# The Gut Microbiome Strains: Non-Small Cell Lung Cancer Treatment: Deciphering the Connection

By

Muhammad Faheem Raziq

(Registration No: 00000402314)

Department of Sciences

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(July 2024)

# The Gut Microbiome Strains: Non-Small Cell Lung Cancer Treatment: Deciphering the Connection

By

Muhammad Faheem Raziq

(Registration No: 00000402314)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in

Bioinformatics

Supervisor: Dr. Masood Ur Rehman Kayani

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(July 2024)

# THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS Thesis written by Mr. Muhammad Faheem Raziq (Registration No. 00000402314), of School of Interdisciplinary Engineering & Sciences (SINES) (School/College/Institute) has been vetted by undersigned, found complete in all respects as per NUST Statutes/ Regulations/ Masters Policy, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of master's degree. It is further certified that necessary amendments as pointed out by GEC members and foreign/ local evaluators of the scholar have also been incorporated in the said thesis.

Signature: ______________________

Name of Supervisor Dr. Masood Ur Rehman Kayani

Date: 25/07/24

Signature (HOD): ______________________

Date: 29-7-2024

Dr. Fouzia Malik
HoD Sciences
Professor
SINES NUST, Sector H 12
Islamabad

Signature (Dean/ Principal) ______________________

Dr. SYED IRTIZA ALI SHAH
Principal & Dean
SINES - NUST, Sector H-12
Islamabad

Assistant Professor
SINES - NUST, Sector H-12
Islamabad

Date: 29/7/24

# AUTHOR'S DECLARATION

I Muhammad Faheem Raziq hereby state that my MS thesis titled "**The Gut Microbiome strains: Non-Small Cell Lung Cancer Treatment: Deciphering the Connection**" is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Muhammad Faheem Raziq

Date: 31/07/2024

# PLAGIARISM UNDERTAKING

I solemnly declare that the research work presented in the thesis titled "**The Gut Microbiome Strains: Non-Small Cell Lung Cancer Treatment: Deciphering the Connection**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and Muhammad Faheem Raziq has written that complete thesis under the supervision of **Dr. Masood Ur Rehman Kayani.**

I understand the zero-tolerance policy of the HEC and National University of Sciences and Technology (NUST), Islamabad towards plagiarism. Therefore, as an author of the above-titled thesis, I declare that no portion of my thesis has been plagiarized and any material used as a reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above-titled thesis even after the award of the MS degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and NUST, Islamabad has the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.

Student Signature: _______________

Name: Muhammad Faheem Raziq

# DEDICATION

I dedicate my thesis to my dear father, ***Muhammad Qayyum Raziq***, with deep respect and enduring affection. The consistent support, limitless compassion, and enduring knowledge he has provided have been the guiding force throughout my academic career.

# ACKNOWLEDGEMENTS

All praise is for **Almighty Allah**, the ultimate source of all knowledge. By His grace, I have reached this stage of knowledge with the ability to contribute something beneficial to His creation. My deepest respects are to the **Holy Prophet Hazrat Muhammad (PBUH)**, the symbol of guidance and fountain of knowledge.

I would like to express my sincere gratitude to my supervisor, **Dr. Masood Ur Rehman Kayani**, for his interest, invaluable guidance, encouragement, and continuous support throughout my research journey. I deeply appreciate his thought-provoking discussions, sound advice, and valuable suggestions, which have greatly enhanced my ability to tackle problems meaningfully and pursue my objectives eagerly.

I also extend my thanks to my GEC committee members, **Dr. Zamir Hussain** and **Dr. Mehak Rafiq**, for their guidance, valuable feedback, and suggestions that have helped refine my thesis. Additionally, I am grateful to all the faculty members of SINES for their assistance at various stages of this research.

My appreciation goes to my colleagues in the **Microbiome Research Group**, especially **Haseeb Manzoor**, **Umme Habiba**, **Muhammad Shahzaib**, and **Adnan Tariq**, for their continuous help and feedback throughout the research. I am also thankful to my friends, **Arisha Wasim** and **Muqadas Chaudhary**, for their determined support. Finally, I am deeply grateful to my parents and siblings for their support throughout this journey.

# TABLE OF CONTENTS

# LIST OF TABLES

**Page No.**

# LIST OF FIGURES

**Page No.**

# LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| NSCLC | Non-Small Cell Lung Cancer |
| SCLC | Small Cell Lung Cancer |
| ICIs | Immune Checkpoint Inhibitors |
| PD-1 | Programmed Cell Death Protein 1 |
| PD-L1 | Programmed Death-Ligand 1 |
| R | Responders |
| NR | Non-Responders |
| RECIST | Response Evaluation Criteria in Solid Tumors |
| CT | Computer Tomography |
| PFS | Progression-Free Survival |
| IHMS | International Human Microbiome Standards |
| DNA | Deoxyribonucleic Acid |
| SRA | Sequence Read Archive |
| ENA | European Nucleotide Archive |
| NCBI | National Center for Biotechnology Information |
| DDBJ | DNA Data Bank of Japan |
| TKI | Tyrosine Kinase Inhibitors |
| GO | Gene Ontology |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| STRING | Search Tool for the Retrieval of Interacting Genes/Proteins |
| FDR | False Discovery Rate |
| EGA | European Genome-phenome Archive |
| SNP | Single Nucleotide Polymorphism |

| | |
|---|---|
| MNP | Multiple Nucleotide Polymorphisms |
| INS | Insertions |
| DEL | Deletions |
| PCA | Principal Component Analysis |
| AUC | Area Under the Curve |
| SVM | Support Vector Machine |
| ROC | Receiver Operating Characteristic |
| RCC | Renal Cell Carcinoma |
| MSA | Multiple Sequence Alignment |

# ABSTRACT

Non-small cell lung cancer (NSCLC) is the second most frequently diagnosed cancer worldwide and the leading cause of cancer-related mortality, with approximately 1.8 million reported deaths in 2020. NSCLC treatment includes surgery, chemotherapy, radiation, and immunotherapy, with Immune Checkpoint Inhibitors (ICIs) such as PD-1/PD-L1 inhibitors revolutionizing patient outcomes. However, treatment response varies significantly among patients, presenting a substantial challenge. Emerging evidence suggests that the gut microbiome profoundly influences the efficacy of cancer therapies, including ICIs. This research investigates the role of gut microbial species, strains, and genetic variants in modulating NSCLC treatment response. Utilizing metagenomic analysis, taxonomic profiling was conducted to identify microbial species such as *B. uniformis*, *F. prausnitzii*, and *A. muciniphila* present in NSCLC patients' gut microbiomes at various time points and response categories. Strain diversity profiling revealed specific strains consistently present across all time points, including strains of *B. uniformis* and *F. prausnitzii*, while others, such as *L. eligens* and *E. coli*, were unique to patient responses. Variant calling identified 35,615 genetic variations in responders and 47,969 in non-responders, including SNPs, indels, and complex mutations. Notably, NR exhibited a higher number of genetic variations, highlighting potential microbial markers for treatment efficacy. Specific genes, including *ftsA*, *lpdA*, and *sufD*, were associated with treatment response, providing insights into the functional attributes of these variations. Further, gene ontology analysis categorized these genetic variants into biological processes, cellular components, and molecular functions, underscoring the role of microbial genes in influencing treatment outcomes. Machine learning models showed an AUC of 85%, indicating the predictive capabilities for treatment response based on gut microbiome composition.

Our findings emphasize the potential of integrating gut microbiome analysis with NSCLC treatment strategies to enhance the efficacy of immunotherapy. By deciphering the connection between gut microbiome and NSCLC treatment responses, this study may highlight the need for developing microbiome-based interventions to optimize cancer therapy outcomes.

**Keywords:** NSCLC, Gut microbiome, Immunotherapy, Immune checkpoint inhibitors (ICIs), PD-1/PD-L1, Microbial strains, Genetic variants.

# CHAPTER 1: INTRODUCTION

Cancer is characterized by uncontrolled cell growth and the potential to metastasize to different body sites (Brown et al. 2019). This usually occurs due to the overexpression of certain genes, known as oncogenes, or the suppression of protective genes, referred to as tumor suppressor genes (Sinkala 2023). Various factors such as age, gender, race, environment, diet, and genetics can influence the occurrence and type of cancer (Seke Etet et al. 2023). As cancer cells proliferate, they often form clusters known as tumors. Tumors can be benign, remaining in one location without invading nearby tissues, or they can be malignant, spreading and invading surrounding tissues (Boutry et al. 2022). Cancers are categorized based on the type of fluid or tissue they originate from or their initial location in the body, such as breast cancer, prostate cancer, liver cancer, lung cancer, etc. (Rahman et al. 2022).

## 1.1 Understanding Lung Cancer

Lung cancer ranks as the second most diagnosed cancer and is the leading cause of cancer-related deaths globally (Sung et al. 2021). In 2020, ~2.2 million new cases of lung cancer were identified, making it the second most common cancer after breast cancer. Lung cancer also had the highest mortality rate, with approximately 1.8 million deaths, mainly due to late detection (Restrepo et al., 2023). The high mortality rate of lung cancer is often due to late diagnoses, with the disease frequently detected at an advanced stage. Effective management of lung cancer depends on a thorough understanding of cancer development, along with efficient early detection methods and suitable pharmaceutical treatments

(Bertolaccini et al. 2022). Early detection is especially crucial when screening high-risk individuals, such as smokers or those exposed to hazardous environments like fumes, oil fields, or toxic workplaces. The discovery of novel biomarkers is also essential. It is vital to accurately identify and understand each lung cancer patient's specific diagnosis (Nooreldeen & Bach 2021). Often, lung cancer is diagnosed at an advanced stage, with metastasis to other sites such as the brain (Souza et al. 2023), as shown in Figure 1.1. This advanced stage makes targeted therapy and conventional treatments less effective (Restrepo et al. 2023).

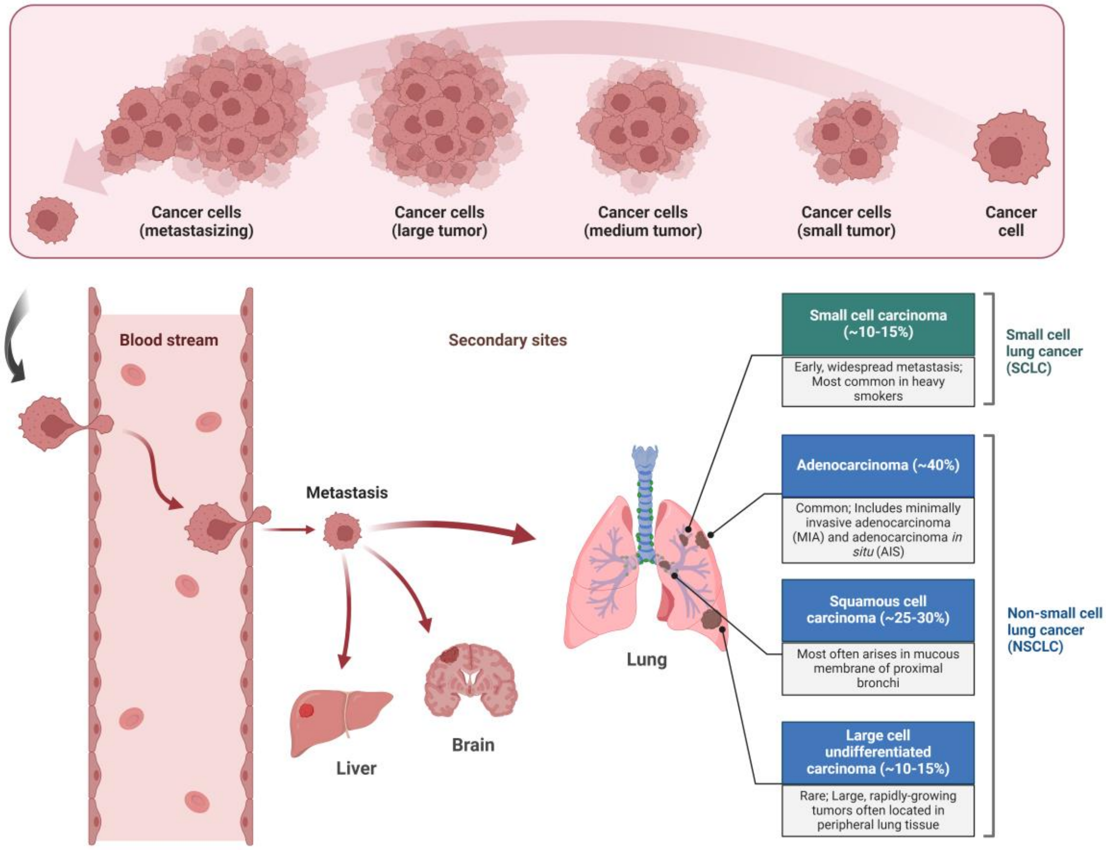

Cancer cells (metastasizing)
Cancer cells (large tumor)
Cancer cells (medium tumor)
Cancer cells (small tumor)
Cancer cell
Blood stream
Secondary sites
Metastasis
Liver
Brain
Lung
Small cell carcinoma (~10-15%)
Early, widespread metastasis; Most common in heavy smokers
Small cell lung cancer (SCLC)
Adenocarcinoma (~40%)
Common; Includes minimally invasive adenocarcinoma (MIA) and adenocarcinoma *in situ* (AIS)
Squamous cell carcinoma (~25-30%)
Most often arises in mucous membrane of proximal bronchi
Large cell undifferentiated carcinoma (~10-15%)
Rare; Large, rapidly-growing tumors often located in peripheral lung tissue
Non-small cell lung cancer (NSCLC)

**Figure 1.1:** Cancer cells and their metastasis via the bloodstream to different body sites

Lung cancers are generally classified into two main histological types: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC) (Howlader et al. 2020). SCLCs are aggressive lung malignancies often associated with smoking, comprising about 15-20% of all primary lung cancer cases. NSCLC, the more prevalent histological subtype, accounts for approximately 85% of all lung cancer cases. NSCLC is often detected at an advanced local stage in about 30% of new cases, presenting a variety of clinical situations with different therapeutic options (Petrella et al. 2023). NSCLC can be further classified into four distinct subtypes: Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Large-Cell Carcinoma, and Bronchial Carcinoid Tumor. LUAD is the most common form of NSCLC and the most frequently occurring primary lung tumor (Nooreldeen & Bach 2021).

### 1.2 Non-Small Cell Lung Cancer (NSCLC)

NSCLC is the leading cause of cancer-related deaths worldwide, resulting in nearly 1.8 million deaths annually (Ibodeng et al., 2023). Early detection of NSCLC and the utilization of diagnostic methods like PET scans and biomarkers are crucial for enhancing patient outcomes and lowering mortality rates (Thakur et al., 2020). NSCLC includes several subtypes such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma, and is often diagnosed at advanced stages, which complicates treatment (Nair et al., 2023). Adenocarcinoma is the most common subtype, accounting for about 40% of cases. It originates from type II alveolar cells that produce mucus and other substances and can affect smokers and non-smokers of all ages. This cancer typically grows more slowly and is often located in the outer regions of the lungs, possibly due to cigarette filters blocking larger particles. Compared to other NSCLC subtypes, adenocarcinoma is more likely to be detected before spreading beyond the lungs (Seguin et al., 2022). Conversely, large cell carcinoma, which lacks specific squamous or glandular features, represents 5-10% of lung cancer cases. It is typically diagnosed by excluding other possibilities and usually originates in the central part of the lungs, with the potential to spread to nearby lymph nodes, the chest wall, and distant organs (Suster & Mino-Kenudson, 2020).

## 1.3 Treatments of NSCLC

The treatment of NSCLC includes various methods such as radiation therapy, surgical intervention, systemic modalities like chemotherapy, targeted molecular therapies, hormone-based regimens, and immunotherapy as illustrated in Figure 1.2 (Alduais et al., 2023). For NSCLC, data indicates that approximately 56% of individuals with early-stage (I and II) disease opt for surgery as their treatment. In contrast, most stage III NSCLC patients (62%) undergo chemotherapy or radiotherapy (Lampridis & Scarci, 2023).

