

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE TOULOUSE  
MATHEMATICAL AND MODELING ENGINEERING DEPARTMENT

October 2017 - January 2018

**Fifth-year project**

---

---

# Identification and characterisation of virome and bacteriome in calves suffering from infectious bronchopneumonia

---

---



AUTHORS :

Gicu Stratan  
Soizick Magon de La Giclais

TUTOR :

Nathalie Villa-Vialaneix



# Abstract

Infectious bronchopneumonia represents, with diarrhea, 30% to 90% of calf breeding mortality but there is almost no information in biology literature on these infections. The purpose of this project was to establish links between bacteria found from calf noses and those found in calf lungs, and also to explain the presence of a list of viruses using the abundance of these bacteria.

Hence, the data set under analysis gives information about infectious agents: some samples were taken from 23 calf breeding farms where the calves were suffering from infectious bronchopneumonia. We implemented statistical methods in order to analyse this data set, such as principal component analysis, sparse partial least squares discriminant analysis, partial least squares in regression mode and random forest. Then, healthy calf breeding data was compared with these results.

[...]

# Contents

Acknowledgment . . . . .	4
Introduction . . . . .	5
<b>1 Context and description of the data set</b>	<b>6</b>
1.1 Context . . . . .	6
1.2 Dataset description . . . . .	7
<b>2 Methods</b>	<b>11</b>
2.1 Normalisation . . . . .	11
2.2 Principal Component Analysis . . . . .	13
2.3 Partial least squares regression . . . . .	13
2.4 Random Forest . . . . .	15
<b>3 Results</b>	<b>16</b>
3.1 Exploratory analysis . . . . .	16
3.1.1 Normalisation . . . . .	16
3.1.2 Principal Component Analysis . . . . .	17
3.2 Differences between EN and LBA samples . . . . .	18
3.2.1 PLS-DA - paired samples . . . . .	18
3.2.2 PLS-DA - unpaired samples . . . . .	22
3.3 Predicting the presence of viruses . . . . .	23
3.3.1 In the nasal cavity (EN) . . . . .	23
3.3.2 In the lungs (LBA) . . . . .	26
Conclusion . . . . .	28
Bibliography . . . . .	29

# Acknowledgment

Nathalie Villa-Vialaneix, [...]

Elias Salem, [...]

# Introduction

# 1 Context and description of the dataset

## 1.1 Context

The provided dataset comes from biological studies performed on calf breeding farms during the acute phases of the respiratory disease complex (BRD). Bovine respiratory disease complex is considered one of the most common and economically important diseases in calves, it is displayed as pneumonia with different states of severity and respiratory signs depending on the causative agents and stressors (host immune reaction, environment...). Furthermore, a control group of healthy calves was sampled as well.

The dataset is divided into three files containing microbiota data; we call microbiota all the microbes that are found in a particular niche or habitat. In our case, the niche is the respiratory tract of the calves, especially in the nasal cavities (upper respiratory tract) and in the lungs (lower respiratory tract).

**First part:** Samples originating from 23 breeding farms taken from calves suffering from BRD, the calves were sampled during the acute phase (less than 3 days since the onset of symptoms) and were not vaccinated nor treated with antibiotics (in order not to intervene with the present microbiota). In each breeding farm, a group of 4-5 calves was studied, from which 2 kinds of samples were taken:

- A nasal swab, taken from the nasal cavity and representing the upper respiratory tract (called "EN");
- Bronchoalveolar lavage fluid, taken from the lungs and representing the lower respiratory tract (called "LBA").

For every pool of calves in the breeding farms and each of the sample types, we have the abundance of each of the detected bacteria, *i.e.* how many times the sequence of specific bacteria was detected.

The set also contains data indicating the presence/absence of seven major pathogens known to cause the respiratory disease; these data were generated by RT-PCR (specific and direct test).

**Second part:** The same kind of samples was taken from 6 breeding farms where the calves did not have any sign of respiratory disease.

In the next section, the details of each one of these data will be provided.

## 1.2 Dataset description

As previously explained, three text files are available for the study:

- **abundances:** bacteria counts where calves are suffering from infectious bronchopneumonia;
- **pathogenes:** presence/absence of seven pathogens targets;
- **abundances\_ctrl:** bacteria counts where calves are healthy.

They are all under *Comma-Separated Values* (CSV) format. In this chapter will be seen what they contain and what type of transformations have been performed, in order to make them suitable for further study.

### The abundances file

In this file, there is information (counts) about every bacteria found in every samples.

At the begining, the file contained 99 columns and 406 rows. A certain number of pretreatments have been performed in response to some findings on the dataset:

- **First nine columns were removed.** They contained the name of each bacteria, its type, its family, its "blast taxonomy" and technical information that we will not use. We are interested in the name of each bacteria (an identifier) and the measurements made in each sample;
- **Columns A and B were summed.** For each sample, there were two columns, identified by the letters *A* and *B*. It corresponds to the two technical replicates of all the farms involved in the study. These columns were merged (simple sums as the counts have already been normalized to identical library sizes), which leads to 45 columns;
- **Duplicates of bacteria were merged.** The list of bacteria names was not unique. We could have used the blast identifier but the problem would have remained the same. First, let us keep only the species name. It is the last one on each row. But sometimes the species is unknown, as we can see above. In this case, the name that appears before will be kept. Then, the replicates are merged by adding together each one of them, which lead us to a list of 270 unique bacteria.
- **Name of one bacteria was corrected.** One name of the bacteria was "&". It has been replaced by the correct one.
- **29<sup>th</sup> column was removed** (only for paired studies) . The number of columns was odd (45), but two samples were supposed to have been taken from each farm: one in the nose and the other in the lungs. It will be removed from the study when a paired analyse will be made.

These pretreatments lead us to a data frame with 270 rows and 45 columns. A part of this data frame can be seen in Figure 1.1.



	10_EN	11_EN	15_EN	16_EN	17_EN	18_EN	19_EN	01_EN	20_EN	21_EN	23_EN	24_EN	25_EN	26_EN	28_EN
&	39	0	6	15	7	9	26	16	39	8	0	0	112	178	
[Eubacterium] coprostanoligenes group	18	0	1	2	70	28	52	54	38	74	27	27	0	11	
[Ruminococcus] gauvreuii group	9	14	0	1	0	27	39	32	0	0	0	0	27	87	5
Acetobacteraceae bacterium SAP1007.2	15	0	0	1	3	0	6	9	0	0	0	0	0	0	
Acholeplasma laidlawii PG-8A	0	0	2	2	15	16	14	31	61	21	0	0	182	84	
Achromobacter sp.	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
Achromobacter spanius	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
Actinoalloteichus cyanogriseus	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
Aerococcaceae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Aerococcus sp.	25	0	2	0	0	4	0	0	96	102	10	5	30	44	
Aeromicrobium	0	0	2	2	0	0	0	2	0	0	1	0	0	0	
Agreia	0	0	0	0	0	0	0	0	0	0	0	1	7	0	
Agrobacterium tumefaciens	0	0	2	0	0	0	0	0	0	0	2	2	0	2	
AKAU3644	0	1	0	0	0	0	0	6	0	1	0	0	0	1	
Alcaligenes faecalis	1	44	0	0	0	2	16	8	23	38	5	14	10	21	
Alcaligenes sp.	88	67	0	0	23	13	19	17	24	23	43	43	67	60	
Alloprevotella	0	0	1	0	0	0	0	14	32	0	3	0	0	1	228
Alysiella crassa	0	5	0	1	0	0	0	0	2	0	1	0	0	0	
Anaerosporebacter	0	0	0	0	0	0	3	0	0	0	0	0	0	0	
Aquamicrobium	0	0	43	46	14	31	71	96	48	20	0	24	33	85	1
Arcobacter	12	21	1	2	0	0	0	79	109	11	0	11	59	2	
Arcobacter cryaerophilus	0	0	7	3	9	9	35	0	22	9	0	0	0	9	1
Arenimonas	0	0	0	0	1	0	0	0	0	0	1	0	2	0	
Arthrobacter	183	154	4719	4920	6	25	188	261	456	531	9	112	273	398	2
Arthrobacter arilaitensis	0	0	8	2	30	7	19	6	12	0	0	0	0	11	

Figure 1.1: *Abundances data*

## The pathogenes file

The second file is composed of 9 columns and 46 rows. This file contains the presence or not of seven viruses for each farm in the two condition. Some pretreatments have been performed:

- **Farm identifier and condition were merged.** The first and the last columns stand for the identifier of each farm and the sampling condition. A unique identifier is needed, hence these two columns were merged.
- **Binary code for presence/absence of virus.** The presence of a virus was coded with a letter, and the absence with 0. In order to simplify the analysis, it will be replaced as 1 and 0 respectively and considered as a factor variable.

