UNIVERSITY OF ROME TOR VERGATA

MACROAREA OF MATHEMATICAL, PHYSICAL AND NATURAL SCIENCES

COURSE OF STUDY IN

Biological Sciences

SCIENTIFIC MEMORY

Bioinformatics

Modulation of the Response to Immunotherapy in Melanoma by the Intestinal Microbiota

Supervisor:
Prof.ssa Manuela Helmer Citterich

Candidate:
matricola: 0295384
Andrea Nardoni

Co-supervisor:
Dr. *Adelaide Teofani*

Academic Year 2022/2023

Index

Index 1

1. Introduction 2
   - 1.1 Melanoma 2
   - 1.2 Metagenomics 2
   - 1.3 Modulation of the Response to Anti-PD-1 Immunotherapy 3
2. Aim of the Thesis 4
3. Materials e Methods 5
   - 3.1 European Nucleotide Archive 5
   - 3.2 Paired End Sequencing 5
   - 3.3 Pandas 6
   - 3.4 FastQC 6
   - 3.5 MetaPhlAn 6
   - 3.6 R and R studio 7
4. Results 9
5. Conclusions 11

References 12

1. Introduction

1.1 Melanoma

Melanoma is a relatively frequent neoplasm, but its incidence is continuously increasing (Schadendorf *et al.,* 2015). This skin cancer arises due to an acquired genetic mutation in melanocytes, the cells responsible for producing melanin (Schadendorf *et al.,* 2015). It usually appears on healthy skin where no nevi are present, or it can form from an existing nevus where melanocytes are densely clustered. Any area of the skin, especially those commonly exposed to the sun, can be affected by melanoma. However, it can also occur more rarely in other locations, such as the eye and the mucous membranes of the oral cavity (Schadendorf *et al.,* 2015). The primary cause of melanoma is often excessive exposure to ultraviolet (UV) radiation from the sun, although genetic factors may contribute to its development. Typically, melanoma progression is counteracted through immunotherapy, which includes treatments such as immune checkpoint inhibitors (e.g., *anti-PD-1*, *anti-PD-L1*, and *anti-CTLA-4*) (Schadendorf *et al.,* 2015). These therapies work by stimulating the immune system to recognize and effectively eliminate cancer cells. However, some patients do not respond to these treatments or develop resistance after an initial response (Schadendorf *et al.,* 2015).

1.2 Metagenomics

Microbial communities of bacteria, archaea, viruses, and unicellular eukaryotes play crucial roles in the environment and human health. One of the most complex habitats where many microbial communities reside is the *gut microbiota*. This is made up of the collection of microorganisms that colonize the gastrointestinal tract (Quince *et al.,* 2017). The gut microbiota plays a fundamental role in multiple physiological processes, including protection against pathogens, as well as the modulation of immune system functions and the development of the gastrointestinal tract.

Metagenomics is the study of the *microbiome* (i.e., the total genetic material possessed by the microbiota) through sequencing techniques and the isolation of DNA from an entire microbial community (Quince *et al.,* 2017). Unlike traditional genomics, which focuses on a single organism, metagenomics allows for the analysis of the genetic material of entire microbial communities without the need to isolate and culture individual microorganisms. This approach is particularly useful for identifying microbial species that cannot be cultured in the laboratory, and therefore would not be identified through conventional approaches.

The study of the gut microbiota through metagenomics provides detailed information about the composition of the microbial community, allowing for the identification of dysbiosis, or imbalances in the microbiota composition, and an understanding of the interactions between different microorganisms in the gut environment (Quince *et al.*, 2017).

Two measures used to characterize and compare microbial communities are *α-diversity* and *β-diversity*. Specifically:

- "*α-diversity*" refers to the diversity of species within a single specific site. In other words, it is an intraspecific measure of the number and variety of species present.
- "*β-diversity*" refers to the diversity between groups, measuring how much microbial communities differ from each other under two different conditions or groups (interspecific measure).

