BINF7001 Metagenomics M4

Practical 1: Gene centric metagenomics analysis of the superworm gut microbiome

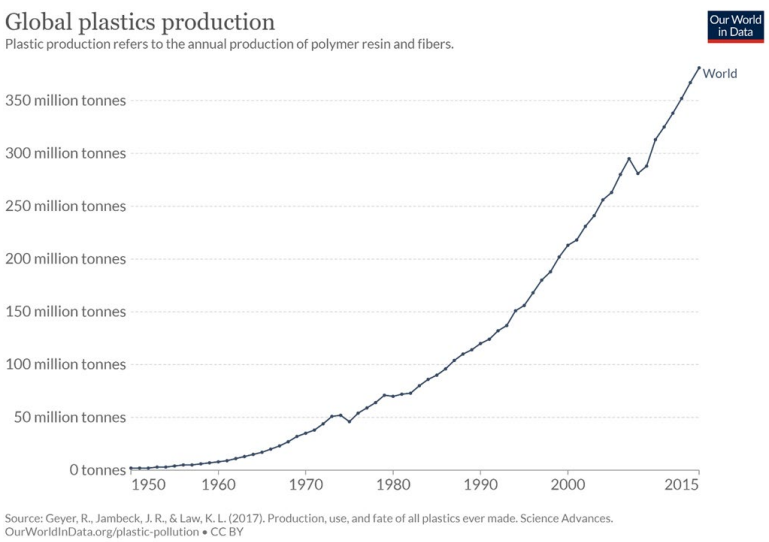
Learning objectives

In this practical, we will work through some steps involved in performing gene centric metagenomic analysis. The learning objectives are:

1) Learn how to remove host reads from a metagenome
2) Learn how to analyse taxonomic profiles from metagenomic data
3) Learn statistical techniques such as exploratory analysis and differential abundance
4) Learn some **R** and bash

1.1 Introduction

Plastics are inexpensive and widely used organic polymers, but their high durability hinders biodegradation. We are currently producing over 370 million tons of plastic globally each year, and the prediction is that plastic production will continue to increase, reaching 1 billion tons per year by 2050. Most of this plastic is used for disposable products that end up as trash. The vast majority of this waste goes straight to landfills or is discarded into the environment. Only a tiny fraction of 9% has been recycled from all the plastic that is no longer in use since 1950 (Geyer et al. 2017).



A major obstacle in increasing the rate of plastic recycling is that the current methods are mainly based on **mechanical recycling**, which includes processes such as sorting, washing, grinding, regranulating and extruding (heating and shaping into pellets). During this process, a plastic polymer degrades each time it is recycled (Ballerstedt et al. 2021). New sustainable biodegradation approaches are needed to recycle plastic waste. Microbes harbour a wide range of enzymes, some of which might be able to break down the long hydrocarbon polymers making up the backbone of most synthetic plastics. These enzymes can be connected to microbial biosynthesis of high-value chemicals, such as bioplastics, to produce higher value products from low value plastic waste. This **bio-upcycling** can increase the economic feasibility of plastic recycling and will incentivise plastic recycling globally.

Polystyrene, including expanded polystyrene foam (commonly known as styrofoam), is among the most commonly produced plastics worldwide and is considered to be recalcitrant to microbial degradation. However, superworms and their gut microbes might be able to break down polystyrene

foam after all.

Superworms are not worms, as their common name suggests, but rather the larvae of a species of darkling beetle (*Zophobas morio*). Superworms are commonly sold as pet food for reptiles, and are, in some countries, used for human consumption (see image).

Chocolate covered superworms are a delicacy in Thailand (source: Thailand Unique)

Aims and objectives of our investigation

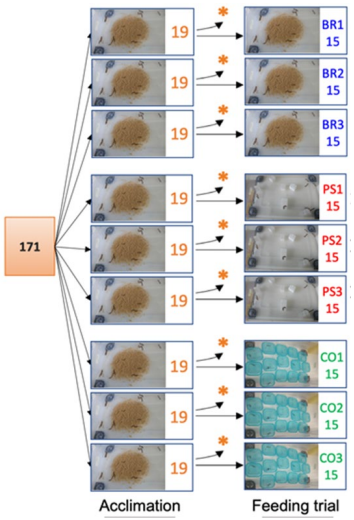
The **aims** of our study and of the bioinformatics analysis in this Practical are:

> ***1) Explore differences in the gut microbiome of superworms reared on wheat bran, polystyrene, and the starvation group.***
> ***2) Identify microbes that are enriched in the gut of superworms reared on a pure polystyrene diet***

The **objectives** to achieve these aims are:

1) Remove host sequences from the samples (read mapping)
3) Determine ratios of microbial to host reads (R)
4) Perform a community diversity analysis (diversity index; R)
5) Explore microbial abundances with DESeq2 (PCA, heatmap; R)
6) Extract differential abundant taxa (DESeq2; R)