Patients diagnosed with stages I, II, and IIIA of NSCLC typically undergo surgery to remove the tumor if it is operable and the patient can withstand the procedure. Post-surgery, some patients may benefit from additional therapy known as adjuvant therapy, aimed at reducing the risk of cancer recurrence. Adjuvant therapy can include radiation, chemotherapy, and targeted therapy. For patients with various advanced stages of NSCLC, chemotherapy is often administered post-surgery to eliminate any remaining cancer cells and improve survival chances (Lim & Yeo, 2022).

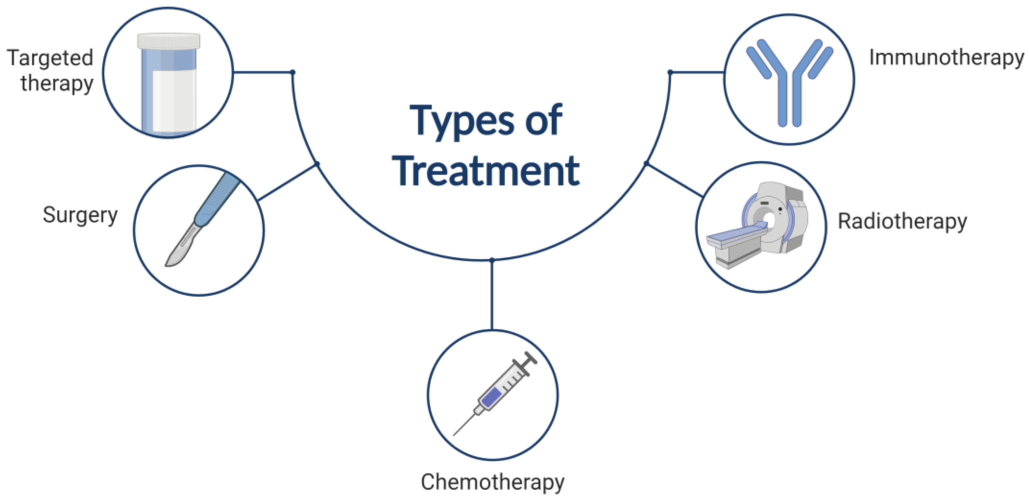


**Figure 1.2:** Different types of NSCLC treatment

Radiation therapy uses powerful beams of energy to damage the DNA in cancer cells, effectively killing them. This approach is effective in managing or eradicating tumors located in specific parts of the body. Patients with chest-localized NSCLC who are not suitable for surgery may benefit from this treatment. Additionally, radiation therapy can be used in palliative care to enhance the quality of life for NSCLC patients who do not respond to surgery or chemotherapy (Alduais et al., 2023). Immunotherapy, a groundbreaking cancer treatment, utilizes the body’s natural defense mechanisms to fight cancer. Some cancer cells closely resemble healthy cells, making it difficult for the immune system to differentiate between them. Immunotherapy works by boosting the immune system's ability to target cancer cells, slow their growth, prevent their spread, or increase its overall effectiveness in combating cancer (Mamdani et al., 2022).

### *1.3.1 Immunotherapy for NSCLC*

Immunotherapies work by removing the constraints on the immune system, exposing the tumor, and enhancing the recognition of tumor-associated neoantigens. This action stimulates an immune response that leads to tumor suppression. This cutting-edge approach empowers the immune system of the host to respond effectively, regardless of the tumor-specific histology or underlying driver mutations. Several strategies have emerged within cancer immunotherapies, focusing on boosting effector mechanisms and reducing inhibitory and suppressive pathways (Yao et al., 2023), as demonstrated in Figure 1.3.

One such strategy involves neutralizing suppressive mechanisms using antibodies against immune checkpoint proteins. Tumors often exploit immune checkpoints to avoid immune detection (Marei et al., 2023). To counteract this, immune checkpoint inhibitors (ICIs) are used therapeutically. They stimulate immune responses against tumor cells within the tumor microenvironment (TME), which includes various immune cell populations and the extracellular matrix (ECM) intricately linked with tumor cells (Shiravand et al., 2022). Significant progress has been made with agents such as pembrolizumab and nivolumab, both inhibitors of the programmed death-1 (PD-1) pathway, and atezolizumab, an inhibitor of its primary ligand, programmed death ligand-1 (PD-L1). These inhibitors have shown superior responses compared to conventional chemotherapy, leading to their endorsement as second-line treatments for patients with metastatic NSCLC (Punekar et al., 2022).

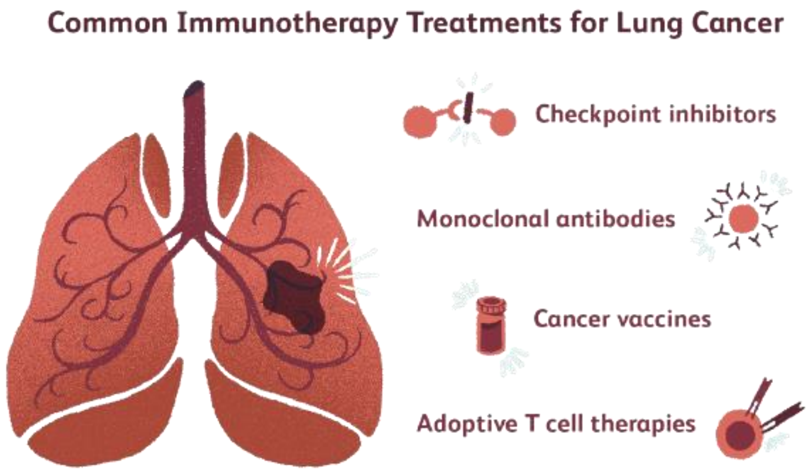

Common Immunotherapy Treatments for Lung Cancer
Checkpoint inhibitors
Monoclonal antibodies
Cancer vaccines
Adoptive T cell therapies

**Figure 1.3:** The common immunotherapy treatments for lung cancer

### *1.3.2 Immune Checkpoint Inhibitors*

The effectiveness of ICIs has significantly expanded the options for cancer treatment. Immune checkpoints are molecules on cell membranes that regulate T-cell responses to prevent overactivation. Unfortunately, cancer cells exploit this system to evade immune detection. ICIs can reactivate previously ineffective T-cells, restoring their ability to respond to tumor-related substances (Naimi et al., 2022). Lung cancer immunotherapy has recently gathered considerable attention for its role in enabling the immune system to detect and eliminate cancer cells. A pivotal milestone in immunotherapy was the discovery of immune checkpoints (ICPs), proteins produced by certain immune cells like T-cells and by cancer cells themselves (Starzer et al., 2022). Under normal conditions, these checkpoints engage with their partner proteins through receptor-ligand interactions, sending inhibitory signals that deactivate T-cell responses to prevent unintended attacks on healthy cells. These checkpoints are crucial for maintaining selftolerance, regulating the immune system, and ensuring overall immune balance (Dutta et al., 2023).

Tumor cells exploit this regulatory mechanism by using ICP proteins to evade destruction by immune cells. Targeting these immune checkpoints with checkpoint inhibitors (CKIs) has shown potential for achieving sustained clinical responses and even curative outcomes in cancer treatment (Marei et al., 2023). From the initial discovery of CTLA-4, various immune checkpoints, including PD-1, have been identified. The interaction between PD-1 on effector T-cells and PD-L1 on tumor cells and myeloid cells within the tumor microenvironment acts as an inhibitory signal, leading to effector T-cell exhaustion (Yi et al., 2022). Similarly, CTLA-4, upregulated in activated T-cells, competes with co-stimulatory molecules CD80/86 on antigen-presenting cells (APCs), dampening T-cell activation and function. While PD-1 and CTLA-4 are the most extensively studied immune checkpoint proteins, other immune checkpoint proteins also hold therapeutic potential (Goleva et al., 2021).

In 2015, the United States Food and Drug Administration (FDA) approved nivolumab for advanced LUSC, later extending its use to all histological types of NSCLC following the failure of initial platinum doublet chemotherapy (Choi & Chang, 2023). Antibodies targeting the ICI mechanism protect tumor cells from immune attacks. In particular, the inhibition of immune checkpoint proteins through the blockade of CTLA-4, PD-1, and PD-L1 has proven especially effective as an immunotherapeutic strategy for NSCLC (Tang et al., 2022). Antibodies targeting the PD-1 protein have shown significant therapeutic potential in NSCLC by counteracting the suppression of T-cell functions.

### *1.3.3 Anti-PD-1/PD-L1 Therapy*

PD-1 and its ligands, PD-L1 and PD-L2, are crucial ICP proteins. Their primary function is to prevent T-cell effector activity in peripheral tissues during inflammatory responses, thus preventing autoimmunity. However, in the tumor microenvironment, these proteins facilitate tumor suppression of the immune response (Waldman et al., 2020).

Over recent decades, immunotherapy has been a focus for treating NSCLC (Dantoing et al., 2021). Data indicates that cancer often arises when the immune system malfunctions. Proteins like PD-1 and PD-L1, which usually help maintain immune balance, instead help tumors evade the immune system in cancer (Davies, 2019). Blocking PD-1 and PD-L1 can enhance the immune system's ability to combat cancer. PD-1 is a receptor on immune cells, while PD-L1 is a ligand on cancer cells. When PD-1 on immune cells binds to PD-L1 on cancer cells, it prevents the immune cells from attacking, allowing cancer to proliferate unhindered. Consequently, scientists have developed drugs that block the PD-1/PD-L1 interaction, enabling immune cells to target and destroy cancer cells more effectively (Lin, 2023). Figure 1.4 shows the mechanism of action of PD-1/PD-L1.

## **1.4 The Human Gut Microbiome**

The human microbiome is a diverse community of microorganisms including bacteria, archaea, viruses, and other microbes that inhabit our bodies both externally and internally. These microorganisms have the potential to significantly affect our bodily functions, influencing our health and disease states (Xia et al., 2023). They contribute to various aspects of our metabolism, protect us from harmful pathogens, guide our immune system, and consequently affect nearly all body functions, either directly or indirectly

(Colella et al., 2023). To understand the impact of gut microbiome on health and disease, it is essential to first study the microorganisms present in healthy individuals. Healthy adults host over a thousand different bacterial species, with Bacteroidetes and Firmicutes being the dominant groups. The gut has an exceptionally diverse microbial population, though the exact composition can vary widely among individuals (Hou et al., 2022).

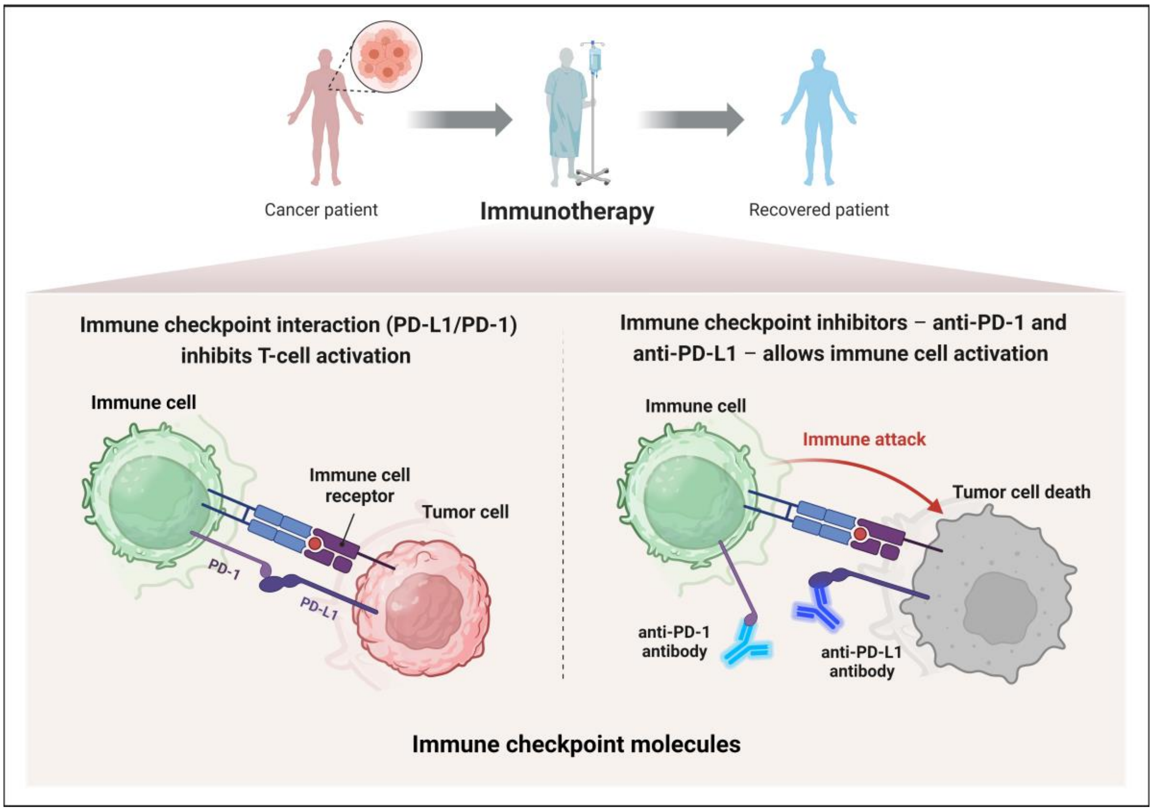

Cancer patient
Immunotherapy
Recovered patient
Immune checkpoint interaction (PD-L1/PD-1) inhibits T-cell activation
Immune cell
Immune cell receptor
Tumor cell
PD-1
PD-L1
Immune checkpoint inhibitors – anti-PD-1 and anti-PD-L1 – allows immune cell activation
Immune cell
Immune attack
Tumor cell death
anti-PD-1 antibody
anti-PD-L1 antibody
Immune checkpoint molecules

**Figure 1.4:** Mechanism of Immunotherapy particularly immune checkpoint inhibitors (PD-1/PD-L1)

The relationship between the host immune system and the gut microbiome is complex, bidirectional, and extensive such as the gut-lung axis, shown in Figure 1.5. The immune system must tolerate harmless microbiota while effectively responding to harmful pathogens. Conversely, the gut microbiome plays a crucial role in developing the immune system to function properly (Yoo et al., 2020). There is significant interest in studying how changes in the gut microbiome are associated with disease. However, it is often unclear whether these changes are a cause or a consequence of disease (Yoo et al., 2020). Diseases can alter the gut microbiome due to various factors such as diet changes, gastrointestinal function alterations, and medication use like antibiotics (Zheng et al., 2020).

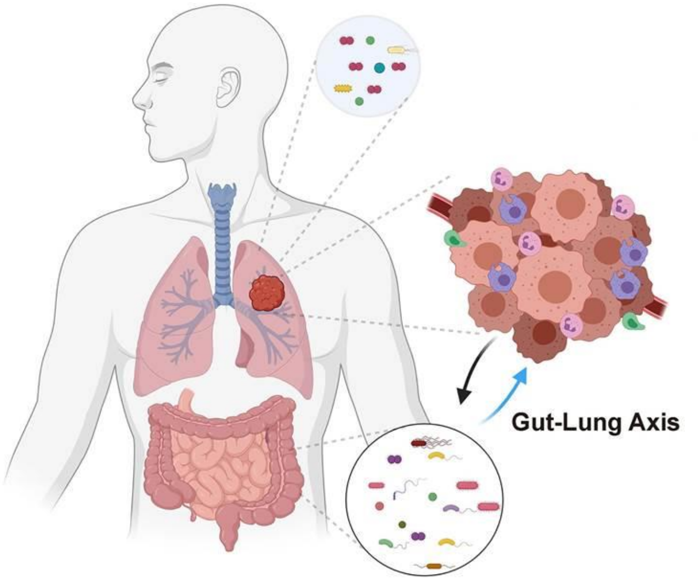


**Figure 1.5:** The gut microbiome can influence the treatment response and treatment can influence the gut microbiome composition

The gut microbiome has been linked to the onset and progression of various cancers, affecting both the epithelial barrier and sterile tissues (El Tekle et al., 2023). The gut microbiome can directly cause cancer by producing harmful metabolites, such as lithocholic acid (LCA) (Yang et al., 2023), or substances with carcinogenic properties like *H. pylori*, classified as a class I carcinogen by the International Agency for Research on Cancer (IARC) (Garg et al., 2023). It can also promote cancer indirectly by causing inflammation, as seen with Campylobacter species (Xia et al., 2023). Emerging evidence shows bacteria can enhance the body's immune response against distant tumors (Jain et al., 2021). Antibiotic use is linked to cancer risk and is influenced by dosage (Simin et al., 2020). The effectiveness of some therapies can be reduced due to the absence or alteration of the gut microbiome. The role of the gut microbiome in enhancing the immune response to cancer treatment varies with the treatment method (Sadrekarimi et al., 2022).

