
g__Streptococcus
o__Chloroplast
g__Micrococcus
g__Corynebacterium
g__Staphylococcus

Beta diversity

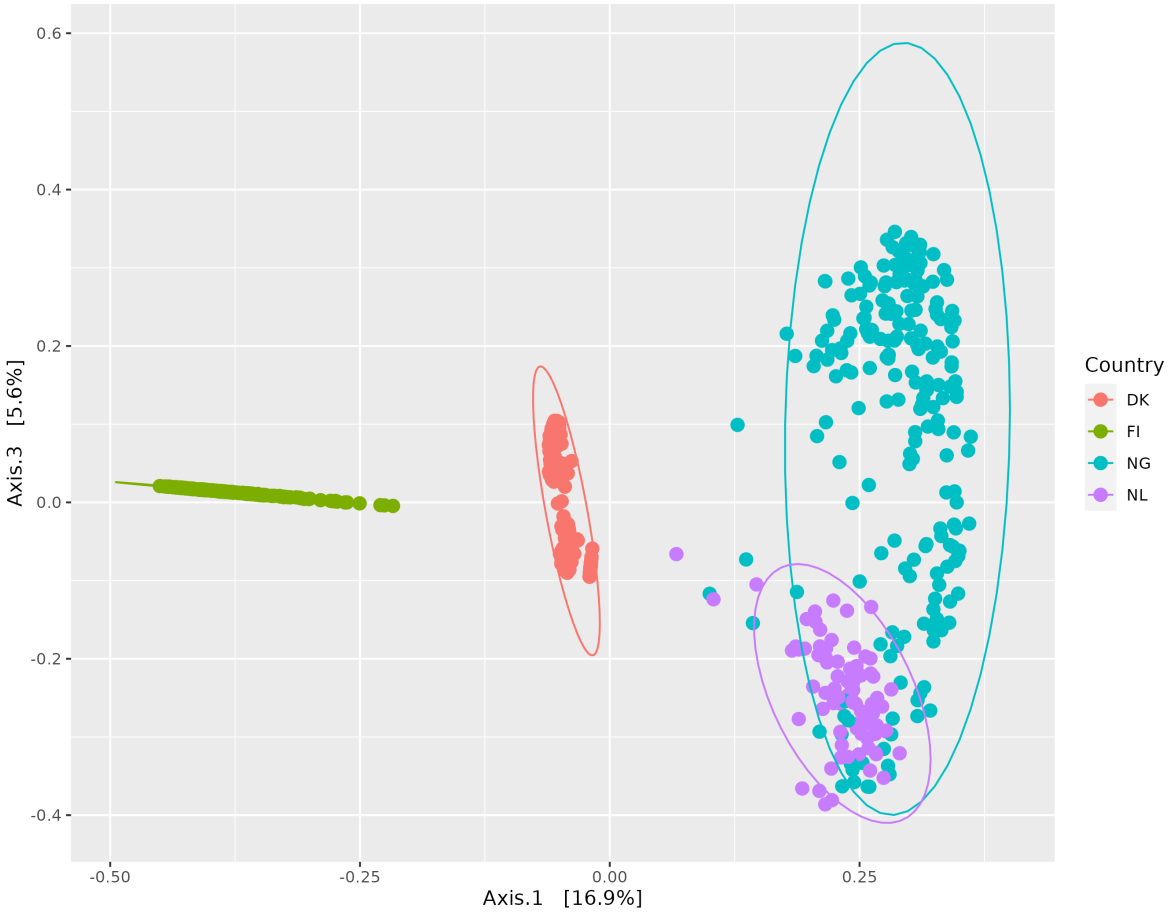
PCoA plot will be generated with the following code:

```
source("scripts/ordination_plots_fungi.R")
```