Experimental design



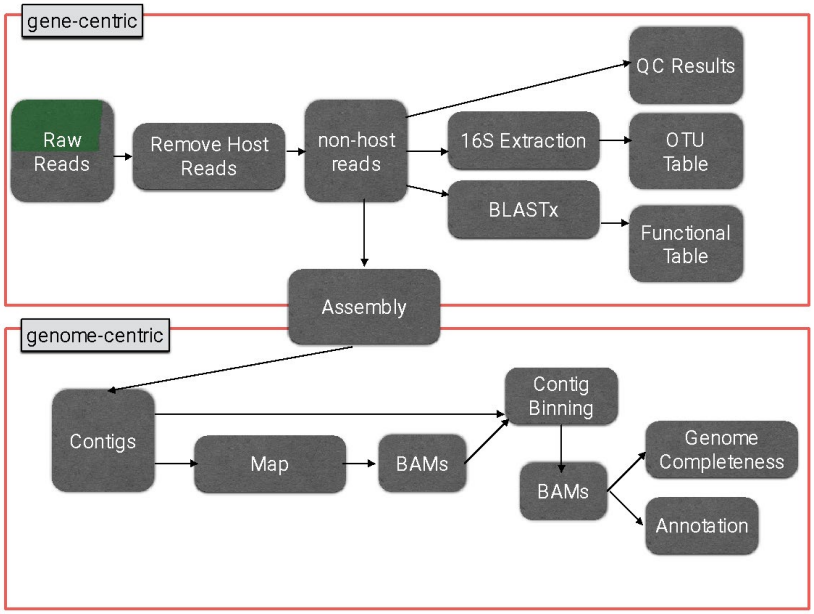
We set up an experiment to assess changes in the gut microbiome of superworms reared on bran, polystyrene or under starvation conditions over a 3 week period.

After an initial one-week acclimatisation period, to allow the superworms to adjust to the conditions in the lab, the worms were separated into three feeding trials: wheat bran (BR) the regular feed for superworms, polystyrene (PS) without any additional food sources, and a starvation control group (CO) which didn't get feed at all.

Q: The superworms in the starvation control group (CO) were separated into individual containers. What might be the reason why such a separation was necessary? To prevent eating each other ie cannibalism

After 3 weeks, the feeding trial ended and the worms were frozen followed by DNA extraction of the superworms guts.

In practicals, you will perform both gene-based and genome-based metagenomic analyses. Many steps in metagenomics take many hours (or days) to process, so some of the data processing has been already performed for you. The figure below illustrates the steps that the analysis pipeline for these practicals.

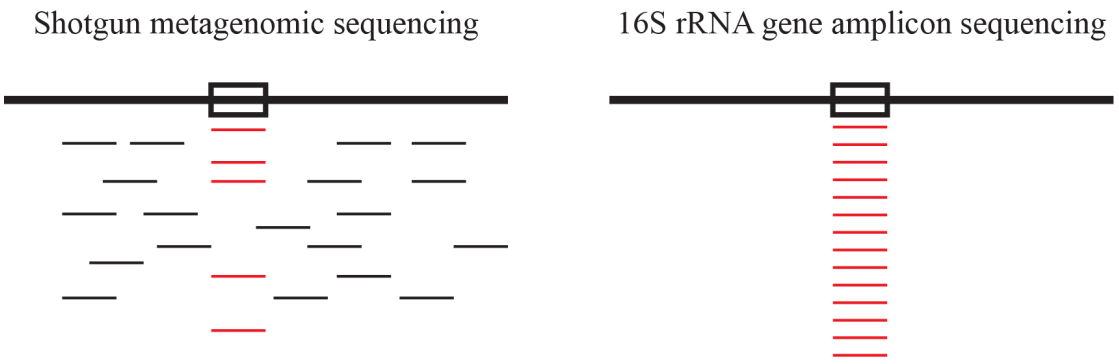


Because we have sequenced the bulk DNA in your samples, the metagenomes will contain host data, i.e. reads from superworm DNA. You will begin by identifying these reads in a subset of the samples by aligning to the host genome and then removing them from the samples before any further processing. Within R, you will explore the proportion of host reads in the samples.

Q: *Why is identifying and removing host data (reads) important?*

Next, you will explore the taxonomic structure of the microbial gut communities. These taxonomic profiles were generated by searching the non-host metagenomic reads for 16S rRNA gene sequences using Hidden Markov Models (HMMs), and taxonomically classifying these against a reference database (in this case the SILVA database). HMMs are beyond the scope of this practical, but don't worry – you just need to know that we are extracting the subset of metagenomic reads that encode **small subunit (SSU) rRNA genes**. Study the figure below to understand the difference between taxonomic profiling from 16S rRNA gene amplicon sequencing (last Prac) and shotgun sequencing (this Prac).