The technique commonly used in metagenomics to identify species in a microbial community is sequencing of the variable regions of the *16S marker gene.* Although this technique is fast and very cost-effective, it is less accurate and reliable compared to *shotgun metagenomics* (Quince *et al.*, 2017). In shotgun sequencing, the term "*shotgun*" refers to the fact that the DNA is randomly fragmented into small pieces, which are then sequenced. Unlike sequencing the genome of a single organism (where the DNA is isolated from a specific organism and sequenced in a targeted manner), shotgun metagenomics allows for the sequencing and analysis of all DNA fragments present in a sample. This technique, unlike 16S sequencing, provides detailed information about the genetic diversity of the microbial communities under study. It not only phylogenetically characterizes the microorganisms present but also enables functional characterization.

1.3 Modulation of the Response to Anti-PD-1 Immunotherapy

Immune checkpoints, such as *CTLA-4* (Cytotoxic T-Lymphocyte-Associated protein 4) and *PD-1* (Programmed Death-1), are proteins that regulate the immune response and are exploited by tumor cells to evade the immune system. In certain cancers, such as melanoma, an interaction occurs between *PD-1* proteins, located on immune cells, and their ligands, *PD-L1* or *PD-L2*, expressed on tumor cells. This interaction effectively "turns off" the immune response, preventing immune cells from attacking the tumor. *Immune checkpoint blockade therapy* uses drugs that prevent these tumor proteins from inhibiting the immune response, allowing the immune system to recognize and fight tumor cells more effectively.

Experiments in both murine and human models have shown that the gut microbiota influences the regulation of patients' responses to immune checkpoint blockade therapy (Gopalakrishnan *et al.*, 2018). A recent study on melanoma patients undergoing *anti-PD-1 immunotherapy* revealed significant differences in the gut microbiota between patients who respond to immunotherapy and those who do not (Gopalakrishnan *et al.,* 2018). The study, analyzing *16S sequencing data* from 43 fecal samples of melanoma patients, found that responsive patients exhibited higher *α-diversity* and a greater *relative abundance* of "*Ruminococcaceae*" bacteria. In contrast, non-responsive patients showed a greater *relative abundance* of "*Bacteroidetes*". Additionally, in the same study, metagenomic analysis revealed functional differences in the gut bacteria of responsive patients, indicating an enhancement of systemic and antitumor immunity.

Transplanting fecal microbiota from responsive patients into germ-free mice improved tumor response to anti-PD-1 therapy and significantly slowed tumor growth compared to mice transplanted with the microbiota of non-responsive patients. The abundance of specific bacteria in the patients' gut microbiome is correlated with systemic and antitumor immune responses. Patients with a higher abundance of "*Ruminococcaceae*" bacteria show better *progression-free survival (PFS),* while those with a higher abundance of "*Bacteroidetes*" bacteria show reduced PFS.

Another important study worth mentioning is the one conducted by Davar *et al.* in 2021 on a small group of melanoma patients with advanced disease who exhibited resistance to *anti-PD-1*. This study also observed that fecal microbiota transplantation effectively overcame resistance to immunotherapy (Davar *et al.,* 2021). These studies suggest that the gut microbiome could play a key role in modulating the response to immune checkpoint blockade therapy in melanoma patients.

2. Aim of the Thesis

Many studies highlight the relationship between the response to anti-PD-1 immunotherapy in melanoma and the composition of the microbiota. Among them, the study by Gopalakrishnan *et al.* (2018) is particularly relevant, as it demonstrates the influence of the gut microbiota on the regulation of tumor response to immune checkpoint blockade therapy in murine models (Gopalakrishnan *et al.,* 2018). The aim of this thesis is to analyze shotgun sequencing data from fecal samples of 11 resistant and 14 non-resistant patients to anti-PD-1 immunotherapy, collected in the work by Gopalakrishnan *et al.* and publicly available in the *Sequence Read Archive (SRA)* database.

Unlike the study by Gopalakrishnan *et al.*, which uses shotgun data solely for functional analysis and phylogenetically characterizes microbial communities based on 16S sequencing data, this work uses *shotgun sequencing data* for phylogenetic characterization. This thesis is part of a larger project aimed at comparing and integrating results from different studies exploring the relationship between immune response and gut microbiota. The comparison and integration of the various data available in public databases will be carried out using meta-analysis techniques, allowing for a comprehensive and detailed evaluation of the collected information.

