Analysis of the Pulmonary Microbiome Composition of Legionella pneumophila-Infected Patients

Ana Elena Pérez-Cobas, Carmen Buchrieser

ANALYSES OF THE PULMONARY MICROBIOME COMPOSITION OF *LEGIONELLA PNEUMOPHILA*-INFECTED PATIENTS

ANA ELENA PÉREZ-COBAS[1, 2] and CARMEN BUCHRIESER[1, 2]*

1 Institut Pasteur, Biologie des Bactéries Intracellulaires, Paris, France

2 CNRS UMR 3525, 75724, Paris, France

Running title: Pulmonary microbiome of Legionnaires' disease patients

*Corresponding author: cbuch@pasteur.fr

Abstract

The analysis of lung microbiome composition is a field of research that recently emerged. It gained great interest in pulmonary diseases such as pneumonia since the microbiome seems to be involved in host immune responses, inflammation and protection against pathogens. Thus, it is possible that the microbial communities living in the lungs play a role in the outcome and severity of lung infections such as *Legionella*-caused pneumonia and in the response to the antibiotic therapy. In this chapter, all steps necessary for the characterization of the bacterial and fungal fraction of the lung microbiome using high-throughput sequencing approaches are explained, starting from the selection of clinical samples to the analysis of taxonomic composition, diversity and ecology of the microbiome.

Key Words: Pulmonary microbiome, *Legionella pneumophila*, 16S rRNA gene, ITS, high-throughput sequencing, bioinformatics, microbial ecology

1. Introduction

In this chapter, we explain the methods used to characterize the bacterial (microbiome) and fungal (mycobiome) communities present in the lungs, based on the high-throughput sequencing of the 16S rRNA gene and the Internal transcribed spacer (ITS). The protocol that we propose is divided in two main parts that are presented in **Figure 1:** the experimental work (pink frame) and the bioinformatics analysis (gray frame). First, we present a detailed experimental protocol for sequencing of the 16SrRNA gene and the ITS of the lung microbiome from clinical samples to the preparation of the sequencing libraries. We have chosen Illumina sequencing since it is a suitable technology for the study of microbial communities and it has also been commonly used for the characterization of the human microbiome. Briefly, from a pulmonary sample such as a bronchoalveolar lavage (BAL) sample, the total DNA extraction is done followed by a specific PCR (16SrRNA gene and ITS in this protocol). After cleaning of the PCR products, the index PCR is performed to allow multiplexing several samples that are each indexed differently in the same sequence run. A final cleaning step is followed by the quantification, normalization and pooling of the samples for sequencing. The sequences obtained are then analysed using different bioinformatics methods. Bioinformatics is a field that has been changing very rapidly in the last years. Many different methods, algorithms, databases and softwares have been developed to deal with the challenges associated with new sequencing technologies. We present a guide to the main bioinformatics analyses tools that one can apply to the 16SrRNA gene and ITS high-throughput sequencing data. In summary, raw sequencing data should be cleaned and filtered using different quality parameters. Since most of the diversity in the lungs remains undiscovered, it is necessary to define Operational taxonomic units (OTUs) to perform the ecological analysis in a correct way. The sequences should be clustered in OTUs according a sequence identity level and when it is possible be taxonomically assigned. With the

OTU/taxon information it is possible to predict the alpha diversity (diversity within a given community) and the beta diversity (diversity among communities or along an environmental gradient) ***(1)***. The final step is to perform statistical analysis, a step that depends on the aim and the questions asked in the project. Here, we suggest some of the most common statistical analysis that could be used for typical comparisons undertaken in microbiome studies.

2. Materials

2.1 DNA Extraction

1. PowerSoil DNA Isolation Kit (MoBio).

2. PowerLyzer 24 Bench Top Bead-Based Homogenizer (MoBio).

3. Qubit dsDNA HS (High Sensitivity) Assay Kit (Thermo Fisher Scientific).

4. Qubit Fluorometer (Thermo Fisher Scientific).

5. Sterilized water.

2.2 16SrRNA and ITS PCR

1. Buffer Taq (10X) (Thermo Fisher Scientific).

2. $MgCl_2$ (25 mM).

3. dNTPs (10 mM).

4. 16SrRNA gene Forward primer including Illumina sequencing adaptor (10 mM): 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3'. Specific primer sequence: 5'-CCTACGGGNGGCWGCAG-3'.