### *1.4.1 Gut Microbiome and Immunotherapy*

Increasing evidence suggests the gut microbiome significantly affects responses to immunotherapy (Shi et al., 2023). Studies on patients undergoing immunotherapy, especially ICIs, show that disruptions in the gut microbiome composition and function are linked to immune-related disorders like inflammatory bowel disease, autoimmune diseases, chronic inflammation, and cancer. Recent research highlights the correlation between the gut microbiome and the effectiveness and side effects of ICI-based immunotherapy (Lu et al., 2022).

A pivotal preclinical study by Sivan et al. demonstrated the interplay between specific gut commensals, such as Bifidobacterium, and outcomes like reduced tumor growth, increased T-cell infiltration into tumors, and enhanced anti-tumor immune responses, supporting PD-L1 blockade effectiveness. Following this, many studies have explored the connection between microbial signatures and ICI treatment responses. Interventional studies aim to manipulate the gut microbiome to enhance positive outcomes and reduce adverse events in patients with solid tumors undergoing ICI therapies (Yi et al., 2018).

Gopalakrishnan et al. examined the link between gut bacteria and anti-PD-1 immunotherapy response in melanoma patients, dividing them into R and NR. R had a richer diversity of gut bacteria compared to NR, suggesting the role of microbial diversity in treatment efficacy. Specific bacteria, like *Faecalibacterium,* were abundant in R and linked to longer progression-free survival, while NR had bacteria like *B. thetaiotaomicron*, *E. coli,* and *A. colihominis*. Increased *Faecalibacterium* abundance correlated with better responses and longer survival in R, indicating the crucial role of gut bacteria composition and diversity in melanoma patients' response to anti-PD-1 therapy (Gopalakrishnan et al., 2018).

A meta-analysis of four shotgun metagenomic studies on microbiome composition between R and NR to immunotherapy revealed that *Faecalibacterium* was common in responders. Additionally, *B. intestinihominis* was more abundant in responders (Limeta et al., 2020). Maia et al. demonstrated a correlation between microbiome composition and response to nivolumab or nivolumab plus ipilimumab in patients with metastatic Renal Cell Carcinoma (RCC). R had higher alpha diversity and more *Roseburia* and *Faecalibacterium* species than NR. There were also reports of a temporal increase in *A. muciniphila* (Maia et al., 2018).

Routy et al. conducted a shotgun metagenomic study to determine gut composition differences between R and NR to PD-1 inhibition in patients with RCC and NSCLC. They found a high abundance of *A. muciniphila* in R patients. Fecal Microbiota Transplantation (FMT) in germ-free mice validated that only stool from R patients enhanced the anticancer effects of PD-1 inhibitors. Moreover, oral supplementation with *A. muciniphila,* alone or with *E. hirae*, restored anticancer effects in antibiotic-treated mice (Routy et al., 2018). Jin et al. reported that NSCLC patients with high gut microbiome α-diversity, enriched in *B. longum*, *A. putredinis*, and *P. copri*, had significantly longer Progression-Free Survival (PFS) compared to those with low diversity and abundant *Ruminococcus*. This suggests that gut microbiome diversity influences immunotherapy response by enhancing antitumor immunity (Mao et al., 2021; Jin et al., 2019).

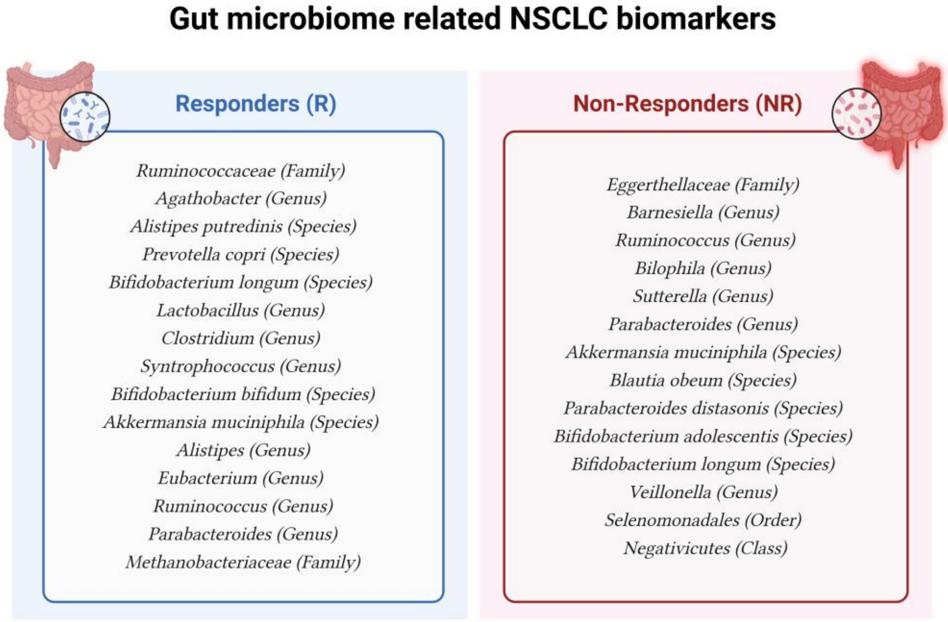
Hakozaki et al. found that in 70 Japanese NSCLC patients treated with anti-PD-1/PD-L1 antibodies, Ruminococcaceae were linked with favorable prognosis, likely due to high colon IFNγ production from CD8+ T-cells. In contrast, butyrate-producing *Agathobacter* was linked with poor prognosis, while *Eggerthellaceae* and *Barnesiella*, promoting IFN-γ-producing γδ T-cells in cancer lesions, were associated with NR (Tanoue et al., 2019; Hakozaki et al., 2020). Jin et al. also found that high-diversity microbiomes in Chinese NSCLC patients correlated with extended PFS, with significant differences in gut microbiome composition between R and NR patients. R patients were enriched in *A. putredinis*, *B. longum*, and *P. copri*, whereas NR patients had more *Ruminococcus_unclassified* (Jin et al., 2019; Abdelsalam et al., 2023).

Katayama et al. analyzed fecal samples from 17 Japanese NSCLC patients treated with ICIs and found that R patients had more *Lactobacillus* and *Clostridium*, which stimulate T-cell mobilization to tumors, correlating with longer time to treatment failure (TTF) (Katayama et al., 2019). Lee et al. sequenced stool samples from five cohorts of ICI-naive advanced melanoma patients, identifying species like *B. pseudocatenulatum* linked to response, highlighting a cohort-dependent association between the gut microbiome and ICI response (S.-H. Lee et al., 2021). The taxonomic composition associated with favorable vs. unfavorable responses to ICI therapy is summarized in Table 1.1.

**Table 1.1:** Major gut microbial taxa associated with response to NSCLC treatment

| **Source** | **Major Taxa associated with favorable response** | **Major Taxa associated with an unfavorable response** |
|---|---|---|
| Hakozaki et al., 2020 | Ruminococcaceae | Eggerthellaceae |
| | *Agathobacter* | *Barnesiella* |
| Jin et al., 2019 | *Alistipes putredinis* | *Ruminococcus* |
| | *Prevotella copri* | |
| | *Bifidobacterium longum* | |
| Katayama et al., 2019 | *Lactobacillus* | *Bilophila* |
| | *Clostridium* | *Sutterella* |
| | *Syntrophococcus* | *Parabacteroides* |
| Lee et al., 2021 | *Bifidobacterium bifidum* | *Akkermansia muciniphila* |
| | | *Blautia obeum* |
| Routy et al., 2018 | *Akkermansia muciniphila* | *Parabacteroides distasonis* |
| | *Alistipes* | *Bifidobacterium adolescentis* |
| | *Eubacterium* | *Bifidobacterium longum* |
| | *Ruminococcus* | |
| Song et al., 2020 | *Parabacteroides* | *Veillonella* |
| | Methanobacteriaceae | Selenomonadales |
| | | Negativicutes |

Song et al. analyzed samples from 63 advanced NSCLC patients on PD-1 inhibitors, finding higher β-diversity in patients with PFS ≥ six months. These patients were rich in *Parabacteroides* and *Methanobrevibacter*, while those with PFS < six months had more *Veillonella*, Selenomonadales (modulating tumor cell properties and DNA processes), and Negativicutes (inducing Treg cells and IL-10 to suppress immune responses) (Song et al., 2020; Ehudin et al., 2022; W. Y. Cheng et al., 2020). The major gut microbiome-related biomarkers of R and NR for the treatment of NSCLC are shown in Figure 1.6.



Gut microbiome related NSCLC biomarkers

Responders (R)

*Ruminococcaceae (Family)*
*Agathobacter (Genus)*
*Alistipes putredinis (Species)*
*Prevotella copri (Species)*
*Bifidobacterium longum (Species)*
*Lactobacillus (Genus)*
*Clostridium (Genus)*
*Syntrophococcus (Genus)*
*Bifidobacterium bifidum (Species)*
*Akkermansia muciniphila (Species)*
*Alistipes (Genus)*
*Eubacterium (Genus)*
*Ruminococcus (Genus)*
*Parabacteroides (Genus)*
*Methanobacteriaceae (Family)*

Non-Responders (NR)

*Eggerthellaceae (Family)*
*Barnesiella (Genus)*
*Ruminococcus (Genus)*
*Bilophila (Genus)*
*Sutterella (Genus)*
*Parabacteroides (Genus)*
*Akkermansia muciniphila (Species)*
*Blautia obeum (Species)*
*Parabacteroides distasonis (Species)*
*Bifidobacterium adolescentis (Species)*
*Bifidobacterium longum (Species)*
*Veillonella (Genus)*
*Selenomonadales (Order)*
*Negativicutes (Class)*

**Figure 1.6:** The gut microbial biomarkers associated with NSCLC treatment responses

### *1.4.2 Gut Microbiome and Microbial Strains*

The human gut microbiome hosts a diverse range of microbial strains, each uniquely contributing to the host's health and disease states (Bou Zerdan et al., 2022). In NSCLC treatment, specific bacterial strains have been identified that significantly influence therapeutic outcomes. For instance, *B. longum* and *L. rhamnosus* have shown promise in enhancing the efficacy of ICIs (Sun et al., 2023). These probiotics are known to modulate the immune system by promoting the production of beneficial cytokines and enhancing the activity of dendritic cells and T-cells, which are crucial for an effective anti-tumor response.

Moreover, the presence of *A. muciniphila* has been correlated with improved responses to PD-1 blockade therapy in NSCLC patients. This strain is particularly effective in strengthening the gut barrier and reducing systemic inflammation, thereby supporting a more robust immune response against cancer cells (Jin et al., 2019). Studies have indicated that patients with higher levels of *A. muciniphila* in their gut microbiome tend to experience better clinical outcomes with fewer adverse effects during ICI treatment. The beneficial effects of *A. muciniphila* are thought to be mediated through its ability to produce short-chain fatty acids and other metabolites with immunomodulatory properties (Souza et al., 2023).

Another notable strain is *F. prausnitzii*, which has been linked to reduced toxicity and enhanced therapeutic efficacy. *F. prausnitzii* is renowned for its anti-inflammatory properties and ability to produce butyrate, a short-chain fatty acid that serves as a critical energy source for colonocytes and helps maintain gut homeostasis. The presence of *F. prausnitzii* in the gut microbiome is associated with a balanced immune response and a lower risk of treatment-related complications, making it a potential target for microbiome-based interventions aimed at improving NSCLC treatment outcomes (Parsaei et al., 2021).

### *1.4.3 Gut Microbiome and Genetic Variations*

Genetic variations within the human gut microbiome significantly influence the efficacy and toxicity of NSCLC treatments. The composition and genetic diversity of the gut microbiome can impact drug metabolism, immune modulation, and inflammation, all of which are critical in cancer therapy (Liu et al., 2023). Certain bacterial strains in the gut microbiome, for example, can activate or deactivate chemotherapeutic agents, thereby affecting their therapeutic effectiveness. Moreover, variations in the gut microbiome can modulate the host's immune response, influencing the success of immunotherapies, which are increasingly utilized in NSCLC treatment. Strains such as *B. fragilis* and *A. muciniphila* have been shown to enhance immune responses, potentially improving the outcomes of immunotherapies (Liu et al., 2023). Understanding these genetic variations helps in predicting patient responses to treatments and in developing personalized therapeutic strategies.

## 1.5 Research Gap and Problem Statement

In NSCLC, the disease is often diagnosed in advanced stages, making traditional treatments less effective. Addressing, the need to make the available treatment options more effective to control cancer growth and progression to improve patient outcomes.

## 1.6 Objectives

- To identify gut microbiome strains and genetic variations linked to favorable or unfavorable treatment responses in NSCLC.
- To explore potential mechanisms by which gut microbiome strains and genetic variations influence treatment efficacy in NSCLC patients.

# CHAPTER 2: MATERIAL AND METHODS

In this study, bioinformatics analysis was conducted on metagenomic samples from NSCLC patients, following several key steps to derive meaningful insights from the data. Initially, the metagenomic data was retrieved and preprocessed to ensure its quality and integrity. Subsequently, taxonomic profiling was performed to determine the taxonomic composition of the samples. Strain profiling of the most abundant species was then conducted to characterize the microbial strains present. Additionally, genetic variants were identified through variant calling of microbial strains, focusing on the genes containing these variants and exploring their potential functional attributes using gene ontology. To further elucidate the findings, statistical analyses were conducted to identify associations among species, genes, and variants with treatment response and time points. This comprehensive approach enabled a deeper understanding of the microbial community and the functional potential encoded within the metagenomic data from NSCLC patients.

## 2.1 Data Acquisition

The metagenomic data along with metadata was retrieved from a study by Routy et al., 2018b titled "Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors,". The study involved 118 samples from 87 NSCLC patients. The metagenomic shotgun sequencing was retrieved from the European Nucleotide Archive (EMBL-EBI) under the accession number PRJEB22863.

Patients eligible for the study had advanced stage IIIA-IV NSCLC with either squamous or non-squamous histology and had documented recurrence or progression after at least one prior line of treatment. The study also included patients with known ALK or EGFR mutations, who had received prior tyrosine kinase inhibitors (TKI). The treatment involved administering the anti-PD-1 monoclonal antibody, nivolumab, intravenously every two weeks until disease progression or intolerable side effects. Between August 2015 and September 2016, 60 NSCLC patients were enrolled, and an additional 27 patients were enrolled in the validation cohort from October 2016 to April 2017. Tumor response was assessed by the Response Evaluation Criteria in Solid Tumors version 1.1 (RECIST1.1), and Computer tomography (CT) scans were performed at baseline and every 8 to 12 weeks for the first year and then every 12 to 15 weeks until disease progression. Data were collected from a case report form (CRF) at each site and evaluated an objective response and considered R those in complete response, partial response, or stable disease compared to non-responders NR, who either progressed or died. Progression-free survival (PFS) at 3 and 6 months was also defined as an endpoint using RECIST 1.1 criteria. Feces were collected according to International Human Microbiome Standards (IHMS) guidelines (SOP 03 V1) before the first injection (T0) and after the 2nd (T1- 1 month), 4th (T2- 2 months) and 12th (T3- 6 months) injection.