In parallel, bacterial gene variants will be evaluated to explore how they may influence patients' differential response to immunotherapy. Future analyses will aim to better clarify how interactions between the immune system and the microbiota influence the response. By analyzing shotgun metagenomic data, it was possible to obtain highly accurate information about the composition of the microbial communities under study. Using specific bioinformatics pipelines for processing and analyzing sequencing data, detailed information on the taxonomy and abundance of bacterial species present in the analyzed samples was obtained.

3. Materials e Methods

3.1 European Nucleotide Archive

The data used in this work were downloaded from the *European Nucleotide Archive (ENA),* an archive that provides free and unlimited access to annotated DNA and RNA sequences. In addition, ENA stores complementary information such as experimental procedures, details on sequence assembly, and other metadata related to sequencing projects. The archive is composed of three main databases: the *Sequence Read Archive (SRA),* the *Trace Archive,* and *EMBL*. SRA is a specific resource within ENA that manages high-throughput sequencing data (*NGS*, Next-Generation Sequencing). SRA contains raw data from sequencing experiments, including next-generation sequencing data generated by platforms such as Illumina.

3.2 Paired End Sequencing

The genetic material isolated from the fecal samples of the 25 patients was sequenced using the "*paired-end*" method, producing a total of 50 ".fastq" files (fast = FASTA; q = quality). Paired-end sequencing involves reading each fragment from both ends (from 5' to 3' and from 3' to 5') to increase sequencing accuracy. For each fragment, a forward read (from 5' to 3') and a reverse read (from 3' to 5') are produced. The output of paired-end sequencing thus generates two files per sample: one fastq file for the forward reads and one fastq file for the reverse reads.

In a ".fastq" file, each sequence is characterized by 4 lines:

- Line 1: Always starts with @ and is a sequence identifier containing sequencing characteristics.

- Line 2: Contains the read sequence.
- Line 3: Always starts with the "+" symbol and may contain a description.
- Line 4: Contains the base quality score, or the Phred Score, encoded in ASCII (a character encoding system). It is a probabilistic value indicating the likelihood that the nucleotide base was sequenced correctly. The Phred Score is calculated using the following formula:

$$Q = -10 \cdot \log_{10}(P)$$

In this formula, Q represents the Phred quality score, while P is the probability that the base is the result of a sequencing error.

3.3 Pandas

The information related to each downloaded sample was extracted from the metadata provided by the authors of the study that produced the sequencing data. Using the *Pandas 1.5.3* library (McKinney *et al.,* 2015) in the *Python 3.10.12* programming environment, the metadata were manipulated to extract the relevant information for each sample. This allowed for the association of the sample identification code from the *SRA* database with the R (responsive) and NR (non-responsive) phenotype.

3.4 FastQC

The quality of the sequencing data was analyzed using the bioinformatics tool "*FastQC 0.11.9*" (Brown *et al.*, 2017), which is useful for examining various aspects of raw data quality before proceeding with further analysis steps. This tool generates a report file for each fastq file, which includes graphs such as base quality distributions, sequence lengths, GC content, and other important aspects of sequencing data quality. The report files produced by FastQC were then input into *MultiQC*, an open-source tool designed to visualize and comprehensively analyze the quality of the data. Subsequently, the data were processed using the "*Fastp*" tool to remove adapters used in library preparation and low-quality bases.

3.5 MetaPhlAn

The output from "*Fastp*," containing the filtered data, was used as input for the bioinformatics tool "*MetaPhlAn 3.0.14*" (Blanco-Míguez *et al.,* 2023). MetaPhlAn is widely used to profile the composition of microbial communities present in biological samples, such as those obtained from shotgun metagenomic sequencing data. This tool allows for the analysis of metagenomic sequencing data, providing detailed information on the presence and relative abundance of organisms within a microbial community. Specifically, MetaPhlAn generates a microbial community composition profile by classifying and quantifying organisms at different taxonomic levels (kingdom, phylum, class, order, family, genus, species, and strain). The output of MetaPhlAn is a tabular text file that reports the relative abundance of the microbial clades identified in the metagenomic samples.

