MiDAS_course_2025

Miriam Peces, Sofie Zacho Vestergaard, Marta Nierychlo

2025-05-21, Aalborg, Denmark

> Note: This is an R markdown report see this cheat sheet for more information on how to make nice Rmarkdown documents.

Background to the project

The dataset contains timeseries data from 4 Danish Wastewater Treatment Plants (WWTPs) from year 2020 collected in the frame of MiDAS project. The following excersices aim to get you familiar with the main microbial community analysis using the `ampvis2` package, and other auxiliary packages such as `ggplot2` for nice plots and `dplyr` for data wrangling.

Install ampvis2

Installation instructions and guides for `ampvis2` can be found on the associated homepage. If you need to install it, please execute the code from the first chunk below (for windows users you need to start Rstudio as administrator).

Note: `eval=F` in the `code chunk` header means that it is not evaluated (run) when the html is build.

Load the packages

Remember to load the packages before you try to use the associated functions.

1. Load and prepare the data for analysis

1.1 Set your working directory

1.2 Load the data into R

With ampvis2 we can load all the necessary files into a single object, to make the further analysis easier and ordered. You can call it anything, but it's best to keep it short but meaningful To make the analysis easy the metadata, taxonomy and asvtable is combined into a single object `d`. You could name it however you like.

```
## Warning: Only 141 of 161 unique sample names match between metadata and otutable. The following unma
## metadata (20):
##  "MQ221006-200", "MQ221006-201", "MQ221006-202", "MQ221006-203", "MQ221006-204", "MQ221006-205", "MQ
```

Q: What does the warning mean?

1.3 Let's explore the original data sets:

ASV-table. It contains the ASV ID (e.g. "ASV15", even though the column name says OTU, it contain ASVs), which can be linked to the original DNA sequence; the sample-identifier (e.g. "MQ201118-152"); the number of reads associated to each ASV in each sample (e.g. "ASV1" is seen 1539 times in sample "MQ201118-152").

```
## # A tibble: 3 x 4
##   ‘#OTU ID‘ ‘MQ201118-152‘ ‘MQ201118-153‘ ‘MQ211213-113‘
##   <chr>              <dbl>          <dbl>          <dbl>
## 1 ASV2                1539            998           3016
## 2 ASV4                 863            533           1901
## 3 ASV5                 404            246           2204
```

We also load the metadata which contains information on each sample. Note that the **SampleID** is what connects your metadata to your ASV-table.

```
##       SampleID Line SampleContent SampleSite SampleDate
## 1 MQ201030-215   LT            AS Kalundborg 15/01/2020
## 2 MQ201030-216   LT            AS Kalundborg 24/01/2020
## 3 MQ201030-217   LT            AS Kalundborg 27/01/2020
## 4 MQ201030-218   LT            AS Kalundborg 03/02/2020
## 5 MQ201030-219   LT            AS Kalundborg 10/02/2020
## 6 MQ201030-220   LT            AS Kalundborg 17/02/2020
```

We can also check that everything in the ampvis2 object is correct, and get an overview of the object. Look at the first rows of the ASV table inside the ampvis2 object:

```
##      MQ201118-152 MQ201118-153 MQ201118-154 MQ201118-155 MQ201118-156
## ASV1          174          147          155          173          232
## ASV2         1539          998          990         2486         1542
## ASV3          430          312          477          498          514
## ASV4          863          533          452          614          597
## ASV5          404          246          257          583          355
## ASV6         1107         1067         1284         1341         1312
##      MQ201118-157 MQ201118-158 MQ201118-159 MQ201118-160 MQ201118-161
## ASV1          249          305          259          232          255
## ASV2         2476         1504         2466         1873         2146
## ASV3          554          705          720          773          854
## ASV4          561          572          658          562          571
## ASV5          533          414          638          501          599
## ASV6         1091         1450         1181         1008          920
##      MQ201118-162 MQ201118-163 MQ201118-164 MQ201118-165 MQ201118-166
## ASV1          280          326          324          285          308
## ASV2         2113         2366         2316         2423         1833
## ASV3         1267         1074         1301         1166         1478
## ASV4          690          575          487          469          514
## ASV5          638          696          898          918          839
## ASV6         1026         1057         1156          986          830
##      MQ201118-167 MQ201118-168 MQ201118-169 MQ201009-287 MQ201009-288
## ASV1          508          431          792          996          633
## ASV2          798         1059         3593          957         1354
## ASV3         1308         1905         4163          428          712
## ASV4          316          518          449          363          303
## ASV5          279          429         1677          888          797
## ASV6          561          463          399          170          101
##      MQ201009-289 MQ201009-290 MQ201009-291 MQ201009-292 MQ201009-293
## ASV1          552          406          241          241          177
## ASV2         1134         1270          718          541          464
## ASV3          666          695          817         1016         1613
## ASV4          239          274          229          179          153
## ASV5          812          709          375          290          217
## ASV6          113           73           64           68           52
##      MQ201009-294 MQ201009-295 MQ210618-86 MQ210618-87 MQ210618-88 MQ210618-89
## ASV1          251          212         615        1064         613        1541
## ASV2          400          263         547         671         469         535
## ASV3         1576         1773        1085        1640         709         477
## ASV4          142          152         159         232         309         360
## ASV5          257          171         512         750         477         520
## ASV6           57           60          79         103          79          70
##      MQ210618-90 MQ210618-91 MQ201118-151 MQ201118-170 MQ201118-171
## ASV1        1553        1939          128          485          484
## ASV2         626         507          849         1156          738
## ASV3         457         416          290         1582         1706
## ASV4         367         595          652          220          238
## ASV5         683         599          196          924          417
## ASV6          61          63         1321          387          439
##      MQ201118-172 MQ201118-173 MQ201118-174 MQ201118-175 MQ201118-176
## ASV1          647          442          391          427          374
## ASV2          862          401          234          188          169
## ASV3         2856         1363         1127         1163         1229
## ASV4          380          160          126          130          183
## ASV5          582          315          219          196          231
## ASV6          931          407          349          260          398
##      MQ201118-177 MQ201118-178 MQ201118-179 MQ201118-180 MQ201118-181
## ASV1          377          370          307          564            0
## ASV2          227          184          223          176            2
## ASV3         1097         1038         1023         1198            0
## ASV4          156          117          137          151            0
## ASV5          338          382          365          323            0
## ASV6          405          432          501          724            0
##      MQ201118-182 MQ201118-183 MQ211213-101 MQ211213-102 MQ211213-103
## ASV1           56            0         1115          934          846
## ASV2          670            1         1145         1356         1146
## ASV3          108            0         4085         4038         3220
## ASV4          143            0          226          440          364
## ASV5          282            0         2103         2660         2130
## ASV6          133            0          699          527          397
##      MQ201110-288 MQ201110-289 MQ201110-290 MQ201110-291 MQ201110-292
## ASV1         1190          957         1022          836         1305
## ASV2           21           12           18           13           29
## ASV3            0            0            0            0            0
## ASV4          326          414          295          425          423
## ASV5           87           82           56           51           87
## ASV6           40           41           36           45           47
##      MQ201110-293 MQ201110-294 MQ201110-295 MQ201110-296 MQ201110-297
## ASV1         1117         1606         2811         1176          822
## ASV2           38           56           45           65           92
## ASV3            0            0            0            0            0
## ASV4          459         1117          581          934         1044
## ASV5          133          206          195          235          245
## ASV6           31           62           35           40           39
##      MQ201110-298 MQ201110-299 MQ201110-300 MQ201110-301 MQ201110-302
## ASV1          855          585          589          276          288
## ASV2           97           92          122           62           55
## ASV3            0            0            0            0            0
## ASV4          770          706          813          400          643
## ASV5          227          214          220          143          139
## ASV6           36           38           39           16           15
##      MQ201110-303 MQ201110-304 MQ201110-305 MQ201110-306 MQ201110-307
## ASV1          274          347          286          682          357
## ASV2           55           58           46           56           44
## ASV3            0            0            0            0            0
## ASV4          519          490          516          374          498
## ASV5          128          179          180          300          220
## ASV6           13            9           17           29           19
##      MQ201110-308 MQ201030-232 MQ201030-233 MQ201030-234 MQ201030-235
## ASV1          360            0            0            1            1
## ASV2           41            0            1            0            0
## ASV3            0            0            0            0            0
## ASV4          469          147          122          270          122
## ASV5          248          122          100          146           77
## ASV6           17            8            4            8            2
##      MQ201030-236 MQ201030-237 MQ201030-238 MQ201030-239 MQ201030-240
## ASV1            0            0            0            1            0
## ASV2            0            0            0            0            0
## ASV3            0            0            0            0            0
## ASV4          200          206          289          176          192
## ASV5           73           75           77          102           77
## ASV6            5            2            2            7            1
##      MQ201030-241 MQ220601-127 MQ220601-128 MQ220601-129 MQ220601-130
## ASV1            0            0            0            0            0
## ASV2            0            0            1            0            0
## ASV3            0            0            0            0            0
## ASV4          169           57           69           58           55
## ASV5           46           14            9           15            5
## ASV6            4            1            1            2            1
##      MQ220601-131 MQ220601-132 MQ220601-133 MQ220601-134 MQ220601-135
## ASV1            0            0            0            0            0
## ASV2            0            1            0            0            1
## ASV3            0            0            0            0            0
## ASV4           63           71           65           83          116
## ASV5            4           10            5            4            2
## ASV6            2            2            3            4            2
##      MQ220601-136 MQ220601-137 MQ220601-138 MQ201009-283 MQ201009-284
## ASV1            0            0            0          444          416
## ASV2            0            0            0          542          450
## ASV3            1            0            0          180           95
## ASV4           93          167           72          451          573
## ASV5            5            0            1          692          548
## ASV6            4            2            0           83          102
##      MQ201009-285 MQ201009-286 MQ211213-114 MQ211213-115 MQ211213-116
## ASV1          512          623         1465          993          866
## ASV2          804          831         3309         2165         1888
## ASV3          308          266         1736         1104          959
## ASV4          401          384         1954          959          654
## ASV5          687          720         2257         1322          922
## ASV6          114          125         1400          944          855
##      MQ211213-117 MQ211213-118 MQ211213-119 MQ211213-120 MQ201110-279
## ASV1         1210          776          960         1156          986
## ASV2         1949         1132         2148         3033            6
## ASV3         1215         1074          786         1097            0
## ASV4          568          258         1080         2654          560
## ASV5          937          422          800          972           91
## ASV6         1137         1284          964         1214           30
##      MQ201110-280 MQ201110-281 MQ201110-282 MQ201110-283 MQ201110-284
## ASV1         1228         1325         1272         1163          816
## ASV2           13           13           10           11           16
## ASV3            0            0            0            0            0
## ASV4          557          453          369          413          255
## ASV5           84           83           86           82           62
## ASV6           36           35           37           47           44
##      MQ201110-285 MQ201110-286 MQ201110-287 MQ201030-215 MQ201030-216
## ASV1          670          812         1007            0            0
## ASV2            9           12           10           10            9
## ASV3            0            0            0            0            0
## ASV4          239          346          348          409          566
## ASV5           50           61           86          304          363
## ASV6           18           37           25           60           62
##      MQ201030-217 MQ201030-218 MQ201030-219 MQ201030-220 MQ201030-221
## ASV1            1            0            1            0            0
## ASV2            5            3            1            3            4
## ASV3            0            0            0            0            0
## ASV4          483          406          431          666          508
## ASV5          376          390          430          469          347
## ASV6           45           50           42           59           44
##      MQ201030-222 MQ201030-223 MQ201030-224 MQ201030-225 MQ201030-226
## ASV1            1            0            1            0            0
## ASV2            2            3            4            1            0
## ASV3            0            0            0            0            0
## ASV4          437          511          541          566          310
## ASV5          303          321          254          271          175
## ASV6           47           35           42           49           28
##      MQ201030-227 MQ201030-228 MQ201030-229 MQ201030-230 MQ201030-231
## ASV1            0            0            1            0            0
## ASV2            4            1            1            1            1
## ASV3            0            0            0            0            0
## ASV4          286          299          264          363          310
## ASV5          129          154          203          194          170
## ASV6           22           15           10           12            6
##      MQ211213-104 MQ211213-105 MQ211213-106 MQ211213-107 MQ211213-108
## ASV1          773          863         1136         1149          953
## ASV2          944          793         1024         1239          791
## ASV3         2238         2072         2146         2327         1806
## ASV4          406          462          473          549          338
## ASV5         1777         1548         1728         1979         1057
## ASV6          445          493          718          909          956
##      MQ211213-109 MQ211213-110 MQ211213-111 MQ211213-112 MQ211213-113
## ASV1          880          766         1248          919         1405
## ASV2          807          863         2018         1957         3016
## ASV3         1314          933         1534         1149         1825
## ASV4          303          420         1132         1179         1901
## ASV5          983         1043         1985         1550         2204
## ASV6          857          808         1181          921         1336
```

Look at the last rows of the metadata inside the ampvis2 object:

```
##                  SampleID Line SampleContent SampleSite SampleDate
## MQ211213-108 MQ211213-108   LT            AS    Randers 05/10/2020
## MQ211213-109 MQ211213-109   LT            AS    Randers 14/10/2020
## MQ211213-110 MQ211213-110   LT            AS    Randers 19/10/2020
## MQ211213-111 MQ211213-111   LT            AS    Randers 26/10/2020
## MQ211213-112 MQ211213-112   LT            AS    Randers 04/11/2020
## MQ211213-113 MQ211213-113   LT            AS    Randers 09/11/2020
```

Look at the first rows of the taxonomy table inside the ampvis2 object:

```
##          Kingdom              Phylum                  Class              Order
## ASV1 k__Bacteria p__Actinobacteriota      c__Actinobacteria   o__Micrococcales
## ASV2 k__Bacteria p__Actinobacteriota      c__Acidimicrobiia  o__Microtrichales
## ASV3 k__Bacteria      p__Chloroflexi        c__Anaerolineae        o__C10-SB1A
## ASV4 k__Bacteria       p__Firmicutes             c__Bacilli o__Lactobacillales
## ASV5 k__Bacteria p__Actinobacteriota      c__Acidimicrobiia  o__Microtrichales
## ASV6 k__Bacteria   p__Proteobacteria c__Alphaproteobacteria o__Rhodobacterales
##                     Family                 Genus                      Species
## ASV1 f__Intrasporangiaceae g__Ca_Phosphoribacter                 s__midas_s_5
## ASV2    f__Microtrichaceae      g__Ca_Microthrix  s__Ca_Microthrix_parvicella
## ASV3     f__Amarolineaceae      g__Ca_Amarolinea    s__Ca_Amarolinea_dominans
## ASV4  f__Carnobacteriaceae       g__Trichococcus                 s__midas_s_4
## ASV5    f__Microtrichaceae      g__Ca_Microthrix s__Ca_Microthrix_subdominans
## ASV6   f__Rhodobacteraceae        g__Rhodobacter
##       OTU
## ASV1 ASV1
## ASV2 ASV2
## ASV3 ASV3
## ASV4 ASV4
## ASV5 ASV5
## ASV6 ASV6
```

Q: What are the minimum and maximum number of reads in the dataset?

```
## ampvis2 object with 4 elements.
## Summary of OTU table:
##      Samples         OTUs  Total#Reads    Min#Reads    Max#Reads Median#Reads
##          141        12229      6316036          886       127366        42719
##    Avg#Reads
##     44794.58
##
## Assigned taxonomy:
##       Kingdom        Phylum         Class         Order        Family
##   12229(100%) 12159(99.43%) 12117(99.08%) 12019(98.28%) 11839(96.81%)
##         Genus       Species
## 11114(90.88%)  8897(72.75%)
##
## Metadata variables: 5
##  SampleID, Line, SampleContent, SampleSite, SampleDate
```