Axes 1 and 2



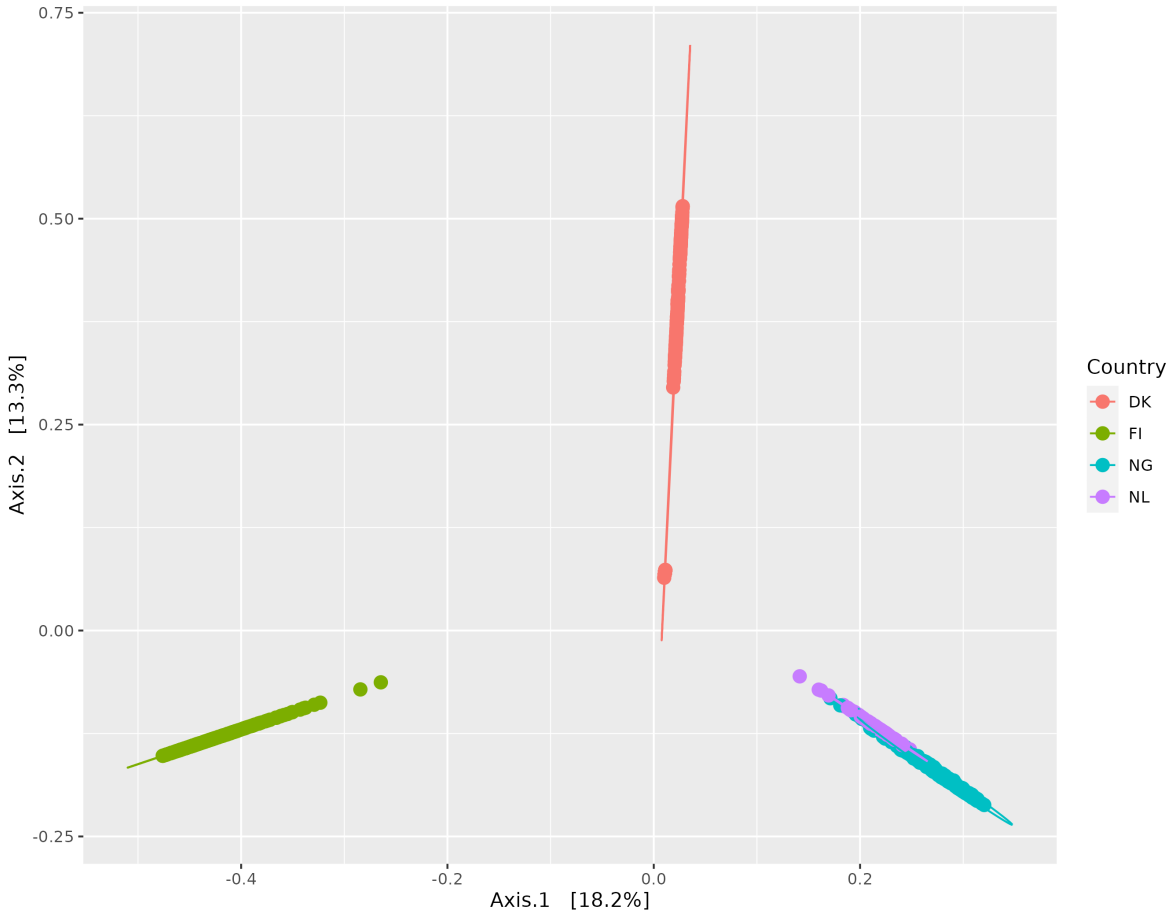
Axes 1 and 3


0.6
0.4
0.2
0.0
-0.2
-0.4
Axis.3 [5.6%]
-0.50
-0.25
0.00
0.25
Axis.1 [16.9%]
Country
DK
FI
NG
NL