Q: What percentage of metagenomic reads would you expect to encode 16S rRNA genes? Q: Would you expect a full 16S rRNA gene to be encoded in a single read ?



In this practical, using the taxa count table we can perform exploratory analysis on the relationships between samples and determine which samples are more similar to each other, and whether sample types cluster together. To do this, we will use Principle Component Analysis (PCA). Other tools we will use include hierarchically organised heatmaps, and differential abundance (DA).

OK let's begin…

***Note:** example figures presented here may not be the same as the ones you produce and are intended to give you an idea of the type of output you should obtain.

1.2) Remove host sequences from the samples

Set your BINF7001 base directory to a bash variable and set up a new project in your BINF7001 directory called '**M3P2**' (e.g. Module 3 Practical 2). Within this create a data and **exp** directory, and within the **exp** directory create a directory called **`1_read_mapping`**.

```
# Set a variable to be the root directory we are working in:
> BINF7001=/home/<username>/BINF7001

# make the directories
> mkdir -p $BINF7001/projects/M4P2/data
> mkdir -p $BINF7001/projects/M4P2/exp/1_read_mapping
> mkdir -p $BINF7001/projects/M4P2/exp/1_read_mapping/index_files

> cd $BINF7001/projects/M4P2

# have a look:
> tree -d

    `-- M4P2
        |-- data
        `-- exp
            `-- read_mapping
                `-- index_files
```

Change directories (**cd**) to the data directory, check your path (**ls**), and symbolically link (**ln -s**) the fastq files.

```
> cd $BINF7001/projects/M4P2/data
# check that your path is correct with an ls
> ls /opt/BINF7001/2023/Prac11_2023/fastq/* -1

# soft link the fastq files
> ln -vs /opt/BINF7001/2023/Prac11_2023/fastq/* .
> ls -l
```

There are **nine samples** (Table 1), each with a forward R1 and reverse R2 read, so 18 files in total. Map these sequences to the *Tenebrio molitor* (yellow mealworm) with **bwa** (index located at /opt/BINF7001/2021/Prac11_2021/hg19.fa*) then extract the unmapped reads with **samtools**, and finally transform the SAM alignments back to fastq files using **bamToFastq**. Do this without writing the BAM file to the file system if you can! Don't worry, we can discuss this in class. Note that these sequence files have been subsampled to 10K reads, so they will map very quickly (comparatively) at 4 threads.

Table 1: Samples for the superworm study

SampleID	Sample description
Bran_1	Fed on bran
Bran_2	Fed on bran
Bran_3	Fed on bran
PS_1	Fed on polystyrene ("styrofoam")
PS_2	Fed on polystyrene ("styrofoam")
PS_3	Fed on polystyrene ("styrofoam")
CO_1	No feed
CO_2	No feed
CO_3	No feed

To map a single sample, e.g. CO_1:

```
# soft link the index files
> cd $BINF7001/projects/M4P2/exp/1_read_mapping/index_files
> ln -vs /opt/BINF7001/2023/Prac11_2023/index_files/* .

# soft link the fastq files
> cd $BINF7001/projects/M4P2/exp/1_read_mapping
> ln -sv $BINF7001/projects/M4P2/data/* .
```

```
# Map one read
> cd $BINF7001/projects/M4P2/exp/1_read_mapping
> bwa mem -t 4 index_files/GCA_014282415.2.fa CO_1_R1.truncated.fastq.gz
CO_1_R2.truncated.fastq.gz | samtools view -b -f 12 -F 256 -F 2048 /dev/stdin |
bamToFastq -i /dev/stdin -fq CO_1_R1_nonhost.fastq -fq2 CO_1_R2_nonhost.fastq
> ls -lh

# Compress the resulting files
> gzip CO_1_R1_nonhost.fastq
> gzip CO_1_R2_nonhost.fastq
> ls -lh
```

Learn more about the samtools view options here: samtools-view manual page
In the command above, *bwa* generates a SAM file, and we use *samtools view* to select reads based on the bitwise flags in the SAM file. Note, that the FLAG field is displayed as a single integer, but is the sum of bitwise flags to denote multiple attributes of a read alignment.

Q1: What output is provided when using the bitwise FLAGs -f and/or -F?
Q2: *Which FLAG attributes does the value of 12 contain, and what does the samtools view option "-f 12" specify?*

The final two **gzip** steps are to reduce the size of the files. As a general rule, it is best to try to avoid leaving fastq files uncompressed, particularly since many bioinformatics tools can accept compressed or uncompressed fastq files.