Q: What are the minimum and maximum number of reads in Ribe?

```
## 111 samples and 4341 OTUs have been filtered
## Before: 141 samples and 12229 OTUs
## After: 30 samples and 7888 OTUs
```

```
## ampvis2 object with 4 elements.
## Summary of OTU table:
##      Samples         OTUs  Total#Reads    Min#Reads    Max#Reads Median#Reads
##           30         7888      1138608        25903        69635      37354.5
##    Avg#Reads
##      37953.6
##
## Assigned taxonomy:
##      Kingdom       Phylum        Class        Order       Family        Genus
##   7888(100%) 7865(99.71%) 7854(99.57%)  7817(99.1%) 7742(98.15%) 7313(92.71%)
##      Species
##  5782(73.3%)
##
## Metadata variables: 5
##  SampleID, Line, SampleContent, SampleSite, SampleDate
```

1.4 Add/modify metadata

Sometimes the data types in the metadata columns are not what we want. For example, we would like *SampleDate* to be a Date column, but now is character. This can create some conflicts later on, so it's better to change upfront. Also, you can create a new column in the metadata with the `Month` information (as a character) based on the `SampleDate` column or as a factor if you'd like to have the months names.

```
## 'data.frame':    141 obs. of  5 variables:
##  $ SampleID     : chr  "MQ201118-152" "MQ201118-153" "MQ201118-154" "MQ201118-155" ...
##  $ Line         : chr  "LT" "LT" "LT" "LT" ...
##  $ SampleContent: chr  "AS" "AS" "AS" "AS" ...
##  $ SampleSite   : chr  "Randers" "Randers" "Randers" "Randers" ...
##  $ SampleDate   : chr  "07/01/2020" "17/01/2020" "21/01/2020" "29/01/2020" ...
```

Check the data types in the metadata after modifications

```
## 'data.frame':    141 obs. of  7 variables:
##  $ SampleID     : chr  "MQ201118-152" "MQ201118-153" "MQ201118-154" "MQ201118-155" ...
##  $ Line         : chr  "LT" "LT" "LT" "LT" ...
##  $ SampleContent: chr  "AS" "AS" "AS" "AS" ...
##  $ SampleSite   : chr  "Randers" "Randers" "Randers" "Randers" ...
##  $ SampleDate   : Date, format: "2020-01-07" "2020-01-17" ...
##  $ Month        : chr  "01" "01" "01" "01" ...
##  $ MonthName    : Factor w/ 12 levels "January","February",..: 1 1 1 1 2 2 2 2 3 3 ...
```

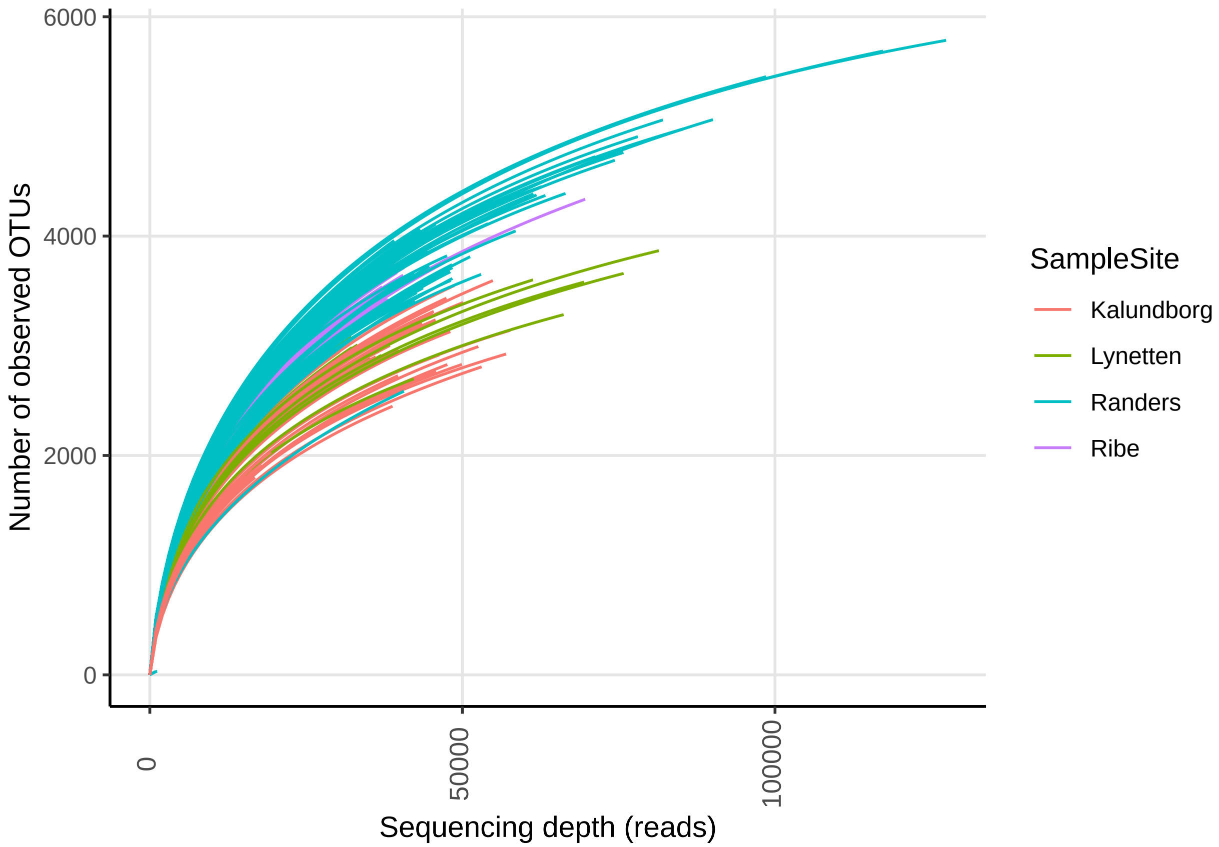
Remove unnecessary files

Basic QC analysis

Evaluation of negative controls: as we often work with tiny amounts of DNA contamination often occur. This could be from other samples, yourself and even the kits/reagents we use. Hence, it is important to take a critical look at the negative controls compared to the real samples. However, in the interest of time you can assume that problematic samples have been removed from the data set.

1. Rarefaction curves

The goal of this analysis is to evaluate if we have sequenced enough reads pr. sample to represent the diversity in the samples. This is often a subjective decision. For every sample, we take 1 read at a time, and evaluate if this belongs to an ASV we have already observed, or if this read represents new diversity (and ASV that has not been observed before). Every time we evaluate a new read we move 1 point on the x-axis and if it is a new ASV we also move 1 point up on the y-axis. When the curve is steep we discover new ASVs often, indicating that we need to sequence more reads to capture the diversity in the sample. When the curve flattens, we rarely observe new ASVs, indicating that we have captured most of the diversity in the sample. Often you have to compromise with the number of reads in order to keep more samples in your analysis.

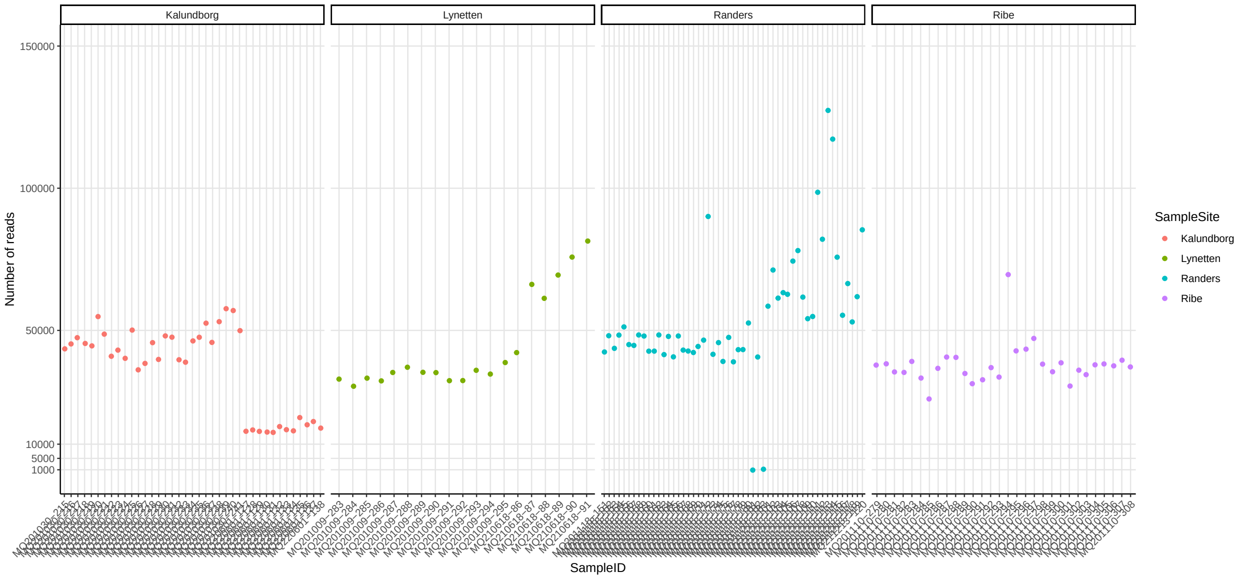


The function `amp_alphadiv` takes the metadata and appends the number of reads and ASVs in each sample which then can be used for further analysis.

```
##                    SampleID Line SampleContent SampleSite SampleDate Month
## MQ201118-181 MQ201118-181   LT            AS    Randers 2020-07-27    07
## MQ201118-183 MQ201118-183   LT            AS    Randers 2020-08-11    08
## MQ220601-131 MQ220601-131   LT            AS Kalundborg 2020-10-16    10
## MQ220601-130 MQ220601-130   LT            AS Kalundborg 2020-10-05    10
## MQ220601-129 MQ220601-129   LT            AS Kalundborg 2020-09-30    09
## MQ220601-127 MQ220601-127   LT            AS Kalundborg 2020-09-09    09
##              MonthName Reads uniqueOTUs  Shannon   Simpson invSimpson
## MQ201118-181      July   886         30 2.864474 0.9282824  13.943586
## MQ201118-183    August  1188         31 2.387568 0.8009968   5.025044
## MQ220601-131   October 14136       1776 5.818483 0.9879157  82.752046
## MQ220601-130   October 14258       1771 5.825115 0.9874007  79.369266
## MQ220601-129 September 14497       1689 5.796092 0.9871528  77.838007
## MQ220601-127 September 14532       1743 5.865044 0.9888748  89.885752
```

2. Check the number of reads produced pr. sample.

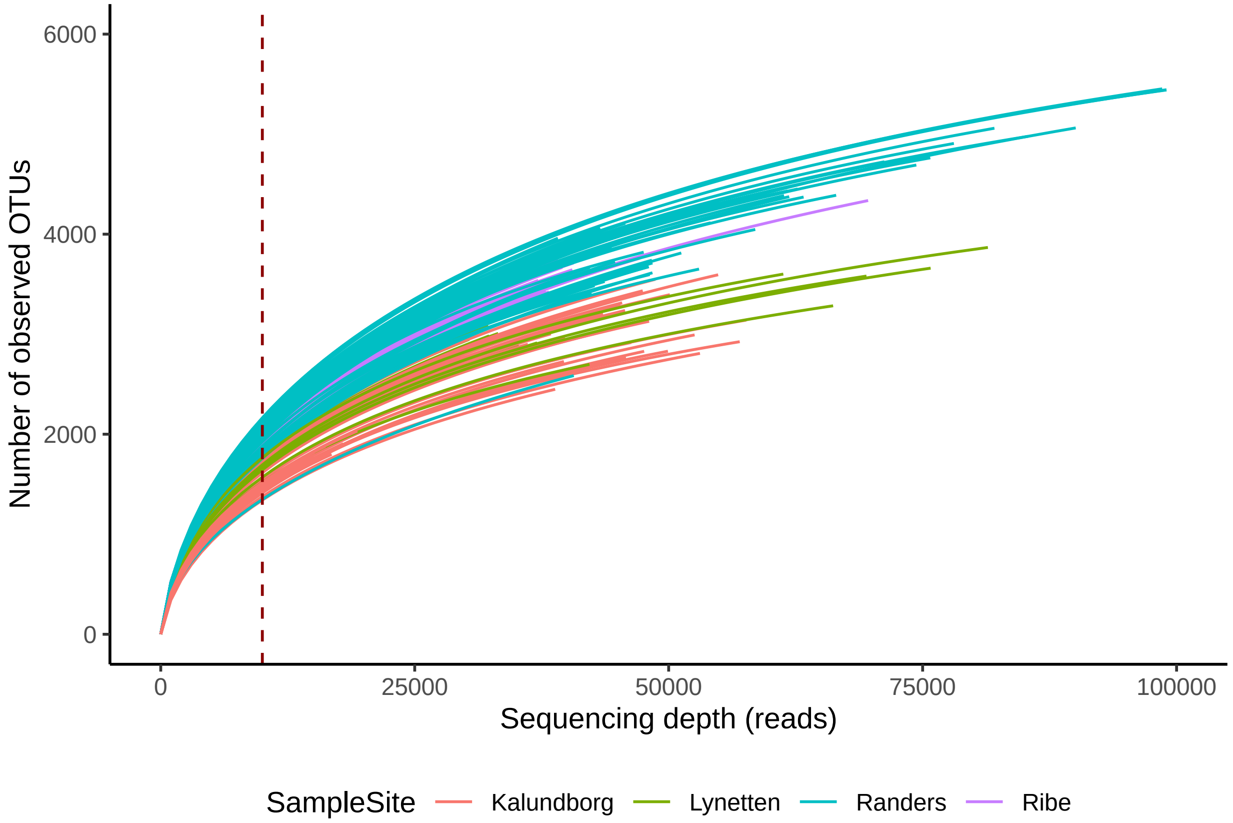
Note: You can use "+" to modify all ampvis2 plots as behind the surface they are just ggplot2 objects. Here we change a number of features (e.g. y axis title or x axis label position).



3. Subset to a miminum number of reads per sample

After we have decided that we don't trust that samples with less than 10000 we remove them from our analysis. We store the subset in the object "ds_midas".