5. 16SrRNA gene Reverse primer including Illumina sequencing adaptor (10 mM): 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'. Specific primer sequence: 5'-GACTACHVGGGTATCTAATCC-3'.

6. ITS Forward primer including Illumina sequencing adaptor (10 mM): 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTGGTCATTTAGAGGAAGTAA-3'. Specific primer sequence: 5'-CTTGGTCATTTAGAGGAAGTAA-3'.

7. ITS Reverse primer including Illumina sequencing adaptor (10 mM): 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTGCGTTCTTCATCGATGC-3'. Specific primer sequence: 5'- GCTGCGTTCTTCATCGATGC-3'.

8. Phusion High-Fidelity DNA Polymerase (5 u/µl) (Thermo Fisher Scientific).

9. DMSO.

10. Nuclease-free water.

11. Bacterial DNA for positive control (ng/µl) (e.g. *Legionella pneumophila* genomic DNA).

12. PCR Machine (Thermal Cycler).

13. Agarose

14. 10X TE Buffer

15. Ethidium Bromide 0.7 mg/ml

16. 10X Loading buffer

17. Double-stranded DNA ladder in the range of 100-2000 bp (*e.g.* 100 bp DNA Ladder, Invitrogen)

18. Agarose gel electrophoresis equipment

2.3. Illumina amplicon library preparation and sequencing

1. 10 mM Tris pH 8.5

2. AMPure XP beads

3. Freshly Prepared 80% Ethanol (EtOH)

4. 96-well 0.2 ml PCR plates

5. Microseal "A".

6. Microseal "B".

7. 2x KAPA HiFi HotStart ReadyMix

8. Nextera XT Index 1 Primers (N7XX)

9. Nextera XT Index 2 Primers (S5XX)

10. TruSeq Index Plate Fixture

11. Magnetic stand-96 (Life Technologies)

2.4 Bioinformatics analysis of sequencing data.

1. Computer and/or computer cluster and Internet connection

3. Methods

It is important to apply the same protocol consistently across all samples in one study to allow correct comparisons.

3.1 Total DNA Extraction

1. Bronchoalveolar lavage (BAL) samples collected from patients should be stored at -80º until further processing (see **Note 1**).

2. Perform the DNA extraction from 1 ml of BAL with the PowerSoil® DNA Isolation Kit following the manufacturer’s instructions using the PowerLyzer 24 Bench Top Bead-Based Homogenizer (see **Note 2**). Include a negative control sample to check eventual contamination arising from the reagents in the kit (see **Note 3**). If possible it is recommended to add a spike-in standard step to evaluate the sequencing data quality and to estimate the absolute microbial abundances for further comparative analysis (see **Note 4**).

3. Quantify the extracted DNA using the Qubit dsDNA HS (High Sensitivity) Assay Kit following the manufacturer’s instructions.

4. Prepare dilutions and adjust the concentration of the samples to 20 ng/µl. If necessary, the DNA samples can be stored at -20Cº until use for PCR.

3.2 Sequence (Illumina) library preparation -16S rRNA gene and ITS specific PCR

1. Prepare a PCR master mix (final volume of 20 µl per sample) by mixing 5 µl of Taq Buffer (10X), 1 µl of 25 mM MgCl2, 0.5 µl of dNTPs (10 mM), 1.25 µl of each primer (10 mM), 0.25 µl of Phusion High-Fidelity DNA Polymerase (5 u/µl), 0.5 µl of DMSO, 8.25 µl of nuclease-free water and 1 µl of DNA template. Include a positive control (bacterial DNA for the 16S rRNA and fungal DNA for the ITS) and a negative control (water) (see **Note 5**).

2. Run the reaction in a PCR machine with the following conditions: 95◦ C for 5 min followed by 20 cycles of 95◦C for 30 s, 55◦C for 1 min and 72◦C for 1 min and a final extension step of 7 minutes at 72◦C. Keep at 4°C (see **Note 6**).

3. Prepare an agarose gel at 1% by mixing 1g agarose in 100mL TE (10 mM Tris, 1mM EDTA, pH 8) and add 100 µl of ethidium bromide (0.7 mg/ml) to the mix. Let the gel solidify for 30 minutes. Mix 1 to 5 µl of sample with 1 to 3 µl of loading buffer. Load the mixed samples and the DNA ladder on the gel. Check the PCR products by running an electrophoresis at 110 V during 1 hour. If it is necessary, the DNA samples can be stored at -20Cº until library preparation.