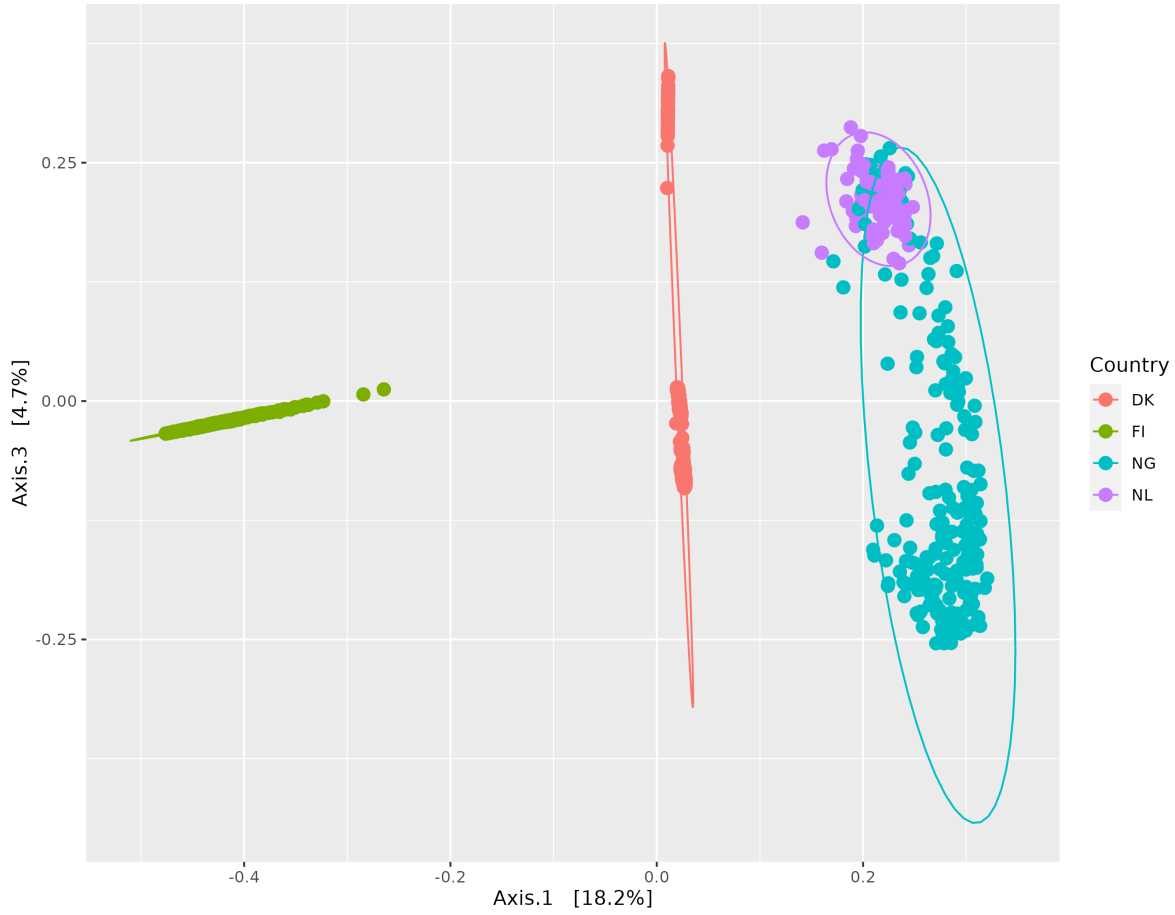
Hellinger transformed

Perhaps there is improvement with Hellinger transformation:

Axes 1 and 2



Axes 1 and 3



Permanova

Permanova done with the following code:

```
source("scripts/permanova_bacteria.R")
```

Results Permanova

Df	SumOfSqs	R2	F	Pr(>F)
3	100.3225	0.3365802	115.6738	0.001
684	197.7417	0.6634198	NA	NA
687	298.0643	1.0000000	NA	NA