Doing all samples using a scripted for loop:

```
>for i in `echo Bran_1 Bran_2 Bran_3 CO_1 CO_2 CO_3 PS_1 PS_2 PS_3`; do echo
mapping $i ;bwa mem -t 4 index_files/GCA_014282415.2.fa ${i}_R1.truncated.fastq.gz
${i}_R2.truncated.fastq.gz | samtools view -b -f 12 -F 256 -F 2048 /dev/stdin |
bamToFastq -i /dev/stdin -fq ${i}_R1_nonhost.fastq -fq2 ${i}_R2_nonhost.fastq;
gzip ${i}_R1_nonhost.fastq ; gzip ${i}_R2_nonhost.fastq ; done

> ls -ltr
```

Have a look at the resulting compressed fastq files. Count the number of lines of each sample using a combination of **zcat** and **wc**, and determine how many reads were host, and how many presumably microbial. Note that a fastq file uses 4 lines per sequence.
Fill out this table:

Sample	Sample type	Total reads	Non-host reads	Non-host reads(%)
Bran_1	Fed on bran	10000		
Bran_2	Fed on bran	10000		
Bran_3	Fed on bran	10000		
PS_1	Fed on polystyrene (styrofoam)	10000		
PS_2	Fed on polystyrene (styrofoam)	10000		

PS_3	Fed on polystyrene (styrofoam)	10000		
CO_1	No food	10000		
CO_2	No food	10000		
CO_3	No food	10000		

***Q:** Do the Bran and PS samples have different fractions of putative microbial reads, on average? How do they compare to the control group?* We will also explore this further in R.

Once the host (insect) reads have been removed, we use a tool to extract the 16S rRNA gene reads from the metagenomic data set and to assign a taxonomy to each identified 16S read. This step takes a while, so the data has been processed for you and is available in the next step.

Now we will move onto the data analysis using **R** and **Rstudio**: https://posit.co/download/rstudio-desktop/

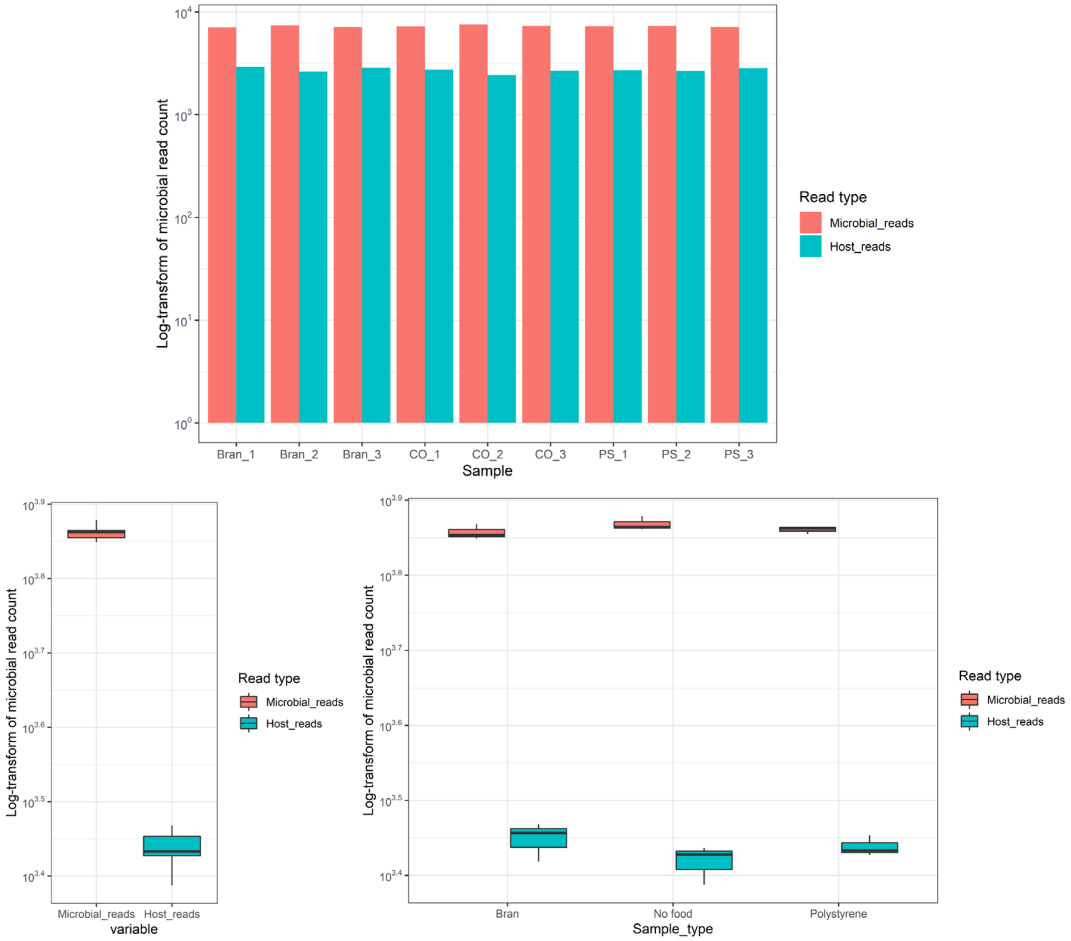
1.3) Determine the ratio of microbial to host reads (R)

Install **Rstudio** on your local machine and create a folder BINF7001_M4P2/Code. This will be your working directory.
Download the zip file "R files" from Blackboard/Learning Resources, and **test** the scripts with your Rstudio installation **before the practical** to make sure all dependencies, i.e. libraries, are installed.