Unlike the previous file, there is no missing farm: there is 23 paired individuals. A part of the data set is displayed in Figure 1.2.

The observed viruses are the following:

- "Ct.RSV" for respiratory syncytial virus
- "Ct.PI.3" for parainfluenza virus
- "Ct.Coronavirus" for coronavirus
- "Ct.P.multocida" for pasteurella multocida virus
- "Ct.M.haemolytica" for mannheimia haemolytica virus
- "Ct.M.bovis" for mycobacterium bovis virus
- "Ct.H.somni" for histophilus somni virus

	Ct.RSV <sup>+</sup>	Ct.PI.3 <sup>+</sup>	Ct.Coronavirus	Ct.P.multocid <sup>+</sup>	Ct.M.haemolytic <sup>+</sup>	Ct.M.bovis	Ct.H.somni
10_EN	0	0	1	1	0	0	1
11_EN	0	1	1	1	0	0	0
15_EN	0	0	1	1	1	0	1
16_EN	0	0	1	0	0	0	0
17_EN	0	0	0	1	0	0	0
18_EN	0	0	0	1	0	0	1
19_EN	0	0	1	1	0	0	0
01_EN	0	0	1	1	0	0	1
20_EN	0	0	1	1	0	0	1
21_EN	0	0	1	1	1	0	0
23_EN	0	0	1	1	0	1	0
24_EN	0	0	0	1	0	0	1
25_EN	0	0	0	1	1	1	1
26_EN	0	0	0	1	1	0	0
28_EN	0	0	0	1	1	0	0
30_EN	0	0	0	1	1	0	0
31_EN	0	0	1	1	0	0	0
03_EN	0	0	1	1	1	0	1
04_EN	1	1	1	1	0	0	0
06_EN	0	0	0	0	0	0	0
07_EN	0	0	0	1	1	0	0
09_EN	0	0	1	1	0	0	1
10_LBA	0	0	1	1	0	0	1
11_LBA	0	1	0	1	0	0	0
15_LBA	0	0	0	1	0	1	1
16_LBA	0	0	1	0	0	0	1

Figure 1.2: *Pathogenes data*

In addition to the first file, this one will allow us to study links between virus and bacteria.

## The abundances\_ctrl file

This file has the same design as the **abundances** file and has been added in the middle of the project. It has been treated the same way. The first difference is that this file corresponds to the samples taken from healthy calves in 6 breeding farms, which gives us 12 samples.

The other difference concerns the list of bacteria. The bacteria found in healthy individuals are not the same as sick individuals. This list is shorter (293 rows) and the bacteria are not the same.

This list is still not composed by unique bacteria names. It will be fixed by taking the last name of each row as explained before.

Now that this list is composed by unique bacteria names, let us see how much they are different from the **abundances** file in Figure 1.3. There is 117 common bacteria, hence this list of these bacteria will be used for comparison analyses.

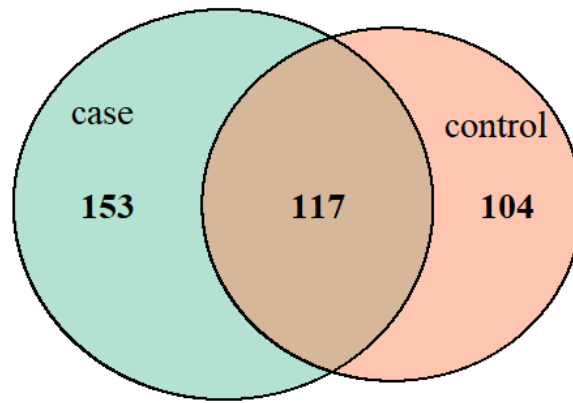


Figure 1.3: *Venn diagram comparing two lists of bacteria: for sick calves and for healthy calves*

Figure 1.4 allows us to compare the total counts for case and control samples (only in common bacteria). Total counts are higher in case samples, hence only the relative abundance will be taken into account for comparison analyses.

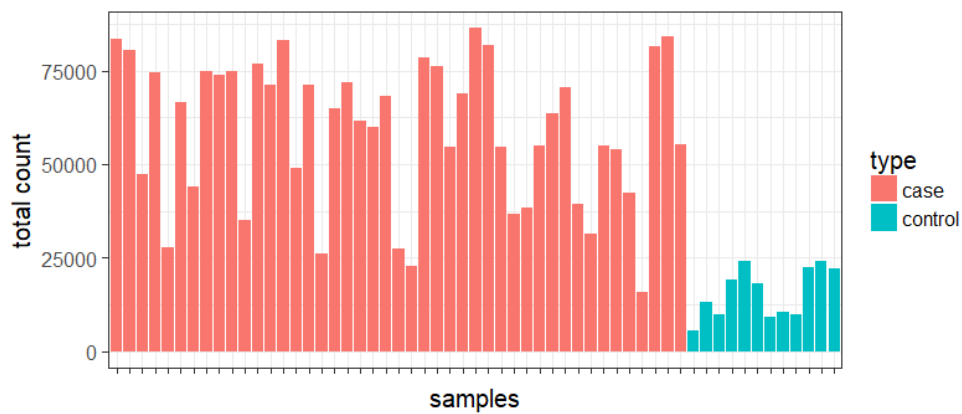


Figure 1.4: *Total counts for control and case sample*

## 2 Methods

### 2.1 Normalisation

The purpose of the normalisation is to make variable distributions more symmetrical. Furthermore, some statistical methods are sensitive to non-normalised data: most of the classification methods compute the distance between two points using eucliden distance. If one of the variable has high values comparing to the others, it will have a greather influence on the final distance, and consequently will lead to a higher variance on prediction quality.

Figure 2.1 represents the distribution of the first simple, with a log scale for the frequencies.

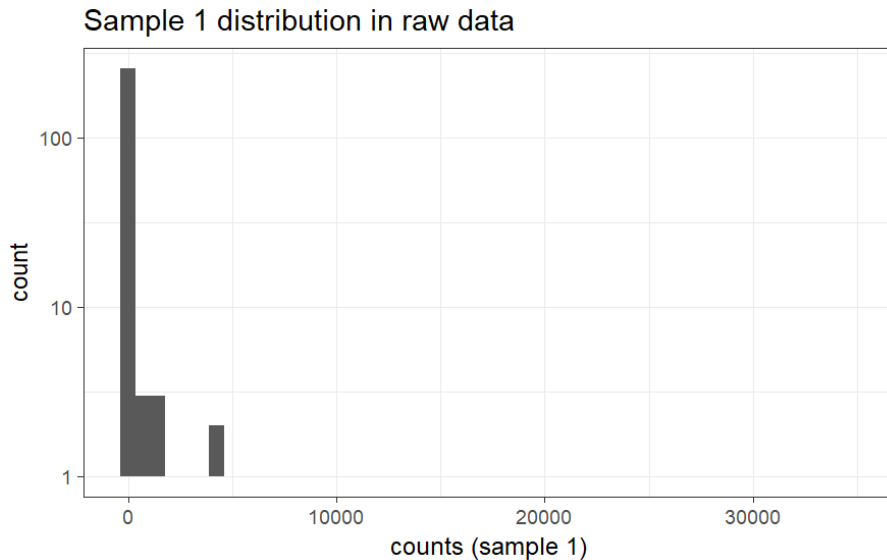


Figure 2.1: *Sample 1 distribution in raw data*

What can be noticed in the figure is that there is a lot of very low values, hence the distribution is highly skewed. It was expected because the observations are sequencing counts. In order to analyse this dataset, we will try several transformations known for microbiote datasets [4].