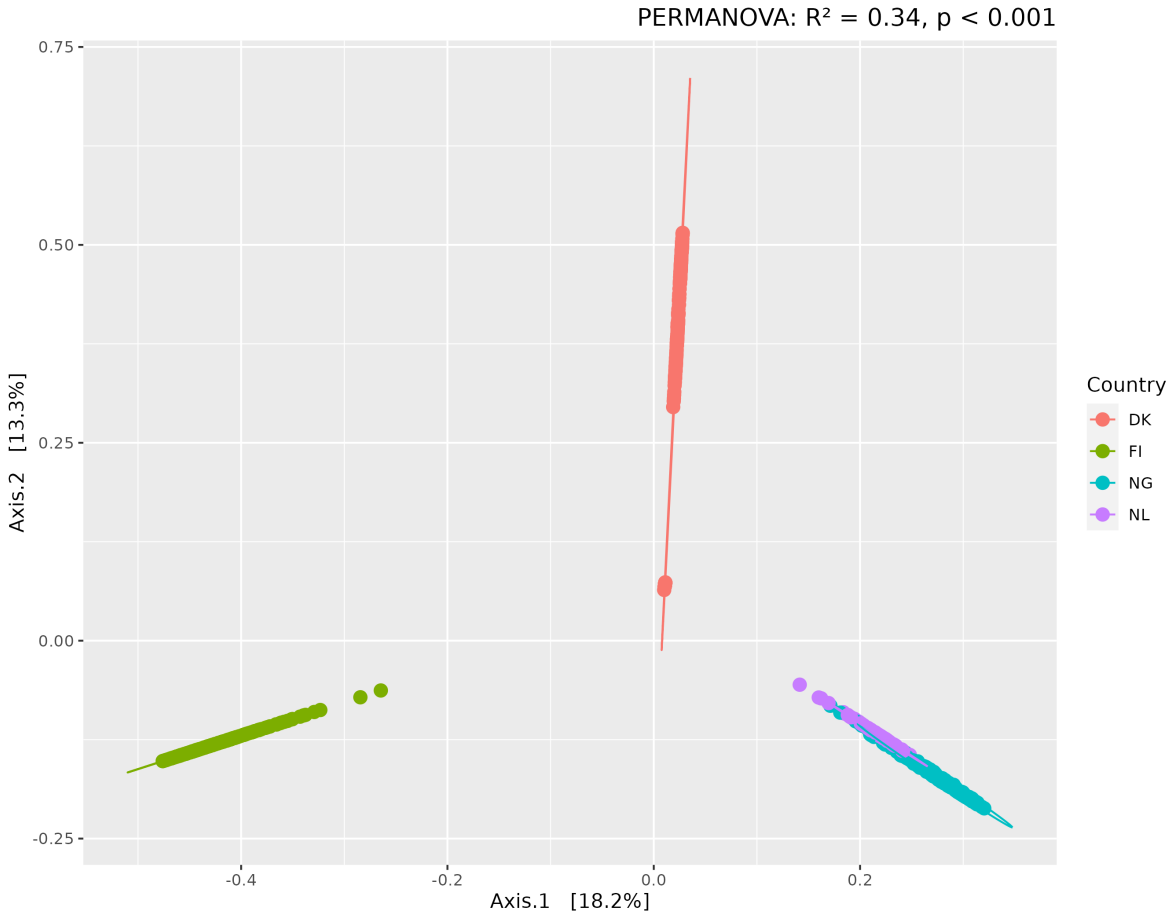
Dispersion Test

```
$Pvalues
            p.values
Country 1.153885e-65

$betadisper
$betadisper[[1]]

    Homogeneity of multivariate dispersions

Call: betadisper(d = dist.mat, group = vars[, i])

No. of Positive Eigenvalues: 685
No. of Negative Eigenvalues: 2

Average distance to median:
    DK     FI     NG     NL
0.5876 0.4569 0.5413 0.5291

Eigenvalues for PCoA axes:
(Showing 8 of 687 eigenvalues)
 PCoA1  PCoA2  PCoA3  PCoA4  PCoA5  PCoA6  PCoA7  PCoA8
54.136 39.690 14.094 12.798 10.206  7.874  3.050  2.864
```

Pairwise Permanova

Pairs	F.Model	R2	Pr(>F)	p.adj	sig
FI vs DK	129.76312	0.2496582	0.001	0.001	**
FI vs NG	167.46582	0.2945926	0.001	0.001	**
FI vs NL	109.23161	0.2872765	0.001	0.001	**
DK vs NG	112.98566	0.2148075	0.001	0.001	**
DK vs NL	67.79775	0.1932673	0.001	0.001	**
NG vs NL	37.26529	0.1124938	0.001	0.001	**

Updated figure:



