PRedictive immunOlogical sIgnatureS in lung cancer (PROMISE) Trial Report

Ece Kartal

6.01.2023

Introduction

Lung cancer is a leading cause of cancer-related deaths worldwide with limited treatment options and poor prognosis. The microbiome has been shown to play a role in lung cancer development and progression. Studies have shown that the microbiome is altered in lung cancer patients compared to healthy individuals. However, the exact role of the microbiome in lung cancer is not fully understood and more research is needed.

The goal of this project is to identify a microbial biomarker signature that can predict whether a patient with Non-Small Cell Lung Cancer (NSCLC) will respond to or resist treatment. Specifically, we aim to identify specific microbial communities or patterns in the microbiome that may be associated with treatment response or resistance in this subset of lung cancer patients over time.

By analyzing the microbial communities present in fecal samples from individuals with NSCLC, we aim to identify specific taxa and pathways that are associated with treatment response. Our findings will provide insight into the role of the microbiome as a potential biomarker in the diagnosis and treatment of lung cancer, and may lead to the development of targeted therapies based on the individual's microbiome profile.

Data analysis pipeline

We used shotgun metagenomics sequencing to analyze the bacterial communities present in fecal samples from these individuals and use statistical and bioinformatic approaches to identify microbial taxa that are differentially abundant in individuals with lung cancer.

The Snakemake pipeline, to process shotgun metagenomics data, consisted of several steps, including quality control, annotation, and functional analysis:

- First, we performed quality control using FastQC to identify any potential issues with the raw sequencing data. Next, we used fastp to perform adapter clipping and merging on the data. We also applied low complexity and quality filtering using Bbduk(Bbmap).
- To further filter the data, we used Bowtie2 to align the reads to a reference genome of the host species (hg19). The output of this step was a single BAM file containing both aligned and unaligned reads. We then used samtools view to filter the BAM file, retaining only the reads that were not aligned to the host genome.
- Next, we used samtools fastq to convert the filtered BAM file back into two separate FASTQ files for the forward and reverse reads. These FASTQ files contained reads that had been filtered for alignment to the host genome and were used for downstream analysis.
- Finally, for taxonomic profiling, we used motus v3.0.3 to process the forward and reverse FASTQ files and a database of marker genes. This step generated a set of two files: a relative abundance file and

a taxa file. The relative abundance file contained the relative abundance of each microbe (defined as a microbial operational taxonomic unit or MOTU) in the sample, while the taxa file contained the taxonomic classification of each MOTU.

- In addition to taxonomic annotation, we used MOTUs to perform functional analysis by generating single nucleotide variant (SNV) profiles. These profiles provided insights into the genetic diversity within the microbial community and helped identify functionally important genetic changes.

Results

In this pilot dataset, I analyzed 48 fecal shotgun samples from NSCLC cancer patients. The total read count in these samples ranged from 6 million to 50 million, with 5 samples having fewer than 10 million reads. The overall host DNA contamination and low quality reads were less than 1%.

```
## [1] 33571    48
```