## 2.2 Preprocessing of Sequence Reads

When analyzing metagenomic sequence data, it is essential to process and analyze the sequence data following the acquisition of the raw reads.

### *2.2.1 Quality Control*

The study sourced its data by extracting and sequencing total faecal DNA using ion-proton technology (ThermoFisher), resulting in an average of 22.7±0.9 million single-end short reads, each 150 bases long. Before commencing the analysis, it is essential to assess the quality of the sequencing output. In the sequencing process, certain DNA fragments may undergo more amplification, leading to an overrepresentation of specific sequences. This imbalance can introduce bias into subsequent analyses. Additionally, reads that are excessively short or long may be less dependable or informative. To ensure optimal data quality, it is recommended to exclude reads that fall outside the specified length range. Furthermore, to minimize biases arising from sequencing errors, it is advisable to remove low-quality reads, duplicates, adapters, and host reads during the analysis process.

The SRA Toolkit comprises a set of tools and libraries designed for interacting with the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/Traces/sra) from NCBI, EBI and DDBJ, a public repository housing raw DNA sequencing data. To acquire a dataset containing 118 single-end samples from SRA, the prefetch command was utilized. Subsequently, the fastq-dump command was employed to convert these single-end SRA samples into the FASTQ format, followed by compression using the “gzip” command. Before applying quality control measures to the compressed FASTQ single-end files, an initial analysis was conducted using FastQC, a tool for quality control visualization (accessible at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

The “fastqc” command was executed to generate quality reports for the single-end files. Upon reviewing the summary reports, it became evident that certain parameters needed adjustment to enhance read quality for subsequent analysis. Adjustments were made, addressing parameters such as per base sequence quality, per base sequence content, and removal of persistent sequencing adapters, utilizing the “fastp” command (Chen et al., 2018). These modifications aimed to improve read quality, facilitating further analysis.

In studying microbial communities in the host gut, it is crucial to filter out non-microbial reads present in samples obtained from the host-associated environment. To achieve this, a two-step process was implemented for the data files. Firstly, pre-processing using the Fastp tool was conducted. Subsequently, the BBDuk tool (*BBMap*, 2023) was applied to eliminate any reads originating from the host.

## 2.3 Taxonomic Profiling

MetaPhlAn3 (Beghini et al., 2021, p. 3), also known as Metagenomic Phylogenetic Analysis, is a bioinformatics tool designed to identify the microbial composition of a sample based on marker genes. Its methodology involves estimating the relative abundance of species by aligning reads against a collection of clade-specific marker sequences. These markers are derived from coding sequences that distinguish microbial clades at the species level or higher taxonomic levels. MetaPhlAn3 utilizes bowtie2 to map reads from a given sample to a comprehensive catalog comprising over 1 million markers associated with 13,475 species. In cases where reads belong to clades lacking available genome data, they are categorized as an 'unclassified' subclade of their closest ancestor with available data. The estimation of clade abundances involves normalizing read-based counts by the average genome size of each clade. Notably, MetaPhlAn3 achieves a classification rate of approximately 10,000 reads per second, ensuring robust and high-throughput assessments of metagenomic data at the species level. For a complete taxonomic profiling run, MetaPhlAn3 requires 2.6 GB of memory.

### 2.3.1 *Visualizing the MetaPhlAn3 Output*

MetaPhlAn3 provides a comprehensive report on the microbial makeup of a sample, detailing the relative abundance of various microbial species. The output includes information on clade-specific marker sequences, enabling users to identify and quantify different microbial taxa within the analyzed sample. The output mainly consists of "bowtie2", "sams" and "profiles". The results generated by MetaPhlAn3 are often presented in a tabular format (.tsv files) stored in the "profiles", providing abundance estimates for each identified species or clade. These tables contain the microbial composition of the samples ranging from kingdom to species in the case of bacteria. The "merge" and "grep" command were used to merge the tables of all the samples into one and extract the species tables with relative abundances, respectively. Bar charts are frequently used to display the relative abundance of different microbial taxa, offering a quick overview of the community structure. Heatmaps are another valuable visualization tool, allowing for the representation of abundance patterns across multiple samples. The MetaPhlAn3 also provides a feature known as "hclust2" (https://github.com/SegataLab/hclust2) that is used to visualize the results from MetaPhlAn3 stored in "profiles" in the form of heatmaps which is based on the relative abundance of the species present in the samples. The top 25 species heatmaps were generated for further analysis. Apart from metaphlan3, the GraPhlAn (Asnicar et al., 2015) is a tool which is built for visualization purposes. We used GraPhlAn to generate the cladograms of the MetaPhlAn3 output which showed the overall microbial composition of the samples.

## 2.4 Strain Diversity Profiling

The analysis of microbial community composition in stool samples involved two main steps: initial estimation using MetaPhlAn3 and subsequent strain-level profiling with the StrainPhlAn3 module. MetaPhlAn3 employed a library of species-specific markers spanning bacterial, archaeal, viral, and eukaryotic phylogenies. These markers, selected for strong conservation within each species and minimal sequence similarity with other species genomic regions, were used to estimate microbial composition. The StrainPhlAn3 (Truong et al., 2017) module then utilized reads mapped against the MetaPhlAn3 database to reconstruct consensus sequences for each detected species-specific marker. Filtering operations followed, including removing markers with over 20% ambiguous bases and trimming the first and last 50 bases. Species presence in a sample was determined based on the number of reconstructed markers exceeding 20% of the total available for that species. Different parameters were used in the StrainPhlAn3 execution command include "--mutation_rates", "--tmp", "--marker_in_n_samples 20" and "--sample_with_n_markers 20". Reconstructed marker sequences for each present species underwent alignment using MUSCLE. The resulting multiple sequence alignment (MSA) was processed to filter poorly covered regions, with subsequent concatenation for each species. Strains with gaps in over 20% of the concatenated alignment were excluded.

### *2.4.1 Visualizing the StrainPhlAn3 Output*

To enhance taxonomic resolution, StrainPhlAn3 was applied to microbial clades identified by MetaPhlAn3, reconstructing specific strains within the metagenome. StrainPhlAn3 allowed the analysis of strains with sufficient sequencing depth, employing output such as phylogenetic trees, MSAs, and temporal folders which contain markers FASTA and alignments. Additionally, the phylogenetic trees resulting from StrainPhlAn3 were visualized using iTOL (Letunic & Bork, 2021), a tool used to visualize the phylogenetic trees showing the distances and variation between samples and references for the particular species. On the other hand, the MSAs resulting from StrainPhlAn3 were visualized via NCBI MSA viewer and UGENE through which the variations were identified at the specific positions in the alignment.

### 2.5 Variant Calling

The identification and analysis of genomic variations such as single nucleotide polymorphisms (SNPs), insertions, and deletions are pivotal in understanding bacterial strain diversity and evolution. This step employs SNIPPY (https://github.com/tseemann/snippy), a rapid variant calling tool, for the analysis of genomic variations. The algorithm of SNIPPY aligns the sample reads against a reference genome, a step crucial in establishing a baseline for identifying genomic variants. It then proceeds to detect SNPs and indels, distinguishing true genetic variations from potential sequencing errors. This precise identification is critical for the reliability of the variant calling process. The input of SNIPPY requires samples in fastq.gz format and reference genomes of that species in GenBank format to identify variant effects and products. SNIPPY applies rigorous filters to these identified variants, ensuring that only those with high confidence are considered in subsequent analyses.

The output from SNIPPY includes detailed information about each variant, such as its genomic location, type, and potential biological impact. The files generated as an output include HTML table summary, VCF file, excel table, and alignments in the form of reports. This data is foundational for in-depth genetic analyses and interpretation. For further analysis, we extracted common and uncommon variations throughout the samples for microbial species found to be abundant and have a link with the NSCLC treatment response.

### 2.6 Gene Ontology

Various terms have been used to describe extensive information about gene products and biological data. Gene Ontology (GO) offers consistent terminology and a controlled vocabulary to describe gene products across different databases, making it useful for various biological communities. Attributes of gene products are categorized into three distinct categories: Molecular Function, Biological Process, and Cellular Component, each with unique information about the genes. GO terms are organized as a network of nodes, with connections based on parent-child relationships. These nodes are linked to many other gene and protein databases such as UniProt, GenBank, and EMBL.

ShinyGO v0.80 (Ge et al., 2020) is a graphical tool for gene-set enrichment analysis, compatible with KEGG and STRING. It uses false discovery rate (FDR) adjustments for p-values of enrichment terms. To identify robust pathways enriched with genes from specific microbial species or strains, we selected the list of genes with identified variations from the variant calling step. Enrichment terms from ShinyGO v0.80 with an adjusted p-value <0.01 were included. The top 10 biological processes, cellular components, and molecular functions were visualized using bar plots, phylogenetic trees, and networks, illustrating the presence and enrichment of genes within various attributes. The "Remove Redundancy" option was unchecked to include all the repeated genes in the list provided because different and various number of variations have occurred in the single gene.

## 2.7 Statistical Analysis

### *2.7.1 Alpha Diversity of Species, Genes Counts & Genetic Variants*

Alpha diversity measures the ecological diversity within a specific sample. The diversity indices used to quantify alpha diversity generally increase with the number of species in the sample (richness) and the evenness of their distribution. Various measures of alpha diversity exist, such as the Shannon index. Species richness and evenness are considered two independent characteristics that together contribute to overall diversity. Species richness accounts for the number of different species in the sample, while species evenness examines how the distribution of individuals among these species affects diversity. Most richness and evenness indices are calculated based on the relative abundance of species.

Additionally, we calculated alpha diversity based on gene count and variation count identified from variants calling on abundant species/strains at different time points and in treatment response. An R script using the "vegan" library was employed to calculate alpha diversity for species abundance, gene count, and variant count, estimating Shannon diversity. The results were visualized through line plots.

### *2.7.2 Median Comparison Using Wilcoxon Test*

The Wilcoxon test was used to assess statistical significance, offering a robust method for evaluating differences within and between groups without relying on strict distributional assumptions. To evaluate and compare species abundance, gene count, and variation count with treatment response, the Wilcoxon test was applied using alpha diversity measures. This analysis involved using the Wilcoxon test to compare the frequency and proportion of species, genes, and variations associated with treatment response alone and across all time points. A p-value of < 0.05 was considered indicative of statistically significant differences.

### *2.7.3 Associations Using MaAsLin2*

This study utilized MaAsLin2 (Mallick et al., 2021) to identify associations between genes and genetic variations from variant calling of abundant species/strains with treatment response, specifically distinguishing between R and NR. For gene associations, we input the gene count matrix and a metadata file containing sample names and their corresponding responses into MaAsLin2. The association analysis between genes and samples was conducted based on treatment response, using R as the reference group. Default parameters were used, except for setting 'normalization' to 'NONE'.

## 2.8 Association via Machine Learning Models

To associate different genes and genetic variations with treatment response (i.e. R and NR), a supervised machine learning approach was employed. The genes and variants data were preprocessed to ensure consistency and quality, addressing any missing or incomplete data points through imputation or removal. Categorical data were converted into numerical form using label encoding to facilitate smooth and accurate association analysis. The dataset was split into training and test sets in an 80:20 ratio, with stratified sampling used to maintain class distribution.

The target variable was "Response," while the features were "Gene" and "Effect." The classification was performed using multiple algorithms, including XGBoost, random forest (RF) classifier, decision tree, gradient boosting machine (GBM), logistic regression, and support vector machine (SVM). K-fold cross-validation was then employed to evaluate model performance and ensure generalizability. Hyper-parameter tuning was conducted using grid search or random search to identify the optimal combination of model parameters, optimizing performance metrics such as accuracy, precision, recall, area under the curve (AUC), and F1-score. The selected model was trained on the preprocessed data using the identified features and optimized hyper-parameters.

### 2.9 Functional Annotation of Genes

The genes found to be associated with treatment response i.e. R and NR from MaAsLin2 were further studied. The functions and pathways of associated genes were analyzed both in the microbial community and the case of NSCLC and its treatment i.e. immunotherapy, particularly, immune checkpoint inhibitors. The genetic variations that occurred in the associated genes were also studied and the effect of these variations on the functional ability of genes making their possible role in treatment response.

# CHAPTER 3: RESULTS AND DISCUSSION

## 3.1 Data Preprocessing

The raw data comprised approximately 2.4 billion reads, with a maximum read length of 26.2 Mbp, a minimum read length of 19.3 Mbp, and an average read length of 20 Mbp. Figure 3.1 shows the distribution of raw reads in the data.

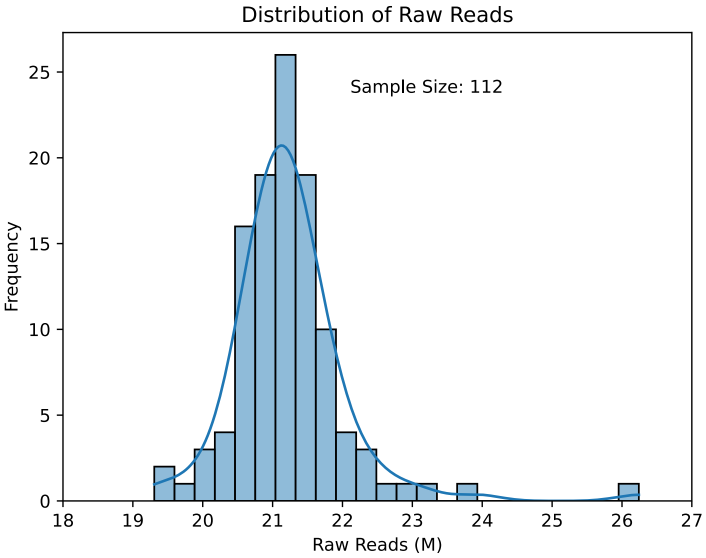


**Figure 3.1:** Distribution of raw reads

During the cleaning stage, which involved removing low-quality reads, adapters, and duplicates, the data count slightly decreased to about 2.1 billion reads. Concurrently, the maximum read length was reduced to 23.5 Mbp, the minimum to 15.6 Mbp, and the average read length to 18.7 Mbp, resulting in a more consistent dataset. Figure 3.2 illustrates the distribution of clean reads in the data.

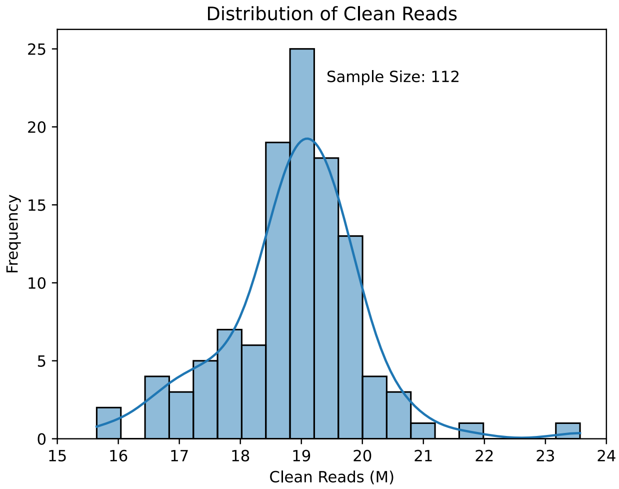


**Figure 3.2:** Distribution of clean reads

Subsequently, during the human genome (hg) trimming stage, where host genome reads were removed, the data count remained at approximately 2.1 billion reads. The maximum and minimum read lengths stayed consistent with the clean stage at 23.5 and 15.6 Mbp, respectively, while the average read length further decreased to 18.5 Mbp. Figure 3.3 shows the distribution of hg-trimmed reads in the data.