3.6 R and R studio

The analysis of the profiling results conducted with MetaPhlAn was performed using the "*R 4.3.2*" programming language (Ariel de Lima *et al.*, 2022), commonly used for statistical analysis and generating high-quality graphics. The R language was used within R Studio, which is an integrated development environment providing a software-user interface for importing and visualizing data, installing packages, and exporting graphics. R Studio is characterized by several sections:

- *Code Editor*: This section is where codes can be written for execution, as well as notes and comments.
- *R Console*: This is where the codes are executed.
- *Workspace and History*: This section displays the various files that have been loaded into the R session or created.
- *Plots and Files*: This section provides access to the folder from which files are being taken and is where the generated plots are displayed.

3.6.1 Vegan

The measurement and visualization of *β-diversity* were performed using the R package “Vegan” (Dixon and Philip, 2003). Vegan is an R package that provides a set of functions for analyzing and interpreting ecological data, particularly for community studies. The package includes functions for diversity analysis, classification and ordination of communities, and similarity and dissimilarity analysis.

β-diversity, which measures interspecific diversity between the responsive and non-responsive patient groups, was calculated using the *Bray-Curtis index.* The dissimilarity between the two groups was analyzed and represented using *Principal Coordinate Analysis (PCoA).* The *Bray-Curtis index* is used to quantify the dissimilarity in species composition between two different sites based on counts at each site. This index is calculated as:

$$BC_{ij} = 1 - \frac{2 * C_{ij}}{S_i + S_j}$$

Where:

- *Cij*: represents the sum of the minimum counts for each species found in both sites.

- *Si*: represents the total number of specimens counted at site i.
- *Sj*: represents the total number of specimens counted at site j.

The *Bray-Curtis index* always ranges from 0 to 1, where:

- 0 indicates that two sites have zero dissimilarity, meaning they share the exact same number of each type of species.
- 1 indicates that two sites have complete dissimilarity, meaning they do not share any of the types of species.

Principal Coordinate Analysis (PCoA) is a statistical technique for dimensionality reduction that converts similarity or dissimilarity information between samples into principal coordinates. *PCoA* can be used in conjunction with *β-diversity* to analyze and represent differences in microbial community composition between two or more sample groups. The result of *β-diversity* measurement is a distance matrix that represents how dissimilar the microbial communities are between the groups under analysis. *PCoA* uses the distance matrix to place each sample into a multidimensional space such that the distances between points in this space reflect the distances in the original matrix. This means that samples with similar microbial communities will be positioned close together in the multidimensional space, while those with different microbial communities will be further apart.

3.6.2 Phyloseq and Microbiome

The analysis of *α-diversity* was conducted in R using the "*Phyloseq*" R package (McMurdie *et al.*, 2013), a software package that integrates taxonomy, phylogeny, and metadata through the creation of so-called "phyloseq objects" and provides tools for managing, analyzing, and visualizing microbiome data in R (Ariel de Lima *et al.*, 2022). *α-diversity* was calculated using the "*alpha()*" function from the "*Microbiome*" R package. The comparison of *α-diversity* between the responder and non-responder groups was made using various indices: *observed species*, *Chao1*, *Shannon*, and *Simpson*. Specifically:

- *Observed species*: represents the total number of species or taxa observed in a sample.
- *Chao1*: is an index that estimates species richness based on the number of rare, unobserved species. This index accounts for species that might be present in the community but have not been observed.
- *Shannon*: considers both species richness (number of species) and evenness (relative distribution of species abundances). Higher Shannon diversity values indicate greater diversity, considering both the presence of different species and their even distribution.

- *Simpson*: unlike Shannon diversity, Simpson diversity focuses more on species dominance. A lower Simpson diversity value indicates higher diversity, accounting for the probability that two randomly selected individuals belong to different species.

