Biobank (you need at least 32GB memory to run this)

init

```
#check for installed packages, install if missing
if (!require("tidyverse"))
  install.packages("tidyverse")
```

```
## Loading required package: tidyverse
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
if (!require("devtools"))
  install.packages("devtools")
```

```
## Loading required package: devtools
```

```
## Loading required package: usethis
```

```
if (!require("openxlsx"))
  install.packages("openxlsx")
```

```
## Loading required package: openxlsx
```

```
if (!require("ampvis2"))
  devtools::install_github("madsalbertsen/ampvis2")
```

```
## Loading required package: ampvis2
```

```r
if (!require("data.table"))
  install.packages("data.table")
```

```
## Loading required package: data.table
```

```
##
## Attaching package: ’data.table’
```

```
## The following objects are masked from ’package:dplyr’:
##
##     between, first, last
```

```
## The following object is masked from ’package:purrr’:
##
##     transpose
```

```r
#load data
library(ampvis2)
library(data.table)
library(tidyverse)

if (interactive()) {
  if (!grepl("data$", getwd())) {
    setwd("data")
  }
}
```

Load metadata

```r
rm(list=ls())
gc()
```

```
##           used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells 2314461 123.7    4158127 222.1  4158127 222.1
## Vcells 3870525  29.6    8388608  64.0  7059126  53.9
```

```r
metadata <- fread("metadata/metadata.txt")

if (any(duplicated(metadata$Sample))) {
  stop("One or more sample name(s) are duplicated in sample metadata. Please fix manually")
}

#filter a few useless samples
metadata <- filter(metadata, !Sample %chin% paste0("MQ201110-", 309:311))
metadata$Date <- lubridate::dmy(metadata$Date)
metadata$Year <- as.character(lubridate::year(metadata$Date))

#### add seasonal period and week number ####
```

```
#extract seasonal periods from dates
WS <- as.Date("2012-12-15", format = "%Y-%m-%d") # Winter Solstice
SE <- as.Date("2012-3-15",  format = "%Y-%m-%d") # Spring Equinox
SS <- as.Date("2012-6-15",  format = "%Y-%m-%d") # Summer Solstice
FE <- as.Date("2012-9-15",  format = "%Y-%m-%d") # Fall Equinox

# Convert dates from any year to 2012 dates
dates <- as.Date(strftime(metadata$Date, format = "2012-%m-%d"))
#extract periods and set factors for correct chronological order
metadata$Period <- ifelse (dates >= WS | dates < SE, "Winter", #winter
                              ifelse (dates >= SE & dates < SS, "Spring", #spring
                                      ifelse (dates >= SS & dates < FE, "Summer", "Fall"))) #summer, fa
metadata$Period <- factor(metadata$Period, levels = c("Spring", "Summer", "Fall", "Winter"))

metadata <- tibble::add_column(metadata,
                               Week = as.character(lubridate::isoweek(metadata$Date)),
                               .after = "Date")

setDT(metadata)
#### fix Plant and ID columns ####
# controls
metadata[grepl("extneg", tolower(LibID)), ID := "EXTNEG"]
metadata[grepl("extneg", tolower(LibID)), Plant := "CTRL"]
metadata[grepl("pcrpor", tolower(LibID)), ID := "PCRPOS"]
metadata[grepl("pcrpos", tolower(LibID)), Plant := "CTRL"]
metadata[grepl("pcrneg", tolower(LibID)), ID := "PCRNEG"]
metadata[grepl("pcrneg", tolower(LibID)), Plant := "CTRL"]
metadata <- metadata[!is.na(Plant)] #this removes the weird LibID samples: MQ181023-148, MQ181203-218,
metadata <- metadata[!Sample %chin% "MQ201110-248"]
metadata[Plant == "Avedoere", Plant := "Avedøre"]
metadata[Plant == "Damhusaaen", Plant := "Damhusåen"]
metadata[Plant == "Ejby Moelle", Plant := "Ejby Mølle"]
metadata[Plant == "Hjoerring", Plant := "Hjørring"]
metadata[Plant == "Naestved", Plant := "Næstved"]
metadata[Plant == "Egaa", Plant := "Egå"]
metadata[grepl("^Dam", ID) & !grepl("CTRL", Plant), Plant := paste0("Damhusåen-", Line)]
metadata[Plant %chin% "Damhusåen", Plant := paste0("Damhusåen-", Line)]
metadata[ID == "Lynetten", Plant := "Lynetten"]
metadata[ID == "Avedøre", Plant := "Avedøre"]
#metadata[is.na(Plant) & is.na(Line), Plant := ID]


#make sure date column is parsed correctly (year-month-day prefered) and
#sort chronologically, abundances will be sorted according to metadata by amp_load
metadata <- arrange(metadata, Plant, Date)
```