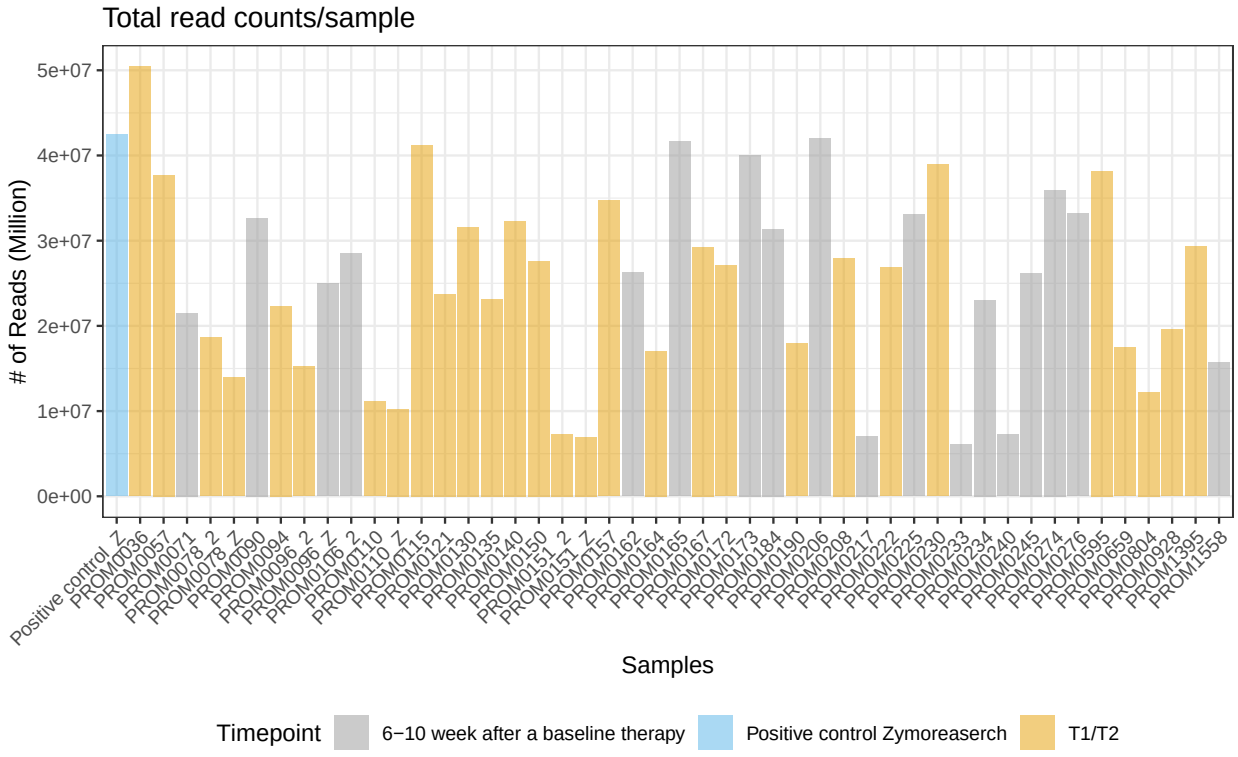


Figure 1. Total read count per sample

Diversity Analysis

Diversity in the ecological sense is intuitively understood as the complexity of a community of organisms.The two main categories of methods are known as **alpha diversity** and **beta diversity**

Rarefaction Rarefaction is used to standardize the number of samples (i.e. reads) across different datasets. This is important when comparing the diversity of samples because it helps to account for differences in sequencing depth. By using this technique to rarefy to a constant depth, we can ensure that the samples

are being compared fairly and that any differences in species detection are not simply due to differences in sequencing depth. This allows for a more accurate comparison of the diversity or similarity of the samples.

When we use rarefied reads, the species detection tends to be lower because we are only considering a subset of the total reads. However, this can be useful for comparing the diversity or similarity of samples because it helps to account for differences in sequencing depth. Due to this reason, I use several diversity measurements to have a consensus.