3.3 Sequence (Illumina) library preparation and PCR Clean-Up 1

1. Centrifuge the Amplicon PCR plate at 1,000× g at 20°C for 1 minute to collect condensation, carefully remove seal.

2. Vortex the AMPure XP beads for 30 seconds to make sure that the beads are evenly dispersed. Add an appropriate volume of beads to a trough depending on the number of samples to process.

3. Using a multichannel pipette, add 20 µl of AMPure XP beads to each well of the Amplicon PCR plate. Change tips between samples.

4. Gently pipette entire volume up and down 10 times if using a 96-well PCR.

5. Incubate at room temperature without shaking for 5 minutes.

6. Place the plate on a magnetic stand for 2 minutes or until the supernatant has cleared.

7. Put the Amplicon PCR plate on the magnetic stand and use a multichannel pipette to remove and discard the supernatant. Change tips between samples.

8. Put the Amplicon PCR plate on the magnetic stand and wash the beads with freshly prepared 80% ethanol as follows:

a) Using a multichannel pipette, add 200 µl of freshly prepared 80% ethanol to each well.

b) Incubate the plate on the magnetic stand for 30 seconds.

c) Carefully remove and discard the supernatant.

9. Put the Amplicon PCR plate on the magnetic stand and perform a second ethanol wash as follows:

a) Using a multichannel pipette, add 200 µl of freshly prepared 80% ethanol to each sample well.

b) Incubate the plate on the magnetic stand for 30 seconds.

c) Carefully remove and discard the supernatant.

d) Use a P20 multichannel pipette with fine pipette tips to remove excess ethanol.

10. Put the Amplicon PCR plate still on the magnetic stand and allow the beads to air-dry for 10 minutes.

11. Remove the Amplicon PCR plate from the magnetic stand. Using a multichannel pipette, add 52.5 µl of 10 mM Tris pH 8.5 to each well of the Amplicon PCR plate.

12. Gently pipette mix up and down 10 times, changing tips after each column (or seal plate and shake at 1800 rpm for 2 minutes). Make sure that beads are fully resuspended.

13. Incubate at room temperature for 2 minutes.

14. Place the plate on the magnetic stand for 2 minutes or until the supernatant has cleared.

15. Using a multichannel pipette, carefully transfer 50 µl of the supernatant from the Amplicon PCR plate to a new 96-well PCR plate. Change tips between samples. If you do not immediately proceed to Index PCR, seal plate with Microseal "B" adhesive seal and store it at -15° to -25°C for up to a week.

3.4 Sequence (Illumina) library preparation - Index PCR

1. Using a multichannel pipette, transfer 5 µl from each well to a new 96-well plate. The remaining 45 µl are not used and can be stored for other purposes.

2. Arrange the Index 1 and 2 primers in a rack (i.e. the TruSeq Index Plate Fixture) using the following arrangements as needed:

a) Arrange Index 2 primer tubes (white caps, clear solution) vertically, aligned with rows A through H (see **Figure 2**).

b) Arrange Index 1 primer tubes (orange caps, yellow solution) horizontally, aligned with columns 1 through 12 (see **Figure 2**).

3. Place the 96-well PCR plate with the 5 µl of resuspended PCR product DNA in the TruSeq Index Plate Fixture.

4. Prepare the following mix for each sample in the plate: 5 µl of DNA, 5 µl of Nextera XT Index Primer 1 (N7 xx), 5 µl of Nextera XT Index Primer 2 (S5xx) 25 µl of 2x KAPA HiFi HotStart ReadyMix, and 10 µl of PCR Grade water.

5. Gently pipette up and down 10 times to mix.

6. Cover the plate with Microseal 'A'.

7. Centrifuge the plate at 1,000 × g at 20°C for 1 minute.

8. Perform PCR on a thermal cycler using the program: 95°C for 3 minutes followed by 8 cycles of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, and a final extension step of 72°C for 5 minutes. Keep at 4°C.

3.5 Sequence (Illumina) library preparation - PCR Clean-Up 2

1. Centrifuge the Index PCR plate at 280 × g at 20°C for 1 minute to collect condensation.

2. Vortex the AMPure XP beads for 30 seconds to make sure that the beads are evenly dispersed. Add an appropriate volume of beads to a trough.