On your local machine do the following:

1. Open RStudio, set the BINF7001_M4P2 folder as your working directory (>Session >Set working Directory), open the BINF7001_M4P2_readcount.R script and install and load the necessary packages/libraries at the top of the script. Please remember to save Rscripts in the "Code" folder to enable complete file paths.
2. The number of host and non-host (putative microbial) reads have been calculated for you and results are listed in the **read_counts.tsv** file. Load the file into the RStudio workspace environment (line 50) and view the data (using R environment) to check that you've loaded it correctly.
3. Next run the code on lines 52-68 to **plot the number of putative microbial and host reads** in each sample (example provided below).
   Q: *How would you describe the distribution of read counts? Note: The y-axis is a log scale. Why would having too few reads confound the analysis?* In addition to the number of microbial reads, the **ratio of host to microbial reads** is also important. *Why?* Run the code on lines 72-81. This will first produce a stacked bar graph that shows proportions (also called a 100% stacked bar graph).
4. Another way to compare the data is to show host vs microbial reads **across all samples**. Run the code starting at line 85 to create a **box plot** to illustrate the microbial to host read counts found in your samples (example output provided below).
5. Now, let's explore the plots and try to answer some more questions:

Q1: Which sample type looks to have the lowest fraction of host reads?
Q2: Why would one sample type have a lower fraction of host reads than the other?
To better answer question 1, let's modify the R code for the box plot: have a look at what is plotted on the x-axis (hint x = Sample_type). **What else can we plot here?** To see what other options you have, check out the variable "samples" in the R environment and choose a different header name to be plotted. Don't forget to save this plot with a different filename or you will overwrite your file. This plot should now tell you microbial vs host read count for each sample type.



1.4) Community diversity analysis

Now that we have performed some basic data assessment and quality control, let's explore the microbial community profiles. One of the first questions you can ask is how diverse is a microbial community? In other words, how many species are present, and at what abundance. Is there a single dominant species, are there several species at high abundance or many species at even abundances?

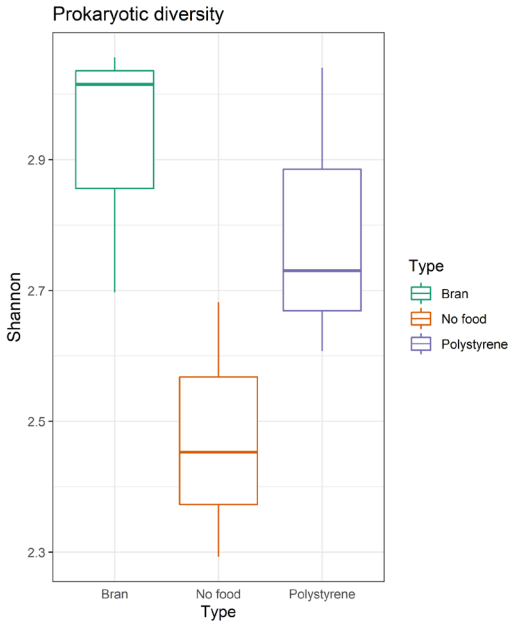
First, we will be analysing the variation of microbes in a single sample, known as the **alpha diversity**. Several commonly used alpha diversity measures exist, including Shannon, Chao, Chao-SE, Simpson and Observed. For this practical, we will calculate species diversity using Shannon's Diversity Index .

The Shannon diversity index combines richness and evenness. It measures both the number of species and the inequality between species abundances.

Let's get started, first, load the next script - ***BINF7001_M4P2_alphadiversity.R***
Install packages and load the libraries.
In this script you will load the **count table** (Input_data/species_counts.csv) which contains counts of taxa identified in each sample type. This count table contains only prokaryotic counts, and has been filtered for chloroplasts, mitochondria and eukaryotes. This script will also load the **metadata file**, so please ensure you have it stored in the Input_data folder as well. The metadata file contains the Sample_ID and Sample type. This file is very useful for carrying out further analyses.

Next, **run** the entire script. If you get a warning message about *missing libraries*, install them, or ask for help.
When running the R script, investigate the data and plot. The script calculates a range of indexes (species_richness_table) including the Shannon's Diversity Index, here the higher the value, the more diverse the community is. The Shannon range goes from 0 to 5, but in real-world ecological data, the Shannon diversity index's range of values is usually 1.5 - 3.5. Note that other indexes are different, e.g., the Simpson index goes from 0 (high diversity) to 1 (low diversity).
Q1: ***Is there a difference in diversity between the two sample types? Why do you think that one sample type may be more diverse than the other?***

Bran is a rich nutrient source, high fibre, cellulose, hemicellulose and lignin



Modify the Rscript and **plot other diversity indices**. Hint, look for *y=Shannon* in the ggplot command.