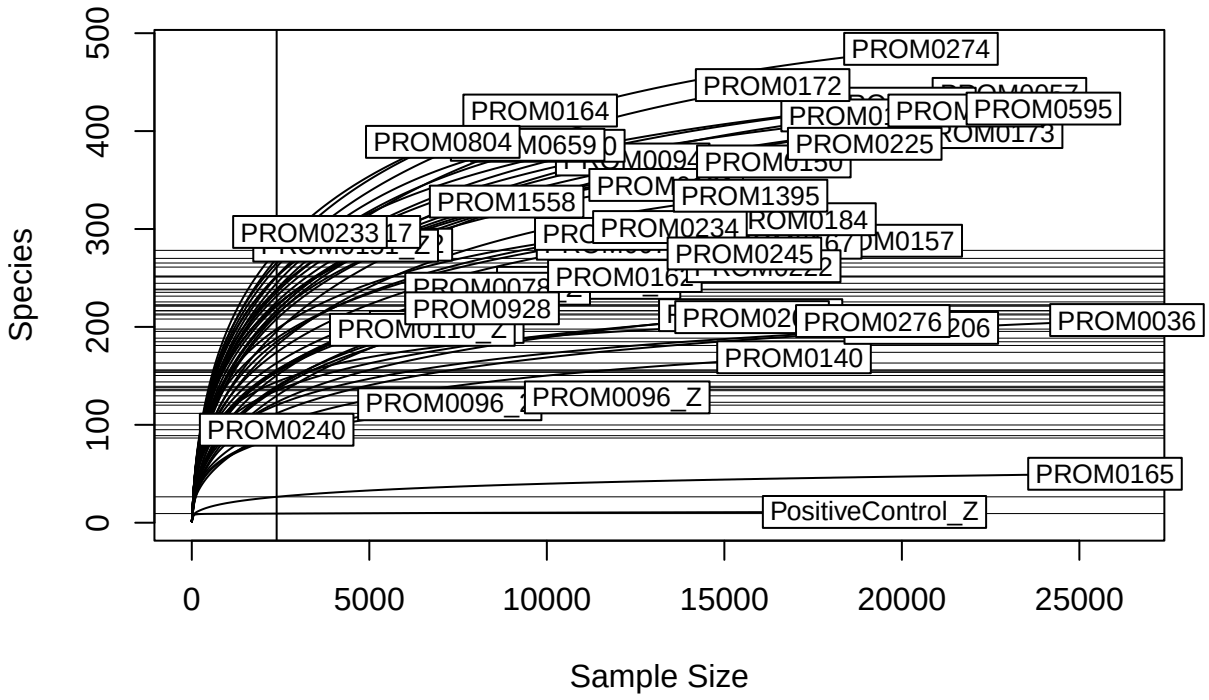


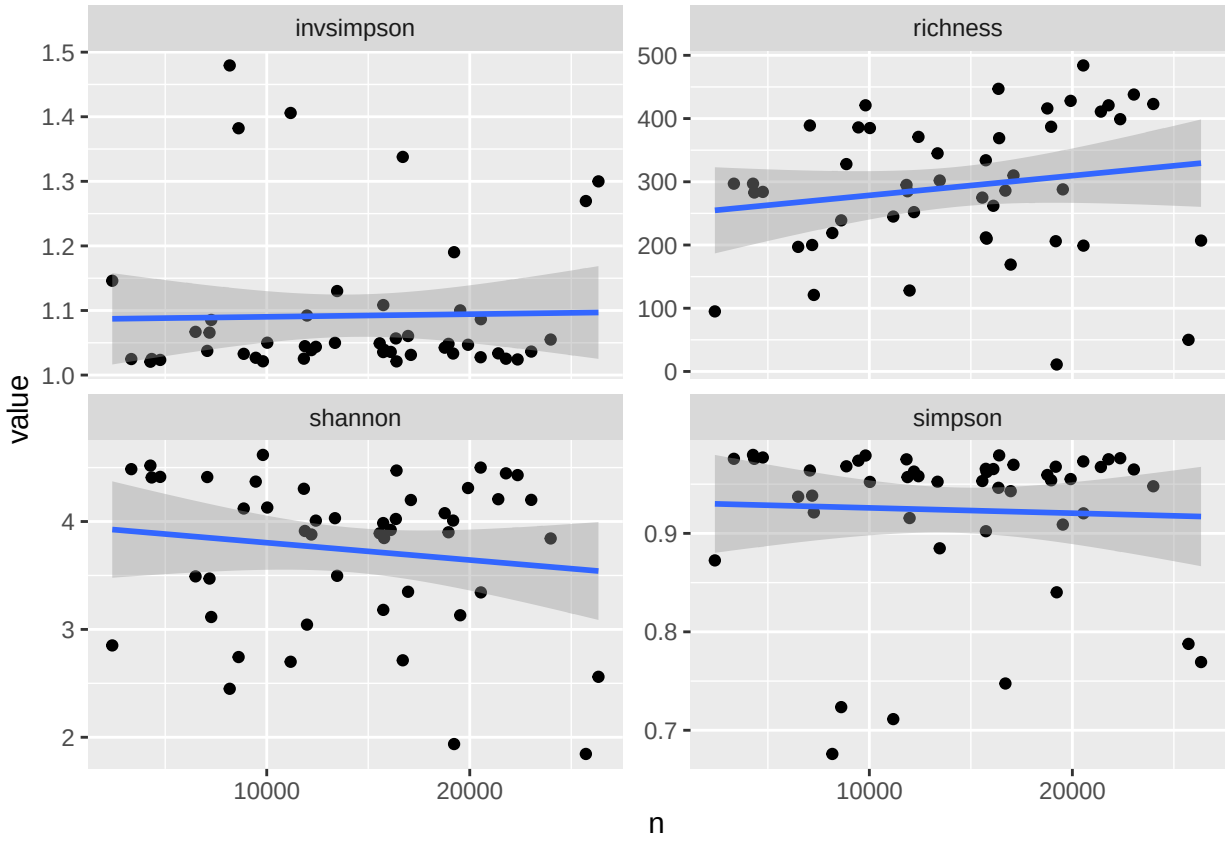
Figure 2. Total marker gene read counts per sample.

