Institut National des Sciences Appliquées de Toulouse
Mathematical and Modeling Engineering Department

October 2017 - January 2018

Fifth-year project

Identification and caracterisation 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 litterature 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

Nathalie Villa-Vialaneix, [...]
Elias Salem, [...]

Introduction

Bovine respiratory disease complex is considered one of the most common and economically important diseases in calf breeding farms. It is displayed as pneumonia with different states of severity and respiratory signs depending on the causative agents and stressors, such as host immune reaction and environment.

In order to study this disease, samplings have been taken in calf breeding farms constituing a dataset to analyze with the help of statistical tools and methods. The purpose of this analysis is to deduce some links and correlations between pathogenic agents that induce infectious bronchopneumiona.

This report is divided in three parts. In a first place the context and the dataset will be introduced in details. Then, the statisticals methods used will be explained, which lead us to the last part where the main results will be presented.

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. These are obtained by Next-generation sequencing (NGS), also known as high-throughput sequencing [7].

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 perfomed, 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 bateria, 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^{th} 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_E
&	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] gauvreauii 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	0	6	0	1	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	
Anaerosporobacter	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	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 by 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	Ct.PI.3	Ct.Coronavirus	Ct.P.multocida	Ct.M.haemolytica	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.

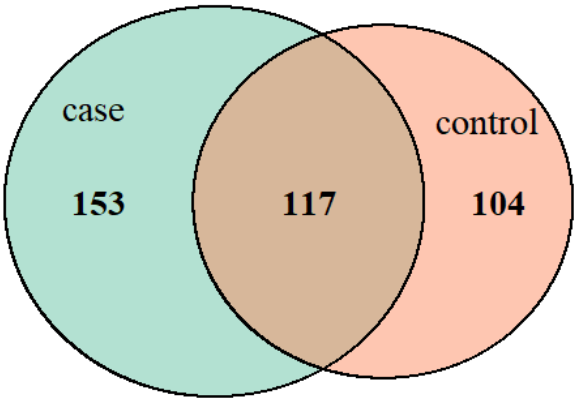

Figure 1.3: *Venn diagram comparing two lists of bacteria: for sick calves and for healthy calves*

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.

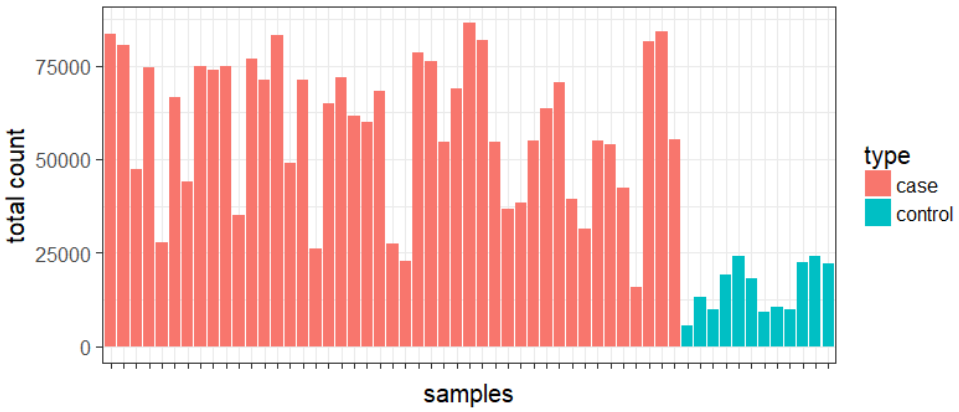

Figure 1.4: *Total counts for control and case sample*

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 euclidien 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.

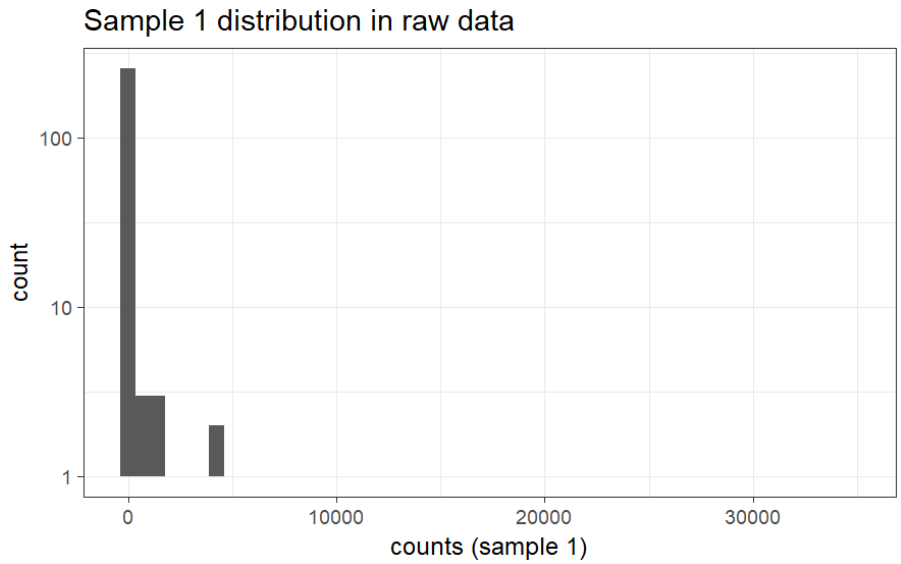

Figure 2.1: *Sample 1 distribution in raw data*

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 [5].