1.5) Explore microbial abundances with DESeq2

Visualising data is great for (amongst other things) data exploration, data comprehension, hypothesis generation and data communication. If we want to pinpoint specific taxa of interest that change in abundance between sample groups or are associated with a specific phenotype then we can run

statistical tests to identify ***differentially abundant* OTUs**. There are many such tests available and this is a very active area of research. When comparing the abundances of two sample groups for a given taxon, it is not sufficient to simply compare the means of the groups, as the variance may be much higher in one group than the other. The actual abundance may be very low and are subject to sampling issues, or the taxa abundances are not normally distributed (as it is almost always the case with ecological community data).

Calculation of differential abundance is statistically equivalent to the calculation of measurable changes of other variables, such as differentially expressed genes in a mouse model, or differential methylation of plant genes during development. It's fantastic when statistical methods can be applied to multiple different problems. Here, we will use a statistical method called DESeq2, which was originally developed for calculating differentially abundant genes from RNA-seq data.

1.5.1) Normalise the count table.

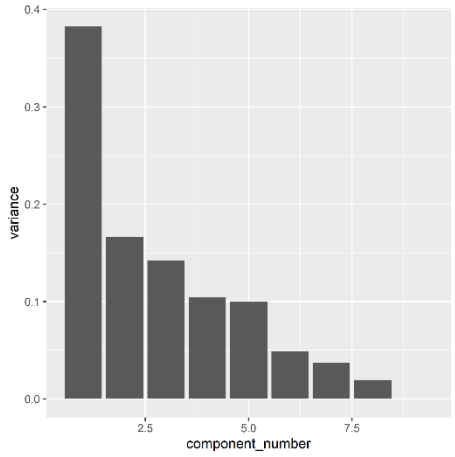
Run the file **BINF7001_M4P2_betadiversity.R** This calls a function which reads the count table file, extracts the taxonomy, and normalises the taxa counts with DESeq2 (for detail of this process see "DESeq2-normalized counts: Median of ratios method"). The count file contains the taxa counts at genus level, and we have prepared this file for you.

Run the script up to **line 59**, and check if the output file "*normalized_data.csv*" has been created.

1.5.2) Exploratory analysis with Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimension reduction technique and is especially useful when the data set contains more variables than samples. *Is this the case with our dataset?* PCA transforms the data into a new coordinate system and finds a linear line through the new transformation such that the **highest amount of variance between samples is captured**. This is referred to as the first principal component. The second principal component is a second line that explains the second most amount of variance, the third is the third most, and so on. In this way, lower components can typically be discarded as they explain very small amounts of variance. PCA has the desirable effect of **reducing the complexity** of the data into interpretable and manageable dimensions, and is frequently used for initial exploration of quantitative data and to search for possible correlations between samples. PCA results can be viewed in scatter plots at two or three dimensions. In this practical we will generate a PCA plot of the OTU table and determine overall similarity between samples.

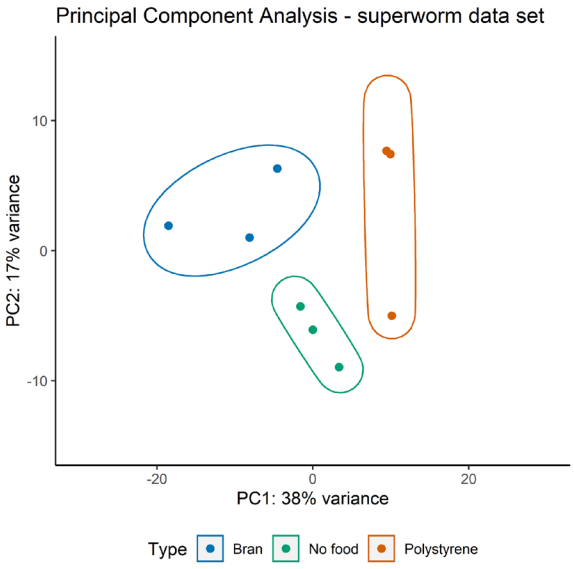
(1) In R, a **PCA object** can be created using the base level function **prcomp()**. Execute the commands on line 60-81 to remove low abundance taxa, compute the PCA result, and draw the scree plot. Scree plots show the proportion of the variance between the samples that each component explains. The first component (PC1) is on the far left. You can see each component explains less variance.