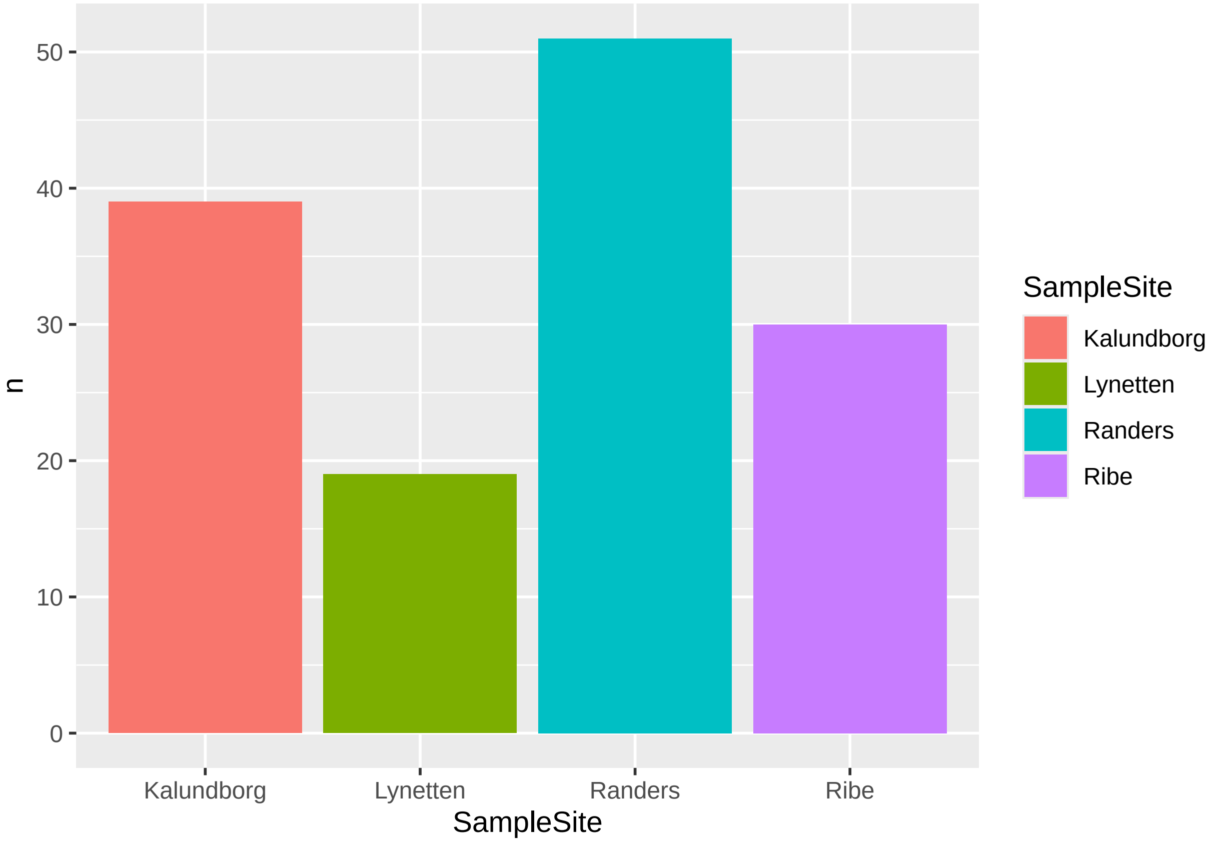
```
## 2 samples and 11 OTUs have been filtered
## Before: 141 samples and 12229 OTUs
## After: 139 samples and 12218 OTUs
```

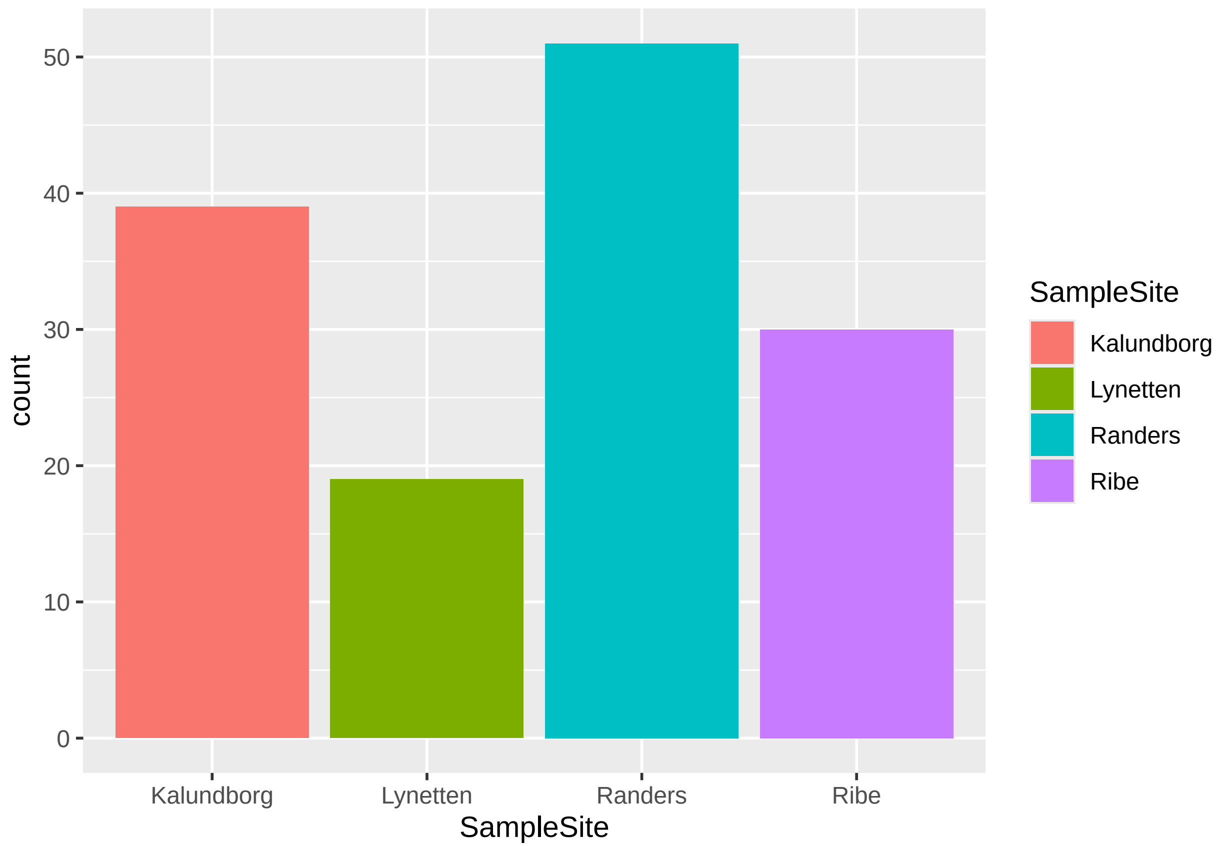


4. Count the number of samples per plant

Use the function `count()` from the tidyverse package to summarize how many samples were taken at each WWTP as a simple table. You could also visualize it using e.g. `geom_col()` or `geom_bar()` from the ggplot2 package.

```
## # A tibble: 4 x 2
## # Groups:   SampleSite [4]
##   SampleSite     n
##   <chr>      <int>
## 1 Kalundborg    39
## 2 Lynetten      19
## 3 Randers       51
## 4 Ribe          30
```


50
40
30
20
10
0
n
Kalundborg
Lynetten
Randers
Ribe
SampleSite
SampleSite
Kalundborg
Lynetten
Randers
Ribe



5. Rarefy

For some analyses it is preferable to rarefy the dataset, in other words standardise sequencing depth across samples, to make fair comparisons. It is a topic that gets highly debated in literature since it produces a "data-loss", but in general it is advisable to use it when performing alpha diversity comparisons.

```
## Warning: The chosen rarefy size (10000) is smaller than the smallest amount of
## reads in any sample (14136).
```

```
## 0 samples have been filtered.
```

Q: Should we worry about the warning?

6. Normalise

For some analyses we may want to have our ampvis2 object to normalise the ASv read counts to 100 (relative abundance). Many ampvis functions have the option to normalise when calling it.

```
## 0 samples have been filtered.
```

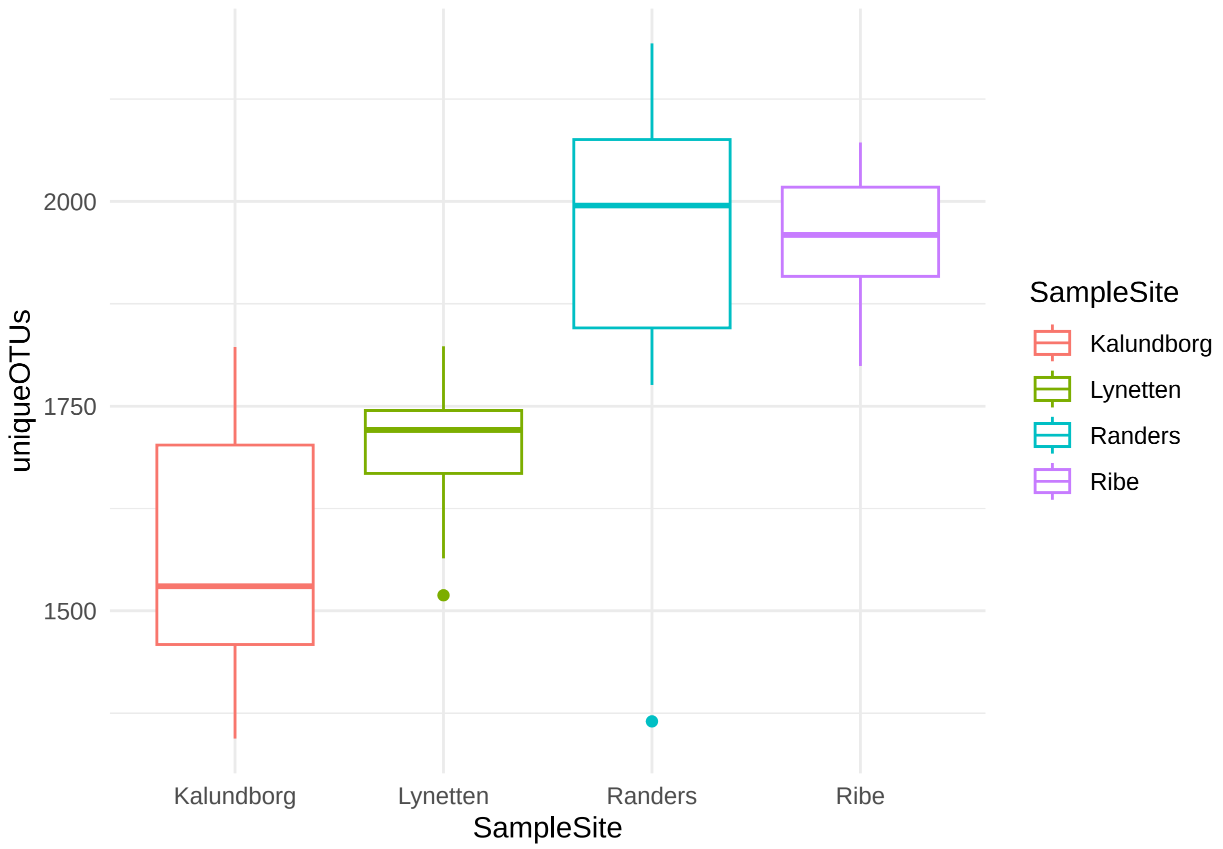
Data analysis

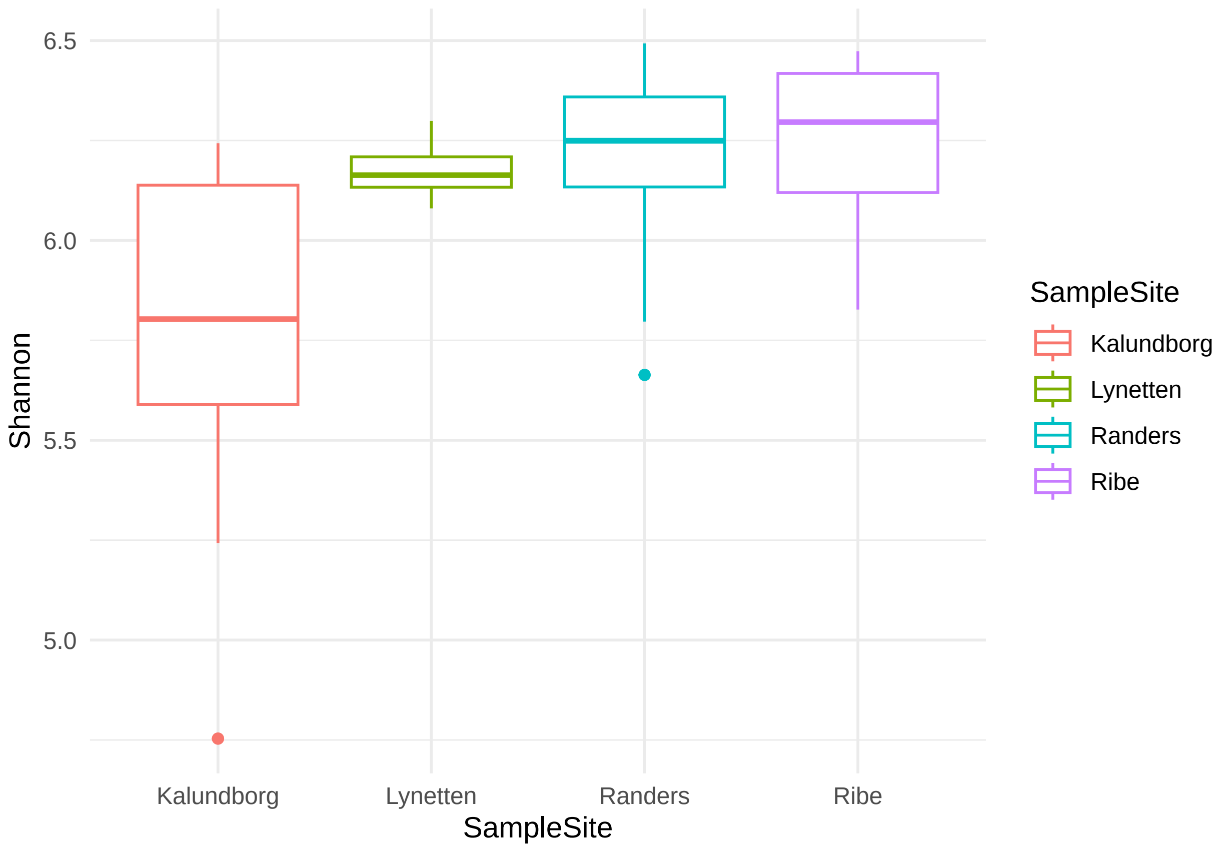
1. Alpha diversity

In microbial community analysis we often also quantify the diversity of the community in any single sample. This can be quantified with a single number that takes into account the number and abundance of the individual taxa. There are numerous ways of calculating these `diversity indices`, and many can be calculated using the `amp_alpha_diversity()` function. The results are appended to the end of metadata as simple columns.

```
##                  SampleID Line SampleContent SampleSite SampleDate Month
## MQ201118-152 MQ201118-152   LT            AS    Randers 2020-01-07    01
## MQ201118-153 MQ201118-153   LT            AS    Randers 2020-01-17    01
## MQ201118-154 MQ201118-154   LT            AS    Randers 2020-01-21    01
## MQ201118-155 MQ201118-155   LT            AS    Randers 2020-01-29    01
## MQ201118-156 MQ201118-156   LT            AS    Randers 2020-02-06    02
## MQ201118-157 MQ201118-157   LT            AS    Randers 2020-02-14    02
##              MonthName Reads uniqueOTUs  Shannon   Simpson invSimpson
## MQ201118-152   January 10000       1827 6.198710 0.9932242   147.5832
## MQ201118-153   January 10000       1872 6.196973 0.9928831   140.5098
## MQ201118-154   January 10000       1882 6.109789 0.9916055   119.1262
## MQ201118-155   January 10000       1856 6.126986 0.9917604   121.3657
## MQ201118-156  February 10000       1797 6.125199 0.9923990   131.5613
## MQ201118-157  February 10000       1915 6.161394 0.9916501   119.7613
```