Let  $y_{i,j}$  be the counts of a bacteria  $j$  found in the sample  $i$ , where  $i \in [1, \dots, n]$  and  $j \in [1, \dots, p]$ ,  $n$  and  $p$  standing for the number of samples and the number of bacteria respectively. Let  $\hat{y}$  be the data after transformation.

## Log transformation

$$\tilde{y}_{ij} = \log(y_{ij} + 1)$$

## TSS transformation

It is common in metagenomic datasets to perform TSS (*Total Sum Scaling*) before further normalization. TSS transformation computes relative abundances:

$$y_{ij} = \frac{n_{ij}}{\sum_{k=1}^p n_{ik}}$$

for  $n_{ij}$  the counts of species  $j$  in sample  $i$ ,  $p$  the number of species and  $n$  the number of individuals.

## TSS+CLR transformation

CLR (*Centered Log Ratio*) transformation:

$$\tilde{y}_{ij} = \log \frac{y_{ij}}{\sqrt[p]{\prod_{k=1}^p y_{ik}}}.$$

## TSS+ILR transformation

ILR (*Isometric Log Ratio*) transformation:

$$\tilde{\mathbf{Y}}' = \tilde{\mathbf{Y}} \times \mathbf{V}$$

for  $\tilde{\mathbf{Y}}$  the matrix of CLR transformed data and a given matrix  $\mathbf{V}$  with  $p$  rows and  $p - 1$  columns such that  $\mathbf{V}\mathbf{V}^\top = \mathbb{I}_{p-1}$  and  $\mathbf{V}^\top \mathbf{V} = \mathbb{I} + a\mathbf{1}$ ,  $a$  being any positive number and  $\mathbf{1}$  a vector full of 1.

## CSS transformation

The CSS (*Cumulative Sum Scaling*) normalisation is an adaptative extension for metagenomic data of the quantile normalisation used in microarray expression datasets. It is designed so as to account for technical differences between samples.

## 2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a multidimensional descriptive method that allows to explore the links between variables and the similarities between individuals.

The objective of the PCA is to identify the largest sources of variation and to project the data into a space of reduce dimension (for example 2 or 3) by maximizing the variability of the projection. In other words, PCA reduces the number of variables, in this way the information becomes less redundant.

From a mathematical point of view, the PCA corresponds to the approximation of a matrix  $(n, p)$  by a matrix of the same dimensions but of rank  $q < p$ . In fact, an orthogonal linear transformation is applied on the data to convert a set of observations of possibly correlated variables into uncorrelated principal components.  $q$  is often of small value 2, 3 and contributes to the construction of easily understandable graphs. The interpretation of these graphs helps to understand the structure of the analyzed data.

## 2.3 Partial least squares regression

Partial least squares regression (PLS) [1] is a fast, efficient and optimal statistical method that is widely used to deal with situations with high multicollinearity and when a large number of variables have to be taken into account. Its use is recommended in the case where the number of variables  $p$  is much greater than the number of individuals  $n$ :  $p \gg n$ , which is relevant for our dataset.

There are different versions of PLS regression depending on the objective:

- **PLS1**: A quantitative target variable  $Y$  is to be explained, modeled, predicted by  $p$  quantitative explanatory variables  $X^j$
- **PLS2**: Canonical version. It relates a set of  $q$  quantitative variables  $Y^k$  and a set of  $p$  quantitative variables  $X^j$ .
- **PLS2**: Regression version. It tries to explain, model a set of  $q$  quantitative variables  $Y^k$  by a set of  $p$  quantitative explanatory variables  $X^j$ .
- **PLS-DA**: Discriminant version. Special case of the previous case. The qualitative variable  $Y$  with  $q$  classes is replaced by  $q$  dummy variables of these classes.

During this project, only the PLS-regression mode and PLS-DA versions were used because they are adapted to our dataset and the pursued objective.

## PLS-regression algorithm

$X$  is the matrix of the explanatory centered variables

$Y$  is the centered variable to explain, uni or multidimensional

Initialization of the latent variable  $\omega_1$  by the first column of  $Y$

Set  $r$  the number of iterations

For  $h = 1$  to  $r$  do:

While Convergence Not Achieved do:

$$u_h = X' \omega_h / \omega_h' \omega_h$$

$u_h = u_h / u_h' u_h$  where  $u_h$  is the loading vector associated with  $X$

$\xi_h = X u_h$  where  $\xi_h$  is the latent variable associated with  $X$

$$v_h = Y' \xi_h / (\xi_h' \xi_h)$$

$v_h = v_h / v_h' v_h$  where  $v_h$  is the loading vector associated with  $Y$

$\omega_h = Y' v_h$  where  $\omega_h$  is the latent variable associated with  $Y$

end while

$c_h = X' \xi_h / \xi_h' \xi_h$  partial regression of  $X$  on  $\xi$

$d_h = Y' \omega_h / \omega_h' \omega_h$  partial regression of  $Y$  on  $\omega$

Deflation  $X_h = X_{h-1} - \xi_h c_h'$

Deflation  $Y_h = Y_{h-1} - \xi_h v_h'$

end for

Convergence is reached when the following vectors verify:

$$YY'XX'u = \lambda u$$

$$Y'XX'Y\omega = \lambda \omega$$

$$XX'YY'v = \lambda v$$

$$X'YY'X\xi = \lambda \xi$$

where  $u, \omega, v, \xi$  are the respective eigenvectors of the matrices  $YY'XX'$ ,  $Y'XX'Y$ ,  $XX'YY'$ ,  $X'YY'X$  associated with the same greater eigenvalue  $\lambda$ .

## PLS-Discriminant Analysis algorithm

The only difference compared to the PLS-regression algorithm consists in the variable to explain  $Y$  which is this time qualitative with  $m$  modalities. It is enough to transform the variable  $Y$  in  $m$  dummy variables and to apply on this the algorithm above.

## sPLS-regression and sPLS-DA methods

PLS-regression is a method that effectively resolves problems of multicollinearity or too many variables. The price to pay is often the increase in the complexity of the interpretation of the results. To make interpretation easier, it is necessary to limit the number of variables participating in each linear combination. sparse PLS fulfills this objective and is capable to perform a variable selection in the data set by introducing LASSO penalization on the pair of loading vectors.

sparse PLS-DA is a special case of sparse PLS and allows variable selection with respect to different classes of the samples.

## 2.4 Random Forest

Random Forest is a machine learning method for classification and regression generally. This algorithm is based on the aggregation of decision, regression or classification binary trees, depending on the type of the target variable  $Y$ .

It consists of drawing  $n\_estimators$  bootstrap samples. A bootstrap sample is a sample constructed by  $n$  random draws with replacement among the  $n$  initial observations. The distribution of this sample is the empirical one which gives a  $\frac{1}{n}$  weight to each successful draw. A tree is estimated for each bootstrap sample and the prediction is obtained by averaging (regression) or vote (classification) the individual predictions of each tree. During the construction of each tree's node in the learning process, a random subset of the features is selected, that allows to build less correlated trees and make aggregation more efficient [1]. It is generally an effective method, which could give satisfying predictions for presence/absence of viruses.