Q: *Roughly, how much variance is explained by the first two components? Is there a big difference between the second and third components? What does this result mean?*

(2) Next, we can **plot** the position of each sample in terms of PC1 and PC2 in a two dimensional plot, with the x-axis showing the first principal component and the y-axis the second principal component. We will also colour the samples according to sample type and group them. Run the code on lines 83-103 and explore the plot. *Do the groups (Bran, PS and CO) cluster together? Are there any taxa that are associated with one of the sample types? Are there any samples that are outliers? Why might this be?*

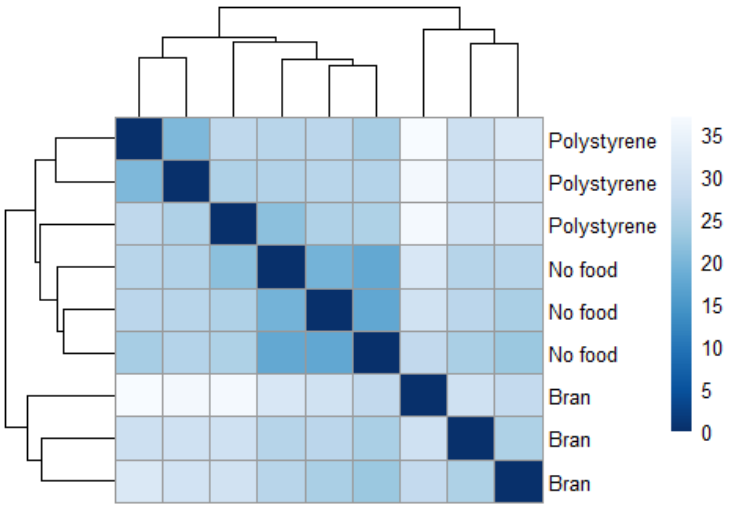
Within groups cluster, yes
PS has one outlier
Reason could be low number of reads, not effectively capturing community diversity



1.5.3) Exploratory analysis with Heatmap (pHeatmap package)

Another use of the transformed data is sample clustering. Here, we apply the *dist* function to the transformed count matrix to get sample-to-sample distances. Run the code from **line 106 to 118.**

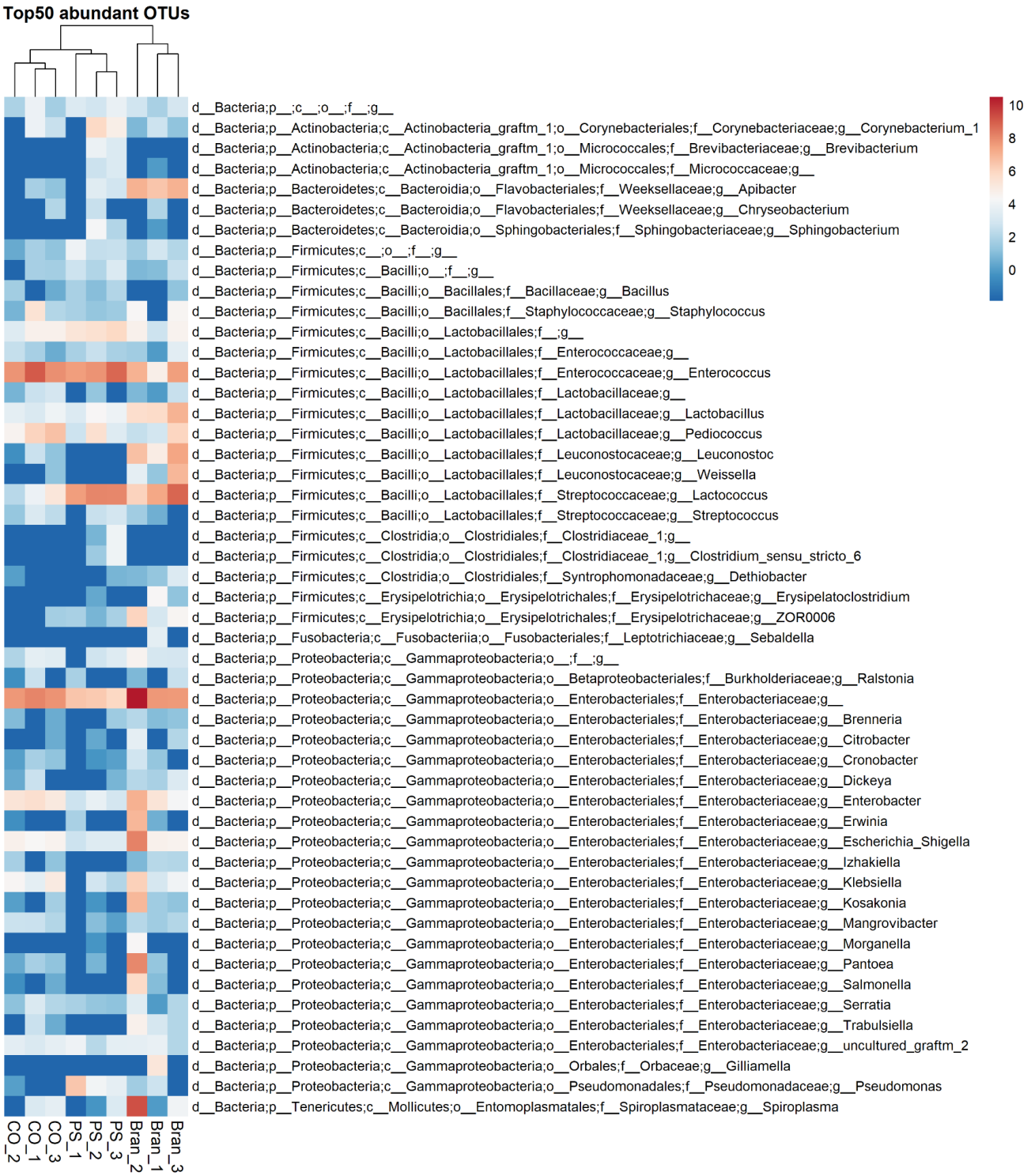
A heatmap of this **distance matrix** gives us an overview over similarities and dissimilarities between samples. We have to provide a hierarchical clustering hc to the heatmap function based on the sample distances, or else the heatmap function would calculate a clustering based on the distances between the rows/columns of the distance matrix.