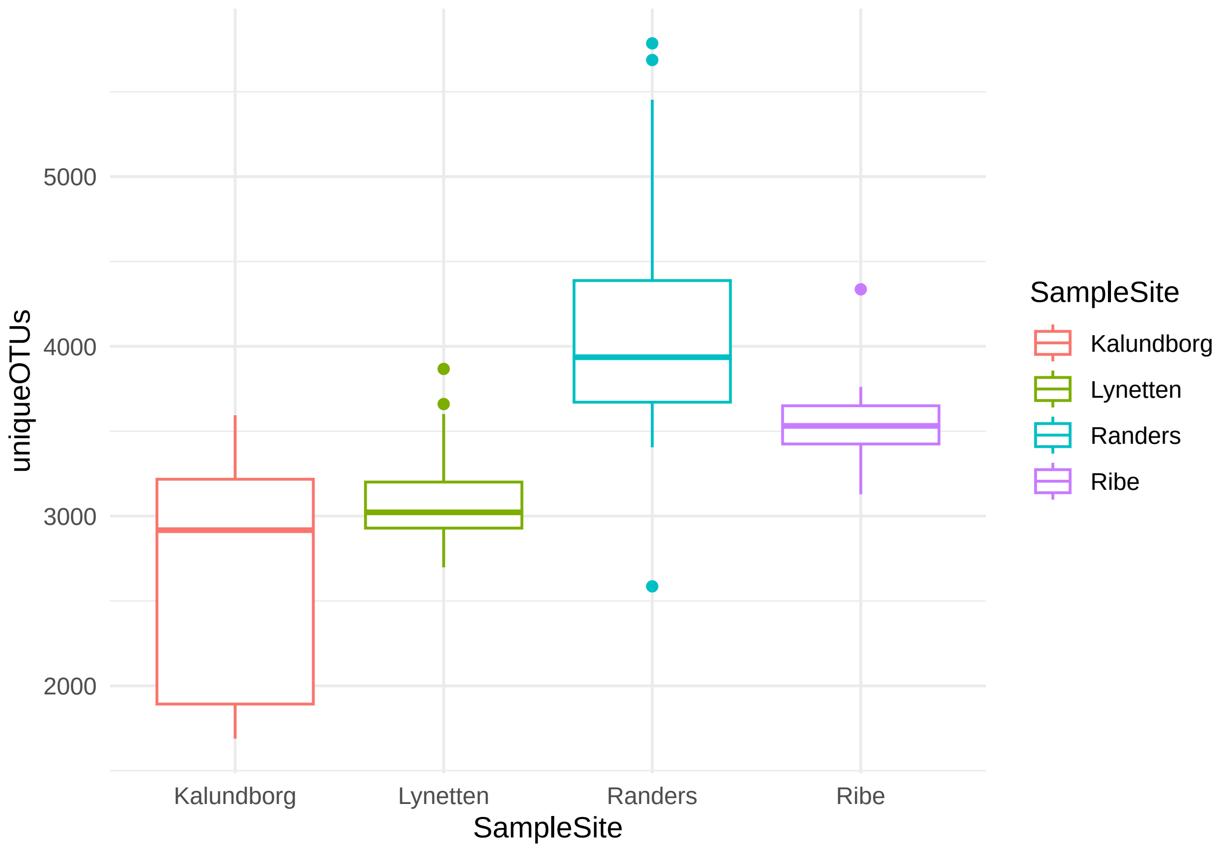
Plot the species richness - `ObservedOTUs` and Shannon diversity of the `alpha` data set using the `geom_boxplot()` from the ggplot2 package. See e.g. [this example](http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization. In which plant the microbial community is least diverse?

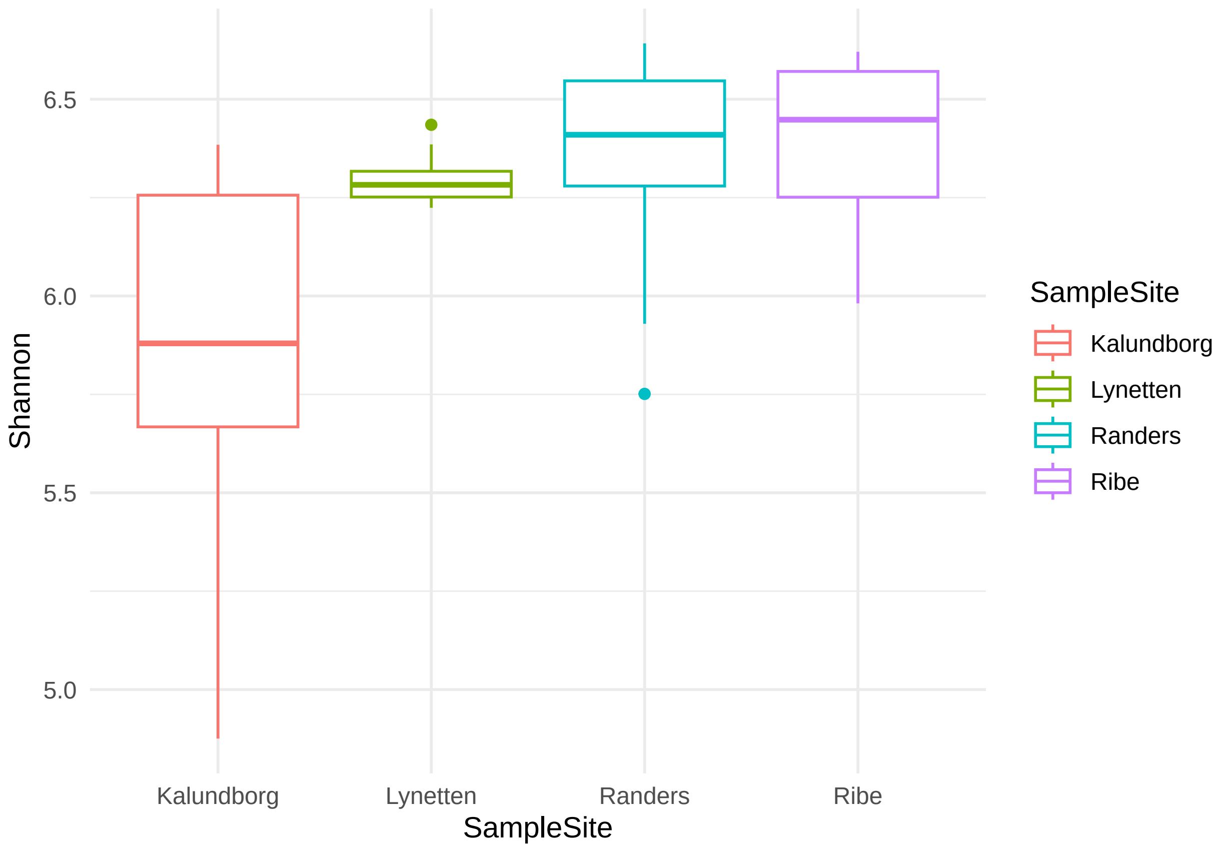

uniqueOTUs
2000
1750
1500
Kalundborg
Lynetten
Randers
Ribe
SampleSite
SampleSite
Kalundborg
Lynetten
Randers
Ribe



1.1 Alpha-diversity additional tasks

Compare the alpha diversity results using the non-rarefied dataset. What are your observations?


5000
4000
3000
2000
uniqueOTUs
Kalundborg
Lynetten
Randers
Ribe
SampleSite
SampleSite
Kalundborg
Lynetten
Randers
Ribe



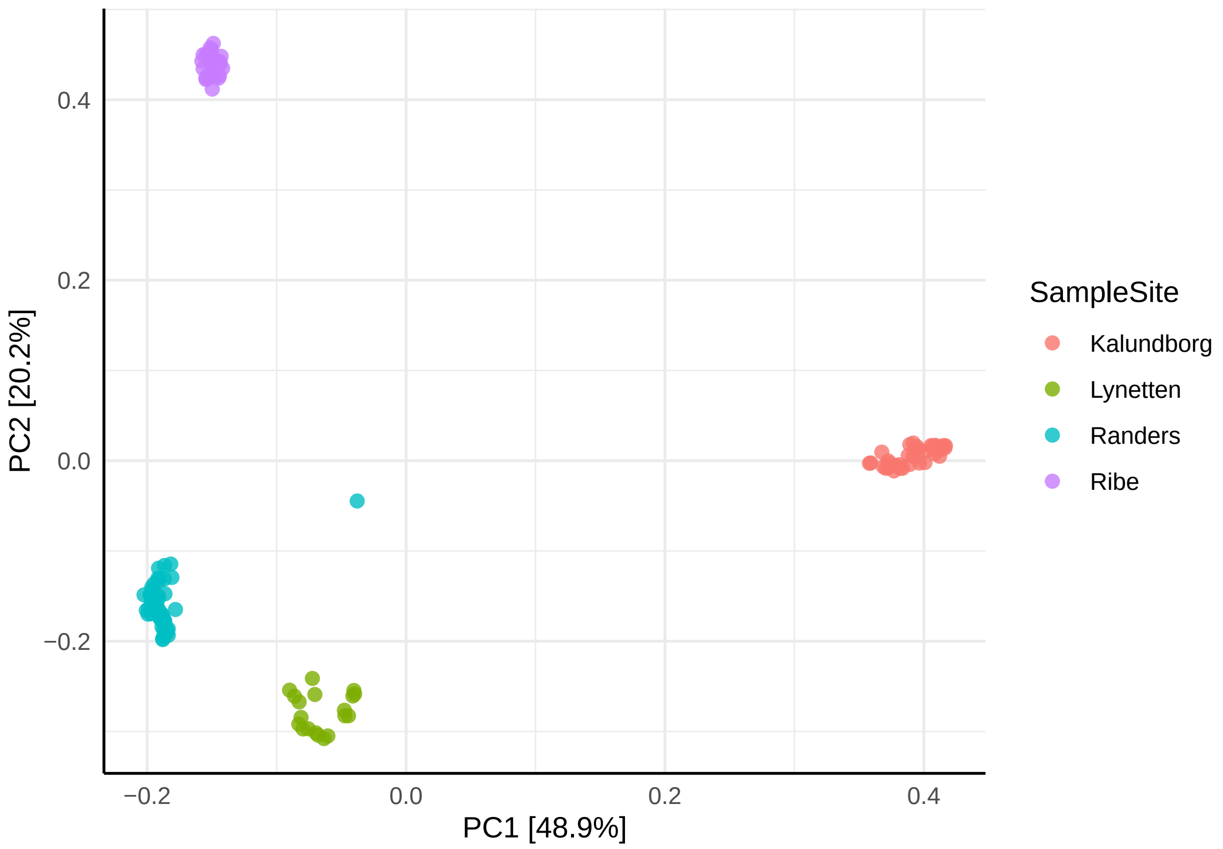
2. Beta diversity

When we are comparing between samples we call it beta-diversity. One of the most common ways or comparing large data sets and identify similarities and differences are using ordination. See this guide for an introduction to the topic. Ordination is trying to show you the largest differences between samples. In ordination we take the ASV table with 1000's of bacteria and try to visualize which samples have similar microbial communities. Samples (colored dots) located close together have similar microbial communities, while samples located far apart have different microbial communities. There are many versions of the ordination. One of the most simple and commonly used is `PCA` where the raw ASV counts are often transformed using `hellinger` transformation that takes the square root of the relative abundance. See this guide for short intro on `Hellinger` and other data transformations. In addition to transforming the data, different types of ordination can be made (PCoA or NMDS are also often used).

2.1 Perform a PCA

Q: What can you say about the similarity of microbial communities in the 4 WWTPs based on PCA plot?

```
## 10898 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before: 12218 OTUs
## After: 1320 OTUs
```