The *α-diversity* indices were represented using a *boxplot*, a statistical graph that depicts the distribution of a data set. Specifically, it allows visualization of the median, first and third quartiles, and provides information on deviations of any outliers. It consists of several components:

1. *Box*: represents the interquartile range (IQR), the difference between the third quartile (Q3) and the first quartile (Q1). The line inside the box indicates the median, or the central value of the data set.
2. *Whiskers*: are lines extending from the sides of the box to the most extreme data points.
3. *Outliers*: are individual points outside the whiskers that indicate extreme values or anomalies in the data set.

The "*aggregate_taxa*" function from the "*microbiome*" R package was used to group microorganisms at the *phylum* taxonomic level. Differences in the relative abundances of phyla were represented using a "*barplot*," a type of graph used to show and compare quantities or frequencies between different classes (in our case, R and NR), using the "*plot_bar*" function from the "phyloseq" R package.

4. Results

The paired-end sequencing data from the 25 samples were downloaded from the *European Nucleotide Archive (ENA)* along with a metadata file containing clinical and demographic data related to each patient. The metadata were processed using Python's "*Pandas*" library to select only the identification codes of the samples subjected to shotgun metagenomics and to determine which samples belonged to the responsive group and which to the non-responsive group. For each of the 25 samples, sequencing quality was assessed using the *FastQC* tool. Adapter trimming and low-quality base removal were performed using the *Fastp* tool. The filtered data were then analyzed using the bioinformatics tool "*MetaPhlAn*," which allows the microbial composition of each sample to be determined. The profiles obtained with MetaPhlAn were imported into the *R Studio* environment using the R package "*Phyloseq*." The data analysis performed in R consisted of comparing relative abundances and measuring *α* and *β-diversity*. Relative abundances were compared using a bar plot, in which microbial species were grouped at the *phylum* taxonomic level. In the bar plot, shown in Figure 1, responsive patients show a higher relative abundance of *Firmicutes* bacteria (20% more abundant compared to NR), while non-responsive patients show a higher relative abundance of *Bacteroidetes* (20% more abundant compared to R).

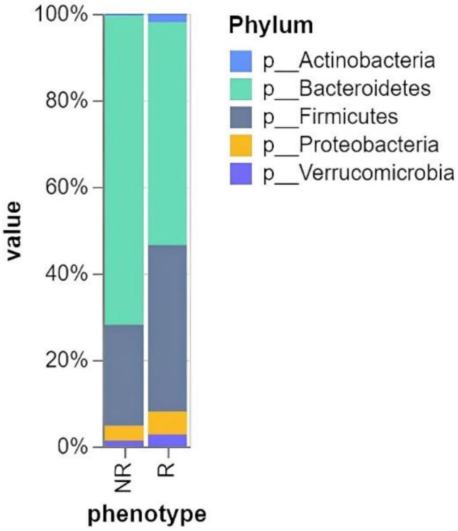

Figure 1. Comparison of relative abundance between the group of responsive patients (R) and non-responsive patients (NR).

The analysis of *α-diversity* was performed using four indices: *observed species*, *Shannon*, *Simpson*, and *Chao1*. The four measures of *α-diversity* were compared between the R and NR groups using boxplots, as shown in Figure 2. These graphs show that, according to all four analyzed indices, patients who respond to therapy have *higher α-diversity* compared to patients who do not respond to therapy.