merge Aalborg West+East temperatures (by weekly average) with metadata

```
# Aalborg East
AAEtemps <- data.table::fread("metadata/AalborgEastTemperatures.csv")
colnames(AAEtemps) <- c("DateTime.Temperature", "Temperature")
AAEtemps <- AAEtemps[!is.na(DateTime.Temperature) & !is.na(Temperature)] #filter empty ones
AAEtemps[,DateTime.Temperature := lubridate::mdy_hm(DateTime.Temperature)] #parse dates
AAEtemps[,DateTime.Temperature := lubridate::floor_date(DateTime.Temperature, unit = "day")] #floor to
AAEtemps[,Year := as.character(lubridate::year(DateTime.Temperature))] #extract year
AAEtemps[,Week := as.character(lubridate::isoweek(DateTime.Temperature))] #extract week (ISO standard)
AAEtemps[,DateTime.Temperature := as.character(DateTime.Temperature)] #coerce back to character
AAEtemps <- AAEtemps[,.(week_mean_temperature = mean(Temperature)),keyby=.(Year, Week)] #sometimes mult
AAEtemps[,Plant := "Aalborg E"]

# Aalborg West
AAWtemps <- data.table::fread("metadata/AalborgWestTemperatures.csv")
colnames(AAWtemps) <- c("DateTime.Temperature", "Temperature")
AAWtemps <- AAWtemps[!is.na(DateTime.Temperature) & !is.na(Temperature)] #filter empty ones
AAWtemps[,DateTime.Temperature := lubridate::mdy_hm(DateTime.Temperature)] #parse dates
AAWtemps[,DateTime.Temperature := lubridate::floor_date(DateTime.Temperature, unit = "day")] #floor to
AAWtemps[,Year := as.character(lubridate::year(DateTime.Temperature))] #extract year
AAWtemps[,Week := as.character(lubridate::isoweek(DateTime.Temperature))] #extract week (ISO standard)
AAWtemps[,DateTime.Temperature := as.character(DateTime.Temperature)] #coerce back to character
AAWtemps <- AAWtemps[,.(week_mean_temperature = mean(Temperature)),keyby=.(Year, Week)] #sometimes mult
AAWtemps[,Plant := "Aalborg W"]

temps <- data.table::rbindlist(list(AAEtemps, AAWtemps))

metadata_merged <- dplyr::left_join(
  metadata,
  temps,
  by = c("Plant", "Year", "Week")
)
metadata_out <- dplyr::filter(
  metadata_merged,
  !Plant %chin% c("EXTNEG", "PCRPOS", "PCRNEG", "", "CTRL")
)
data.table::fwrite(metadata_out, "metadata.csv")
```

load amplicon data

```
d <- amp_load(
  otutable = "amplicon_data/ASVtable.tsv.zip",
  metadata = metadata_out,
  taxonomy = "amplicon_data/ASVs.R1.midas481.sintax.zip")
```

```
## Warning: Only 3846 of 4075 unique sample names match between metadata and otutable. The following unm
## otutable (229):
##  "MQ181116-127", "MQ181116-128", "MQ181116-129", "MQ181116-130", "MQ181116-124", "MQ181116-125", "MQ
```

remove samples with few reads, and normalise reads to sample total

```r
ds <- d %>%
  amp_subset_samples(minreads = 1000, removeAbsents = TRUE, normalise = TRUE)
```

```
## 97 samples and 4 OTUs have been filtered
## Before: 3846 samples and 90123 OTUs
## After: 3749 samples and 90119 OTUs
```