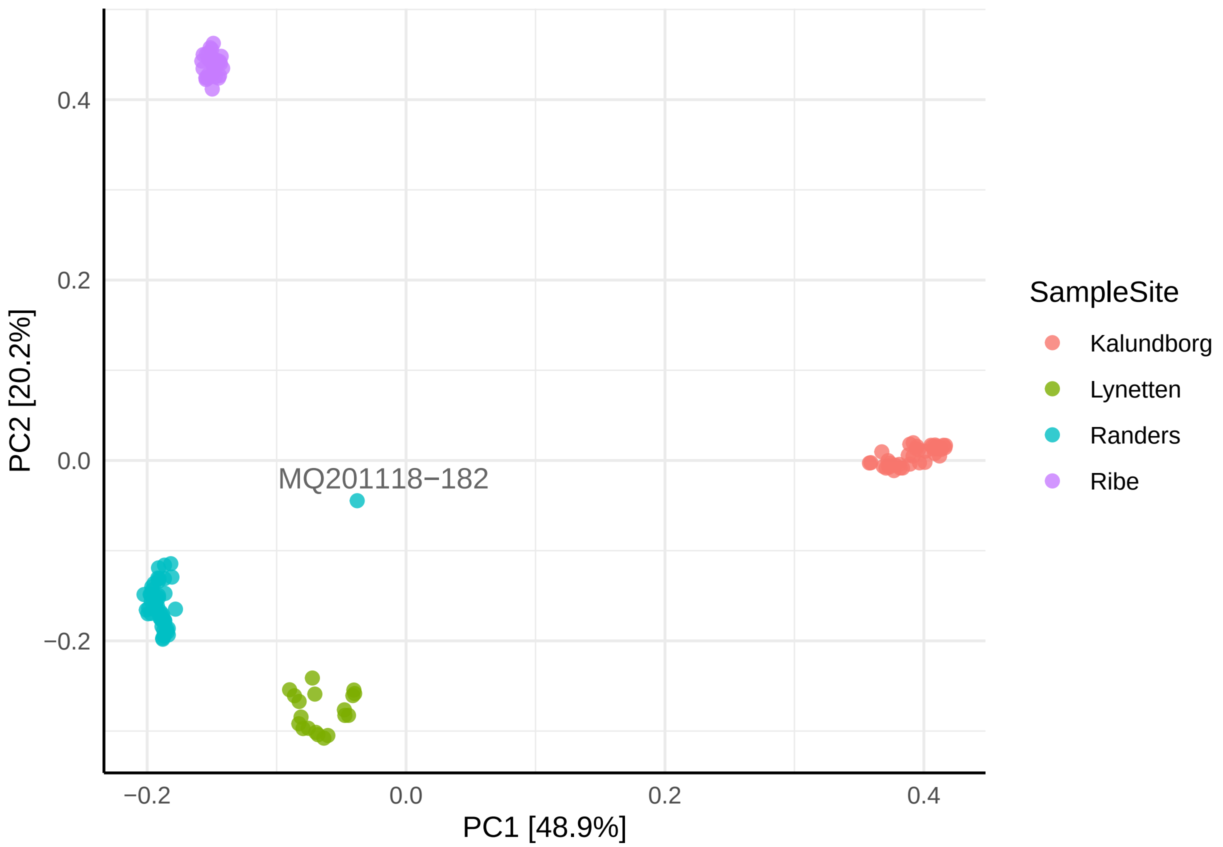


2.2 Problematic samples

Identify the outlier (by e.g. using amp_ordination & sample_label_by option or amp_heatmap & adjusting the "Group_by" parameter to show the "Sample"), subset the dataset to remove the outlier sample and replot the ordination and heatmap.

```
## 10898 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before: 12218 OTUs
## After: 1320 OTUs
```

```
## Warning: ggrepel: 138 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

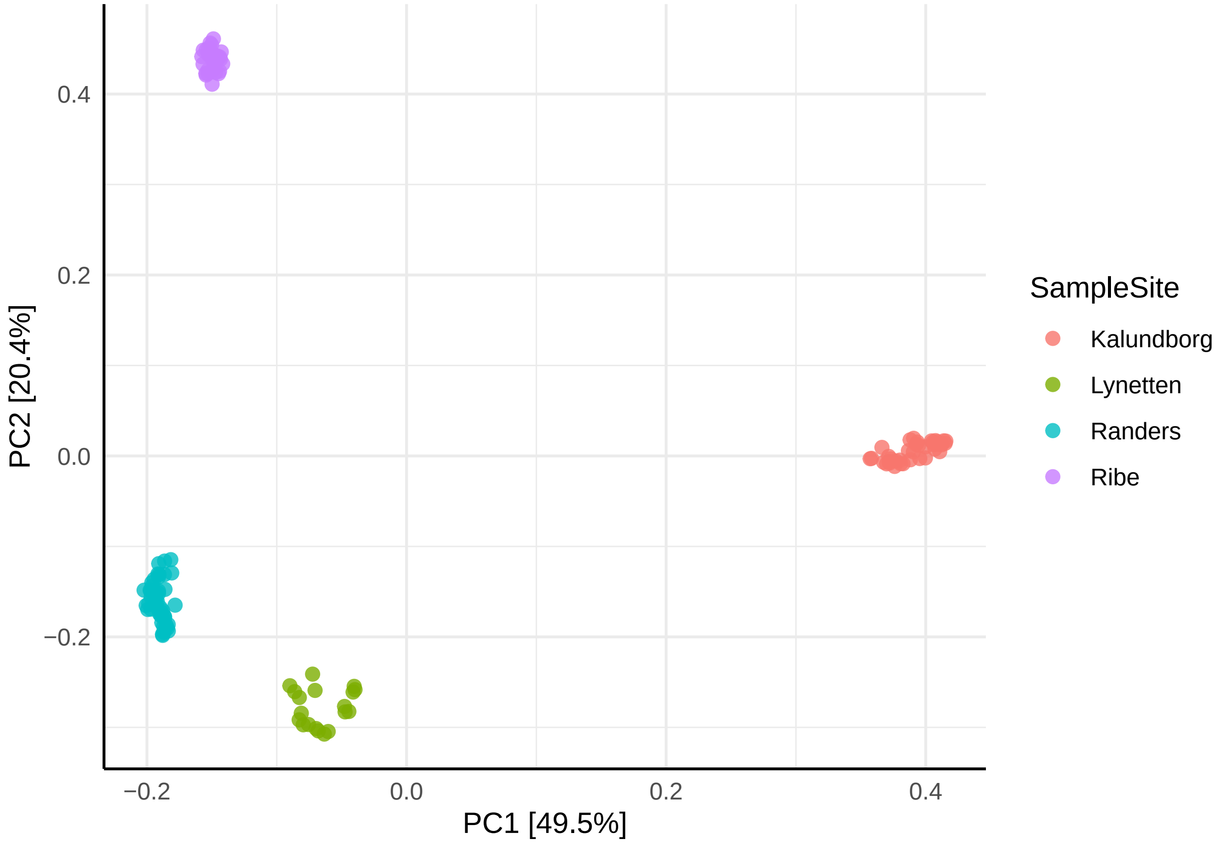


Remove outlier. Depending on which stage of your analysis, it can be necessary to re-do it with the outlier(s) removed from the beginning. For this exercise we will move from here on without the outlier: **Create new ampvis2 object**

```
## 1 samples and 25 OTUs have been filtered
## Before: 139 samples and 12218 OTUs
## After: 138 samples and 12193 OTUs
## 1 samples and 25 OTUs have been filtered
## Before: 139 samples and 12218 OTUs
## After: 138 samples and 12193 OTUs
```

Check that we have effectively removed the outlier

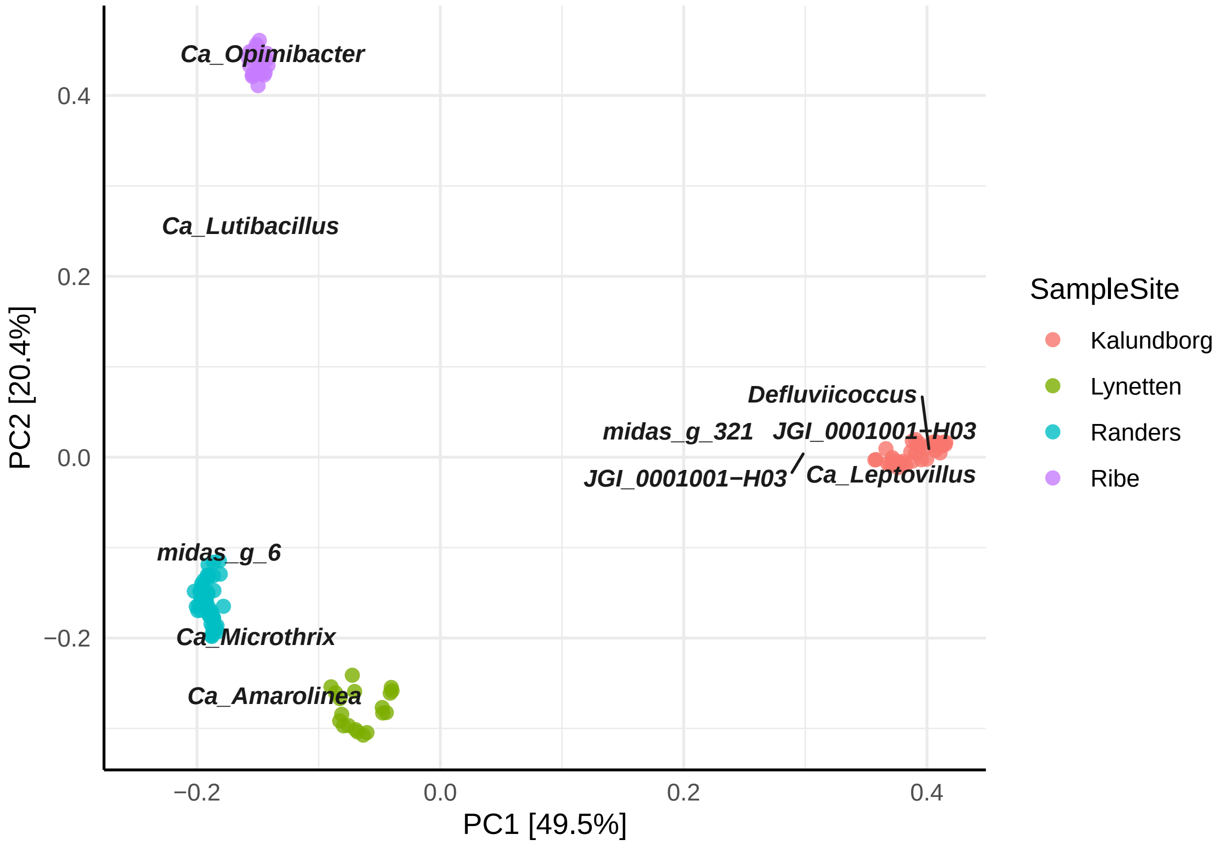
```
## 10986 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before: 12193 OTUs
## After: 1207 OTUs
```



2.3 Explore ordinations

Which bacteria are mainly causing the differences among the observed clusters (hint: try using `species_nlabels` and `species_label_taxonomy`). Keep the ordination results handy, how do the ordination relate the ordination results to your `heatmap`?

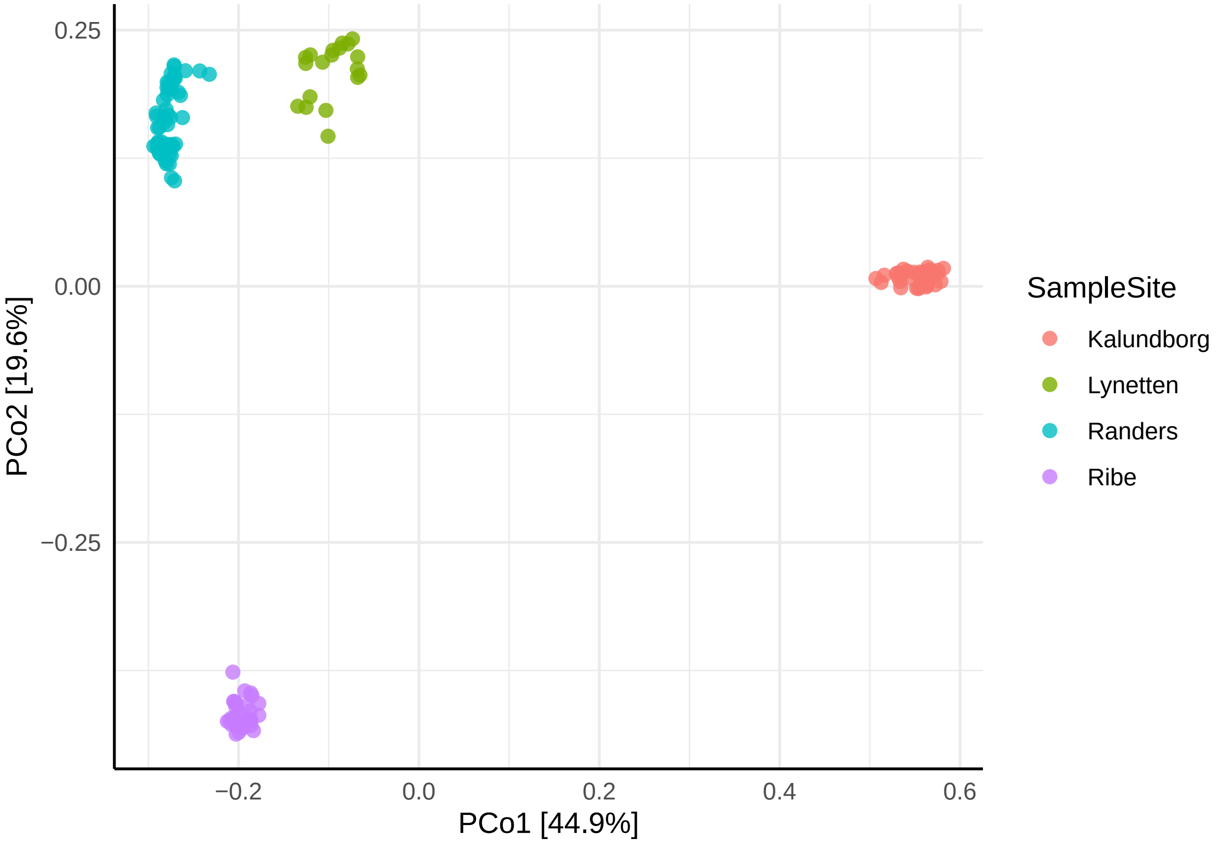
```
## 10986 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before: 12193 OTUs
## After: 1207 OTUs
```



2.4. Beta-diveristy addtional tasks

Try different ordinations, e.g PCoA based on bray-curtis distance

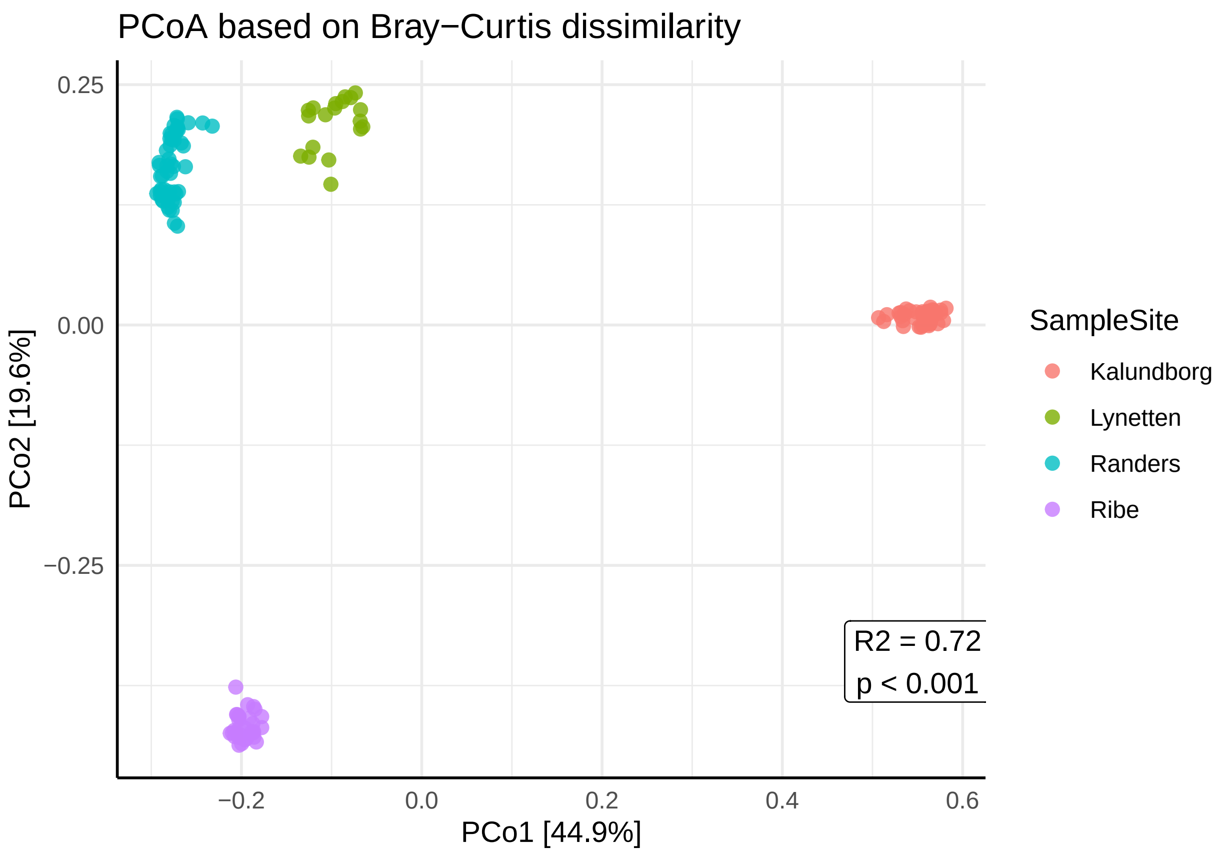
```
## 10986 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before: 12193 OTUs
## After: 1207 OTUs
```



(Advanced) Evaluate the statistical significance Install vegan package (install only if you don't have it)

Plot adding the statistical significance

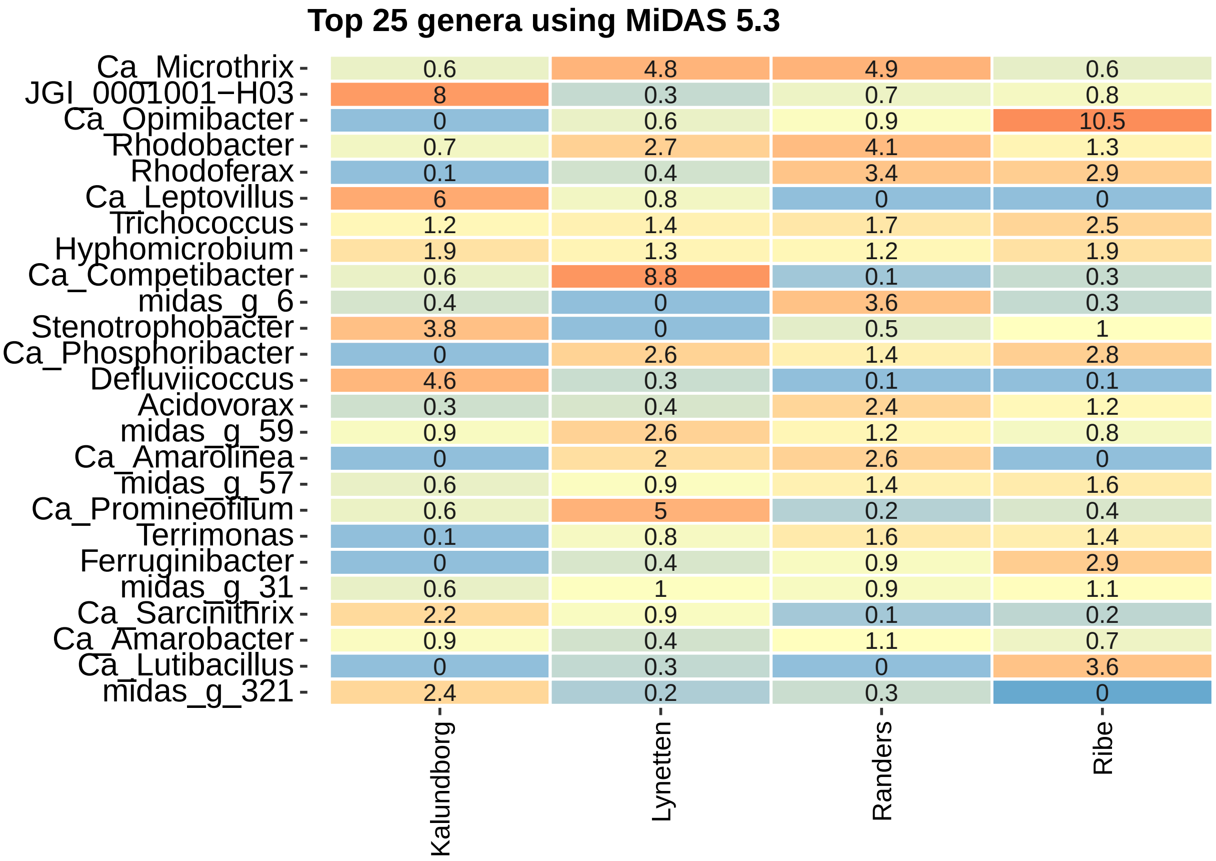
```
## 10986 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before: 12193 OTUs
## After: 1207 OTUs
```



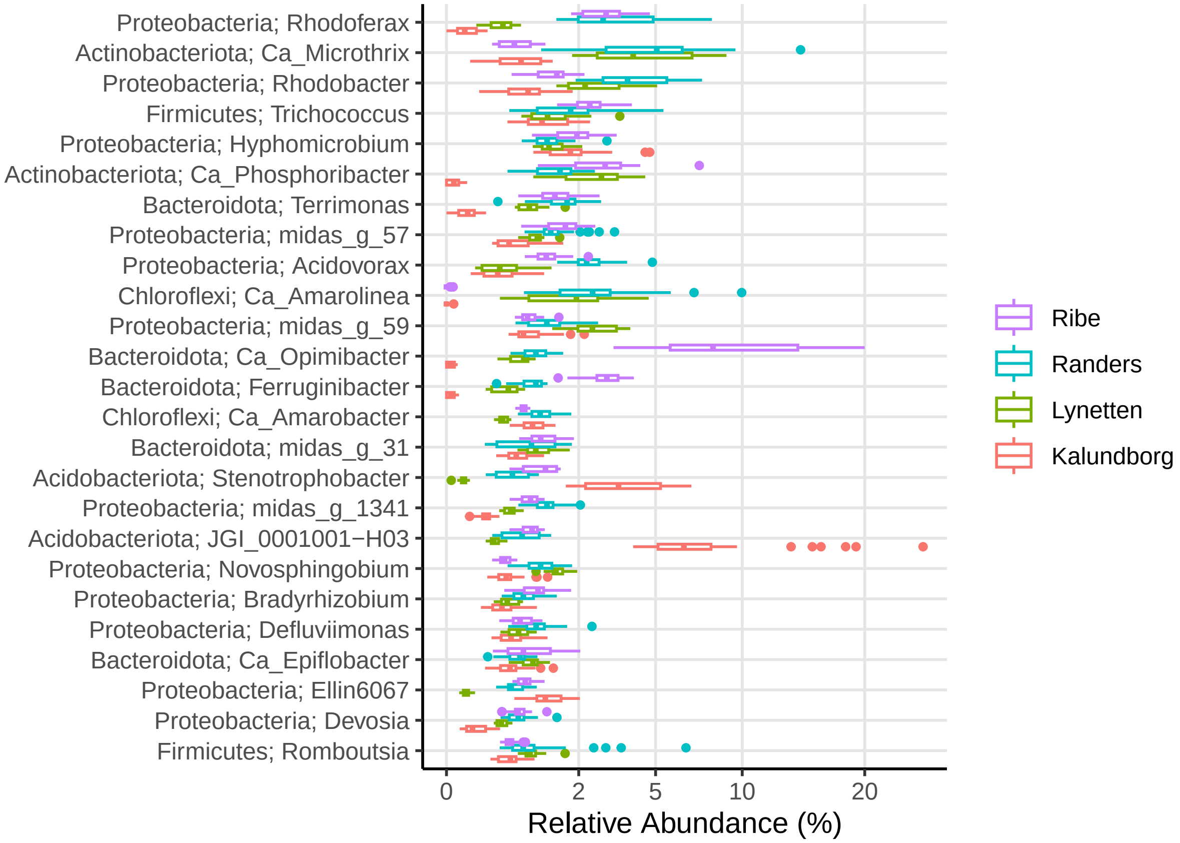
3. Microbial abundance

3.1 Which are the 25 most abundant genera in each Plant?

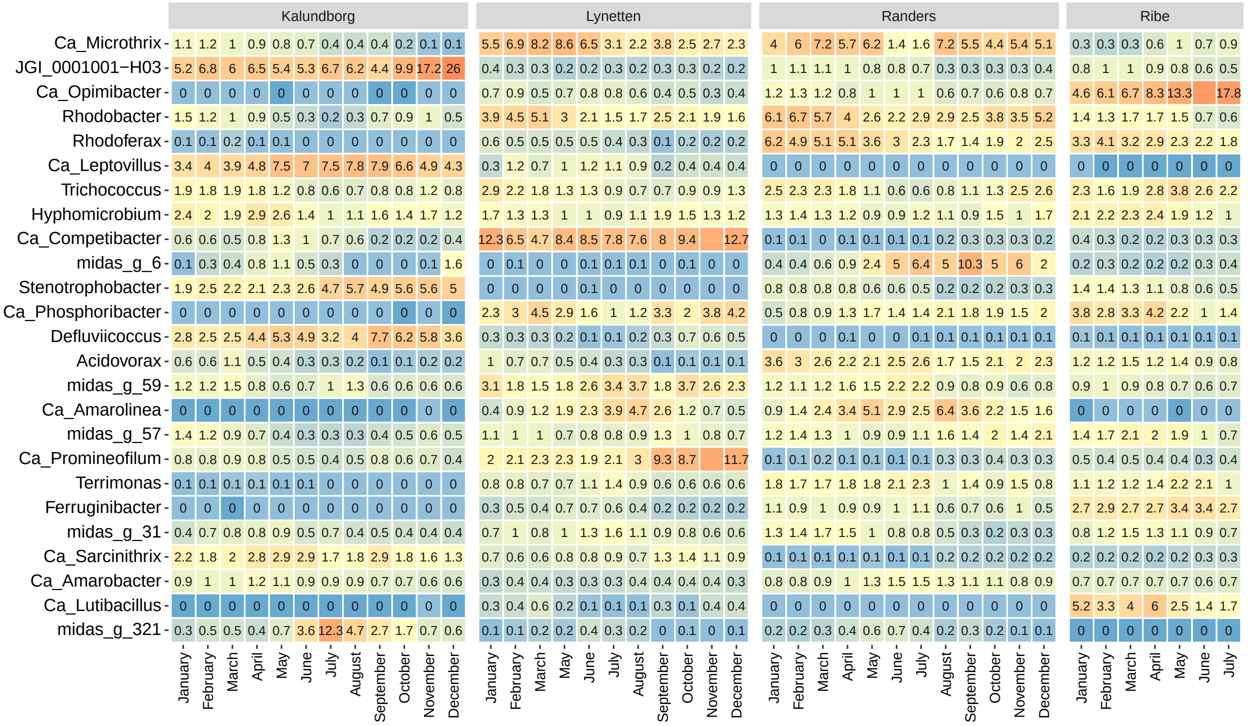
We normally start data analysis by making overview using the `amp_heatmap()` function. Modify the heatmap below using the relevant options - inspirations can be found in the Get started guide.


Top 25 genera using MiDAS 5.3
Ca_Microthrix 0.6 4.8 4.9 0.6
JGI_0001001−H03 8 0.3 0.7 0.8
Ca_Opimibacter 0 0.6 0.9 10.5
Rhodobacter 0.7 2.7 4.1 1.3
Rhodoferax 0.1 0.4 3.4 2.9
Ca_Leptovillus 6 0.8 0 0
Trichococcus 1.2 1.4 1.7 2.5
Hyphomicrobium 1.9 1.3 1.2 1.9
Ca_Competibacter 0.6 8.8 0.1 0.3
midas_g_6 0.4 0 3.6 0.3
Stenotrophobacter 3.8 0 0.5 1
Ca_Phosphoribacter 0 2.6 1.4 2.8
Defluviicoccus 4.6 0.3 0.1 0.1
Acidovorax 0.3 0.4 2.4 1.2
midas_g_59 0.9 2.6 1.2 0.8
Ca_Amarolinea 0 2 2.6 0
midas_g_57 0.6 0.9 1.4 1.6
Ca_Promineofilum 0.6 5 0.2 0.4
Terrimonas 0.1 0.8 1.6 1.4
Ferruginibacter 0 0.4 0.9 2.9
midas_g_31 0.6 1 0.9 1.1
Ca_Sarcinithrix 2.2 0.9 0.1 0.2
Ca_Amarobacter 0.9 0.4 1.1 0.7
Ca_Lutibacillus 0 0.3 0 3.6
midas_g_321 2.4 0.2 0.3 0
Kalundborg
Lynetten
Randers
Ribe

3.2 Try visualising with a boxplot

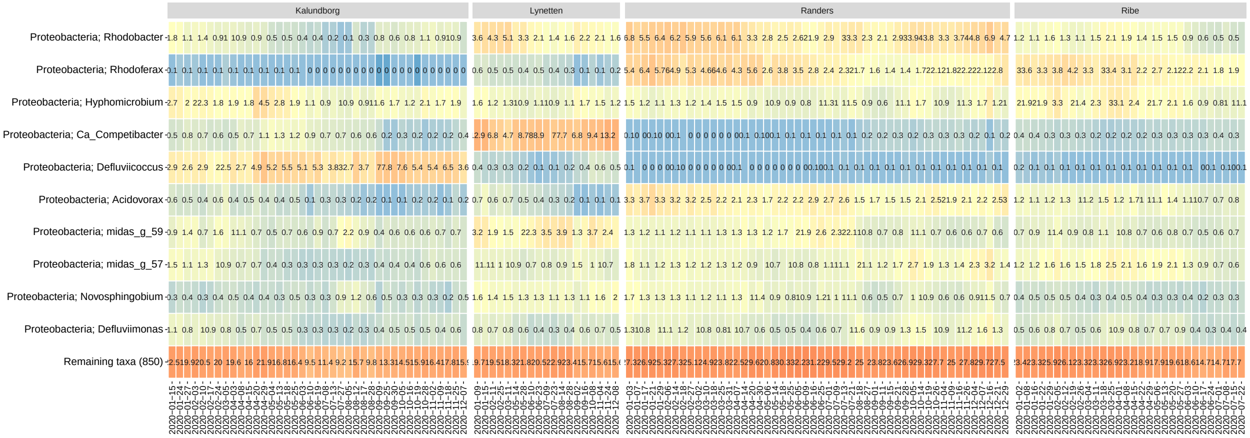


3.3 Which are the 25 most abundant genera in each WWTP and month?



3.4 Which genera within Proteobacteria are the most abundant across all samples?