Let $y_{i,j}$ be the counts of a bacteria j found in the sample i, where $i \in [1, .., n]$ and $j \in [1, .., p]$, 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

The log transformation [4] is such that:

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

TSS transformation

It is commun in metagenomic datasets to perform TSS [5] (*Total Sum Scaling*) before further normalization. TSS transformation computes relative abundances, it divides each count by the total number of observed species for each sample.

$$\tilde{y}_{ij} = \frac{n_{ij}}{\sum_{k=1}^{p} n_{ik}}$$

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

The TSS transformation is quite limited, however it can be projected in an euclidien space using a log centered transformation, called CLR (*Centered Log Ratio*) [5]:

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

TSS+ILR transformation

ILR (*Isometric Log Ratio*) transformation [5]:

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

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 [5] 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. It is optained by dividing counts by the cumulative sum of counts until some percentile defined from the data.

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 [1].

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 ω_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
    $\xi_h = Xu_h$ where ξ_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 ω_h is the latent variable associated with Y
  end while
  $c_h = X^{'}\xi_h / \xi_h^{'}\xi_h$ partial regression of X on ξ
  $d_h = Y^{'}\omega_h / \omega_h^{'}\omega_h$ partial regression of Y on ω
  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, ω, v, ξ are the respective eigenvectors of the matrices $YY^{'}XX^{'}$, $Y^{'}XX^{'}Y$, $XX^{'}YY^{'}$, $X^{'}YY^{'}X$ associated with the same greater eigenvalue λ.

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.

As is stated above, the supervised models obtained by the PLS-DA work with a dummy indicator matrix of Y to indicate the class membership of each sample. The prediction of a new observation results in either a predicted dummy variable or a predicted variate. Therefore, an appropriate distance needs to be applied to those predicted values to assign the predicted class. Several distances can be used to predict the class of each sample and the distance chosen is the one that minimizes the classification error rate.

It is necessary to note that the PLS-DA method can take into account the structure of repeated measurements where different treatments are applied on the same subjects. This allows to enables the selection of features separating the different treatments. The multilevel function first decomposes the variance in the data set X and applies A multilevel decomposition can help unravelling subtile differences hidden by individual variation.

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 able to perform a variable selection in the dataset by introducing LASSO penalization on the pair of loading vectors [2].

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

Student test

Moreover, Student test can be performed to assess whether the selected variables are differentially abundant between the two conditions tested. We will also have a look at the ajusted p-values using the Benjamini & Hochberg method [6] which controls the false discovery rate.

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 $z = (x_1, y_1), ..., (x_n, y_n)$ 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 $\widehat{f}_B(Y) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}_{z_b}(Y)$ 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 count data.

If we look at the graphs in the Figure 3.1, the less asymetric 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.

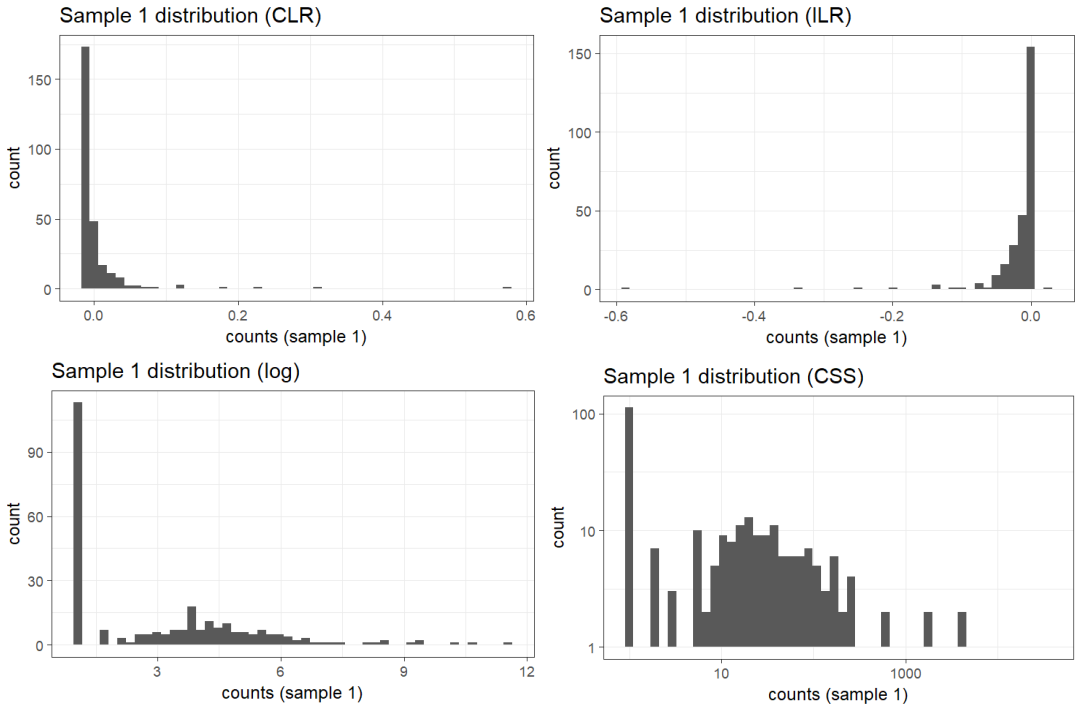

Figure 3.1: *Sample 1 distribution according to the type of transformation*

Let us have a look at the distribution by boxplots of all the samples displayed in 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 the log-transformation will be used.