3. Using a multichannel pipette, add 56 µl of AMPure XP beads to each well of the Index PCR plate.

4. Gently pipette mix up and down 10 times if using a 96-well PCR plate

5. Incubate at room temperature without shaking for 5 minutes.

6. Place the plate on a magnetic stand for 2 minutes or until the supernatant has cleared.

7. Put the Index PCR plate on the magnetic stand and use a multichannel pipette to remove and discard the supernatant. Change tips between samples.

8. Put the Index PCR plate on the magnetic stand and wash the beads with freshly prepared 80% ethanol as follows:

a) Using a multichannel pipette, add 200 µl of freshly prepared 80% ethanol to each sample well.

b) Incubate the plate on the magnetic stand for 30 seconds.

c) Carefully remove and discard the supernatant.

9. Put the Index PCR plate on the magnetic stand, perform a second ethanol wash as follows:

a) Using a multichannel pipette, add 200 µl of freshly prepared 80% ethanol to each sample well.

b) Incubate the plate on the magnetic stand for 30 seconds.

c) Carefully remove and discard the supernatant.

d) Use a P20 multichannel pipette with fine pipette tips to remove excess ethanol.

10. With the Index PCR plate still on the magnetic stand, allow the beads to air-dry for 10 minutes.

11. Remove the Index PCR plate from the magnetic stand. Using a multichannel pipette, add 27.5 µl of 10 mM Tris pH 8.5 to each well of the Index PCR plate.

12. If using a 96-well PCR plate, gently pipette mix up and down 10 times until beads are fully resuspended, changing tips after each column.

13. Incubate at room temperature for 2 minutes.

14. Place the plate on the magnetic stand for 2 minutes or until the supernatant has cleared.

15. Using a multichannel pipette, carefully transfer 25 µl of the supernatant from the Index PCR plate to a new 96-well PCR plate. Change tips between samples. If you do not plan to proceed to Library Quantification, Normalization, and Pooling, seal the plate with Microseal "B" adhesive seal. Store the plate at -15° to -25°C for up to a week.

16. Prepare an agarose gel at 1% by mixing 1g agarose in 100mL TE 10X and add 100 µl of ethidium bromide (0.7 mg/ml) to the mix. Let the gel solidify for 30 minutes. Mix from 10 µl of sample with 5 µl of loading buffer. Load the mixed samples and the DNA ladder in the gel. Check the PCR products by running an electrophoresis at 90 V during 45 minutes.

3.6 Quantification, Normalization and Pooling of the sequence library and Sequencing

1. Quantify the extracted DNA using the Qubit dsDNA HS (High Sensitivity) Assay Kit following the manufacturer's instructions.

2. Calculate the DNA concentration in nM, based on the size of DNA amplicons as determined by an Agilent Technologies 2100 Bioanalyzer trace:

$$\frac{(\text{concentration in ng/µl}) \times 106}{(660\text{ g/mol} \times \text{average library size})} = \text{concentration in nM}$$

3. Dilute concentrated final library using Resuspension Buffer (RSB) or 10 mM Tris pH 8.5 to 4 nM.

4. Aliquot 5 µl of diluted DNA from each library and mix aliquots for pooling libraries with unique indices.

5. Send the samples to a sequencing platform or do the sequencing in your own laboratory if you dispose of an Illumina sequencer (see **Note 7**).

3.7 Bioinformatics analysis of the 16S rRNA gene and ITS sequencing data: Raw data processing

1. For the raw high throughput sequencing data use a software for quality control checks and to detect whether your data has any problems of which you should be aware before starting with the analysis (e.g. FASTQC software, ***(2)***).

2. The data quality control is a critical step to obtain meaningful analyses from the sequencing data, especially Illumina data that are characterized by a very high number of short reads. Erroneous reads can lead to an over-estimation of the alpha-diversity, as well as to wrong taxonomic annotations and loss of specific microbial groups. Thus, it is critical to discard all erroneous, short (<50 bp) and low-quality reads (Q < 33) and also to trim the sequencing adapters and low-quality extremes. Different softwares have been developed to achieve this task such as the FASTX-Toolkit ***(3)*** or PRINSEQ ***(4)***.

3. If the sequences are paired-end (*e.g.* MiSeq paired-end sequencing), the pairs of reads should be joined to get longer and higher-quality reads. The fastq-join script ***(5)*** can be used for that task.