```
## 9123 OTUs have been filtered
## Before: 12193 OTUs
## After: 3070 OTUs
```



3.5 Taxonomic reference database

Compare the results of microbial composition based on MiDAS 5 vs SiLVA 138.2

Q: Does the choice of taxonomic database influence alpha and beta-diversity analysis? When does it matter?

Create an ampvis2 file using Silva taxonomy

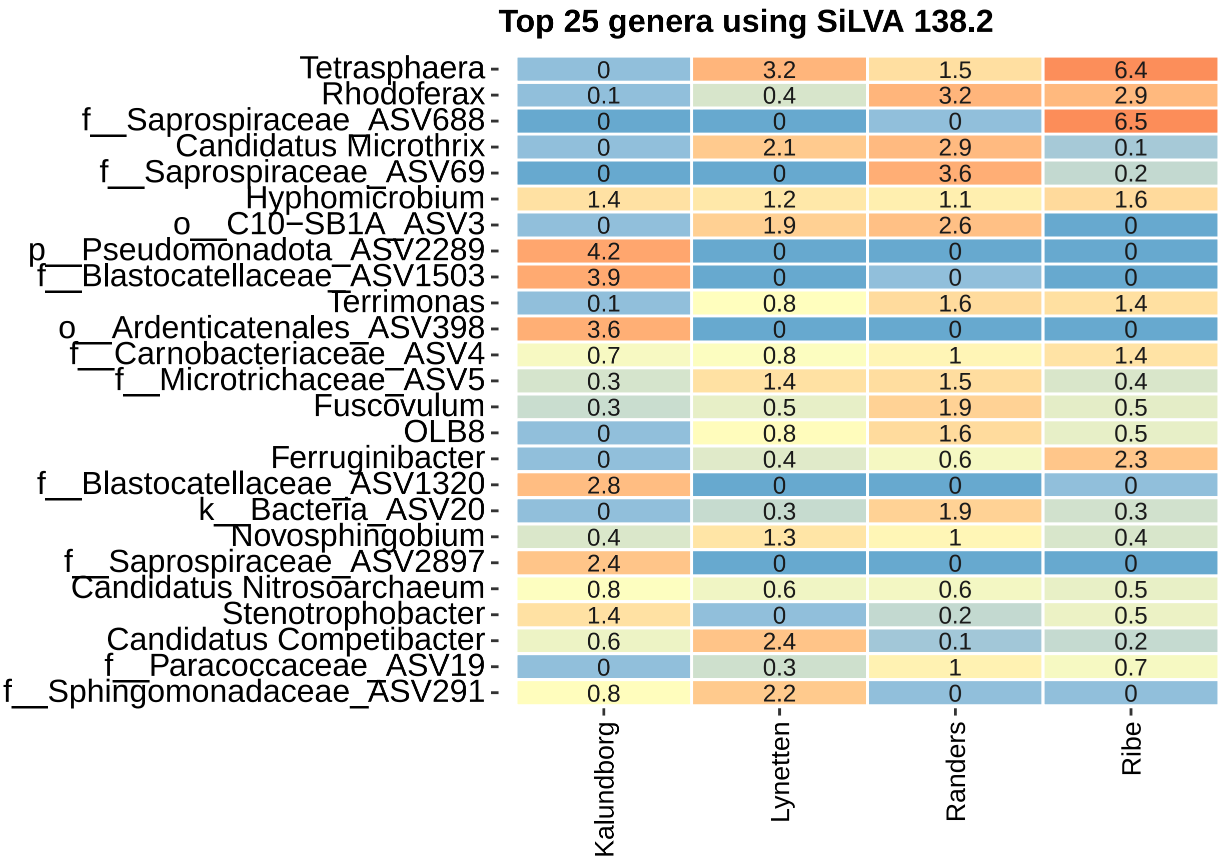
```
## Warning: Only 141 of 161 unique sample names match between metadata and otutable. The following unma
## metadata (20):
##  "MQ221006-200", "MQ221006-201", "MQ221006-202", "MQ221006-203", "MQ221006-204", "MQ221006-205", "MQ
```

Create the normalised dataset removing the outlier and subseting for samples with at least 10000 reads (tip: you can do all at once)

```
## 3 samples and 36 OTUs have been filtered
## Before: 141 samples and 12229 OTUs
## After: 138 samples and 12193 OTUs
```

What are the most abundant 25 genera based on silva?