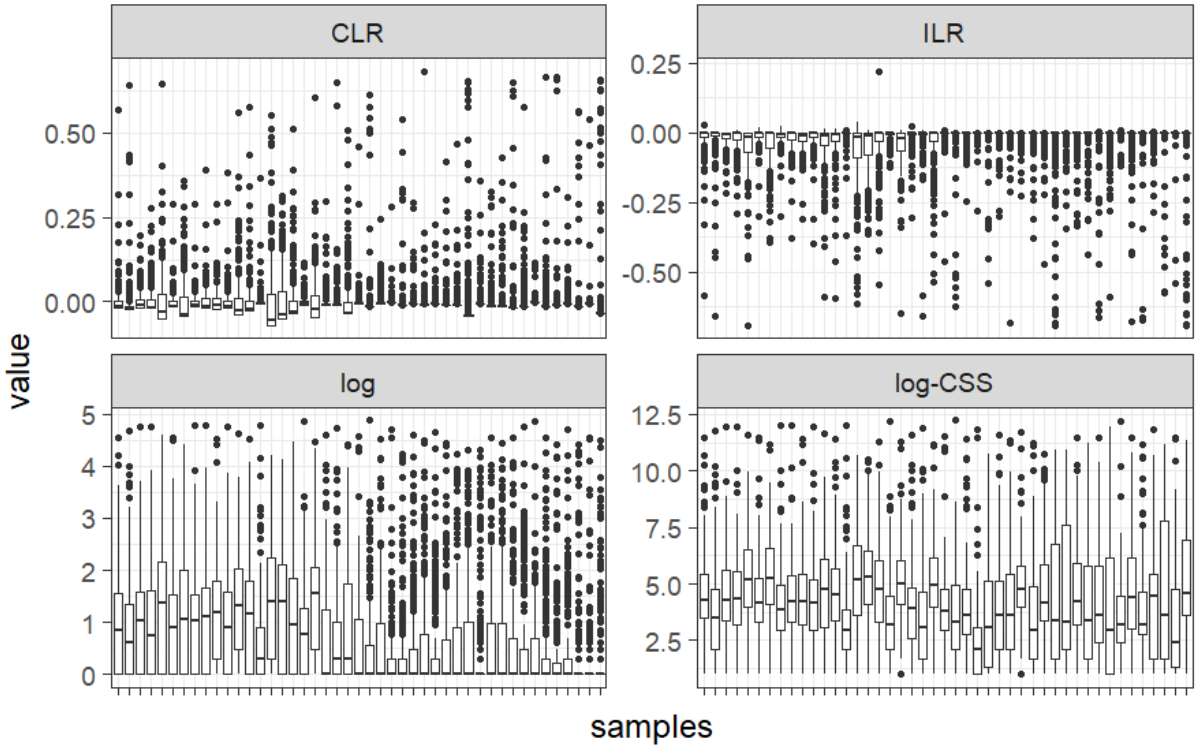

Figure 3.2: *Distributions of all samples according to the type of transformation and the sample*

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 perfomed in `R`, using the `mixOmics` library [2].

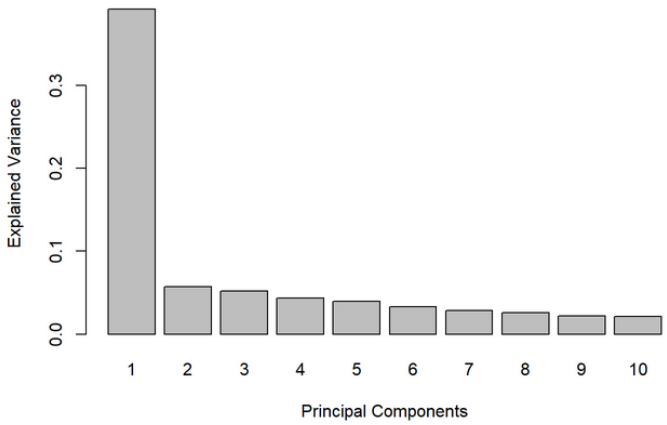

Figure 3.3: *Sample 1: The variance explained by the first 10 components*

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 first 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.

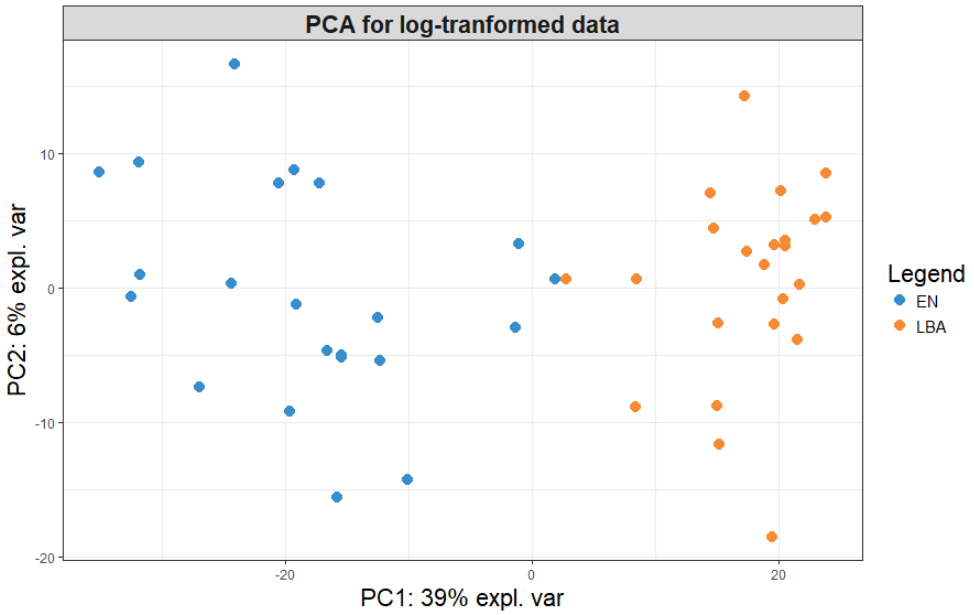

Figure 3.4: *Sample 1: Projection of individuals in the first factorial plan*

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 differences between EN and 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].

We will focus on a multilevel analysis, therefore the 29^{th} column will be removed (section 1.2 page 7) for the multilevel option (in the 4^{th} step below).

- **Step 1**: Performance of the PLS-DA on the first two components. A first PLS-DA is computed and its performances are evaluated (with 10-fold cross-validation repeated 10 times) to choose the model that improves the prediction and which type of distance to use in its computation.