3.8 Bioinformatics analysis of the 16S rRNA gene and ITS sequencing data: OTU clustering and taxonomic annotation

1. A main step is the clustering of reads from the cleaned sequencing files (16SrRNA and ITS) into OTUs and to construct the OTU-abundance-tables. The Quantitative Insights Into Microbial Ecology (QIIME) pipeline ***(6)*** and mothur ***(7)*** are the most used pipelines for OTU analyses in microbiome studies. As an example, the QIIME pipeline has implemented different OTU picking strategies: (1) "de novo OTU picking" where the reads are clustered against one another without any external reference sequence database, (2) "the closed-reference OTU picking" process where reads are clustered against a reference sequence collection and only the positive matches are included and (3) the "open-reference OTU picking process" where the reads are clustered against a reference sequence collection and the not-matching reads are clustered *de novo*. The open-reference OTU picking process method is the most useful for the 16SrRNA and ITS data of the pulmonary microbiome since the reference database allows to classify OTUs belonging to already described OTUs but keeps also the undiscovered diversity which is included in the "*de novo*" OTUs. The OTUs can be defined at different sequence identity levels; the most used value for bacteria is 97%, which represents approximately the species level. The same value has been standardized for the ITS OTU picking.

2. Perform the taxonomic classification of your OTUs and set the abundance-tables at a specific taxonomic level (*e.g.* genus, family) for further comparative analysis. Different classification methods and databases have been developed for the taxonomic assignment of the 16SrRNA and ITS reads. Some common classification methods based on different algorithms are RDP classifier ***(8)***, uclust ***(9)*** or BLAST ***(10)***. The largest databases for the 16SrRNA are the Ribosomal Database Project (RDP) ***(11)*** and Greengenes ***(12)***. Classification of the ITS is more complicated than 16S rRNA due to the high variation in sequence and size of this region and also because a great amount of fungal diversity has not been described yet. Thus, there is a big bias in fungal databases towards some specific phyla.

For the ITS taxonomy, the most up-to-date databases are UNITE ***(13)*** and the Warcup ITS training set ***(14)***. In our experience, fungal classification from BAL samples based on the Warcup ITS training allows to reach lower taxonomic levels (genus, family) with higher accuracy.

3.9 Bioinformatics analysis of the 16S rRNA gene and ITS sequencing data: Estimation of alpha and beta diversity

1. Before starting with the ecological and diversity analyses it is recommended to make rarefactions of the OTU-table and set all samples to the same number of reads to avoid bias in the results due to the different sequencing depth of the samples. If there is a big difference in the number of reads of the samples, rarefaction at the same depth can lead to an underestimation of diversity. Thus, it is recommendable to normalize the table by the number of sequences to base the following analyses on a relative abundance table (%). In addition, it is important to correct with respect to the copy number of the marker gene since different species have different copy numbers. Due to the low probability of having an accurate information of the copy number of all the species present in a sample different software exist (mainly for 16S rRNA gene) that estimate and correct by the copy number such as the rrNDB ***(15)*** and the Copyrighter software ***(16)***.

The alpha-diversity is generally characterized by using the species richness (estimated with the Chao 1, number of OTUs, rarefaction curves), the species richness and evenness (Shannon Index, Simpson Index) and the phylogenetic relationship (Phylogenetic Diversity) ***(1)***. Calculation of the main diversity metrics can be performed with the QIIME pipeline (e.g. core_diversity_analyses.py script) ***(6)*** and also with R software ***(17)*** packages: phyloseq ***(18)*** or vegan ***(19).***

2. Many different approaches allow comparing a set of samples based on the composition (beta-diversity) ***(20)***. The selected techniques depend on the goal of the study, for instance, if you want to evaluate whether two microbial communities differ depending on the disease state (*e.g.* healthy vs. pneumonia) or to evaluate the dynamics of a microbial community over time (*e.g.* evolution of the microbiome composition during antibiotic therapy). Ordination techniques are very useful exploratory approaches, such as principal coordinates analysis (PCoA), canonical correspondence analysis (CCA), or principal component analysis (PCA). These techniques summarize the microbiome variability and help in the identification of patterns in the microbial composition of the samples. Clustering analyses allow also identifying and visualizing clusters of samples in terms of OTU/taxa composition. The clusters can be generated on the basis of ecological metrics as the Bray–Curtis dissimilarity or based on phylogenetic distances as Unifrac ***(21)***. Moreover, heatmaps are used to visualize the relative abundance of the OTUs/taxa in the different samples. It is very useful for identifying those OTU/taxa explaining the differences between the different clusters. QIIME pipeline ***(6)*** and the Vegan package ***(19)*** of R software ***(17)*** can be used to perform these analyses.