Select and create data subsets

```r
datasets <- c(
  "Aalborg E",
  "Aalborg W",
  "Avedøre",
  "Damhusåen-A",
  "Damhusåen-B",
  "Damhusåen-C",
  "Damhusåen-D",
  "Egå",
  "Ejby Mølle",
  "Esbjerg E",
  "Hirtshals",
  "Hjørring",
  "Kalundborg",
  "Lynetten",
  "Mariagerfjord",
  "Marselisborg",
  "Randers",
  "Ribe",
  "Viby"
)

dlist <- lapply(
  datasets,
  function(wwtp) {
    #filter
    dataset <- amp_subset_samples(
      ds,
      Plant %chin% wwtp,
      normalise = FALSE
    ) %>%
      ampvis2:::filter_species(filter_species = 0.1)

    dataset_folder <- paste0("datasets/", wwtp)
    dir.create(dataset_folder, recursive = TRUE, showWarnings = FALSE)

    #write out abundance table
    fwrite(
      data.table(
        ASV = rownames(dataset$abund),
        dataset$abund
      ),
      file = paste0(dataset_folder, "/ASVtable.csv")
    )

    #write out taxonomy
    fwrite(
      dataset$tax,
      file = paste0(dataset_folder, "/taxonomy.csv")
    )

    #write out metadata
    fwrite(
      dataset$metadata,
      file = paste0(dataset_folder, "/metadata.csv")
    )

    dataset
  }
)
```

```
## 3517 samples and 39761 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 232 samples and 50358 OTUs

## 49309 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:50358 OTUs
## After:1049 OTUs

## 3411 samples and 38770 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 338 samples and 51349 OTUs

## 50354 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:51349 OTUs
## After:995 OTUs

## 3662 samples and 53479 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 87 samples and 36640 OTUs

## 35920 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:36640 OTUs
## After:720 OTUs

## 3584 samples and 45382 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 165 samples and 44737 OTUs
## 44095 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:44737 OTUs
## After:642 OTUs

## 3576 samples and 43205 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 173 samples and 46914 OTUs

## 46292 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:46914 OTUs
## After:622 OTUs

## 3588 samples and 45070 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 161 samples and 45049 OTUs

## 44373 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:45049 OTUs
## After:676 OTUs

## 3586 samples and 43707 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 163 samples and 46412 OTUs

## 45669 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:46412 OTUs
## After:743 OTUs

## 3657 samples and 46915 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 92 samples and 43204 OTUs

## 42614 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:43204 OTUs
## After:590 OTUs

## 3618 samples and 50182 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 131 samples and 39937 OTUs

## 39356 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:39937 OTUs
## After:581 OTUs

## 3638 samples and 52734 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 111 samples and 37385 OTUs

## 36767 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:37385 OTUs
## After:618 OTUs
## 3638 samples and 55752 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 111 samples and 34367 OTUs

## 33601 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:34367 OTUs
## After:766 OTUs

## 3623 samples and 54419 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 126 samples and 35700 OTUs

## 34870 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:35700 OTUs
## After:830 OTUs

## 3638 samples and 62462 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 111 samples and 27657 OTUs

## 27009 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:27657 OTUs
## After:648 OTUs

## 3679 samples and 57893 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 70 samples and 32226 OTUs

## 31624 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:32226 OTUs
## After:602 OTUs

## 3601 samples and 50990 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 148 samples and 39129 OTUs

## 37771 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:39129 OTUs
## After:1358 OTUs

## 3679 samples and 56561 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 70 samples and 33558 OTUs

## 33034 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:33558 OTUs
## After:524 OTUs

## 3463 samples and 38872 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 286 samples and 51247 OTUs
## 50280 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:51247 OTUs
## After:967 OTUs

## 3582 samples and 50093 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 167 samples and 40026 OTUs

## 39029 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:40026 OTUs
## After:997 OTUs

## 3595 samples and 47200 OTUs have been filtered
## Before: 3749 samples and 90119 OTUs
## After: 154 samples and 42919 OTUs

## 42235 OTUs not present in more than 0.1% relative abundance in any sample have been filtered
## Before:42919 OTUs
## After:684 OTUs
```