  Figure 3.5 indicates that the classification error rate decreases from one component to two components in the model, for three different distances: Maximum, Centroids and Mahalanobis. BER stands for Balanced Error Rate, and here the groups are

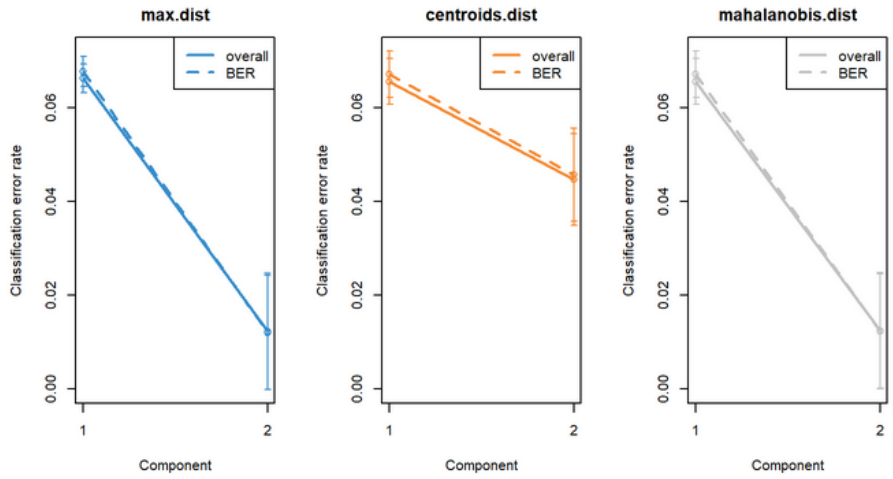

Figure 3.5: *Distances for the first two components*

unbalanced because of the 29$^{\text{th}}$ 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. The separation between the two conditions is clear according to the first component but the LBA ellipse does not contain all the LBA samples.

- **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.

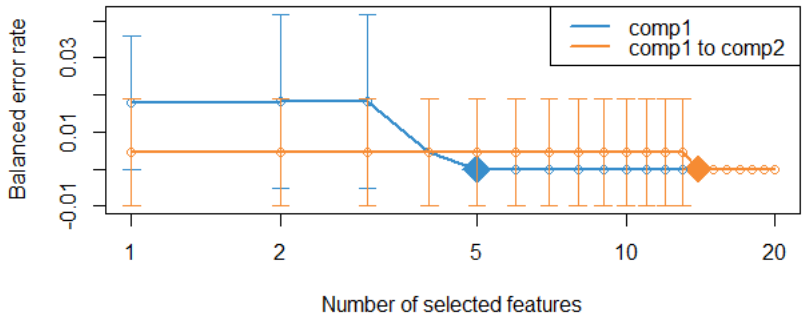

Figure 3.6: *Error rate for the first two components by the number of selected variables*

- **Step 4**: Running sPLS-DA tuned (right graph in Figure 3.7). This step consists 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. Now both of the ellipses contain all of their respectives samples.

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

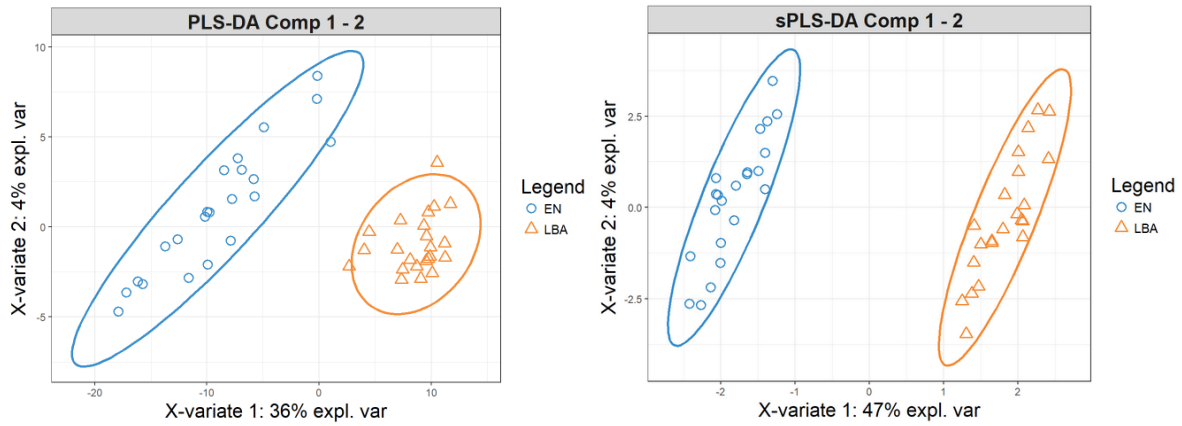

Figure 3.7: *Projection of individuals for PLS-DA and sPLS-DA first two components*

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.

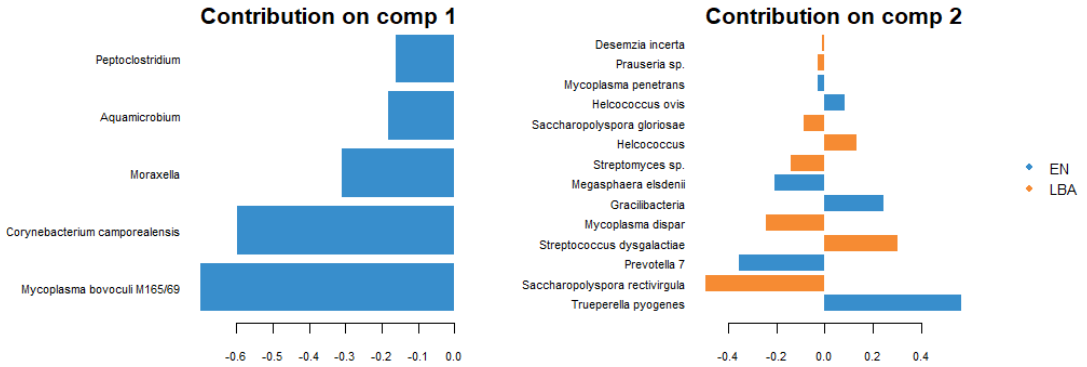

Figure 3.8: *Loadings values of selected bacteria on the first two components*

Biologist suspected this kind of bacteria for having an impact on the differences between EN and LBA.