Let  $Y$  be the variable to predict and  $z = (x_1, y_1), \dots, (x_n, y_n)$  a sample, and  $B$  the number of replicates of bootstrap samples each obtained by  $n$  draws with replacement.

### Random Forest algorithm

For  $b = 1$  to  $B$  do:

    Draw a bootstrap sample  $z_b^*$

    Estimate a tree on this sample with randomization of variables :  
    the search for each optimal division of the tree is preceded by a  
    random draw of a subset of  $m$  predictors

Calculate the average estimate  $\hat{f}_B(Y) = \frac{1}{B} \sum_{b=1}^B \hat{f}_{z_b}(Y)$  or the result of  
the vote.



# 3 Results

## 3.1 Exploratory analysis

### 3.1.1 Normalisation

Previously, we saw that the first sample distribution was highly asymmetrical due to the fact that it is from counts data.

If we look at the graphs in the Figure 3.1, the less asymmetric distribution seems to be the one obtained with the CSS transformation and the log-transformation, but there is still a lot of very low values.

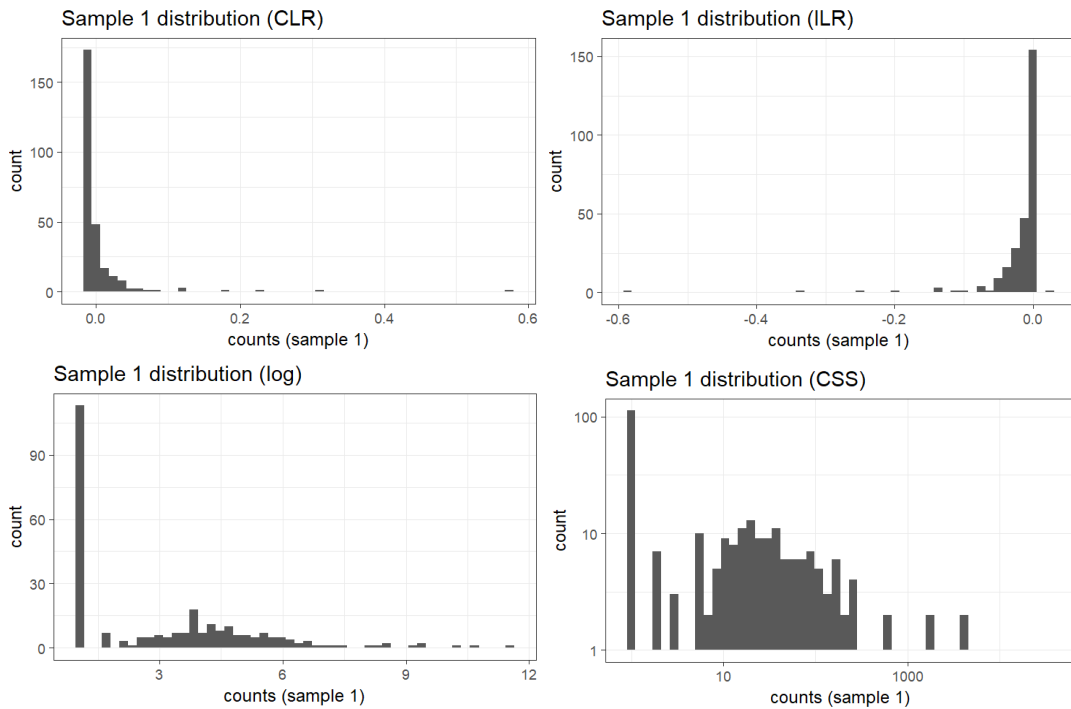


Figure 3.1: *Sample 1 distribution according to the type of transformation*

Let us have a look at the distribution by boxplots of all the samples displayed in the Figure 3.2. It confirms that the best transformations are CSS and logarithmic. It seems that the CSS normalisation treats outliers more efficiently than the log one, but the relative levels are important in our case. In the following will be used the log-transformation.

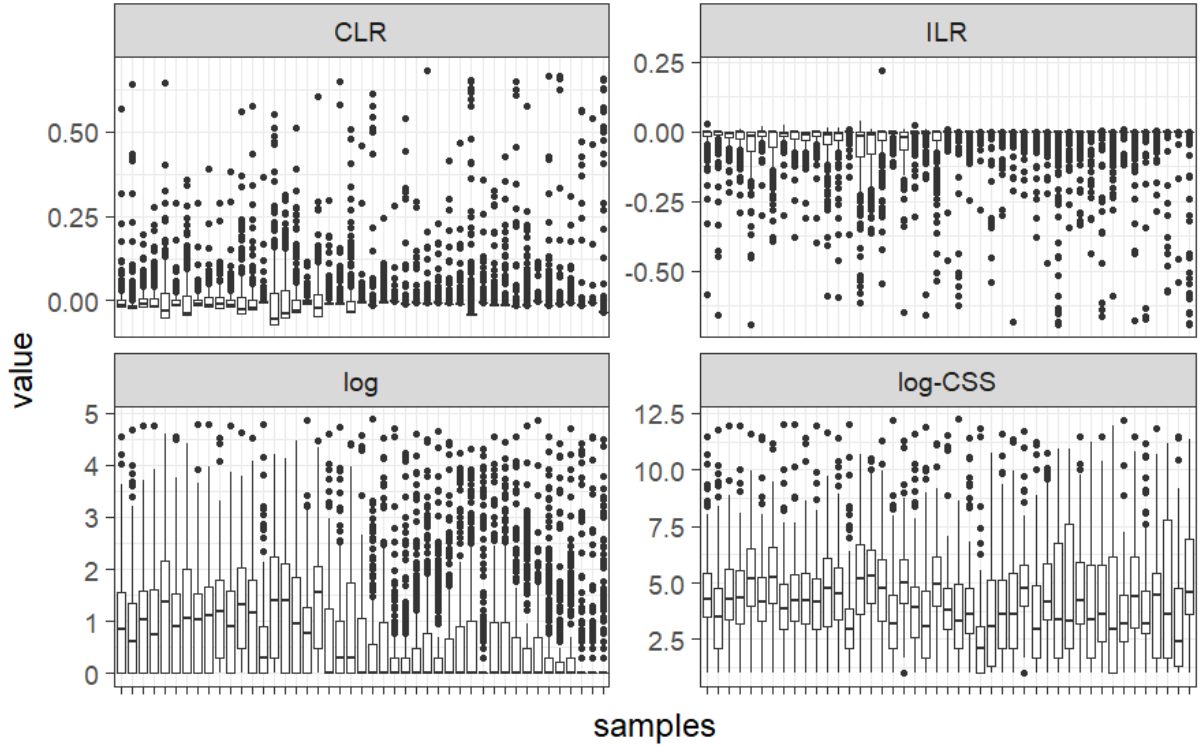


Figure 3.2: *Distributions of all samples according to the type of transformation and the sample*

### 3.1.2 Principal Component Analysis

It is important to have a clear idea of the data structure. A principal component analysis is adapted to this objective. In the following, we will analyze the results of the PCA on the Log and CSS transformations made on the observation matrix because the two others show very similar results. It has been performed in R, using the `mixOmics` library [2].

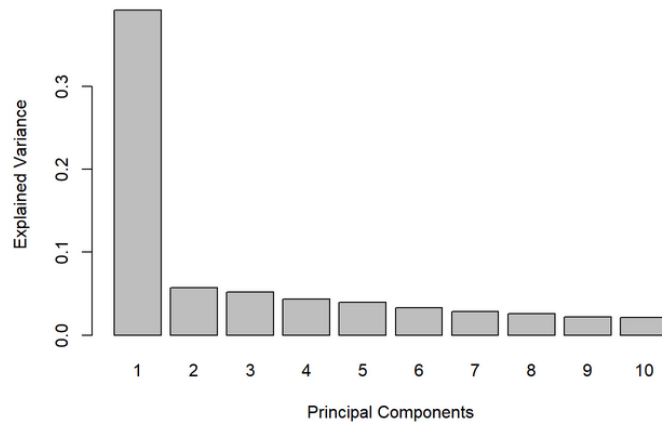


Figure 3.3: *Sample 1: The variance explained by the first 10 components*

Figure 3.3 represents the decay of the eigenvalues associated to the principal components. It seems that the first component on 406 is sufficient to reconstruct the data. This is

explained by the rapid decay of the firsts ten eigenvalues, which are not important beyond that. The first axis, showing the projection of the data on the first principal component seems to illustrate its stability in the construction of the representation subspace.