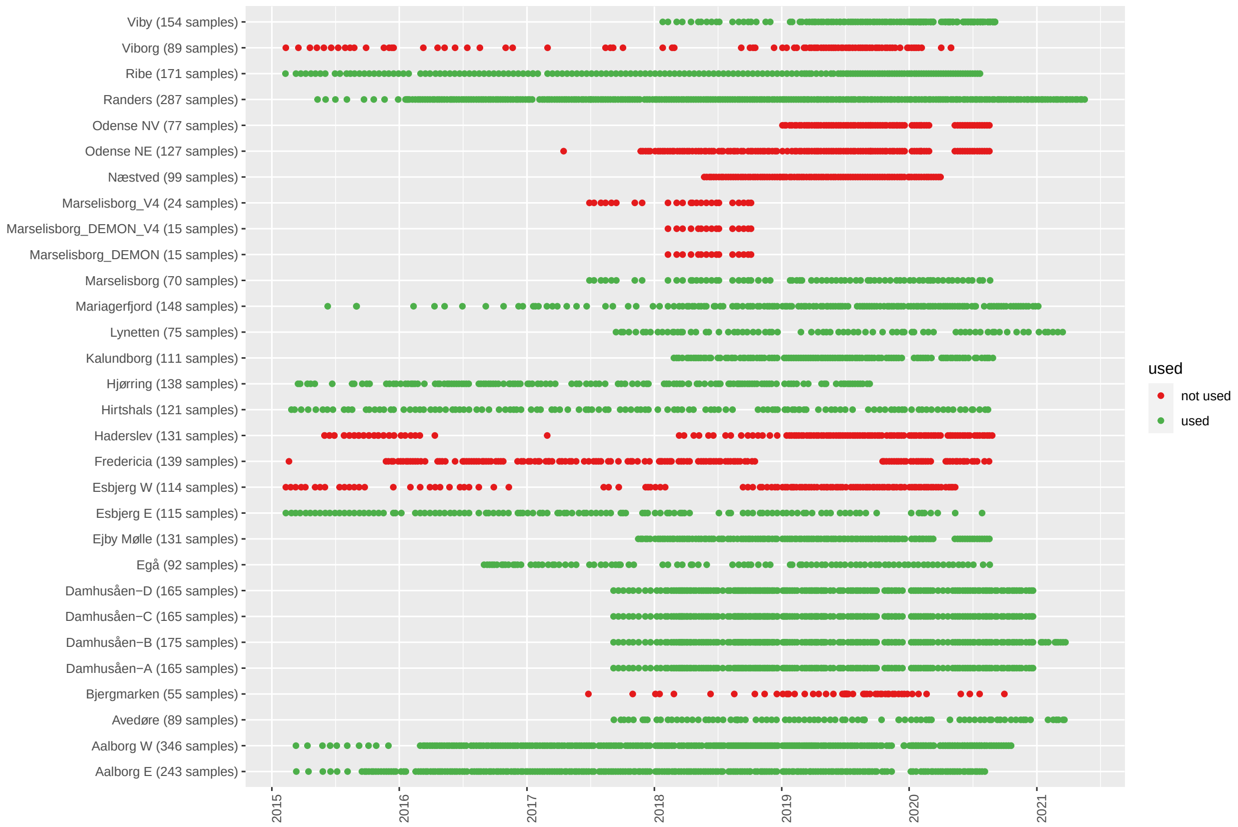
overview of samples over time

```r
m <- metadata_out
m[, used := ifelse(Plant %chin% datasets, "used", "not used")]
```

```
## Warning in ‘[.data.table‘(m, , ‘:=‘(used, ifelse(Plant %chin% datasets, :
## Invalid .internal.selfref detected and fixed by taking a (shallow) copy of the
## data.table so that := can add this new column by reference. At an earlier point,
## this data.table has been copied by R (or was created manually using structure()
## or similar). Avoid names<- and attr<- which in R currently (and oddly) may
## copy the whole data.table. Use set* syntax instead to avoid copying: ?set, ?
## setnames and ?setattr. If this message doesn’t help, please report your use case
## to the data.table issue tracker so the root cause can be fixed or this message
## improved.
```

```r
m[, Plant := paste0(Plant, " (", .N, " samples)"), by = Plant]
samples_overview <- ggplot(
  m,
  aes(x = Date,
      y = Plant,
      color = used)) +
  geom_point() +
  scale_x_date(date_breaks = "year", date_labels = "%Y") +
  scale_color_manual(values = RColorBrewer::brewer.pal(3, "Set1")[c(1, 3)]) +
  theme(
    axis.title = element_blank(),
    axis.text.x = element_text(angle = 90))
samples_overview
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```


Viby (154 samples)
Viborg (89 samples)
Ribe (171 samples)
Randers (287 samples)
Odense NV (77 samples)
Odense NE (127 samples)
Næstved (99 samples)
Marselisborg_V4 (24 samples)
Marselisborg_DEMON_V4 (15 samples)
Marselisborg_DEMON (15 samples)
Marselisborg (70 samples)
Mariagerfjord (148 samples)
Lynetten (75 samples)
Kalundborg (111 samples)
Hjørring (138 samples)
Hirtshals (121 samples)
Haderslev (131 samples)
Fredericia (139 samples)
Esbjerg W (114 samples)
Esbjerg E (115 samples)
Ejby Mølle (131 samples)
Egå (92 samples)
Damhusåen−D (165 samples)
Damhusåen−C (165 samples)
Damhusåen−B (175 samples)
Damhusåen−A (165 samples)
Bjergmarken (55 samples)
Avedøre (89 samples)
Aalborg W (346 samples)
Aalborg E (243 samples)
2015
2016
2017
2018
2019
2020
2021
used
● not used
● used