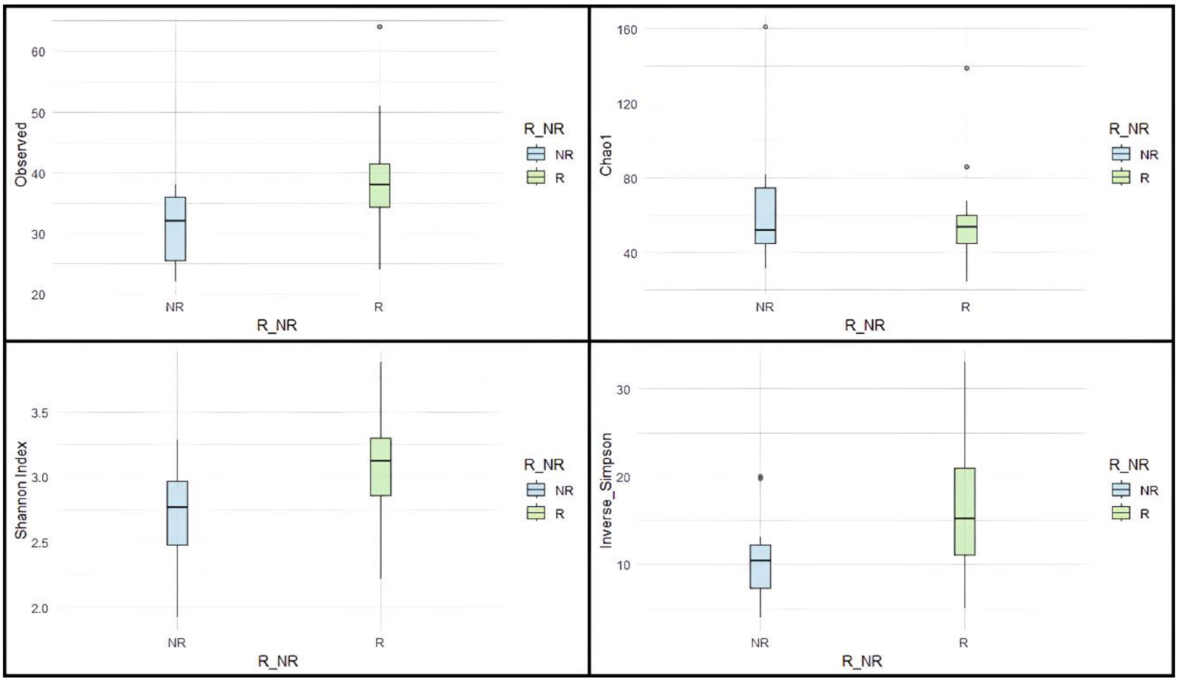

Figure 2. Comparison of α-diversity between the group of responsive patients (R) and non-responsive patients. α-diversity was calculated using the Shannon, Inverse Simpson, Observed species, and Chao1 indices.

The visualization of *β-diversity* using *PCoA*, obtained from the distance matrix calculated with *Bray-Curtis* distance, shows a clear separation between the two groups, R and NR, with the exception of a few samples (Figure 3). This separation indicates the degree of diversity in the structure and composition of microbial communities between the R and NR groups.