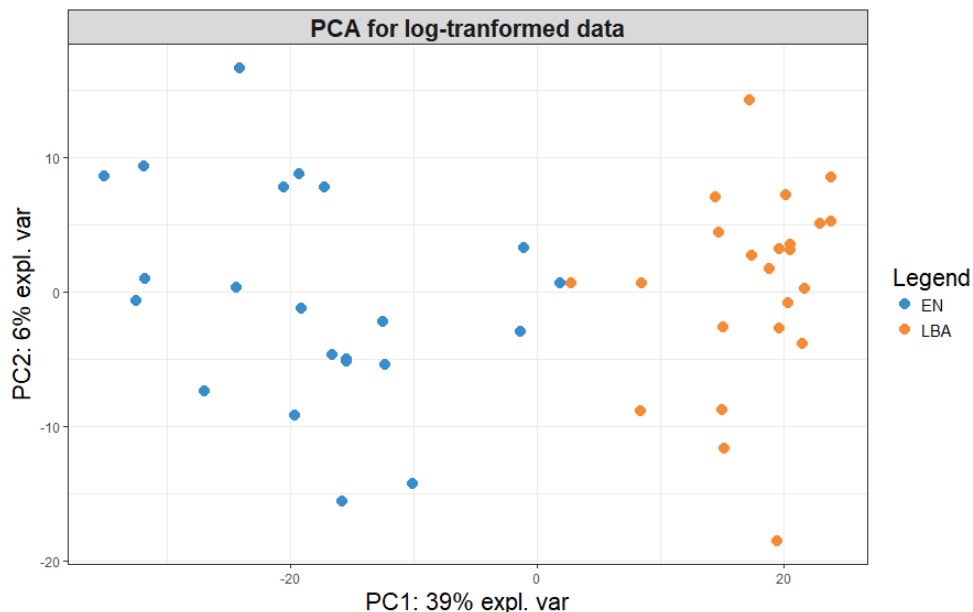


Figure 3.4: *Sample 1: Projection of individuals in the first factorial plan*

The projection of individuals in the first factorial plan displayed in Figure 3.4 shows that the separation of "EN" and "LBA" classes is strong, and can even be divided linearly. Also, it can be seen that the first axis associated to the first principal component explains 39% of variance and with the second, they contribute mostly to the variability of the data.

## 3.2 Differences between EN and LBA samples

The purpose of this part is to select bacteria that have an impact on the EN/LBA samples. The methods used are PLS-DA, and more specifically sparse PLS-DA. They will be performed using paired or unpaired analysis, which means indicating the sparse PLS-DA that our samples are paired or not. The library used to implement this method remains `mixOmics` [2].

### 3.2.1 PLS-DA - paired samples

In this section we will focus on a multilevel analysis, therefore the 29<sup>th</sup> column will be removed (section 1.2 page 7) for the multilevel option (in the 4<sup>th</sup> step below).

- **Step 1:** Performance of the PLS-DA on the first two components. A first PLS-DA is computed (with 10-fold cross-validation repeated 10 times) to check the efficiency of the method and which type of distance to use in its computation.

Figure 3.5 indicates that the classification error rate decrease from one component to two components in the model, for three different distances: Maximum, Centroids

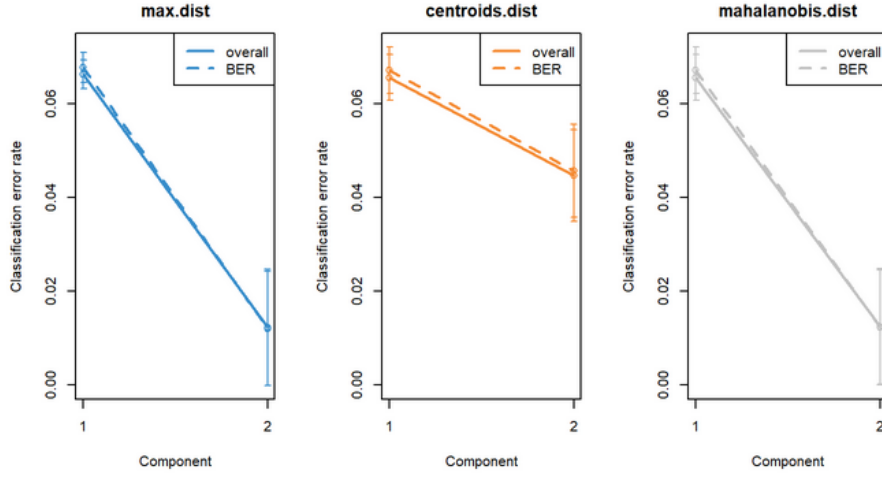


Figure 3.5: *Distances for the first two components*

and Mahalanobis. BER stands for Balanced Error Rate, and here the groups are unbalanced because of the 29<sup>th</sup> column, which explains the small difference between the two lines.

- **Step 2:** Projection of individuals (left graph in Figure 3.7). The sample plot shows the PLS-DA first two components. The ellipse are 0.95 confidence interval ellipse for each condition type.
- **Step 3:** Tuning sparse PLS-DA. In Figure 3.6, the error rate decreases when 2 components are included in sPLS-DA. The diamonds stand for the optimal number of variables to select on components 1 & 2. In this case, 5 variables will be selected for the first component and 15 for the second.

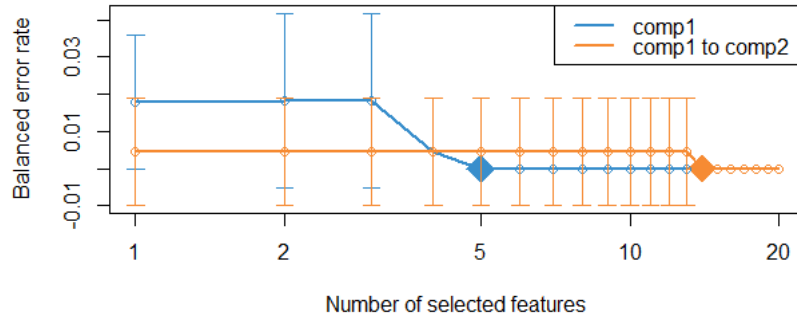


Figure 3.6: *Error rate for the first two components by the number of selected variables*

- **Step 4:** Running sPLS-DA tuned (right graph in Figure 3.7). This step consist of running the sPLS-DA with new parameters such as the number of selected variables and the multilevel option. The multilevel option will indicate that our individuals are paired, *i.e.* that two samples have been taken from the same farm.

What can be noticed is that the clusters are totally separated by the first component on the sPLS-DA graph.