```
## Warning in scale_fill_gradientn(colours = color.pal, trans = plot_colorscale, :
## log-10 transformation introduced infinite values.
```



Q: why is it important to get good taxonomic classifications? What strikes you the most?

4. Additonal tasks, timeseries

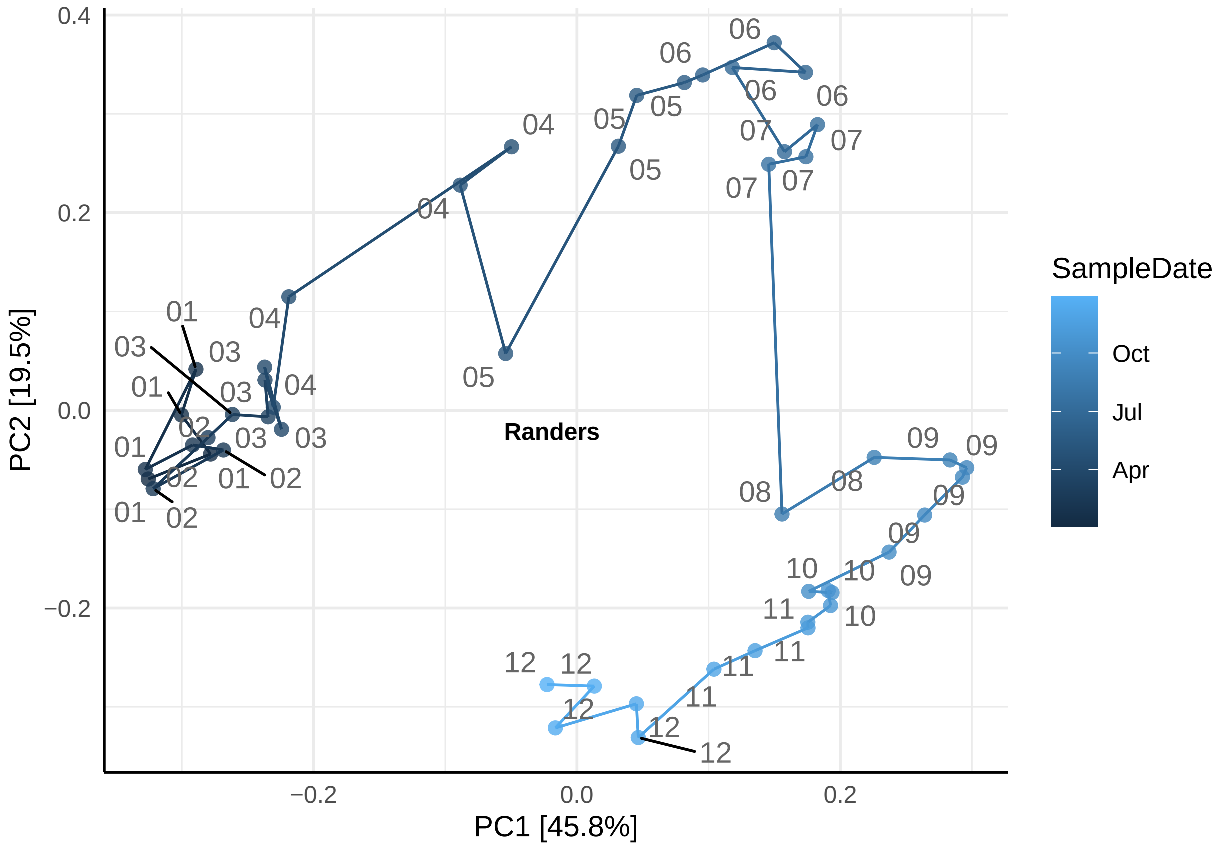
4.1 Time-series ordinations

Make 2 PCA plots for Kalundborg and Randers using `sample_trajectory` option to see the changes in the community over time. Comment on the stability of the communities in the two WWTPs. What do you think may cause the progression of the communities that you see on the plot?

```
## 88 samples and 2843 OTUs have been filtered
## Before: 138 samples and 12193 OTUs
## After: 50 samples and 9350 OTUs
```

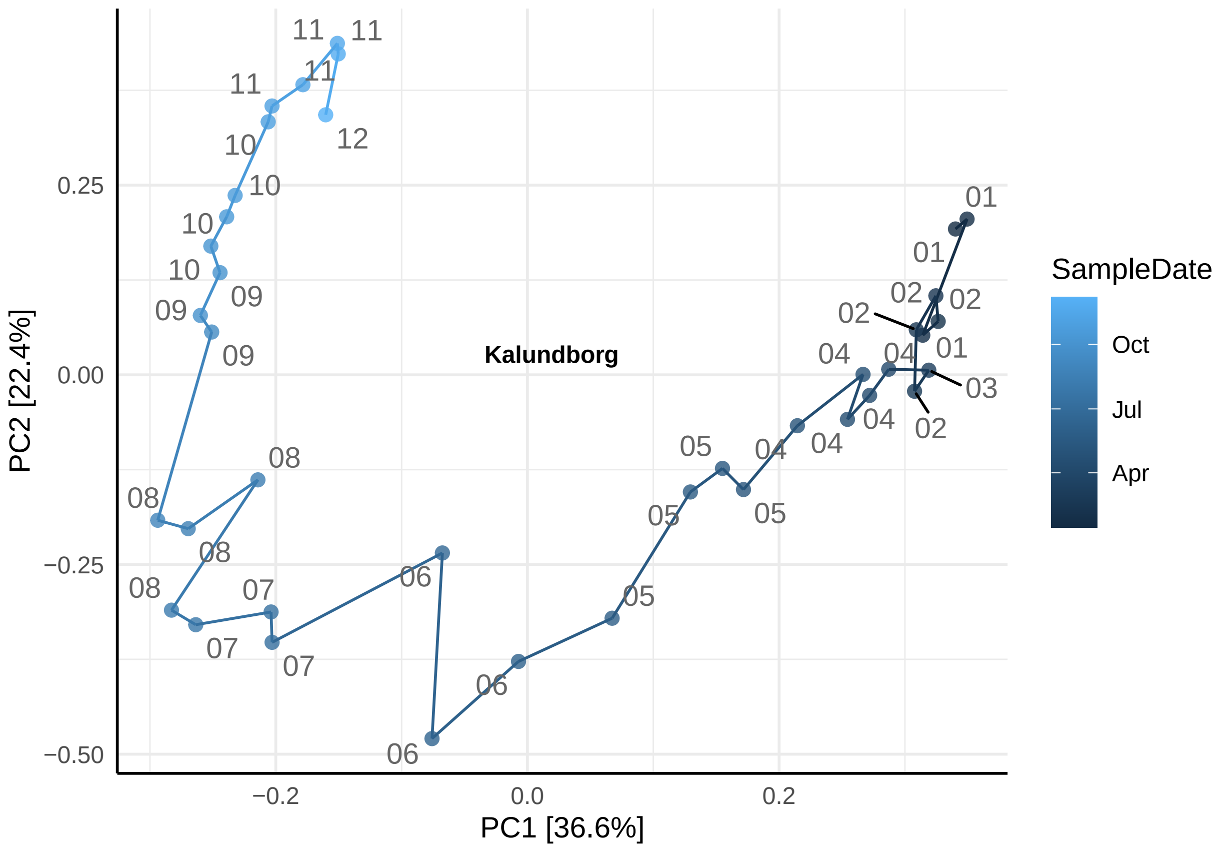
```
## 8906 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before: 9350 OTUs
## After: 444 OTUs
```

```
## Warning: ggrepel: 1 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
## 99 samples and 4759 OTUs have been filtered
## Before: 138 samples and 12193 OTUs
## After: 39 samples and 7434 OTUs
```

```
## 7021 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before: 7434 OTUs
## After: 413 OTUs
```

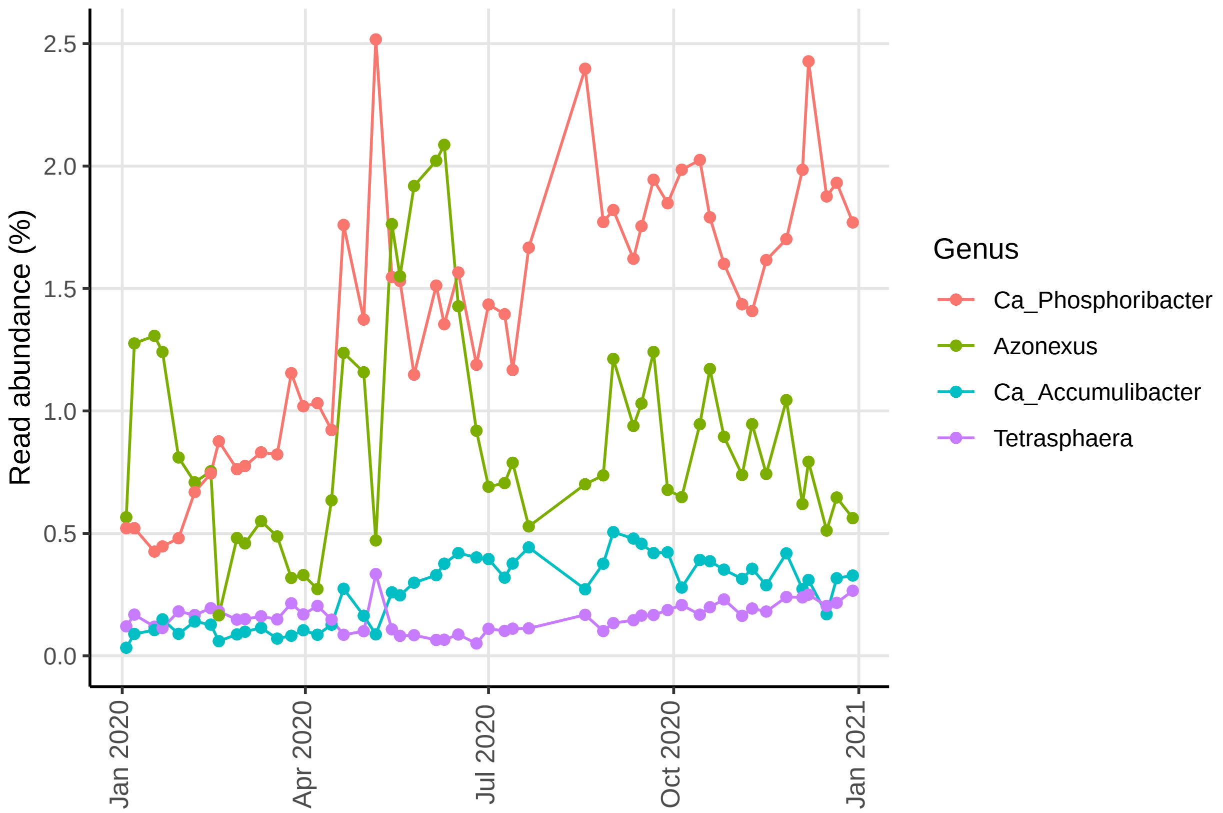


4.2 Timeseries plots

Subset the `Tetrasphaera`, `Ca. Phosphoribacter`, `Azonexus` and `Ca, Accumulibacter`genus data from Randers WWTP using `amp_subset_taxa()` and `amp_timeseries()` functions; and plot the data to identify the temporal dynamics of the different polyphospate accumulating genera.

```
## 12078 OTUs have been filtered
## Before: 12193 OTUs
## After: 115 OTUs
```

```
## 88 samples and 11 OTUs have been filtered
## Before: 138 samples and 115 OTUs
## After: 50 samples and 104 OTUs
```



4.3 Functional information

Subset the data for Lynetten and plot the heatmap showing the 25 most abundant genera. The `amp_heatmap` function offers the possibility of directly linking the genus-level plot with functional information from midas field guide. To do that, use:

option plot_functions = TRUE functions = c("Filamentous", "AOB", "NOB", "PAO", "GAO")

How many of the genera have the functional information available? What is the function of the most abundant bacteria in this WWTP?

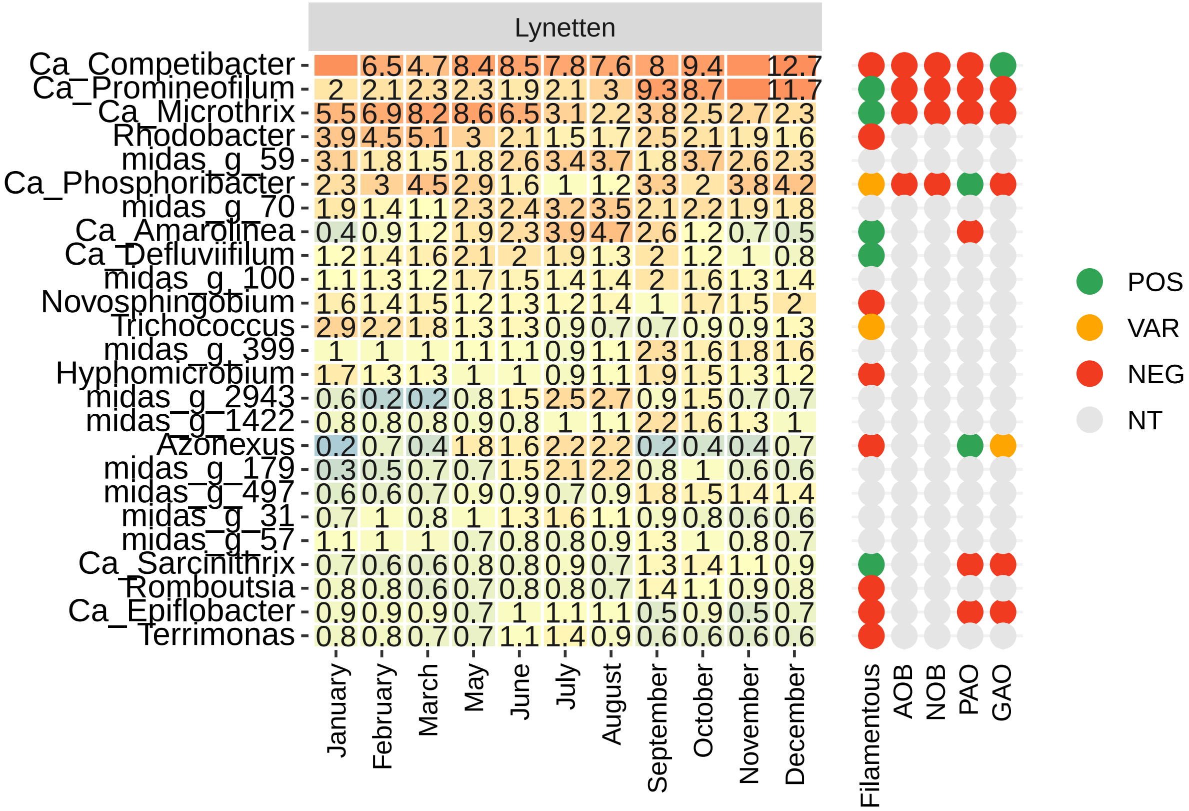
```
## 119 samples and 5157 OTUs have been filtered
## Before: 138 samples and 12193 OTUs
## After: 19 samples and 7036 OTUs
```