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

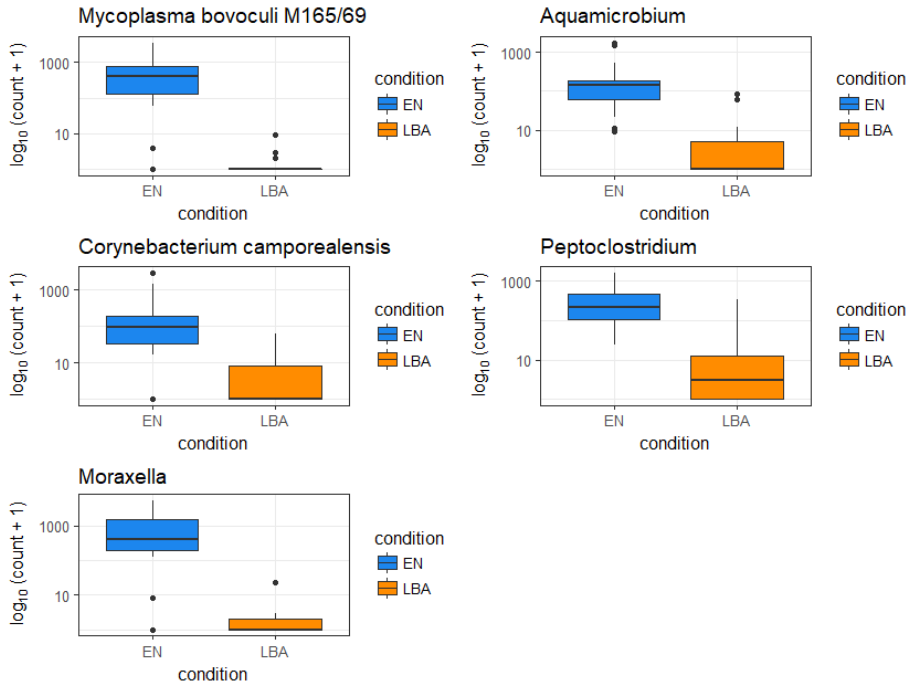

Figure 3.9: *Boxplots of abundance differences between the two conditions*

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 the differences between LBA and EN are not related to the status (case/control) of the sample.

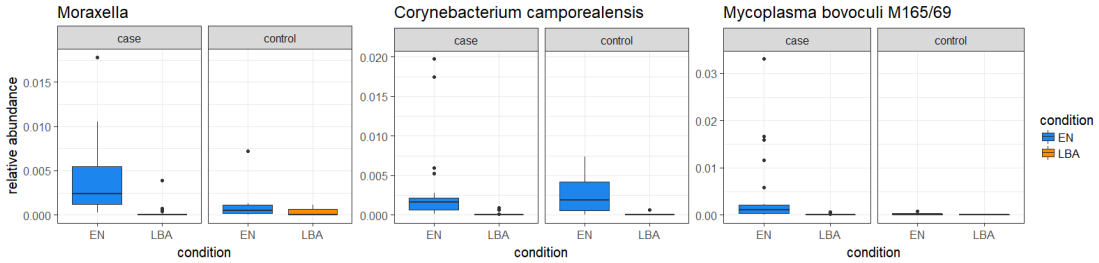

Figure 3.10: *Boxplots of selected bacteria abundance distributions of EN and LBA samples for healthy and sick samples*

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. Each step is compute separatly for EN and LBA.

3.3.1 sPLS in regression mode

To better understand 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.

In the nasal cavity (EN)

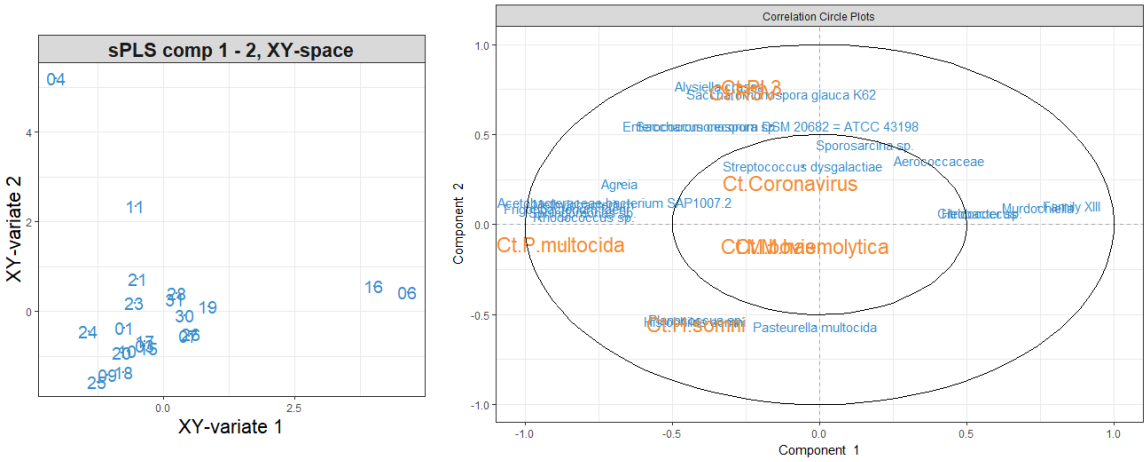

Figure 3.11: *sPLS on XY-space and Correlation circle plot for the first two components*

The first graph in Figure 3.11 shows 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.