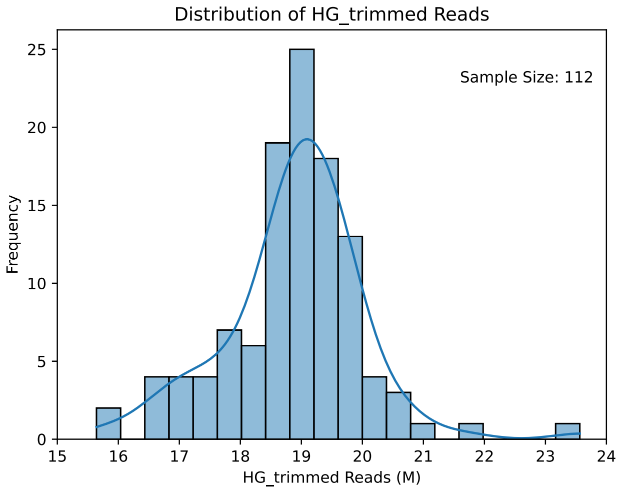


**Figure 3.3:** Distribution of HG-trimmed reads

This progression highlights significant refinement and increased consistency in the dataset through each preprocessing step, particularly in reducing data size and variability. Table 3.1 demonstrates the total number of reads at each stage of preprocessing.

**Table 3.1:** Total number of reads at each preprocessing stage

| | **Raw** | **Clean** | **HG Trimmed** |
|---|---|---|---|
| **Count** | 2.40 B | 2.17 B | 2.15 B |
| **Max** | 26.24 M | 23.56 M | 23.56 M |
| **Min** | 19.30 M | 15.64 M | 15.64 M |
| **Mean** | 20.68 M | 18.74 M | 18.56 M |

### 3.2 Taxonomic Profiling

The MetaPhlAn3 analysis provided a detailed profile of the microbial composition in our samples by aligning reads against a collection of clade-specific marker sequences. This approach enabled MetaPhlAn3 to identify and estimate the relative abundance of microbial species. Cladogram was generated using GraPhlAn to determine the taxonomic classification at all three time points: T0, T1, and T2, as shown in Figure 3.4. At all-time points, the identified phyla included Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria, and Verrucomicrobia.

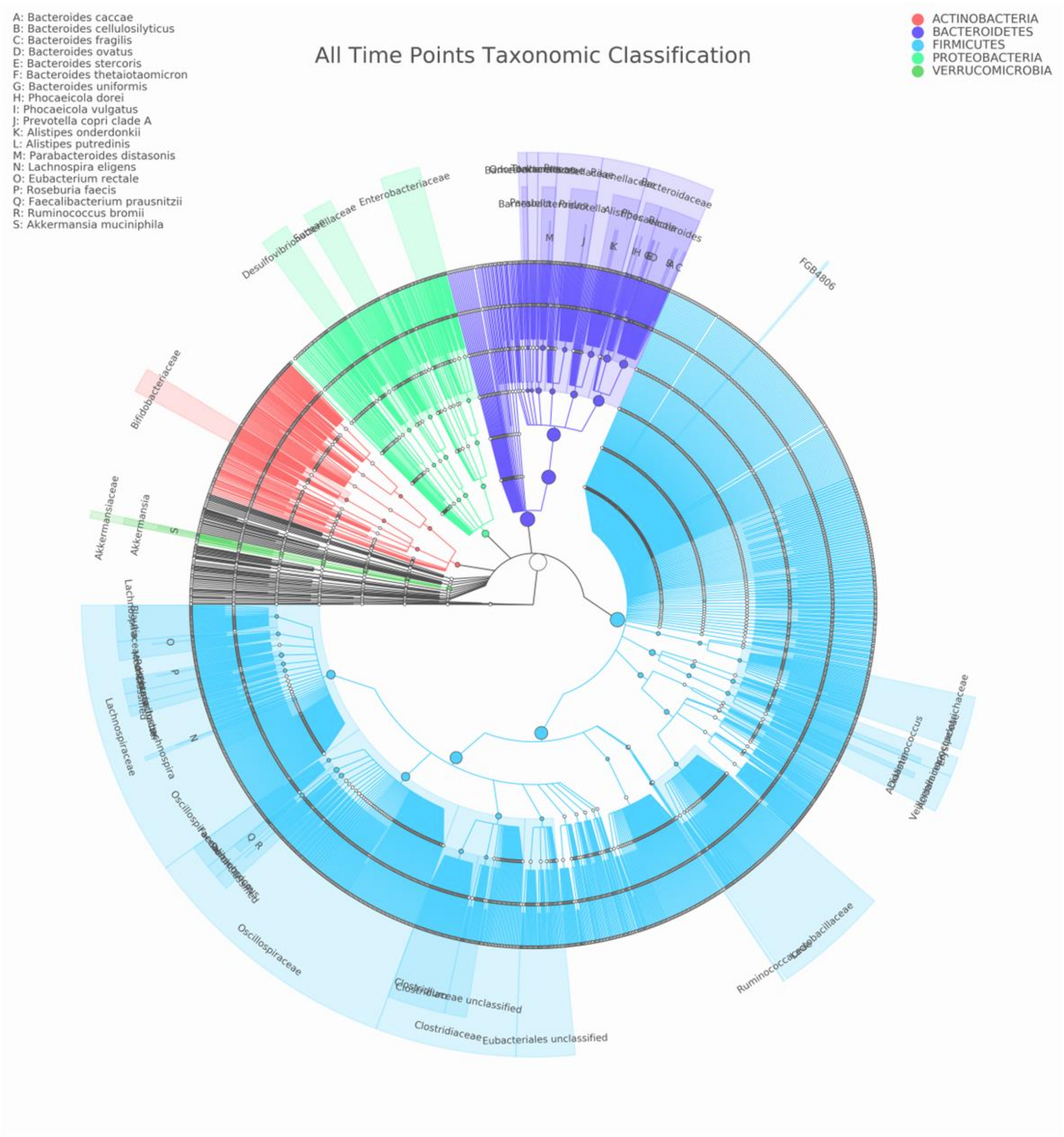


**Figure 3.4:** Taxonomic classification at all time points

The cladogram revealed the top 25 abundant species from these phyla, which included *B. caccae*, *B. cellulosilyticus*, *B. dorei*, *B. ovatus*, *B. stercoris*, *B. thetaiotaomicron*, *B. uniformis*, *B. vulgatus*, *B.intestinihominis*, *P. copri*, *A. finegoldii*, *A. putredinis*, *B. distasonis*, *P. johnsonii*, *P. merdae*, *E. eligens*, *E. sp. CAG 180*, *L.*

*pectinoschiza*, *E. rectale*, *R. faecis*, *F. prausnitzii*, *R. bromii*, *A. intestini*, *E. coli*, and *A. muciniphila*.

Heatmaps were generated for both time points (T0, T1, and T2) and response (R and NR). At T0, the most abundant species were *P. vulgatus*, *B. uniformis*, *F. prausnitzii*, *P. dorei*, and *P. distasonis*. At T1, the abundant species included *B. uniformis*, *B. vulgatus*, *F. prausnitzii*, *A. muciniphila*, and *B. dorei*. At T2, *B. vulgatus*, *B. uniformis*, *F. prausnitzii*, *B. dorei*, and *A. finegoldii* were the most abundant. Figures 3.5-3.7 illustrate the heatmaps of the T0, T1 and T2 time points.

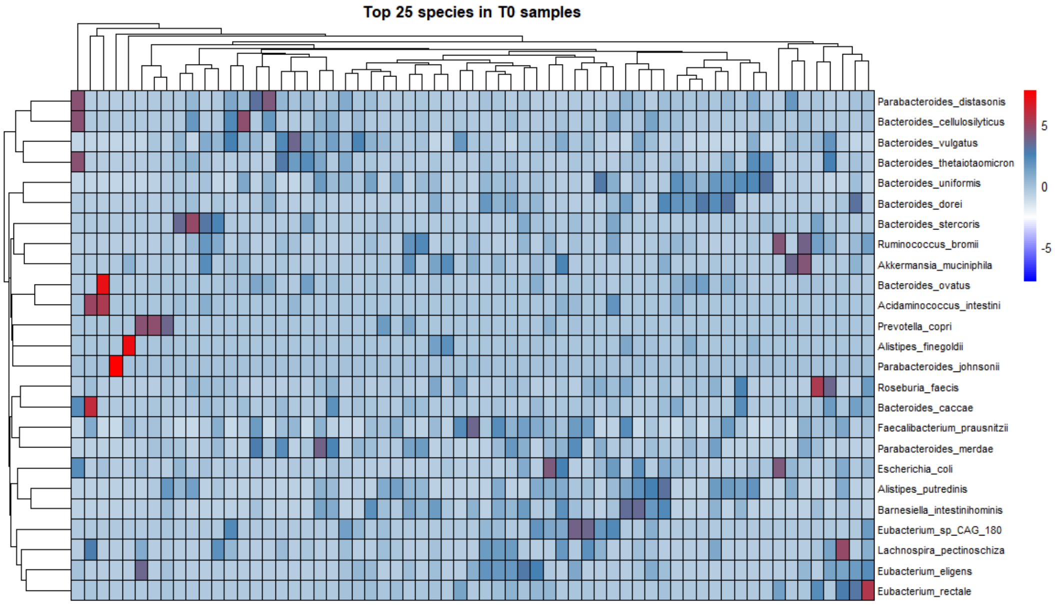


**Figure 3.5:** Heatmaps of the top 25 abundant species in the T0 sample group

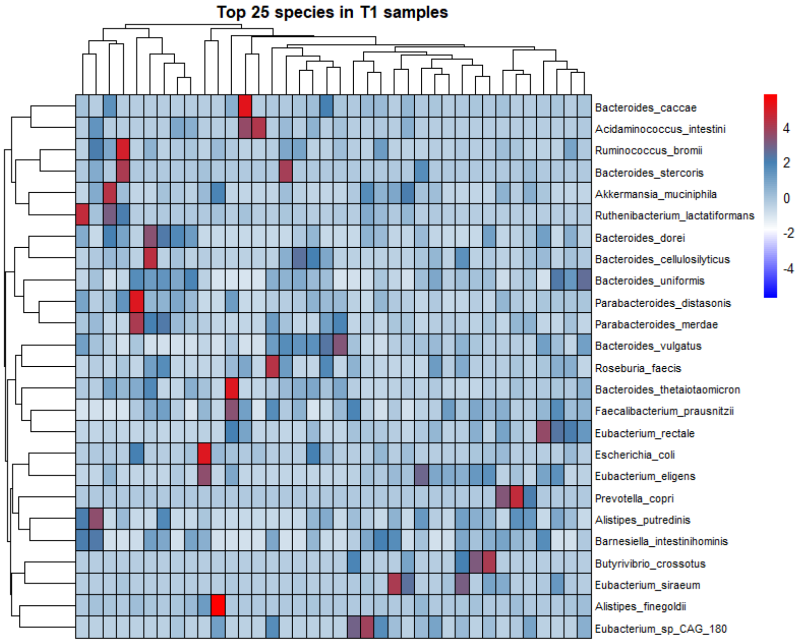


**Figure 3.6:** Heatmaps of the top 25 abundant species in the T1 sample group

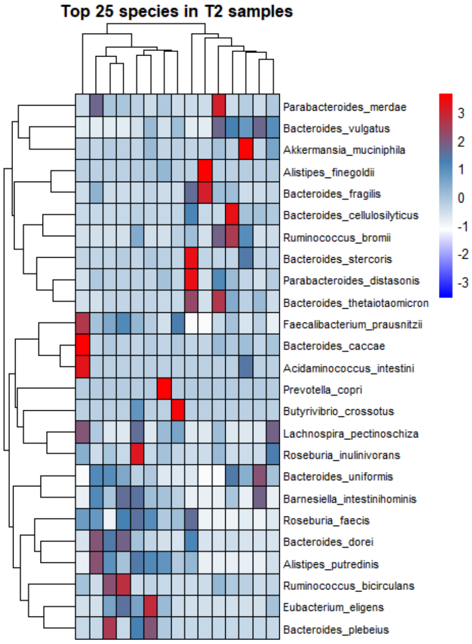


**Figure 3.7:** Heatmaps of the top 25 abundant species in the T2 sample group

Across all timepoints, *P. vulgatus*, *B. uniformis*, *F. prausnitzii*, *P. dorei*, and *P. distasonis* were the most abundant species. Based on response, the most abundant species in R were *P. vulgatus, B. uniformis*, *F. prausnitzii*, *P. dorei*, and *A.muciniphila*, whereas in NR, *B. uniformis*, *P. vulgatus*, *F. prausnitzii*, *P. distasonis*, and *P. dorei* were predominant. Figure 3.8 illustrates the heatmap showing the top 25 most abundant species by response.

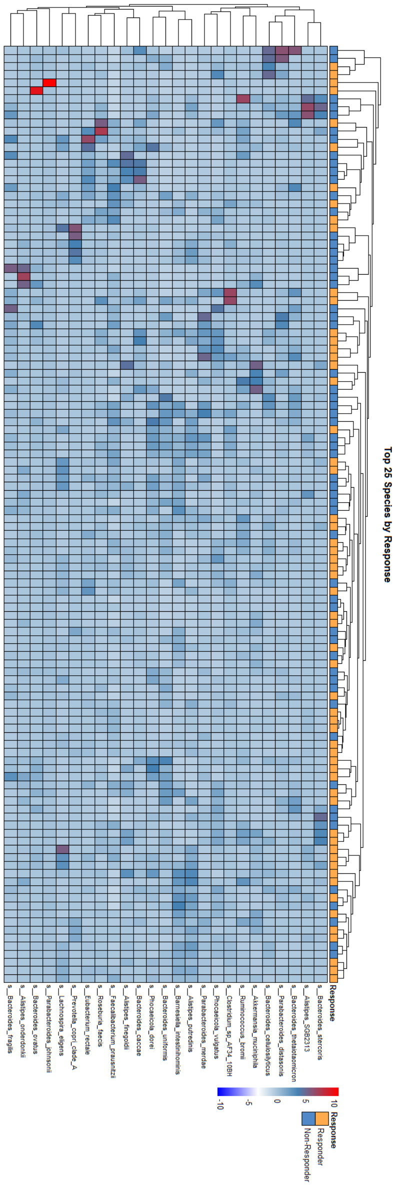


**Figure 3.8:** Heatmap of the top 25 abundant species by Response

Across all time points (T0, T1, and T2), *P. vulgatus, B. uniformis*, and *F. prausnitzii* consistently emerged as the most abundant species, suggesting their pivotal role in maintaining gut health and stability (Zappa & Mousa, 2016). In R, the presence of *P. vulgatus*, *B. uniformis,* and *A. muciniphila* indicates a potential beneficial impact on the efficacy of immunotherapy in NSCLC patients. Conversely, NR showed higher abundances of *P. distasonis* and *P. dorei*, which may be associated with a less favorable response to treatment (Qi et al., 2022).

Strain diversity profiling was further conducted on the top 25 abundant species identified by cladograms and heatmaps at various time points and responses to elucidate complex mechanisms and processes through phylogenetic and functional relationships. Table 3.2 shows the relative abundances of the most abundant species identified.