3.10 Bioinformatics analysis of the 16S rRNA gene and ITS sequencing data: Statistical analysis

1. Different statistical analyses have been developed to test several ecological hypotheses. The statistical comparisons should be based on the biological question and the exploratory analysis results. For example, a very common question in human microbiome studies is whether there are statistically significant differences between two conditions such as healthy and disease status. To know if the alpha-diversity differs between the two groups the Wilcoxon rank-signed test can be used to compare the different diversity parameters (*e.g.* Shannon Index). Moreover, to identify those OTU/taxa under or over-represented in the two conditions (biomarker identification) the linear discriminant analysis (LDA) effect size (LEfSe) analysis is a powerful statistical method ***(22)***. Many statistical models and methods exist for analysing the association of the microbial community composition and covariates (*e.g*. clinical variables), very useful are multivariate statistical methods ***(23)***. A very important aim in clinical studies is to identify possible variables (*e.g.* age, immune system parameters, treatments) associated with a specific microbiome state or some microbial species. In this regard, PERMANOVA is one of the most widely used methods (based on distances) to determine if two conditions differ in a statistically significant way. Also, multivariate ANOVA based on dissimilarities (Adonis) could be used to test the significance of associations between environmental variables (*e.g.* antibiotic usage) and the microbiome composition. These statistical tests and others related are implemented in the Vegan package ***(19)*** of the R program ***(17)***.

4. Notes

1. Different types of clinical samples have been used to characterize the lung microbiome such as sputum, BAL, protected specimen brushings or oral washes. Despite the disadvantage of the invasive character of the bronchoscopy, the microbial composition of BALs is the closest to the real pulmonary microbiome ***(24)***. During *Legionella* infection a very low abundance of the pathogen is detected in the sputum microbiome ***(25)***, while a higher abundance of *Legionella* is expected in BAL samples during infection. The analysis of the BAL microbiome is the best approach so far, to characterize the lung microbiome during *Legionella* infection.

2. Different DNA extraction methods can lead to different results since some types of microorganisms are more resistant to chemical or mechanical lysis. However, protocols based on mechanical lysis (*e.g.* bead beating system) have proven more effective for the extraction of bacterial and fungal DNA from BAL samples. The two major consortia involved in the human microbiome project, the European MetaHIT and the American Human Microbiome Project (HMP) perform DNA extractions including a bead-beating step in their protocols ***(26)***. In our experience, mechanical lysis by bead-beating results in better detection of *Legionella* compared to methods based on chemical lysis.

3. Since PCR-based studies are extremely sensitive to low levels of DNA, it is also necessary to test the commercial reagents in the kit, which may be contaminated with microbial DNA. If there is a low level of contamination it is possible to sequence the negative control and subtract the contaminated reads from the dataset bioinformatically.

4. To evaluate the sequencing bias and to estimate the amount of the different microbial groups in a sample it is recommended to perform a step of spiking-in standards composed of known rations of DNA. Recent studies for the 16S rRNA gene have shown the advantages to spike-in the samples by using different strategies such as spiking the sample with exogenous bacteria ***(27)*** or by adding synthetic 16S rRNA genes ***(28)***.

5. The library preparation protocol is taken from the Illumina 16S Metagenomic Sequencing Library Preparations instructions ***(29)***. We included some modifications mainly in the specific-PCR step to increase the amount of product for the 16SrRNA and ITS amplicons when using BAL samples.

6. Keep the number of PCR cycles as low as possible. If necessary, for samples with very low DNA concentration the cycles can be increased up to 25 for bacterial DNA and 35 for fungal DNA samples at maximum.

7. The MiSeq paired-end sequencing method using the Illumina technology is one of the best approaches until now for the study of the human microbiome (specially tested in 16SrRNA analysis) ***(30)***. Since the paired-end method is based on the sequencing of both ends of a fragment, longer and higher-quality reads are generated compared to the single-end sequencing method, improving the accuracy and quality of the data. For example, MiSeq v2 and v3 Reagent Kits from Illumina allow obtaining up to 7.5 and 15 Gb of data with a read length up to 2 x 250 bp and up to 2 x 300 bp, respectively.