In the lungs (LBA)

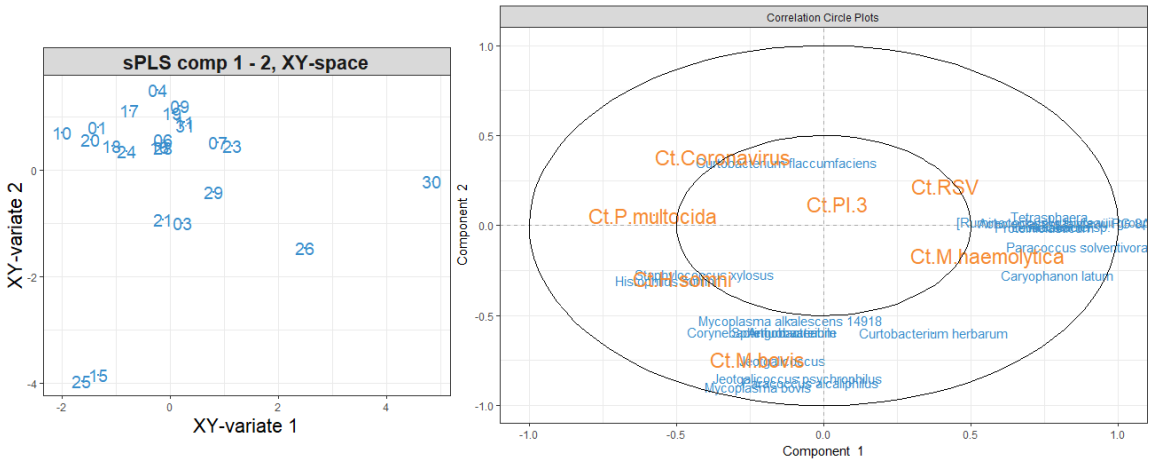

Figure 3.12: *sPLS on XY-space and Correlation circle plot for the first two components*

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 3.12 shows that:

- 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.
- 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.

3.3.2 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.

	RSV	PI.3	Coronavirus	Multocida	Haemolytica	Bovis	Somni	**group**
0	37	41	22	6	31	39	26	**18**
1	7	3	22	38	13	5	18	**27**

For each condition, the steps below will be followed:

- Step 1: PLS-DA and sPLS-DA results
- Step 2: Student test results
- Step 3: Common bacteria results

In the nasal cavity (EN)

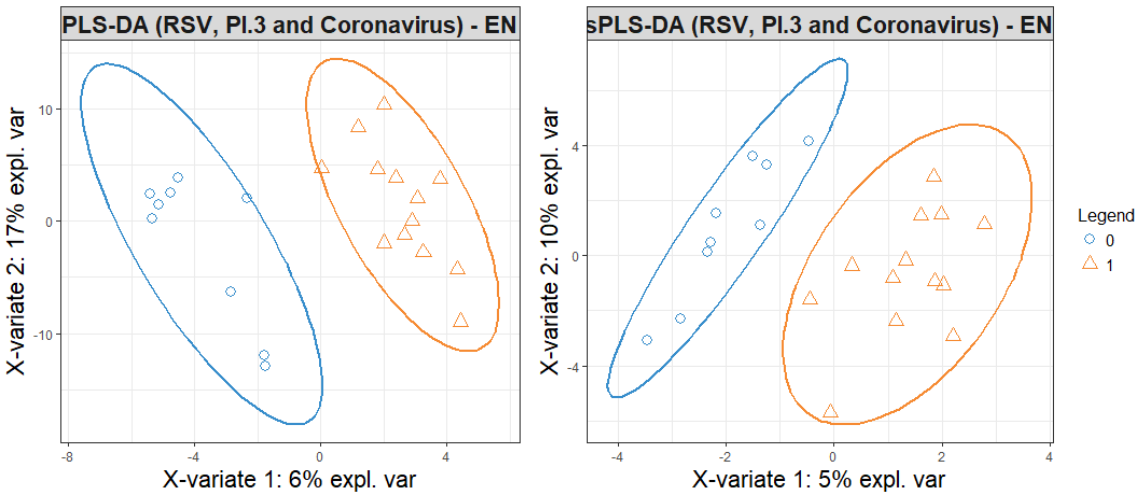

Figure 3.13: *PLS-DA and sPLS-DA on the groupe of three viruses*

- Step 1: In Figure 3.13, 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.

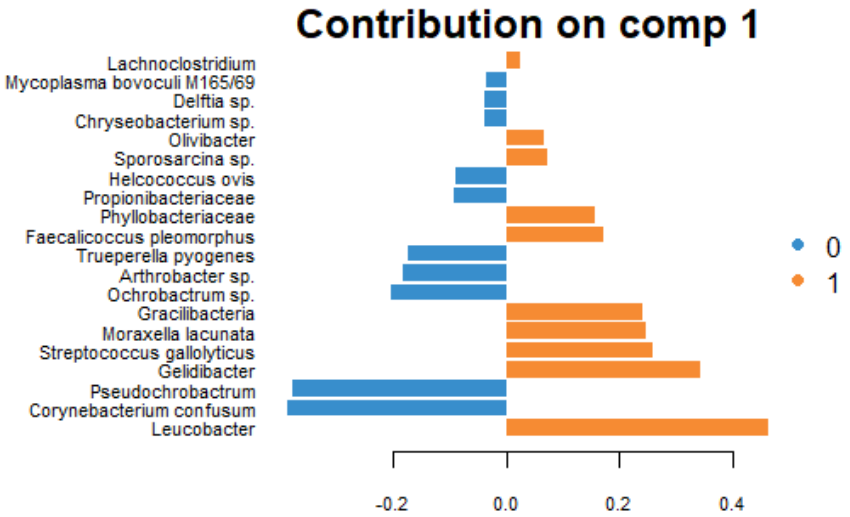

Figure 3.14: *Loading values of selected variables on first component*

The contribution plot (Figure 3.14) above displays the abundance of each bacteria and in which status (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 $Corynebacteriumconfusum$.