**Table 3.2:** Relative abundances of Species abundant in different categories

| **Species** | **Rel. abundance** | **Category** |
|---|---|---|
| *Phocaeicola vulgatus* | 467.75 | T0 |
| *Bacteroides uniformis* | 353.66 | |
| *Faecalibacterium prausnitzii* | 214.62 | |
| *Phocaeicola dorei* | 184 | |
| *Parabacteroides distasonis* | 168.19 | |
| *Bacteroides uniformis* | 344.66608 | T1 |
| *Bacteroides vulgatus* | 339.19621 | |
| *Faecalibacterium prausnitzii* | 206.02129 | |

|  |  |  |
|---|---|---|
| *Akkermansia muciniphila* | 115.76012 |  |
| *Bacteroides dorei* | 115.2153 |  |
| *Bacteroides vulgatus* | 149.89851 | T2 |
| *Bacteroides uniformis* | 141.52046 |  |
| *Faecalibacterium prausnitzii* | 95.08367 |  |
| *Bacteroides dorei* | 77.61672 |  |
| *Alistipes finegoldii* | 57.7487 |  |
| *Phocaeicola vulgatus* | 802.7912 | All Timepoints |
| *Bacteroides uniformis* | 641.7307 |  |
| *Faecalibacterium prausnitzii* | 427.6501 |  |
| *Phocaeicola dorei* | 303.9049 |  |
| *Parabacteroides distasonis* | 261.728 |  |
| *Phocaeicola vulgatus* | 523.58272 | Responder |
| *Bacteroides uniformis* | 300.78456 |  |
| *Faecalibacterium prausnitzii* | 205.7608 |  |
| *Phocaeicola dorei* | 129.44106 |  |
| *Akkermansia muciniphila* | 110.78218 |  |
| *Bacteroides uniformis* | 340.9462 | Non-Responder |
| *Phocaeicola vulgatus* | 279.2084 |  |
| *Faecalibacterium prausnitzii* | 221.8893 |  |
| *Parabacteroides distasonis* | 179.4488 |  |
| *Phocaeicola dorei* | 174.4638 |  |

### 3.3 Strain Diversity Profiling

Diving deeper into the microbiome beyond the species level allows us to characterize specific strains or subspecies via StrainPhlAn3 and link them to the treatment response in NSCLC patients. Strain diversity profiling was performed based on both time points and response, resulting in phylogenetic trees that show variations and relationships between samples and bacterial strains. Table 3.3 demonstrates the abundant and common species whose strains were identified across timepoint samples.

**Table 3.3:** Species with identified strains across time points

| **Species with identified strain** | **Timepoint** |
|---|---|
| *Phocaeicola dorei* | **T0-T1-T2** |
| *Bacteroides uniformis* | |
| *Bacteroides stercoris* | |
| *Parabacteroides distasonis* | |
| *Akkermansia muciniphila* | |
| *Faecalibacterium prausnitzii* | |
| *Phocaeicola vulgatus* | |
| *Bacteroides thetaiotaomicron* | **T0-T1** |
| *Bacteroides ovatus* | |
| *Ruminococcus bromii* | |
| *Bacteroides cellulosilyticus* | |
| *Acidaminococcus intestini* | |
| *Prevotella copri* | |
| *Bacteroides caccae* | |
| *Lachnospira eligens* | **T0-T2** |
| *Parabacteroides merdae* | |
| *Roseburia faecis* | **T0** |
| *Escherichia coli* | |
| *Eubacterium rectale* | |
| *Bacteroides fragilis* | **T2** |
| *Alistipes finegoldii* | |

As shown in Table 3.3, the strains of seven bacterial species, including *P. dorei*, *B. uniformis*, *B. stercoris*, *P. distasonis*, *A. muciniphila*, *F. prausnitzii*, and *P. vulgatus*, were present in samples from all time points (T0, T1, and T2). The strains of seven other bacterial species, including *B. thetaiotaomicron*, *B. ovatus*, *R. bromii*, *B. cellulosilyticus*, *A. intestini*, *P. copri*, and *B. caccae*, were specifically present in T0 and T1 samples. The strains of *L. eligens* and *P. merdae* were uniquely found in T0 and T2 samples. Additionally, strains of *R. faecis*, *E. coli*, and *E. rectale* were found only in T0 samples, while strains of *B. fragilis* and *A. finegoldii* were unique to T2 samples. Figure 3.9 illustrates the Venn diagram and upset plot showing the distribution of microbial strains of abundant species across the time points.

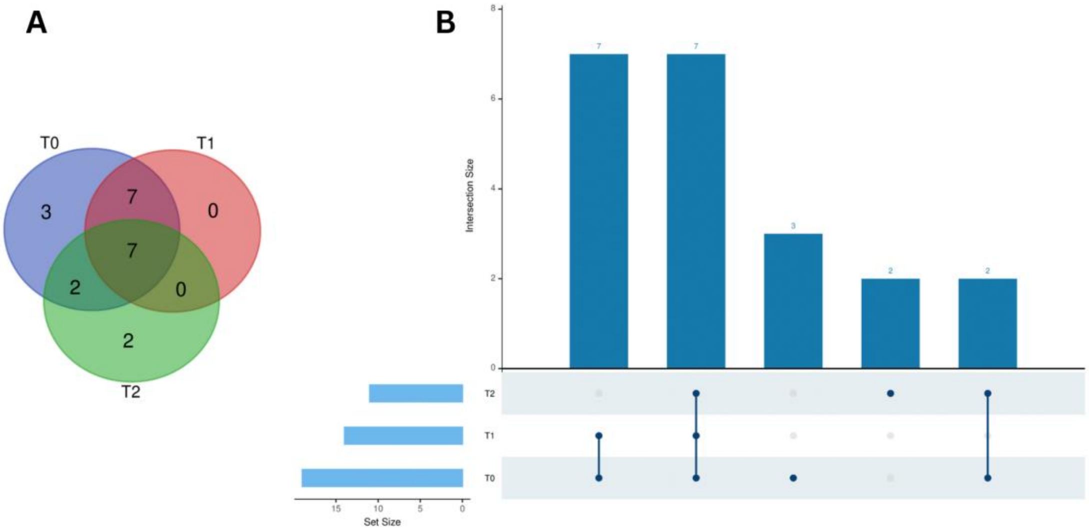


**Figure 3.9: A.** Venn diagram; **B.** UpSet plot showing the distribution of species with identified strains across time points

Based on response shown in Table 3.4, sixteen strains of bacterial species, including *P. dorei*, *B. uniformis*, *B. thetaiotaomicron*, *B. ovatus*, *R. bromii*, *B. stercoris*, *B. cellulosilyticus*, *A. intestini*, *R. faecis*, *A. muciniphila*, *F. prausnitzii*, *P. copri*, *A. finegoldii*, *P. vulgatus*, *E. rectale*, and *B. caccae*, were present in both R and NR.

**Table 3.4:** Species with identified strain identified in R and NR

| **Species with identified strains** | **Response** |
|---|---|
| *Phocaeicola dorei* | **Non-Responder & Responder** |
| *Bacteroides uniformis* | |
| *Bacteroides thetaiotaomicron* | |
| *Bacteroides ovatus* | |
| *Ruminococcus bromii* | |
| *Bacteroides stercoris* | |
| *Bacteroides cellulosilyticus* | |
| *Acidaminococcus intestini* | |
| *Roseburia faecis* | |
| *Akkermansia muciniphila* | |
| *Faecalibacterium prausnitzii* | |
| *Prevotella copri* | |
| *Alistipes finegoldii* | |
| *Phocaeicola vulgatus* | |
| *Eubacterium rectale* | |
| *Bacteroides caccae* | |
| *Lachnospira eligens* | **Responder** |

| | |
|---|---|
| *Escherichia coli* | |
| *Parabacteroides merdae* | |
| *Parabacteroides distasonis* | **Non-Responder** |
| *Bacteroides fragilis* | |

Strains of only three bacterial species were linked exclusively with R which includes *L. eligens*, *E. coli*, and *P. merdae*. Conversely, strains of two bacterial species, *P. distasonis* and *B. fragilis*, were found only in non-responders. Figure 3.10 illustrates Venn diagram and upset plot showing the distribution of microbial strains of abundant species in R and NR.

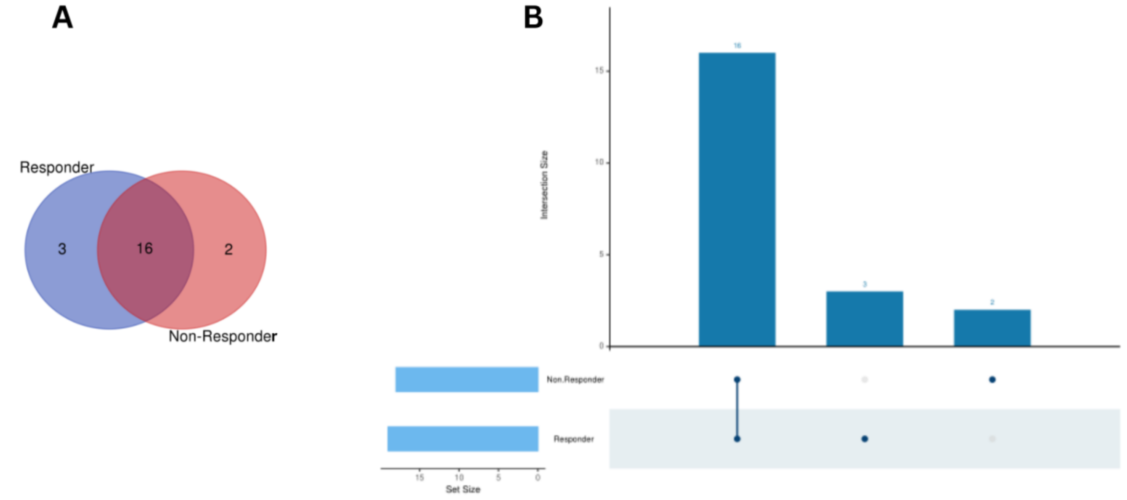


**Figure 3.10: A.** Venn diagram; **B.** UpSet plot showing the distribution of species with identified strains by Response

The analysis revealed that strains of *P. dorei*, *B. uniformis*, *B. stercoris*, *P. distasonis*, *A. muciniphila*, *F. prausnitzii*, and *P. vulgatus* were consistently present across all time points (T0, T1, and T2), indicating their stable role in the microbiome. In both R and NR, sixteen bacterial strains, including *P. dorei*, *B. uniformis*, and *F. prausnitzii*, were common, suggesting their significant influence on treatment response. Strains unique to R were *L. eligens*, *E. coli*, and *P. merdae*, while *P. distasonis* and *B. fragilis* were specific to NR, highlighting potential microbial markers for treatment efficacy (Li et al., 2024). Figure 3.11 illustrates the presence-absence plot showing the presence and absence of strains of abundant microbial species in R and NR.

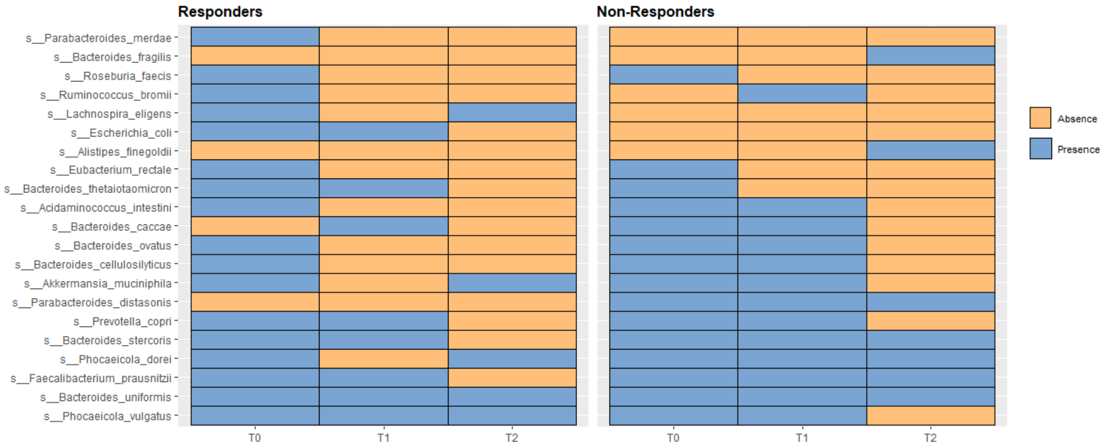


**Figure 3.11:** Presence-absence plot showing species with identified strains in R and NR

## 3.4 Variant Calling

The rapid haploid variant calling and core genome alignment tool Snippy (version 4.4.0) was utilized to identify genetic variations in samples to investigate the link with treatment response. Sequencing reads were aligned to a reference sequence, and polymorphisms were identified. The alignments revealed thousands of genetic variations in samples compared to reference sequences of different bacterial strains as shown in figure

3.12. Among the seven most abundant common species across all time points (T0, T1, and T2) were *A. muciniphila*, *B. dorei*, *B. stercoris*, *B. uniformis*, *F. prausnitzii*, *P. distasonis*, and *P. vulgatus*. A total of 83,583 variations were identified, with 63,737 being SNPs, 17,553 complex mutations, 1,803 multiple nucleotide polymorphisms (MNPs), 278 deletions (del), and 216 insertions (ins). NR exhibited 47,969 genetic variations, while R had 35,615, indicating a higher number of variations in NR, as shown in Figure 3.13.

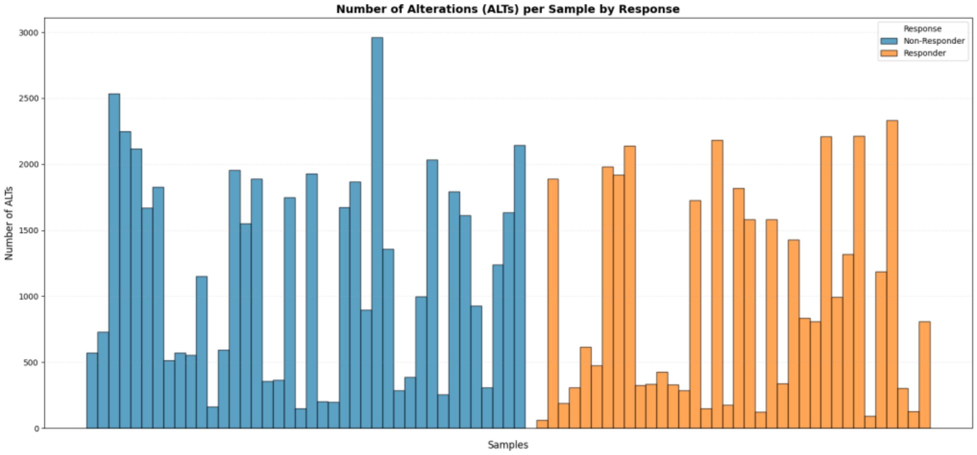


**Figure 3.12:** Number of alterations across samples in R vs. NR

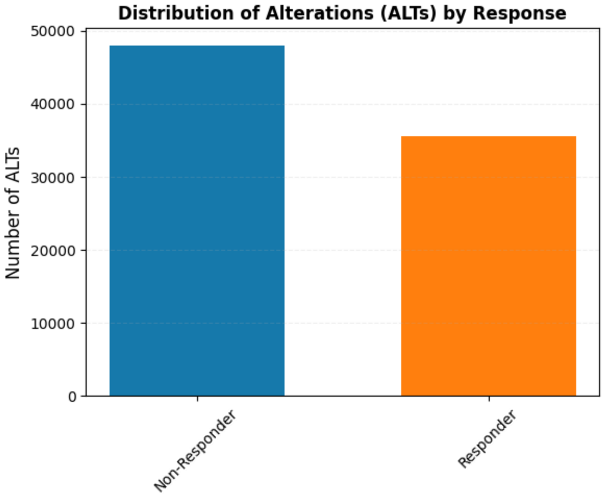


**Figure 3.13:** Total number of alterations in R vs. NR

*A. muciniphila* contained 10,135 variations (5,939 in NR, 4,197 in R), *B. dorei* had 4,689 (2,719 in NR, 1,971 in R), *B. stercoris* had 4,977 (3,177 in NR, 1,801 in R), *B. uniformis* had 48,773 (26,810 in NR, 21,964 in R), *F. prausnitzii* had 2,644 (1,532 in NR, 1,113 in R), *P. distasonis* had 4,507 (all in NR), and *P. vulgatus* had 7,864 (3,291 in NR, 4,574 in R), as illustrated in Figure 3.14.