5. References

1. Lozupone CA, Knight R (2008) Species divergence and the measurement of microbial diversity. FEMS Microbiol Rev. 32(4):557-578

2. Andrews S (2010) FastQC A Quality Control tool for High Throughput Sequence Data http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 26 Abr 2010

3. Hannon Lab (2009) FASTX Toolkit http://hannonlab.cshl.edu/fastx_toolkit/index.html. Accessed 02 Feb 2010

4. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27(6):863-864

5. Aronesty E (2011) ea-utils: "Command-line tools for processing biological sequencing data" https://expressionanalysis.github.io/ea-utils/. Accessed 20 Jun 2017

6. Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335-336

7. Schloss PD, Westcott SL, Ryabin T et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 75(23):7537-7541

8. Wang Q, Garrity GM, Tiedje JM et al (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 73(16):5261-5267

9. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460-2461

10. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol. 215(3):403-410

11. Cole JR, Wang Q, Fish JA et al (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42(Database issue):D633-642

12. DeSantis TZ, Hugenholtz P, Larsen N et al (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 72(7):5069-5072

13. Kõljalg U, Larsson KH, Abarenkov K et al (2005) UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytol. 166(3):1063-1068

14. Deshpande V, Wang Q, Greenfield P et al (2016) Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. Mycologia 108(1):1-5

15. Stoddard SF, Smith BJ, Hein R et al (2015) rrnDB: improved tools for interpreting and rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res 43:D593-598. doi: 10.1093/nar/gku1201

16. Angly FE, Dennis PG, Skarshewski A et al (2014) CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome 2:11. doi: 10.1186/2049-2618-2-11

17. R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/. Accessed 28/09/2017

18. McMurdie PJ, Holmes S (2013) phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8(4):e61217. doi: 10.1371/journal.pone.0061217

19. Oksanen J, Guillaume F, Friendly M et al (2017) vegan: Community Ecology Package. R package version 2.4-4 http://CRAN.R-project.org/package=vegan. Accessed 24/08/2017

20. Goodrich JK, Di Rienzi SC, Poole AC (2014) Conducting a microbiome study. Cell 158(2):250-262

21. Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 71(12):8228-8235

22. Segata N, Izard J, Waldron L et al (2011) Metagenomic biomarker discovery and explanation. Genome Biol. 12(6):R60. doi: 10.1186/gb-2011-12-6-r60

23. Oksanen J (2015) Multivariate Analysis of Ecological Communities in R: vegan tutorial. http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

24. Dickson RP, Erb-Downward JR, Freeman CM et al (2017) Bacterial Topography of the Healthy Human Lower Respiratory Tract. MBio doi: 10.1128/mBio.02287-16

25. Mizrahi H, Peretz A, Lesnik R et al (2017) Comparison of sputum microbiome of legionellosis-associated patients and other pneumonia patients: indications for polybacterial infections. Sci Rep. 7:40114. doi: 10.1038/srep40114

26. Wesolowska-Andersen A, Bahl MI, Carvalho V et al (2014) Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome 2:19. doi: 10.1186/2049-2618-2-19

27. Stämmler F, Gläsner J, Hiergeist A et al (2016) Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome 4(1):28. doi: 10.1186/s40168-016-0175-0

28. Tourlousse DM, Yoshiike S, Ohashi A et al (2017) Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. Nucleic Acids Res. 45(4):e23. doi: 10.1093/nar/gkw984