- Step 2: To evaluate the significance of the selected bacteria in the process of appearance or absence of these three viruses, Student 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.

- Step 3: 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 3.15 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.

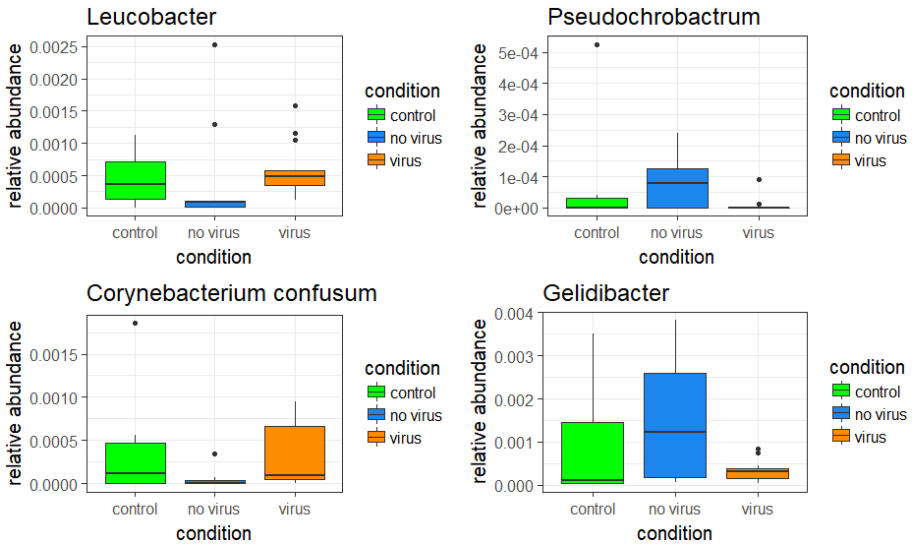

Figure 3.15: *Boxplots on ...*

In the lungs (LBA)

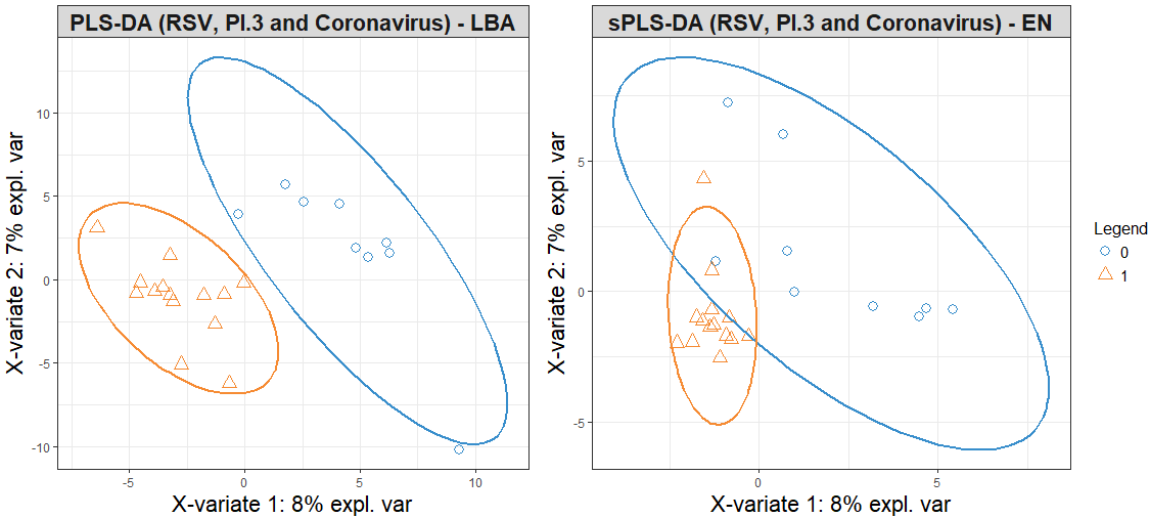

Figure 3.16: *PLS-DA and sPLS-DA on the group of three viruses*

- Step 1: the first graph in Figure 3.16 shows that by using all the features in the PLS-DA, the method succeeds in discriminating the infected and non-infected samples. The first component seems to contribute more to this separation. This is no longer the case after the selection of the optimal number of variables on first component. It can be seen that clusters are overlapping and the separation has a lower discrimination quality. Thus, the following results are less reliables than for the previous condition.

  Figure 3.17 shows that most bacteria seem to be abundant when the group of viruses is not present in LBA.

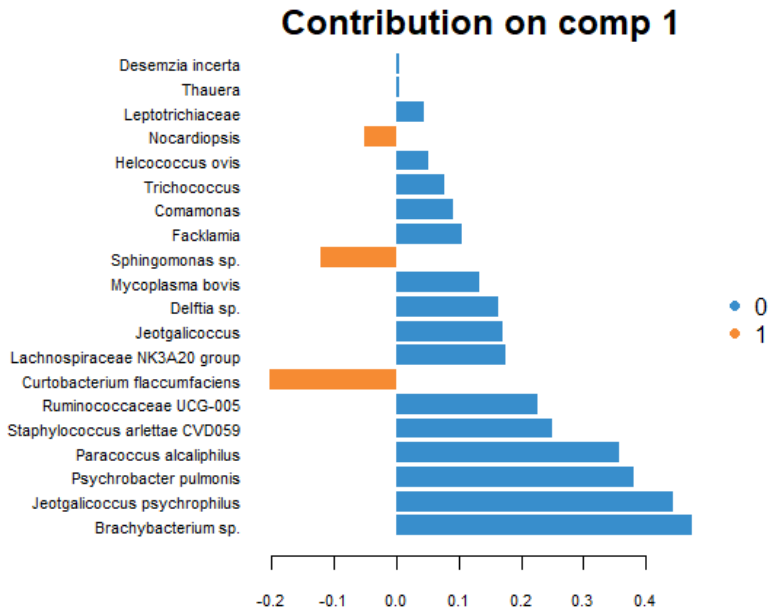

Figure 3.17: *Loading values of selected variables on first component*

- Step 2: as previously, Student tests are performed for each selected bacteria to assess their significance.