Alpha (within sample) Diversity Analysis

Alpha diversity measures the diversity within a single sample and is generally based on the number and relative abundance of taxa at some rank **Shannon**: How difficult it is to predict the identity of a randomly chosen individual. **Simpson**: The probability that two randomly chosen individuals are the same species. **Inverse Simpson**: This is a bit confusing to think about. Assuming a theoretically community where all species were equally abundant, this would be the number of species needed to have the same Simpson index value for the community being analyzed.

The `diversity` function from the vegan package can be used to calculate the alpha diversity of a set of samples.

```
## `geom_smooth()` using formula 'y ~ x'
```




```
## Saving 6.5 x 4.5 in image
```

Figure 3. Comparison of alpha diversity measurements with read counts.

As we can see, a higher read count increases the likelihood of detecting a greater diversity of species in the sample (richness).

In general, you will see roughly normal distribution for Shannon's diversity as well as most richness metrics. Simpson's diversity, on the other hand, is usually skewed. So most will use inverse Simpson (1/Simpson) instead. This not only increases normalcy but also makes the output more logical as a higher inverse Simpson value corresponds to higher diversity.

```
## Saving 6 x 6 in image
```

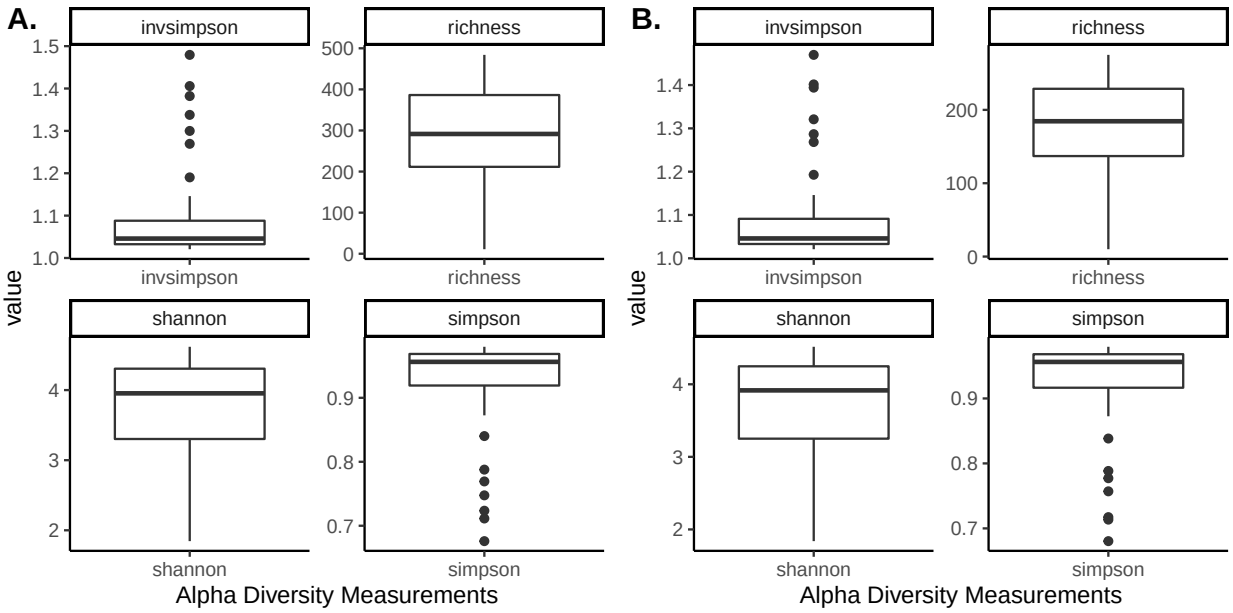


Figure 4. Comparison of alpha diversity measurements between non-rarefied (A) and rarefied reads (B).