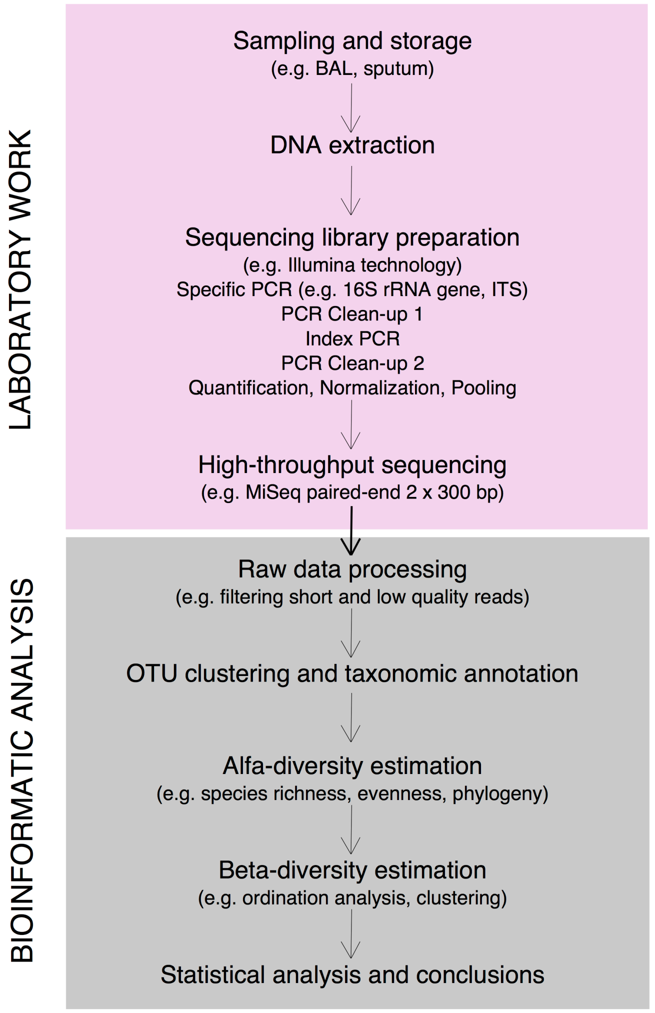
29. Illumina (2013) 16S Metagenomic Sequencing Library Preparation. https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

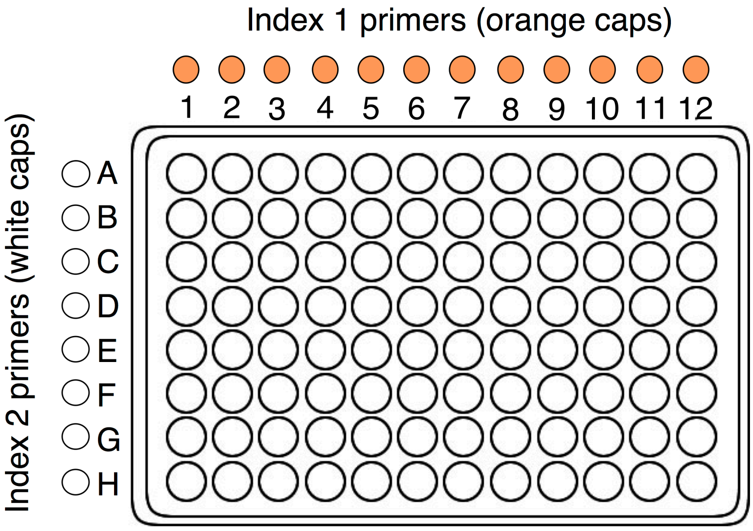
30. Illumina (2017) 16S Metagenomics Studies with the MiSeq System. https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_16s_sequencing.pdf

Figure legends:

Fig. 1. Different steps for performing a pulmonary microbiome study. The analyses are divided in the experimental procedure described within the pink frame (from the lung sampling to the high-throughput sequencing) and the bioinformatics analyses described within the gray frame (from the raw data processing to the statistical analysis and biological conclusions).

Fig 2. Representation of the TruSeq Index Plate. To perform the index PCR it is recommendable to arrange the Index 1 (orange caps) and 2 (white caps) primers in a rack. The Index 2 primers should be aligned vertically with the rows from the A to H and the Index 1 horizontally with the columns from 1 to 12.


LABORATORY WORK
Sampling and storage
(e.g. BAL, sputum)
DNA extraction
Sequencing library preparation
(e.g. Illumina technology)
Specific PCR (e.g. 16S rRNA gene, ITS)
PCR Clean-up 1
Index PCR
PCR Clean-up 2
Quantification, Normalization, Pooling
High-throughput sequencing
(e.g. MiSeq paired-end 2 x 300 bp)
BIOINFORMATIC ANALYSIS
Raw data processing
(e.g. filtering short and low quality reads)
OTU clustering and taxonomic annotation
Alfa-diversity estimation
(e.g. species richness, evenness, phylogeny)
Beta-diversity estimation
(e.g. ordination analysis, clustering)
Statistical analysis and conclusions


Index 1 primers (orange caps)
1
2
3
4
5
6
7
8
9
10
11
12
Index 2 primers (white caps)
A
B
C
D
E
F
G
H