	bacteria	pvalue	FWER
1	Brachybacterium sp.	0.05052448	0.1447919
2	Jeotgalicoccus psychrophilus	0.04775828	0.1447919
3	Psychrobacter pulmonis	0.04029216	0.1447919
4	Paracoccus alcaliphilus	0.06314502	0.1447919
5	Staphylococcus arlettae CVD059	0.09985810	0.1521428
6	Ruminococcaceae UCG-005	0.10649995	0.1521428
7	Curtobacterium flaccumfaciens	0.02126597	0.1447919
8	Lachnospiraceae NK3A20 group	0.06462858	0.1447919
9	Jeotgalicoccus	0.08698581	0.1521428
10	Delftia sp.	0.07341817	0.1468363
11	Mycoplasma bovis	0.13123811	0.1543978
12	Sphingomonas sp.	0.02971395	0.1447919
13	Facklamia	0.10558693	0.1521428
14	Comamonas	0.12787987	0.1543978
15	Trichococcus	0.12570172	0.1543978
16	Helcococcus ovis	0.14864534	0.1588076
17	Nocardiopsis	0.03295577	0.1447919
18	Leptotrichiaceae	0.06515637	0.1447919
19	Thauera	0.18595844	0.1859584
20	Desemzia incerta	0.15086719	0.1588076

The conclusion of this part is the same as for the condition EN. The Student test does not reject the hypothesis H_0: the abundance of the bacteria studied does not change according to the presence or absence of the group of viruses.

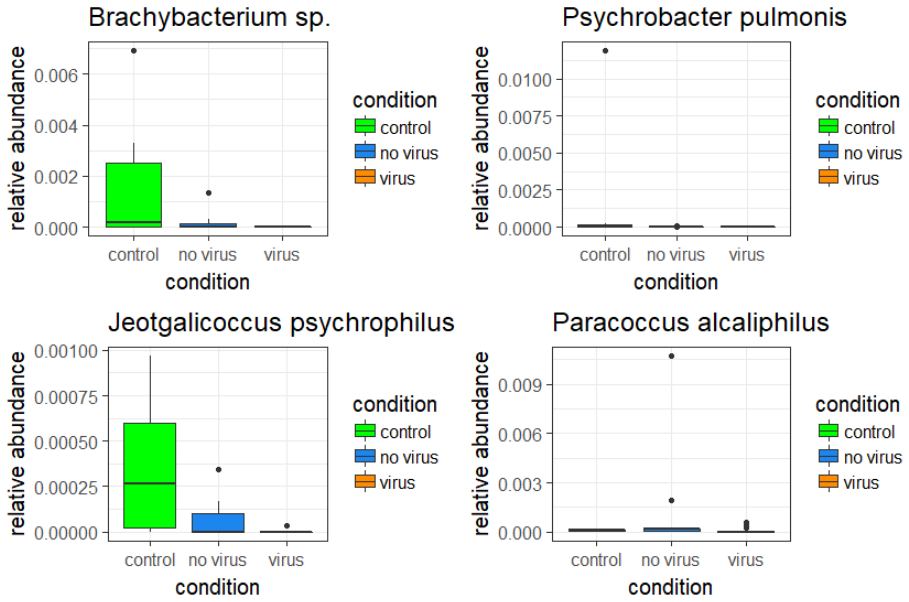

Figure 3.18: *Boxplots on ...*

- Step 3: as Figure 3.18 shows, the same situation is encountered for the common bacteria of the LBA condition. The distribution of variables does not appear to be different between infected and uninfected samples except for *Brachybacterium sp.* and *Jeotgalicoccus Psychrophilus* which are more abundant in the control samples. This could be due to the fact that the results given by the PLS-DA can not be considered reliable because of the unsatisfactory separation.

3.3.3 Random Forest

In the nasal cavity (EN)

In view of the previous results, another method was tested to predict the absence or presence of viruses from abundance data. This is the Random Forest algorithm. The results are displayed bellow:

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

It seems that this method can not correctly predict the status of the group of viruses. The misclassification rate is very important in relation to the sample size. In particular, the algorithm encounters difficulties in predicting the absence of viruses.

In the lungs (LBA)

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

The results for the LBA condition are similar to those for the EN condition.

This behavior can be explained by the modest amount of data on viruses and by the fact that RandomForest is an often inefficient method for linear problems. It can be concluded that this method is not suitable for this dataset.

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 Gautier Ignacio Gonzalez, Francois Bartolo, contributions from Pierre Monget, Jeff Coquery, FangZou Yao, and Benoit Liquet. *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] Feng, Changyong, and al. Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, February 2014.

[5] Jérôme Mariette. *Apprentissage statistique pour l'intégration de données omiques*. PhD thesis, Université Toulouse 3 Paul Sabatier, 2017.

[6] John H. McDonald. *Handbook of Biological Statistics*. 2008. http://www.biostathandbook.com/multiplecomparisons.html.

[7] Reuter, Jason A., Damek Spacek, and Michael P. Snyder. High-throughput sequencing technologies. *Molecular cell*, May 2015.