Package References

- Cheng J, Sievert C, Schloerke B, Chang W, Xie Y, Allen J (2023). *htmltools: Tools for HTML*. R package version 0.5.5, https://CRAN.R-project.org/package=htmltools.
- Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate." *Journal of Statistical Software*, *40*(3), 1-25. https://www.jstatsoft.org/v40/i03/.
- Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J (2024). *gt: Easily Create Presentation-Ready Display Tables*. R package version 0.10.1, https://CRAN.R-project.org/package=gt.
- Makowski D, Lüdecke D, Patil I, Thériault R, Ben-Shachar M, Wiernik B (2023). "Automated Results Reporting as a Practical Tool to Improve Reproducibility and Methodological Best Practices Adoption." *CRAN*. https://easystats.github.io/report/.
- McMurdie PJ, Holmes S (2013). "phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data." *PLoS ONE*, *8*(4), e61217. http://dx.plos.org/10.1371/journal.pone.0061217.
- Microsoft, Weston S (2022). *foreach: Provides Foreach Looping Construct*. R package version 1.5.2, https://CRAN.R-project.org/package=foreach.
- Morgan M, Ramos M (2023). *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.22, https://CRAN.R-project.org/package=BiocManager.
- Müller K, Wickham H (2023). *tibble: Simple Data Frames*. R package version 3.2.1, https://CRAN.R-project.org/package=tibble.
- Neuwirth E (2022). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-3, https://CRAN.R-project.org/package=RColorBrewer.
- Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M, Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J (2022). *vegan: Community Ecology Package*. R package version 2.6-4, https://CRAN.R-project.org/package=vegan.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
- Sarkar D (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5, http://lmdvr.r-forge.r-project.org.
- Simpson G (2022). *permute: Functions for Generating Restricted Permutations of Data*. R package version 0.9-7, https://CRAN.R-project.org/package=permute.
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
- Wickham H (2022). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.0, https://CRAN.R-project.org/package=stringr.
- Wickham H (2023). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 1.0.0, https://CRAN.R-project.org/package=forcats.

- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.
- Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.2, https://CRAN.R-project.org/package=dplyr.
- Wickham H, Henry L (2023). *purrr: Functional Programming Tools*. R package version 1.0.1, https://CRAN.R-project.org/package=purrr.
- Wickham H, Hester J, Bryan J (2023). *readr: Read Rectangular Text Data*. R package version 2.1.4, https://CRAN.R-project.org/package=readr.
- Wickham H, Vaughan D, Girlich M (2023). *tidyr: Tidy Messy Data*. R package version 1.3.0, https://CRAN.R-project.org/package=tidyr.
- Wright K (2021). *pals: Color Palettes, Colormaps, and Tools to Evaluate Them*. R package version 1.7, https://CRAN.R-project.org/package=pals.

Study References

Adams, R.I., Sylvain, I., Spilak, M.P., Taylor, J.W., Waring, M.S., Mendell, M.J., 2020. Fungal Signature of Moisture Damage in Buildings: Identification by Targeted and Untargeted Approaches with Mycobiome Data. Applied and Environmental Microbiology 86, e01047–20. https://doi.org/10.1128/AEM.01047-20

Adams, R.I., Tian, Y., Taylor, J.W., Bruns, T.D., Hyvärinen, A., Täubel, M., 2015. Passive dust collectors for assessing airborne microbial material. Microbiome 3, 46. https://doi.org/10.1186/s40168-015-0112-7

Amin, H., Šantl-Temkiv, T., Cramer, C., Vestergaard, D.V., Holst, G.J., Elholm, G., Finster, K., Bertelsen, R.J., Schlünssen, V., Sigsgaard, T., Marshall, I.P.G., 2022. Cow Farmers' Homes Host More Diverse Airborne Bacterial Communities Than Pig Farmers' Homes and Suburban Homes. Frontiers in Microbiology 13, 883991. https://doi.org/10.3389/fmicb.2022.883991

Fakunle, A.G., Jafta, N., Bossers, A., Wouters, I.M., Kersen, W. van, Naidoo, R.N., Smit, L.A.M., 2023. Childhood lower respiratory tract infections linked to residential airborne bacterial and fungal microbiota. Environmental Research 231, 116063. https://doi.org/10.1016/j.envres.2023.116063

Hickman, B., Kirjavainen, P.V., Täubel, M., Vos, W.M. de, Salonen, A., Korpela, K., 2022. Determinants of bacterial and fungal microbiota in Finnish home dust: Impact of environmental biodiversity, pets, and occupants. Frontiers in Microbiology 13, 1011521. https://doi.org/10.3389/fmicb.2022.1011521

Vestergaard, D.V., Holst, G.J., Basinas, I., Elholm, G., Schlünssen, V., Linneberg, A., Šantl-Temkiv, T., Finster, K., Sigsgaard, T., Marshall, I.P.G., 2018. Pig Farmers' Homes Harbor More Diverse Airborne Bacterial Communities Than Pig Stables or Suburban Homes. Frontiers in Microbiology 9, 870. https://doi.org/10.3389/fmicb.2018.00870