Q: Which replicates cluster together? Do you see a grouping per sample type?

Heatmap of the most abundant community members

Execute lines 120 to 144, to create a pretty heatmap, and to find the most abundant taxa. Note, the output is written to a file. If you want R to display the heatmap: # filename = ... **Q:** Which taxa are highly abundant across all samples? Can you find taxa that are more abundant in any of the sample types?



Close this script, but do **NOT clear the R environment**, we'll need some of the data and values for the next script !

1.6) Determine which taxa are significantly more abundant in a sample type (Differential Abundance)

1. To begin, we have to decide which two 'conditions' we want to compare. The results of this DA (Differential Abundance) comparison will be a table showing for each variable (taxon) the mean abundance of each group (condition), the log2 fold change in abundance between the two groups (for example log2 (group1/group2)), and a p-value, indicating the likelihood of whether the *null hypothesis* is rejected. A common experimental design includes a control group and a treatment. In our experiment, we have polystyrene as a treatment group and two control groups (bran and starvation), so we can compare them to each other. To do this, we create a *design formula* where the dependent variable (i.e. the taxa abundance) is explained by the sample **Type** (; e.g. line 9; in the script, see below).

2. Run the Rscript ***BINF7001_M4P2_differentialabundance.R*** to perform the **differential abundance analysis** - this may take a few minutes (!).

3. Once the DESeq2 analysis has finished, we extract the results and place them into a data frame, then we plot the top differentially abundant taxa. The script produces a plot showing all the taxa on the y-axis with the x-axis being the mean abundance, shown as **log2 fold change** between the groups. The script will produce plots comparing PS vs bran, PS vs starvation - let's have a look at the latter one.

4. We are particularly interested in the taxa that are significant different. The standard practice is to use a cutoff when identifying "significant" taxa. DESeq2 produces a results table with p values, adjusted pvalues andLog2Fold change values for hypothesis testing. This means that, if the value for Log2FC is 1 and -1, the abundance is doubled and halved respectively. Adjusted p-value is the p-value corrected using Bonferroni-Hochberg correction to correct the false discovery rate. The lower the value, higher the significance of the OTU in the sample.
   Why do we log transform the data when comparing fold differences? Why might we use log2 rather than log10?

5. Next, try to extract the **significant taxa** from the results. Copy over the lines to the blank lines in the script and modify. Hint: add **%>% filter(padj < 1)** when converting the results to a dataframe with tibble (e.g. line 85, 86). This adjusted p-value results are 99% significant. *Tip: If you encounter any errors, start the whole script over again (i.e., run from beginning)*

6. Use all these results to answer these questions: **Q1:** Which taxa are significantly more abundant in PS compared to bran? **Q2**: Which are significantly more abundant in PS compared to the starvation group? **Q3:** Are there any taxa that are always more abundant in the PS group?

References

Geyer, Roland, Jenna R. Jambeck, and Kara Lavender Law. 2017. 'Production, Use, and Fate of All Plastics Ever Made'. Science Advances 3 (7): e1700782. https://doi.org/10.1126/sciadv.1700782.

Ballerstedt, Hendrik, Till Tiso, Nick Wierckx, Ren Wei, Luc Averous, Uwe Bornscheuer, Kevin O'Connor, et al. 2021. 'MIXed Plastics Biodegradation and UPcycling Using Microbial Communities: EU Horizon 2020 Project MIX-UP Started January 2020'. Environmental Sciences Europe 33 (1): 99. https://doi.org/10.1186/s12302-021-00536-5.