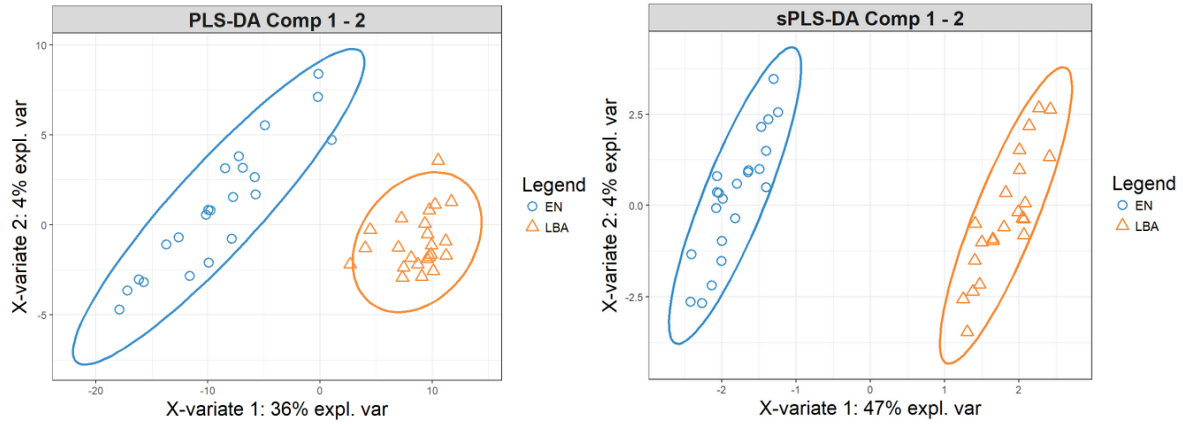


Figure 3.7: *Projection of individuals for PLS-DA and sPLS-DA first two components*

The list of selected variables and their loading values are represented for each component in Figure 3.8 below. All the bacteria describing the first component are highly abundant in EN samples.

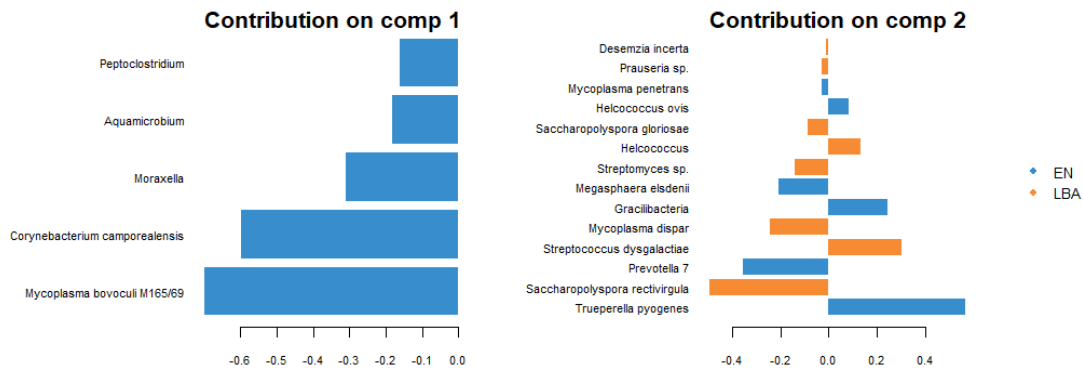


Figure 3.8: *Loadings values of selected bacteria on the first two components*

### Tests on the extracted bacteria

In this section two complementary analyses are performed: Student tests to see whether the difference between EN and LBA is significant for the selected bacteria for the first component, and boxplots to display the abundances difference of these bacteria between EN and LBA samples.

	bacteria	pvalue	FWER
1	Mycoplasma bovoculi M165/69	3.728266e-12	1.553412e-11
2	Corynebacterium camporealensis	6.213648e-12	1.553412e-11
3	Moraxella	2.367632e-11	3.946053e-11
4	Aquamicrobium	4.077402e-11	4.484280e-11
5	Peptoclostridium	4.484280e-11	4.484280e-11

According to the p-values obtained, all of the bacteria selected are significantly different in the two conditions. Moreover, the boxplots (Figure 3.9) reveal that their abundances are higher in the nose than in the lungs.

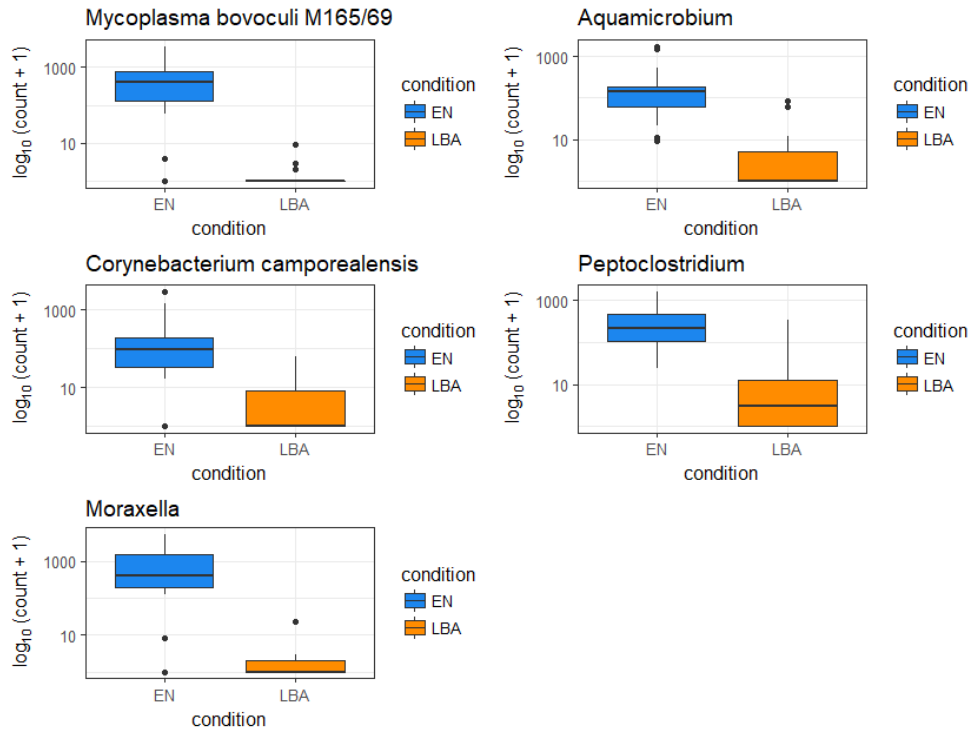


Figure 3.9: *Boxplots of abundance differences between the two conditions*

### Links with control samples

In order to compare infected samples and healthy samples, only the common bacteria are studied. Only three bacteria among the five selected are in the control sample data.

Figure 3.10 shows that, except for *Mycoplasma bovoculi*, the differences between LBA and EN is not related to the status (case/control) of the sample.

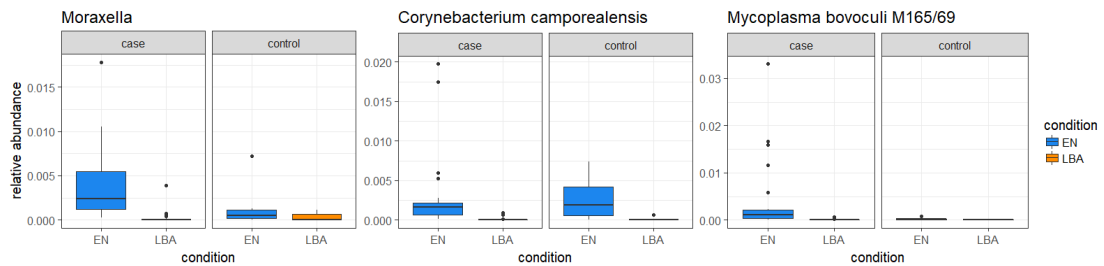


Figure 3.10: *Boxplots of abundance differences between the two conditions for healthy and sick samples*

### 3.2.2 PLS-DA - unpaired samples

The same process was performed removing the multilevel option.

**AJOUTER QQUES GRAPHS QUI MONTRENT QUE C'EST PAS MIEUX**

### 3.3 Predicting the presence of viruses

The purpose of this section is to understand the mechanisms of relationships between the two microbiota (EN and LBA) and several viruses presence.

#### 3.3.1 In the nasal cavity (EN)

In this part, the results presented were obtained from the samples characterizing only the nasal cavity.