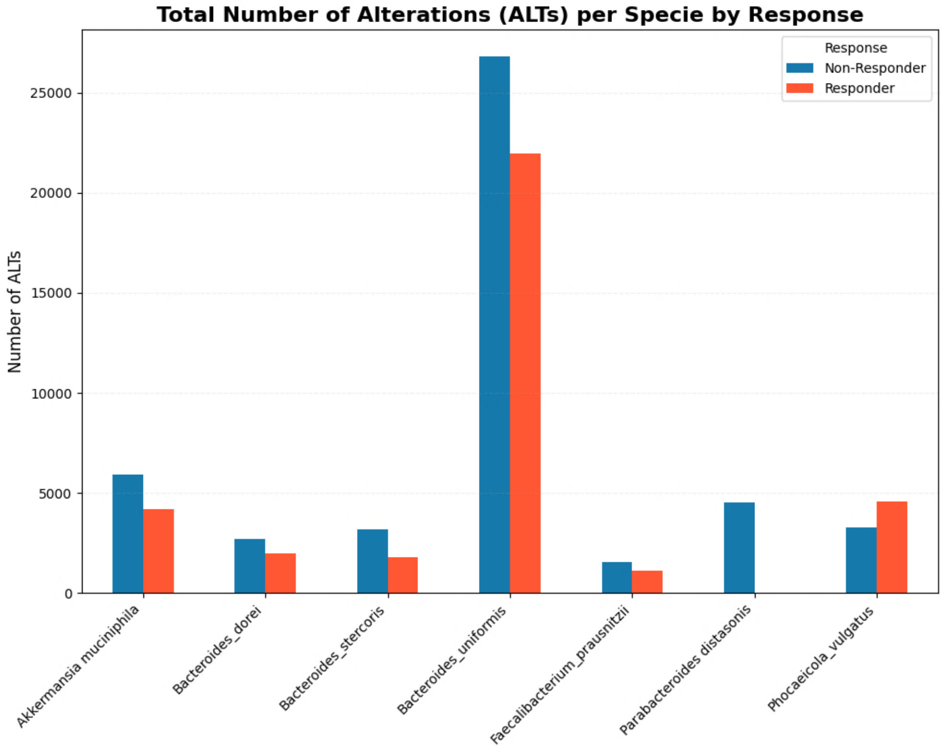


**Figure 3.14:** Number of alterations in species in R and NR

Variant calling identified genes in these species, highlighting 31,827 genes with variations, 17,794 in NR, and 14,034 in R, which is also illustrated in Figure 3.15. Eight genes associated with treatment response were identified: *ftsA* (cell division protein FtsA), *lpdA* (dihydrolipoyl dehydrogenase), *nadB* (L-aspartate oxidase), *obgE* (GTPase ObgE), *rhaT* (L-rhamnose-proton symporter), *sufD* (Fe-S cluster assembly protein SufD), *uxaC* (glucuronate isomerase), and *xylE* (D-xylose transporter XylE). In NR, the number of variations in these genes was 35 (*ftsA*), 76 (*lpdA*), 28 (*nadB*), 16 (*obgE*), 28 (*rhaT*), 119 (*sufD*), 77 (*uxaC*), and 37 (*xylE*). In responders, the number of variations was 7 (*ftsA*), 11 (*lpdA*), 3 (*nadB*), 6 (*rhaT*), 41 (*sufD*), 7 (*uxaC*), and 5 (*xylE*), as shown in Figure 3.16.

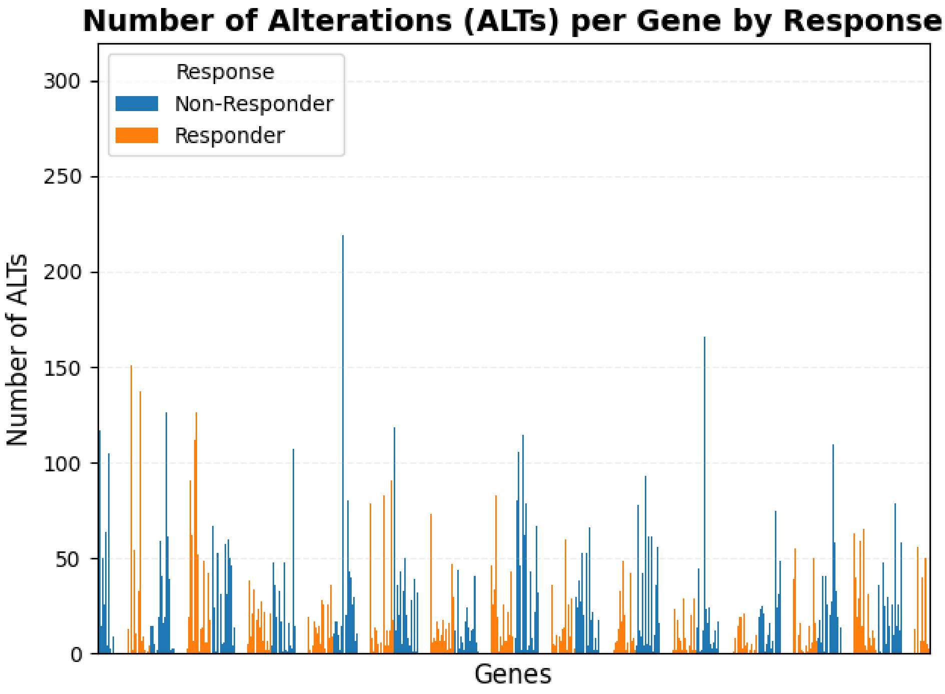


**Figure 3.15:** Number of alterations in total genes in R vs. NR

The study identified a greater number of genetic variations in NR compared to R, with specific genes linked to treatment response showing significant variation. This suggests that genetic variability in common abundant bacterial species may influence treatment outcomes. The genes *ftsA*, *lpdA*, *nadB*, *obgE*, *rhaT*, *sufD*, *uxaC*, and *xylE* were particularly associated with treatment response, their impact is discussed further in the thesis.

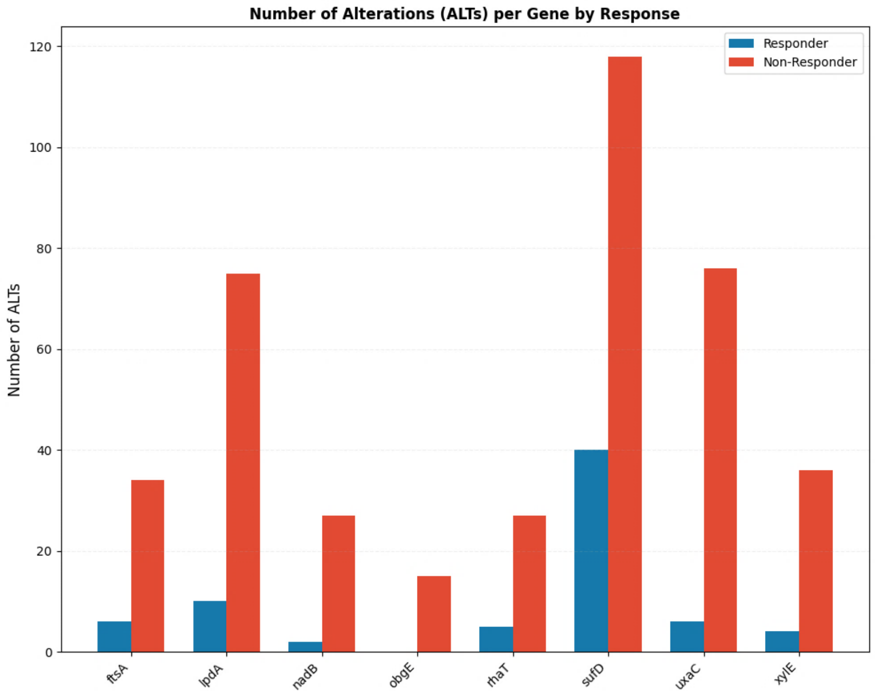


**Figure 3.16:** Number of alterations in eight associated genes in R vs. NR

## 3.5 Gene Ontology

The Gene Ontology (GO) analysis was conducted to establish consistent gene product descriptions across various databases, focusing on three main categories: biological processes, cellular components, and molecular functions. Among the most prevalent species observed across all time points (T0, T1, and T2) were *A. muciniphila*, *B. dorei*, *B. stercoris*, *B. uniformis*, *F. prausnitzii*, *P. distasonis*, and *P. vulgatus*. This GO analysis was performed specifically for these species due to the identification of numerous variations in the genes discussed in the previous section.

### 3.5.1 *Biological Processes*

In *A. muciniphila*, enriched pathways in NR include Organic substance biosynthetic process (FDR = 1.2811e$^{-89}$, nGenes = 149, Fold Enrichment = 5.70), Cellular biosynthetic process (FDR = 1.1329e$^{-86}$, nGenes = 146, Fold Enrichment = 5.65), and Biosynthetic process (FDR = 2.4758e$^{-89}$, nGenes = 150, Fold Enrichment = 5.60) shown in Figure 3.17.

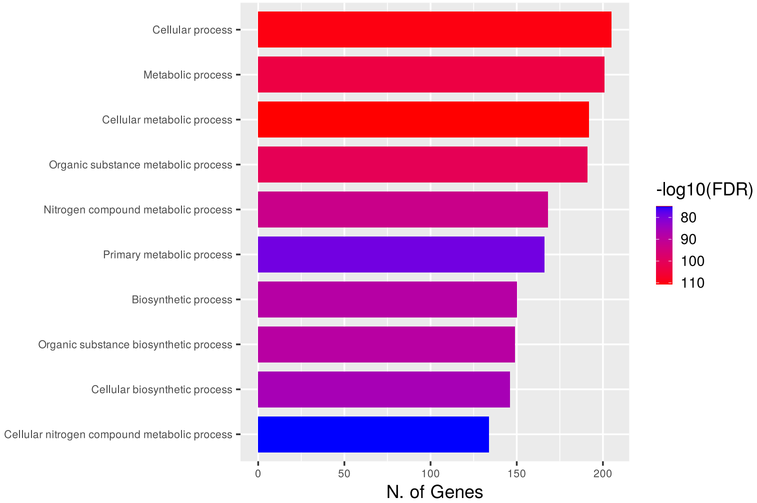


**Figure 3.17:** Top biological processes in *A. muciniphila* in NR

In R, the species showed significant involvement in Cellular nitrogen compound metabolic process (FDR = 8.897e$^{-81}$, nGenes = 138, Fold Enrichment = 5.71), Organic substance biosynthetic process (FDR = 1.279e$^{-89}$, nGenes = 149, Fold Enrichment = 5.70), and Cellular biosynthetic process (FDR = 5.428e$^{-88}$, nGenes = 147, Fold Enrichment = 5.69) illustrated in Figure 3.18.

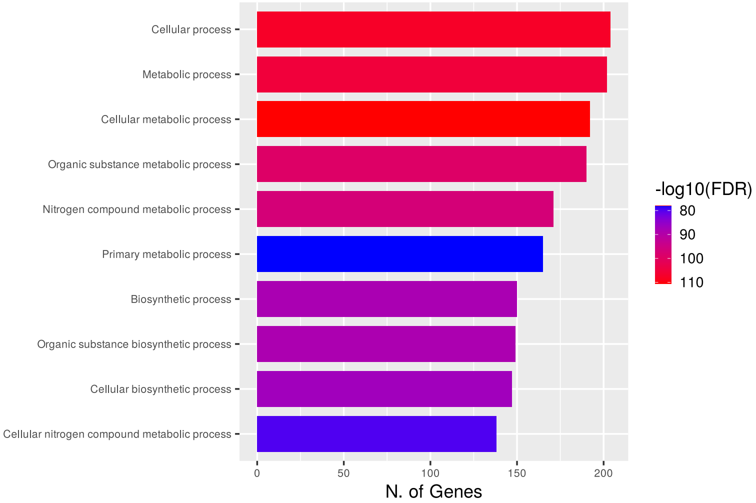


**Figure 3.18:** Top biological processes in *A. muciniphila* in R

For *B. dorei*, enriched pathways in NR include Cellular aromatic compound metabolic process (FDR = 7.78e$^{-31}$, nGenes = 101, Fold Enrichment = 3.34), Heterocycle metabolic process (FDR = 1.63e$^{-31}$, nGenes = 103, Fold Enrichment = 3.33), and Organic cyclic compound metabolic process (FDR = 1.34e$^{-32}$, nGenes = 106, Fold Enrichment = 3.33) shown in Figure 3.19.

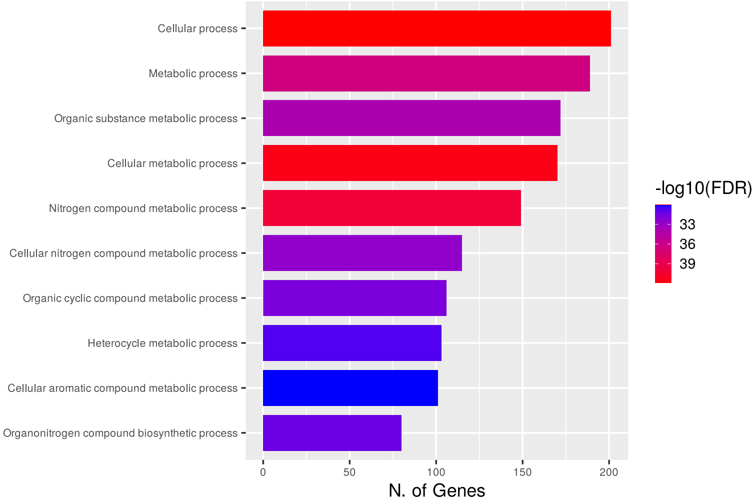


**Figure 3.19:** Top biological processes in *B. dorei* in NR

In R, the species showed significant involvement in small molecule metabolic process (FDR = $1.20e^{-38}$, nGenes = 108, Fold Enrichment = 3.83), Heterocycle metabolic process (FDR = $1.01e^{-38}$, nGenes = 119, Fold Enrichment = 3.45), and Organic cyclic compound metabolic process (FDR = $1.15e^{-39}$, nGenes = 122, Fold Enrichment = 3.44) demonstrated in figure 3.20.

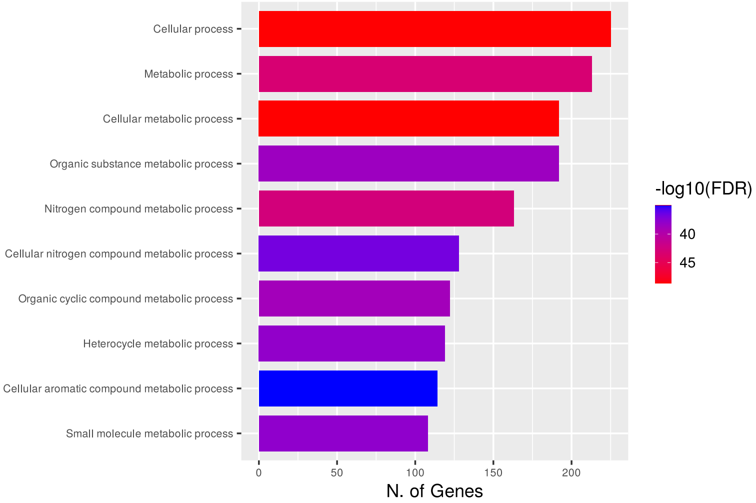


**Figure 3.20:** Top biological processes in *B. dorei* in R

For *B. stercoris*, enriched pathways in NR include Organonitrogen compound biosynthetic process (FDR = $2.08e^{-62}$, nGenes = 108, Fold Enrichment = 6.27), Small molecule metabolic process (FDR = $1.73e^{-62}$, nGenes = 119, Fold Enrichment = 5.46), and Cellular biosynthetic process (FDR = $2.34e^{-66}$, nGenes = 138, Fold Enrichment = 4.73) shown in Figure 3.21.