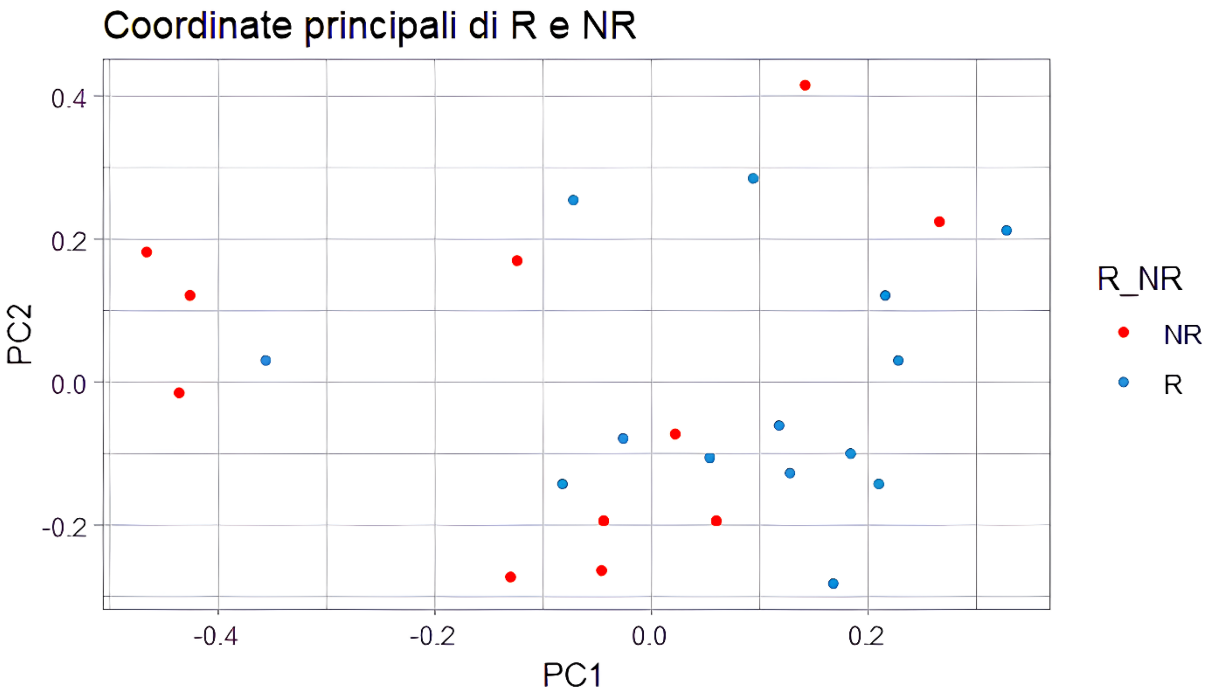

Figure 3. Representation of the dissimilarity between the R and NR groups through PCoA (Principal Coordinate Analysis). The PCoA uses the distance matrix calculated with the Bray-Curtis dissimilarity index to place each sample in a two-dimensional space, so that the distances between points in this space reflect the distances in the original matrix.

5. Conclusions

There is increasing evidence of the relationship between the response to *anti-PD-1* immunotherapy in melanoma and the composition of the microbiota. The results observed in this thesis highlight the existence of a difference in microbiota composition between patients who respond and those who do not respond to *anti-PD-1* immunotherapy. Specifically, a higher *α-diversity* was observed in the group of patients who respond to therapy compared to those who do not. In other words, a greater number of microbial species is associated with a positive response to immunotherapy. This may be because greater microbial diversity can contribute to a more robust and resilient immune system, which can enhance the patient's response to therapy. These findings confirm that the gut microbiome could play a key role in modulating the response to immune checkpoint blockade therapy in melanoma patients, but the specific molecular mechanisms through which this modulation occurs remain to be clarified.

References

Ariel de Lima, D., Helito, C. P., Lima, L. L. D., Clazzer, R., Gonçalves, R. K., & Camargo, O. P. D. (2022). How to perform a meta-analysis: a practical step-by-step guide using r software and rstudio. *Acta Ortopédica Brasileira, 30*, e248775.

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., ... & Segata, N. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, 1-12.

Brown, J., Pirrung, M., & Mccue, L. A. (2017). FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, *33*(19), 3137–3139.

Davar, D., Dzutsev, A. K., McCulloch, J. A., Rodrigues, R. R., Chauvin, J. M., Morrison, R. M., Deblasio, R. N., Menna, C., Ding, Q., Pagliano, O., Zidi, B., Zhang, S., Badger, J. H., Vetizou, M., Cole, A. M., Fernandes, M. R., Prescott, S., Costa, R. G. F., Balaji, A. K., … Zarour, H. M. (2021). Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. *Science*, *371*(6529), 595–602.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of vegetation science*, *14*(6), 927-930.

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., Prieto, P. A., Vicente, D., Hoffman, K., Wei, S. C., Cogdill, A. P., Zhao, L., Hudgens, C. W., Hutchinson, D. S., Manzo, T., Petaccia De Macedo, M., Cotechini, T., Kumar, T., Chen, W. S., … Wargo, J. A. (2018). Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science (New York, N.Y.)*, *359*(6371), 97.

McKinney, W., & Team, P. D. (2015). Pandas-Powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit*, *1625*.

McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, *8*(4), e61217.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology 2017 35:9*, *35*(9), 833–844.

Schadendorf, D., Fisher, D. E., Garbe, C., Gershenwald, J. E., Grob, J. J., Halpern, A., Herlyn, M., Marchetti, M. A., McArthur, G., Ribas, A., Roesch, A., & Hauschild, A. (2015). Melanoma. *Nature Reviews Disease Primers 2015 1:1*, *1*(1), 1–20.