As seen in the figure, when we rarefy samples, we loose a lot of low abundant species, that would change the diversity measurements.

Beta (between sample) Diversity Analysis

Beta diversity is a way to quantify the difference between two communities. There are many metrics that are used for this, but we will only mention a few of the more popular ones.

- Indexes used with presence/absence data: *Jaccard*: the number of species common to both communities divided by the number of species in either community. *Unifrac*: The fraction of the phylogenetic tree branch lengths shared by the two communities.
- Indexes used with count data: *Bray–Curtis*: The sum of lesser counts for species present in both communities divided by the sum of all counts in both communities. This can be thought of as a quantitative version of the Sørensen index. *Weighted Unifrac*: The fraction of the phylogenetic tree branch lengths shared by the two communities, weighted by the counts of organisms, so more abundant organisms have a greater influence.

Calculating Rarefied Beta Diversity `avgdist` function computes the dissimilarity matrix of a dataset multiple times using `vegdist` while randomly subsampling the dataset each time. All of the subsampled iterations are then averaged (mean) to provide a distance matrix that represents the average of multiple subsampling iterations.

Non-metric Multi-dimensional Scaling (NMDS) is a way to condense information from multidimensional data (multiple variables/species/OTUs), into a 2D representation or ordination. The closer the points/samples are together in the ordination space, the more similar their microbial communities.

- NMDS plots are non-metric, meaning that among other things, they use data that is not required to fit a normal distribution (there are only a few abundant species, and many, many species with low abundance).
- What makes an NMDS plot non-metric is that it is rank-based. This means that instead of using the actual values to calculate distances, it uses ranks.

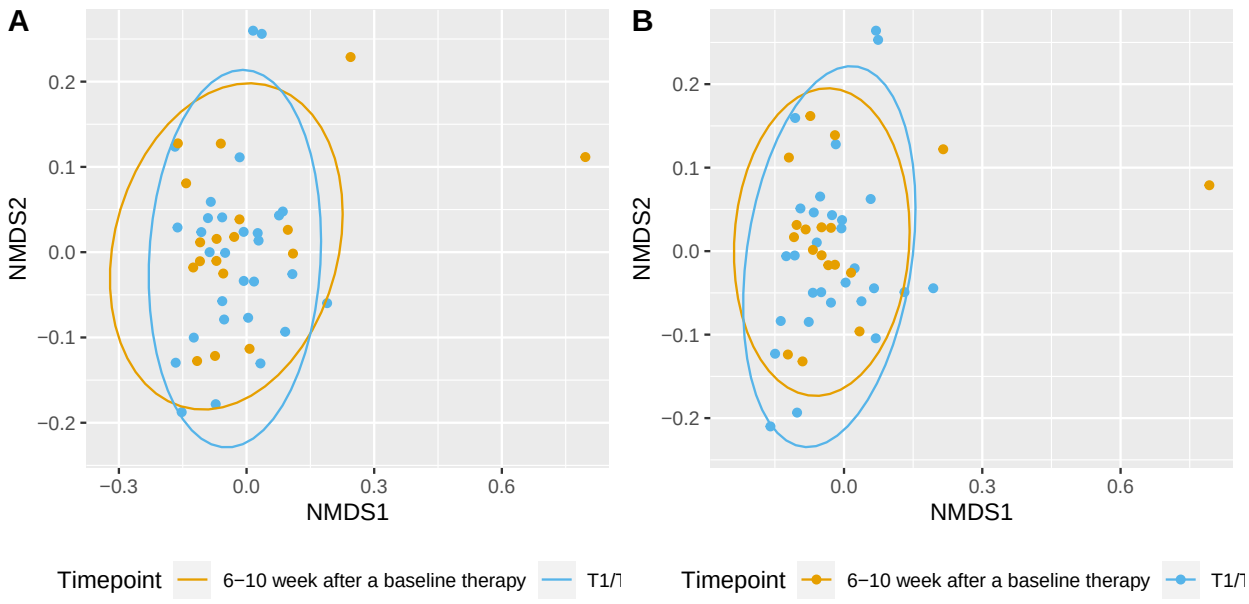


Figure 5. Comparison of beta diversity measurements between non-rarefied (A) and rarefied (B) reads.