```
## Warning: package 'jsonlite' was built under R version 4.4.2
```

```
##
## Attaching package: 'jsonlite'
```

```
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
## Warning: package 'patchwork' was built under R version 4.4.2
```



5. Core communities

(Advanced) We will evaluate the core communities in our dataset, First we need to choose the desired taxonomic level. For this exercise we will choose "Species". The analysis is done outside ampvis2, therefore we will export the ampvis2 object to a long format data.frame

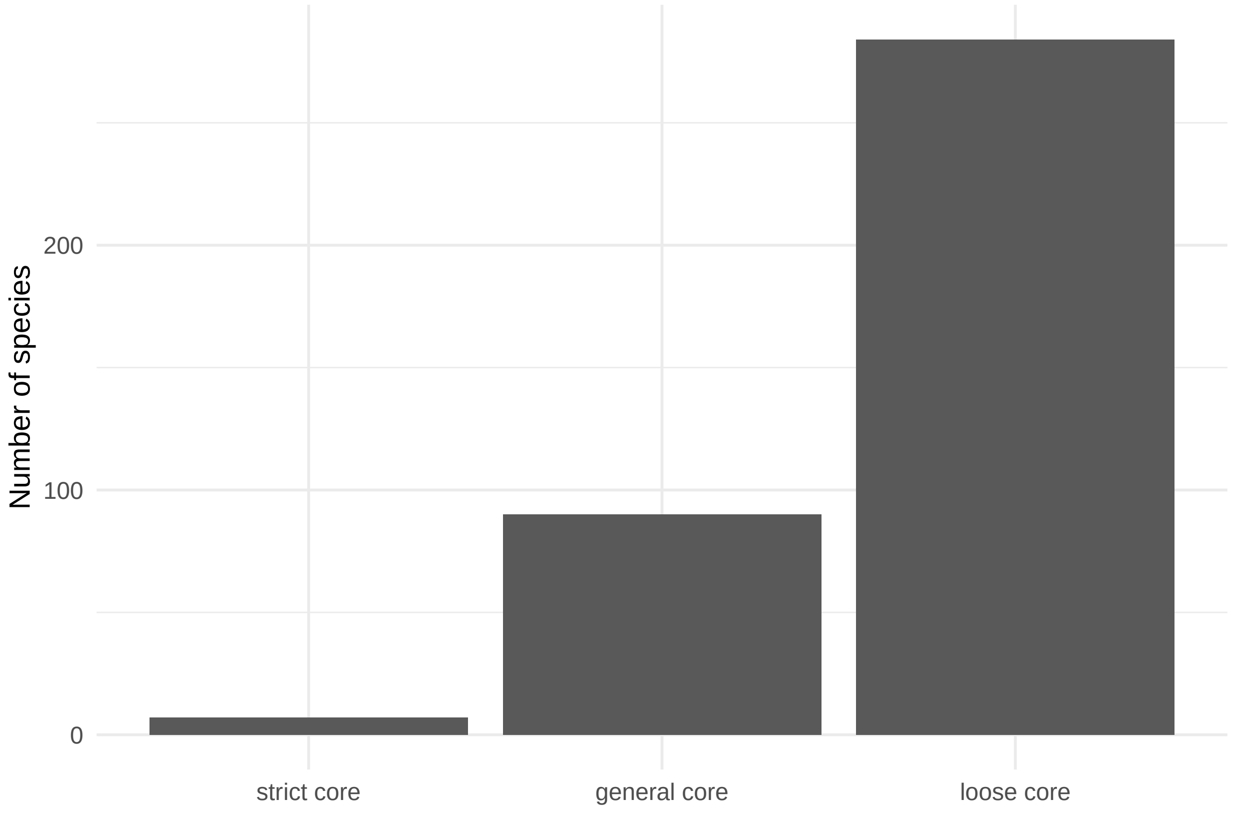
Calculate the relative abundance per species and the mean abundance in each WWTP

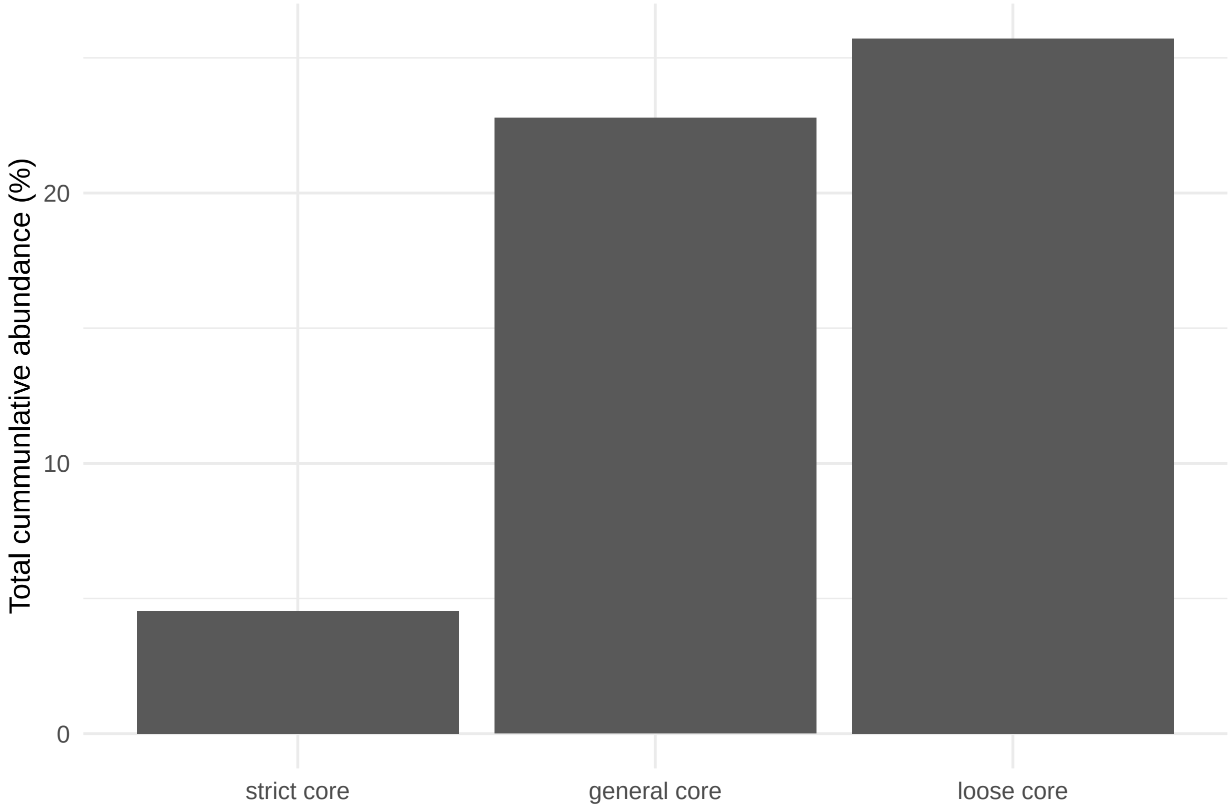
```
## # A tibble: 6 x 7
## # Groups:   SampleSite, Species [6]
##   SampleID     OTU   count SampleSite Species                    sumSpp meanSppSite
##   <chr>        <chr> <dbl> <chr>      <chr>                       <dbl>       <dbl>
## 1 MQ201118-152 ASV1  0.361 Randers    s__midas_s_5                0.364    1.04
## 2 MQ201118-152 ASV2  3.20  Randers    s__Ca_Microthrix_parvi~     3.22     2.65
## 3 MQ201118-152 ASV3  0.893 Randers    s__Ca_Amarolinea_domin~     0.900    2.61
## 4 MQ201118-152 ASV4  1.79  Randers    s__midas_s_4                3.38     1.70
## 5 MQ201118-152 ASV5  0.839 Randers    s__Ca_Microthrix_subdo~     1.09     2.02
## 6 MQ201118-152 ASV7  0     Randers    s__midas_s_220              0        0.000297
```

Define core groups based on abundance and create combined data.frame

```
## # A tibble: 6 x 9
##   Species                 mean_abu Category Kingdom Phylum Class Order Family Genus
##   <chr>                      <dbl> <chr>    <chr>   <chr>  <chr> <chr> <chr>  <chr>
## 1 s__midas_s_4               1.68  strict ~ k__Bac~ p__Fi~ c__B~ o__L~ f__Ca~ g__T~
## 2 s__Ca_Microthrix_su~    1.44  strict ~ k__Bac~ p__Ac~ c__A~ o__M~ f__Mi~ g__C~
## 3 s__midas_s_57          0.573 strict ~ k__Bac~ p__Pr~ c__A~ o__R~ f__Rh~ g__m~
## 4 s__midas_s_1112        0.316 strict ~ k__Bac~ p__Ac~ c__T~ o__T~ f__Th~ g__S~
## 5 s__midas_s_101         0.181 strict ~ k__Bac~ p__Fi~ c__C~ o__C~ f__Cl~ g__C~
## 6 s__midas_s_64          0.179 strict ~ k__Bac~ p__Fi~ c__C~ o__C~ f__Cl~ g__C~
```

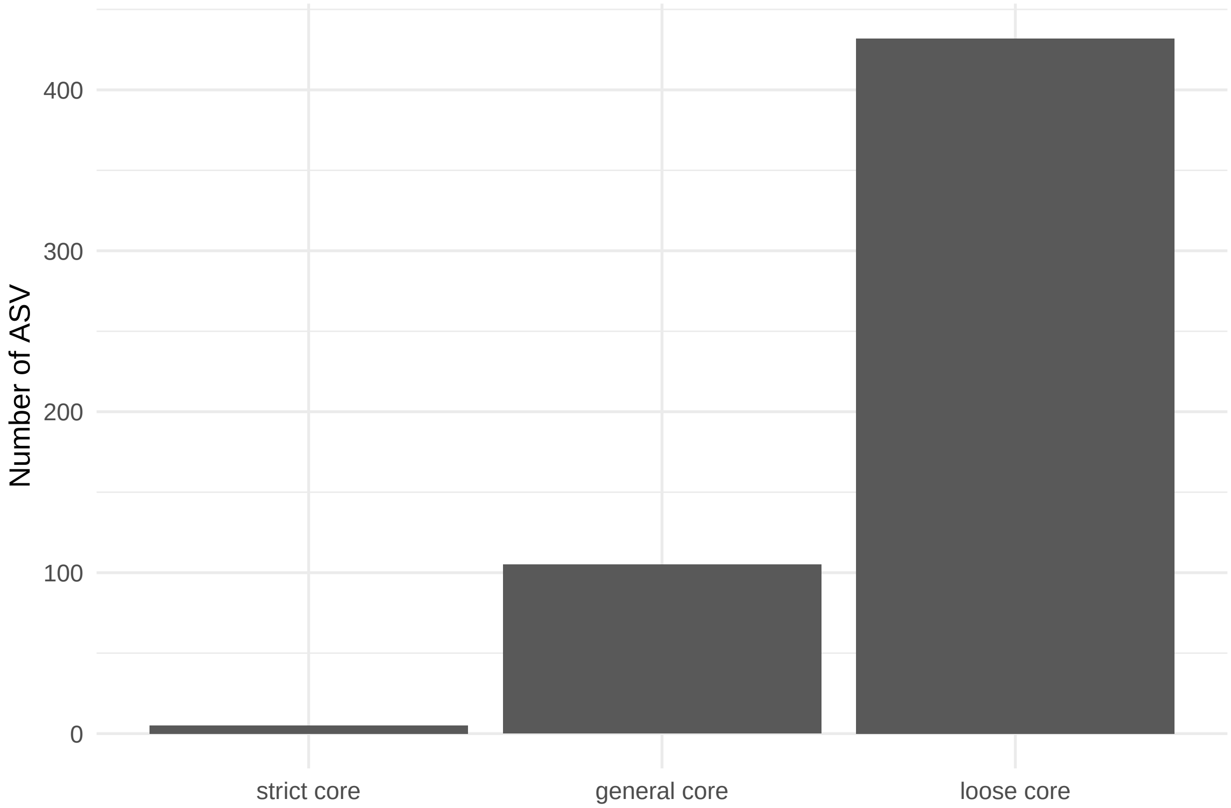
Visualise how many species are per core category and the total mean cummulative abundance the three categories explain

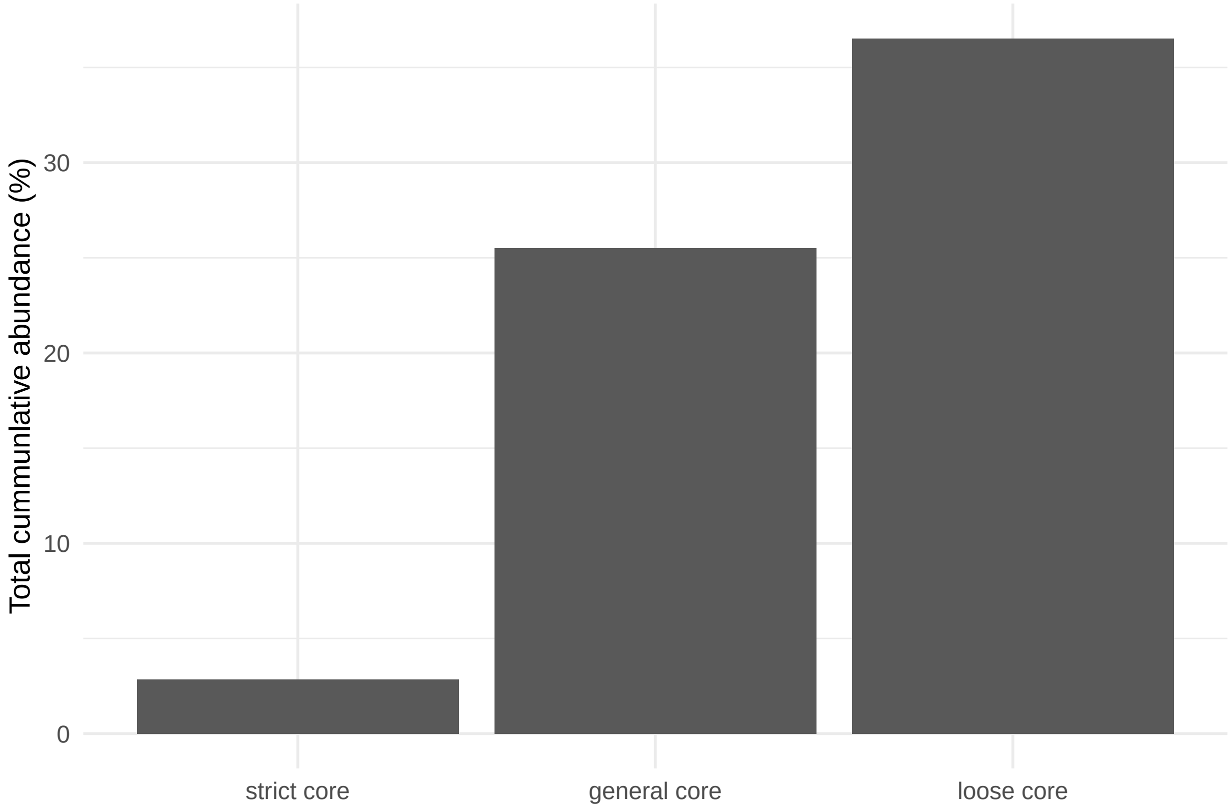




5.1 Core community additional tasks

Find the core communities at ASV level and visualise the outcomes


Number of ASV
0
100
200
300
400
strict core
general core
loose core


Total cummunlative abundance (%)
30
20
10
0
strict core
general core
loose core