##### sPLS in regression mode

To understand better the similarities between samples in body site EN, an exploratory analysis is performed using sPLS-regression mode method with abundance data as explanatory variables and the presence of viruses as variables to predict.

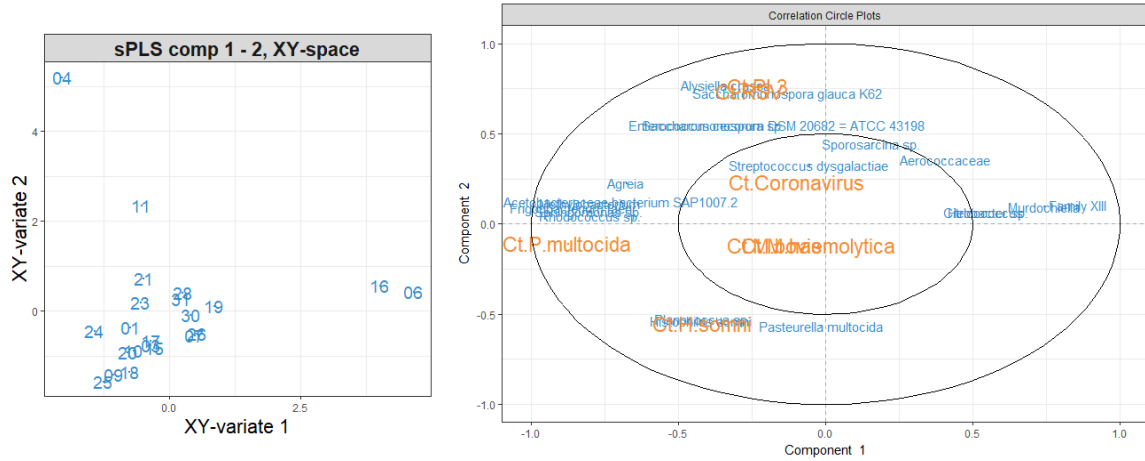


Figure 3.11: *sPLS on XY-space and Correlation circle plot for the first two components*

In the first graph in Figure 3.11, can be seen the samples projected onto the mean subspace in which the coordinates are averaged from the first two subspaces that correspond to the first two components of sPLS. The cloud of points is quite concentrated which seems to explain the similarities between the individuals in EN, but also presents three outliers that draw the definition of the axes. It seems that the individuals 06 and 16 have a strong influence on the first axis and that the individual 04 on the axis 2.

The second graph (where bacteria are colored in blue and viruses in orange) represents the projection of the variables selected by sPLS onto a correlation circle and shows:

- The first axis is mostly driven by the absence of *Ct.P.multocida* that is associated to a small group of bacteria, while its presence is associated to a larger group of bacteria.
- The second axis is mostly driven by the opposition between the presence of *Ct.PI.3* and *CT.RSV* that is opposed to the presence of *Ct.H.somni*. The first two viruses are associated to sample 04 (probably) and to the presence of a small group of bacteria, while *Ct.H.somni* is associated to the presence of another group of bacteria.



## PLS-DA

In this part, the presence or absence of viruses is predicted from abundance data. As the presence of most viruses is very rare (with very unbalanced data sets), a group of three viruses *RSV*, *PI.3*, *Coronavirus* is defined and used as a target for prediction. The union of these viruses provides a more balanced sample.

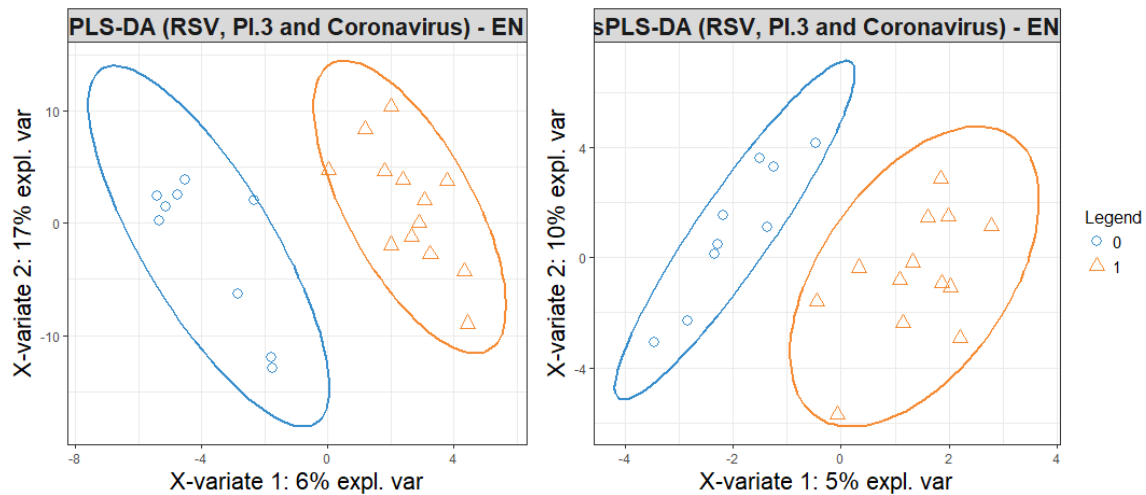


Figure 3.12: *PLS-DA and sPLS-DA on the groupe of three viruses*

In Figure 3.12, the clusters that represent the projection of the individuals on the first two components and that characterize the presence/absence of the group of viruses are separated, but not only according to the first component. This is due to the orientation of the ellipses and the fact that the first two components of the PLS-DA explain very small variance (23% in the first graph and 15% after variables selection). Despite this, the first component seems to contribute more to this separation.

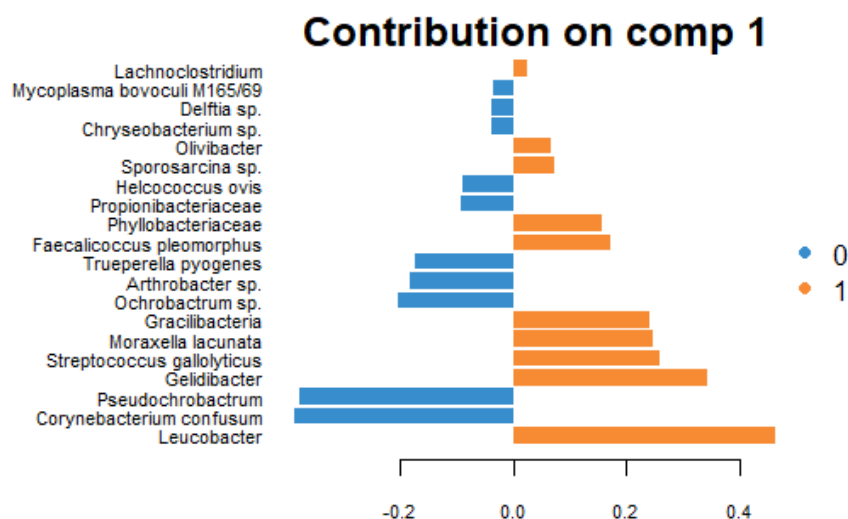


Figure 3.13: *Loading values of selected variables on first component*

The contribution plot (Figure 3.13) above displays the abundance of each bacteria and in which body site **?presence/absence?** they are the most abundant for the first sPLS-

DA component. For example, it seems that *Leucobacter* is abundant when the group of viruses is present in EN and vice versa for *Corynebacteriumconfusum*.