Comparison of Bray-Curtis measurements by using non-rarefied and rarefied reads showed no significant difference between the two sets of measurements. This suggests that the rarefication process did not have a significant impact on the beta diversity calculated using the Bray-Curtis metric.

Biological biases in taxonomic composition

Are there systematic biases between detected gram-positive/ gram-negative bacterial ratio?

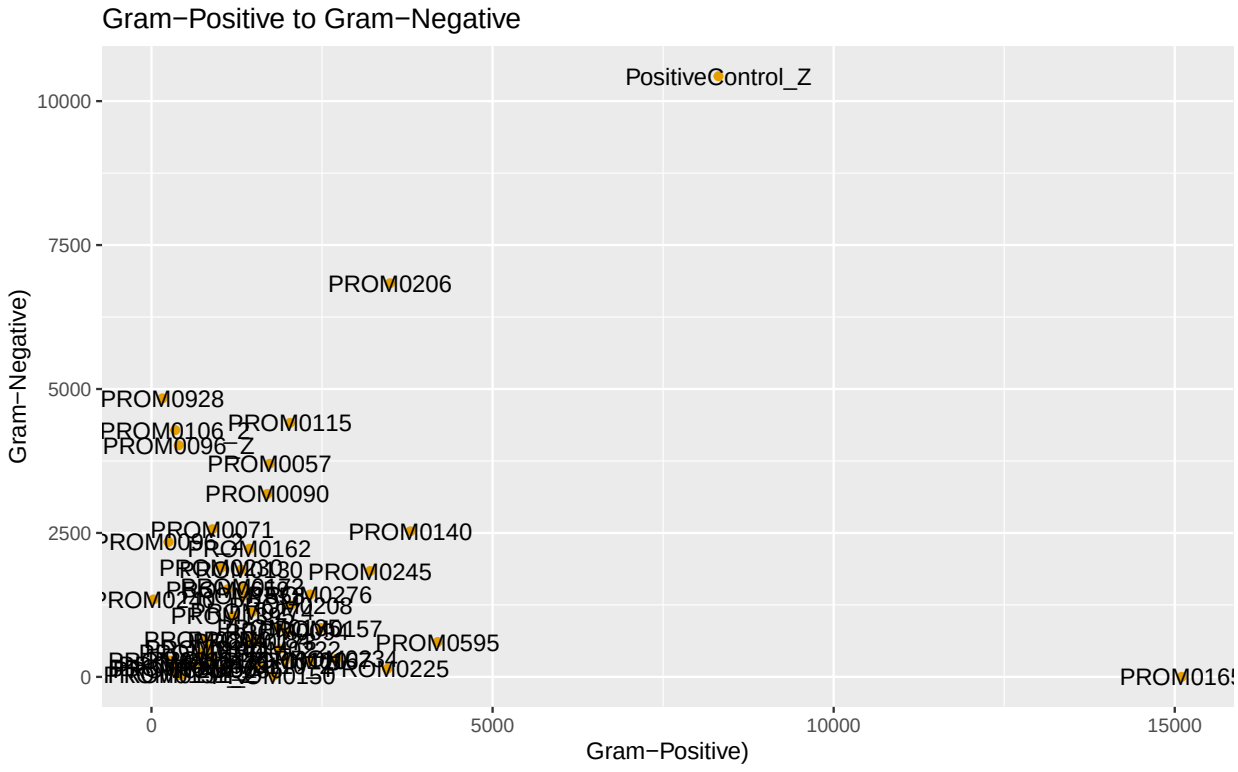


Figure 6. Comparison of most abundant gram positive and gram negative phylum

Majority of the samples in the pilot show a skew towards gram-negative bacteria. It is important to carefully examine any outliers, such as the absence of gram-negative bacteria in sample 165. This could be due to a variety of factors, including errors in the sampling or analysis process, or it could indicate the presence of unique characteristics or conditions in that sample that are not present in the others. We should carefully check the detailed metadata to understand the underlying reasons.

Conclusion and next steps

Overall, the dataset appears to be in good condition and we will be able to move forward with our analysis as soon as we receive the metadata.