### Student tests

To evaluate the significance of the selected bacteria in the process of appearance or absence of these three viruses, Student's tests are performed for each of them and the results obtained are presented bellow.

	bacteria	pvalue	FWER
1	Leucobacter	0.03481991	0.08704976
2	Corynebacterium confusum	0.03045623	0.08704976
3	Pseudochrobactrum	0.03477661	0.08704976
4	Gelidibacter	0.01618017	0.08704976
5	Streptococcus gallolyticus	0.02026699	0.08704976
6	Moraxella lacunata	0.01911170	0.08704976
7	Gracilibacteria	0.02123296	0.08704976
8	Ochrobactrum sp.	0.07461385	0.12304604
9	Arthrobacter sp.	0.09038226	0.12304604
10	Trueperella pyogenes	0.04204057	0.09342350
11	Faecalicoccus pleomorphus	0.02917639	0.08704976
12	Phyllobacteriaceae	0.08753570	0.12304604
13	Propionibacteriaceae	0.15041696	0.16178188
14	Helcococcus ovis	0.13733066	0.16178188
15	Sporosarcina sp.	0.05414143	0.10828286
16	Olivibacter	0.09228453	0.12304604
17	Chryseobacterium sp.	0.14258569	0.16178188
18	Delftia sp.	0.16064938	0.16178188
19	Mycoplasma bovoculi M165/69	0.06573544	0.11951897
20	Lachnoclostridium	0.16178188	0.16178188

The p-values obtained lead to the conclusion that none of the bacteria are found differentially abundant between infected and non infected samples with this group of viruses.

### Common bacteria

To confirm or refute the conclusion of the previous part, a comparison of the relative abundance is performed for the selected species that are also found in the control samples data.

Figure 3.14 shows that the difference between case and control samples in terms of relative abundance of common bacteria seems to be insignificant. It appears that the absence or presence of the group of viruses does not affect the quantity of bacteria studied.

The same study was done on the other viruses separately and did not lead to conclusive results. This is due to unbalanced and small sized samples.

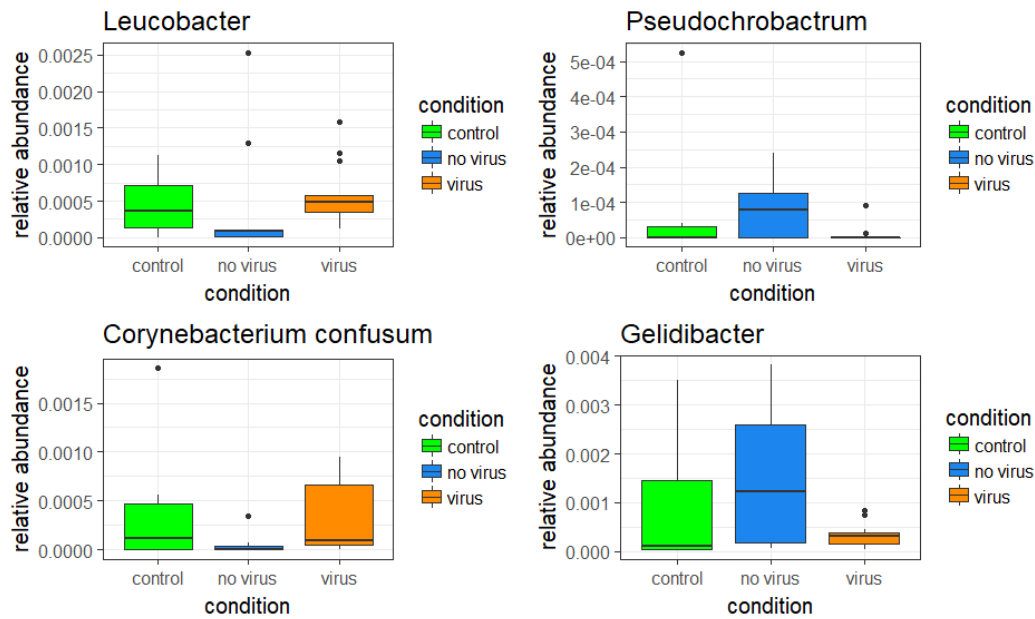


Figure 3.14: *Boxplots on ...*

## Random Forest

In view of the previous results, another method was tested to predict the absence or presence of viruses from abundance data. This is the RandomForest algorithm.

EN:

```
## 0 1 class.error
## 0 1 8 0.88888889
## 1 1 12 0.07692308
```

LBA:

```
## 0 1 class.error
## 0 0 8 1.0000000
## 1 2 12 0.1428571
```

### 3.3.2 In the lungs (LBA)

#### sPLS in regression mode

The same exploratory analysis is performed for LBA samples.

As before, it can be noticed the similarities between most samples in LBA and also four outliers that influence the orientation of the axes. The samples 30 and 26 seems to act on axis 1 while 15 and 25 on axis 2.

The correlation circle in Figure 3.15 shows :

- The first axis is mostly driven by the presence of *Ct.M.haemolytica* (in samples 30 and 26 probably) that is associated to a large group of bacteria, while its absence is associated to the presence of *Ct.P.multocida* and *Ct.H.somni* and a small group of two bacteria.

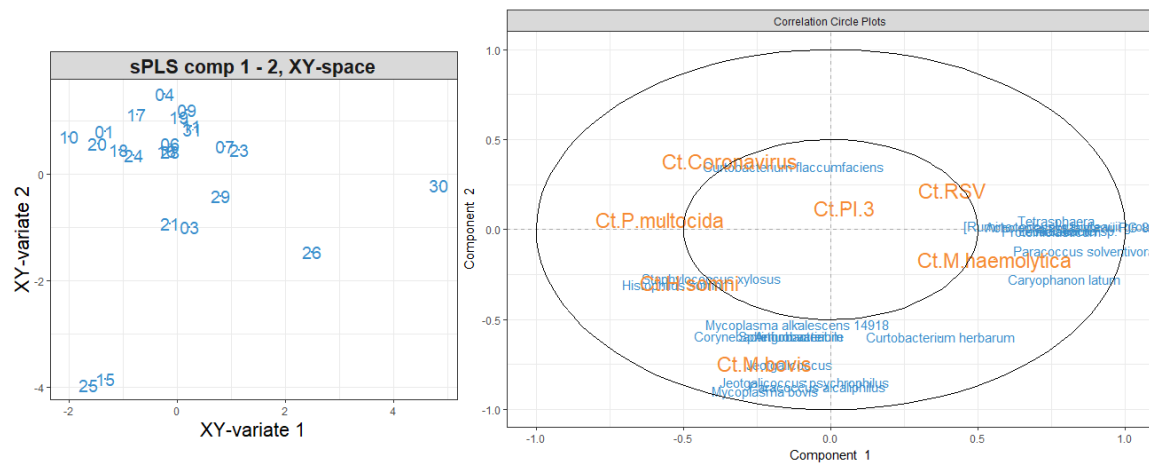


Figure 3.15: *sPLS on XY-space and Correlation circle plot for the first two components*

- The second axis is mostly driven by the presence of *Ct.M.bovis* (probably present in samples 15 and 25) that is associated to a large group of bacteria while its absence is associated to the presence of *Ct.Coronavirus* and only one bacteria.

## PLS-DA

Student tests

Common bacteria

Random Forest

# Conclusion

# Bibliography

- [1] Philippe Besse. *Science des données : Apprentissage Statistique*. 2017.  
<https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-Intro-ApprentStat.pdf>.
- [2] Kim-Anh Le Cao, Florian Rohart, Sebastien Dejean with key contributors Benoit Gauthier Ignacio Gonzalez, Francois Bartolo, contributions from Pierre Monget, Jeff Coquery, FangZou Yao, and Benoit Lique. *Package mixOmics: Omics Data Integration Project*. 2016.  
<https://CRAN.R-project.org/package=mixOmics>.
- [3] Kim-Anh Lê Cao, Mary-Ellen Costello, Vanessa Anne Lakis, François Bartolo, Xin-Yi Chua, Rémi Brazeilles, and Pascale Rondeau. MixMC: A multivariate statistical framework to gain insight into microbial communities. *PLoS ONE*, August 2016.
- [4] Jérôme Mariette. *Apprentissage statistique pour l'intégration de données omiques*. PhD thesis, Université Toulouse 3 Paul Sabatier, 2017.