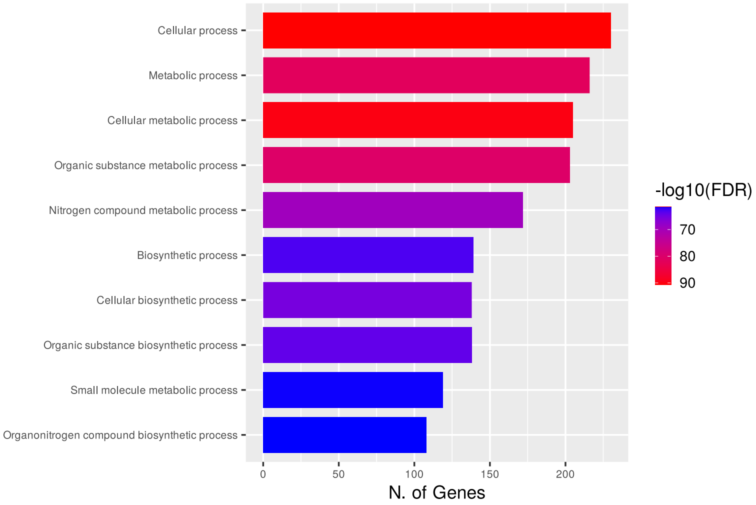


**Figure 3.21:** Top biological processes in *B. stercoris* in NR

In R, the species showed significant involvement in Cellular biosynthetic process (FDR = 2.2647e$^{-61}$, nGenes = 128, Fold Enrichment = 4.746), Organic substance biosynthetic process (FDR = 4.5170e$^{-60}$, nGenes = 128, Fold Enrichment = 4.632), and Biosynthetic process (FDR = 3.3827e$^{-58}$, nGenes = 128, Fold Enrichment = 4.476) illustrated in Figure 3.22.

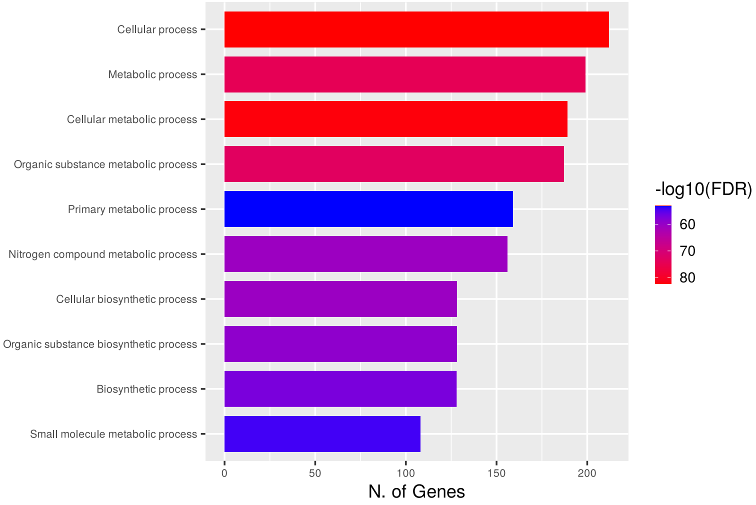


**Figure 3.22:** Top biological processes in *B. stercoris* in R

For *B. uniformis*, enriched pathways in NR include Organonitrogen compound biosynthetic process (FDR = 6.85e$^{-82}$, nGenes = 200, Fold Enrichment = 3.98), Small molecule metabolic process (FDR = 7.55e$^{-98}$, nGenes = 251, Fold Enrichment = 3.71), and Cellular biosynthetic process (FDR = 4.10e$^{-76}$, nGenes = 267, Fold Enrichment = 2.93) shown in Figure 3.23.

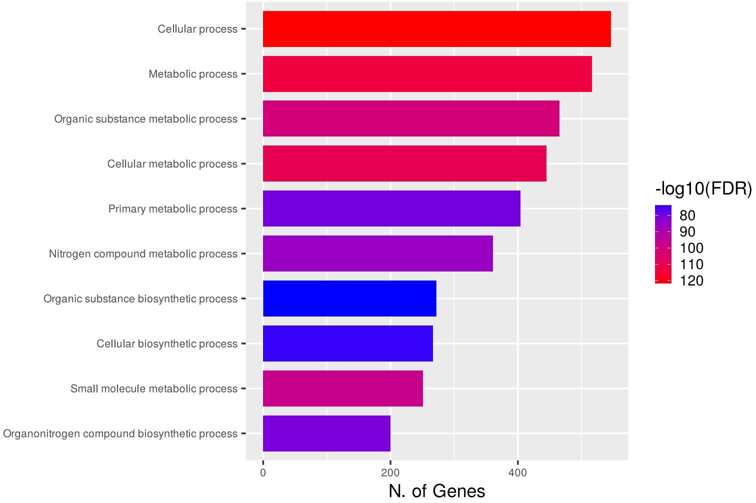


**Figure 3.23:** Top biological processes in *B. uniformis* in NR

In R, the species showed significant involvement in Organonitrogen compound biosynthetic process (FDR = 1.91e$^{-76}$, nGenes = 191, Fold Enrichment = 3.94), Small molecule metabolic process (FDR = 1.34e$^{-90}$, nGenes = 239, Fold Enrichment = 3.67), and Cellular biosynthetic process (FDR = 4.16e$^{-74}$, nGenes = 259, Fold Enrichment = 2.95) demonstrated in Figure 3.24.

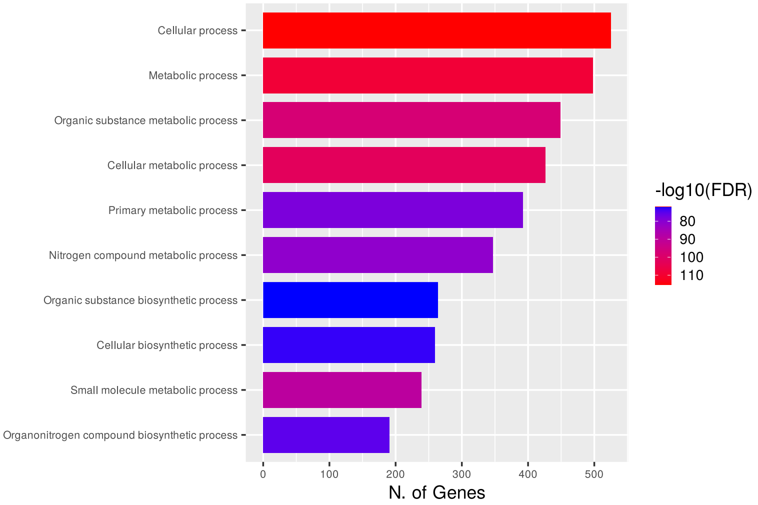


**Figure 3.24:** Top biological processes in *B. uniformis* in R

For *B. vulgatus*, enriched pathways in NR include Organonitrogen compound biosynthetic process (FDR = 1.17e$^{-42}$, nGenes = 105, Fold Enrichment = 4.28), Small molecule metabolic process (FDR = 6.09e$^{-45}$, nGenes = 119, Fold Enrichment = 3.89), and Organic cyclic compound metabolic process (FDR = 7.08e$^{-45}$, nGenes = 133, Fold Enrichment = 3.39) shown in Figure 3.25.

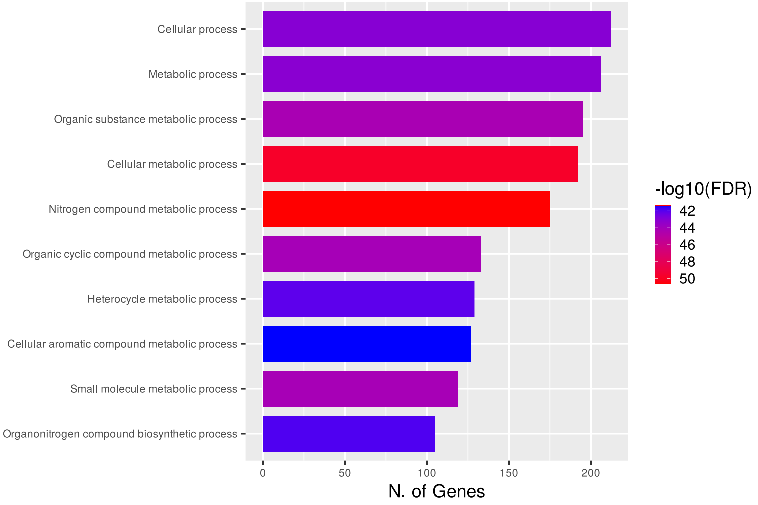


**Figure 3.25:** Top biological processes in *B. vulgatus* in NR

In R, the species showed significant involvement in Organonitrogen compound biosynthetic process (FDR = 2.35e$^{-44}$, nGenes = 106, Fold Enrichment = 4.40), Small molecule metabolic process (FDR = 1.71e$^{-43}$, nGenes = 116, Fold Enrichment = 3.86), and Cellular biosynthetic process (FDR = 2.18e$^{-42}$, nGenes = 133, Fold Enrichment = 3.20) illustrated in Figure 3.26.

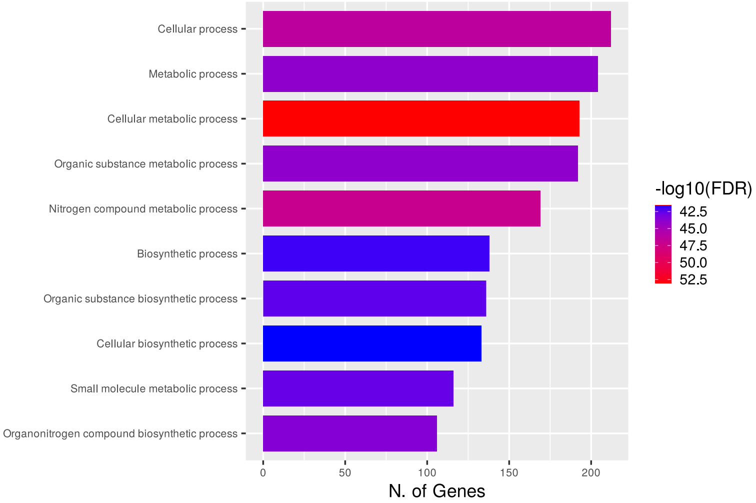


**Figure 3.26:** Top biological processes in *B. vulgatus* in R

For *F. prausnitzii*, enriched pathways in NR include Organic substance biosynthetic process (FDR = 7.07e$^{-23}$, nGenes = 40, Fold Enrichment = 6.08), Cellular biosynthetic process (FDR = 3.80e$^{-22}$, nGenes = 39, Fold Enrichment = 6.06), and Biosynthetic process (FDR = 2.00e$^{-22}$, nGenes = 40, Fold Enrichment = 5.90) shown in Figure 3.27.

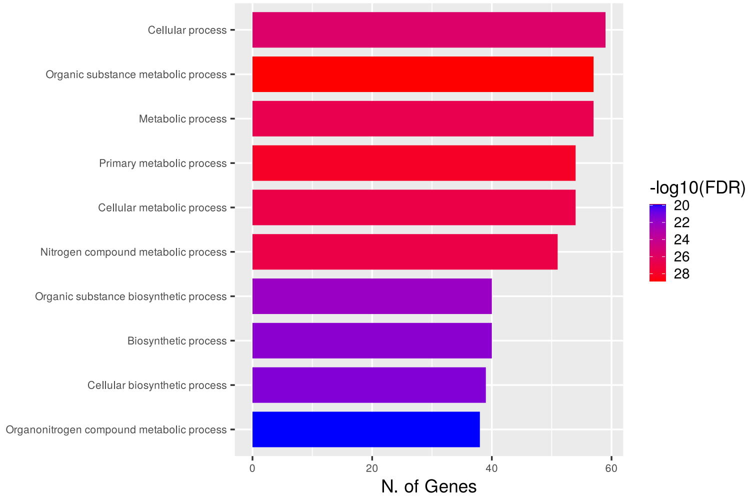


**Figure 3.27:** Top biological processes in *F. prausnitzii* in NR

In R, the species showed significant involvement in Organic substance biosynthetic process (FDR = 2.59e$^{-29}$, nGenes = 54, Fold Enrichment = 5.75), Cellular biosynthetic process (FDR = 1.14e$^{-27}$, nGenes = 52, Fold Enrichment = 5.65), and Organonitrogen compound metabolic process (FDR = 3.09e$^{-28}$, nGenes = 53, Fold Enrichment = 5.63) demonstrated in Figure 3.28.

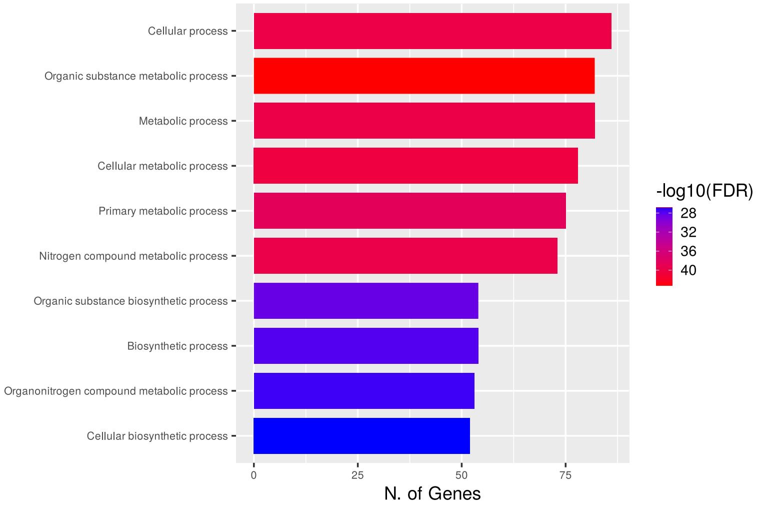


**Figure 3.28:** Top biological processes in *F. prausnitzii* in R

Lastly, *P. distasonis*, in NR, shows significant enrichment in cellular aromatic compound metabolism (FDR = 5.29e$^{-36}$, nGenes = 98, fold enrichment = 3.76), organic cyclic compound metabolism (FDR = 1.97e$^{-37}$, nGenes = 102, fold enrichment = 3.71), and heterocycle metabolism (FDR = 5.60e$^{-36}$, nGenes = 99, fold enrichment = 3.70) as shown in Figure 3.29. These findings underscore the diverse metabolic activities of *P.distasonis* in NR.

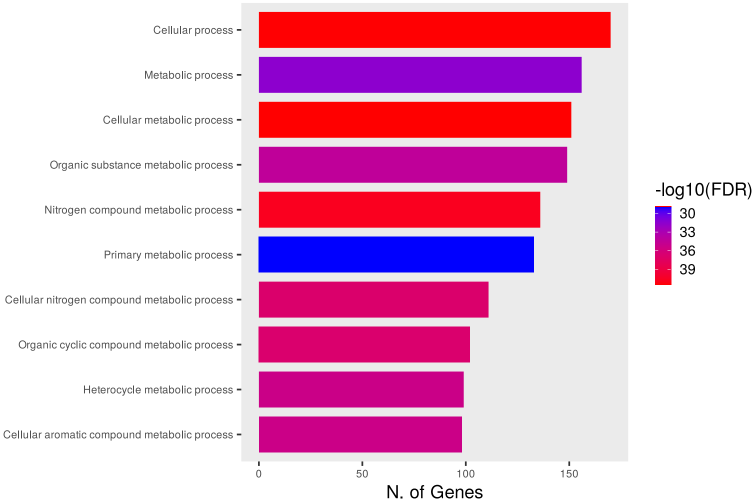


**Figure 3.29:** Top biological processes in *P. distasonis* in NR

Overall, these results highlight the intricate involvement of different genes of abundant microbial species in distinct metabolic pathways, potentially influencing treatment response in patients.

### 3.5.2 *Cellular Processes*

In NR, *A. muciniphila* demonstrated significant enrichment in various cellular components. Among the top three most notable were the Intracellular (FDR = 3.8643e$^{-104}$, nGenes = 176, fold enrichment = 4.85), Cytoplasm (FDR = 2.9588e$^{-83}$, nGenes = 151, fold enrichment = 4.96), and Cytosol (FDR = 1.7047e$^{-11}$, nGenes = 30, fold enrichment = 4.50) shown in Figure 3.30.

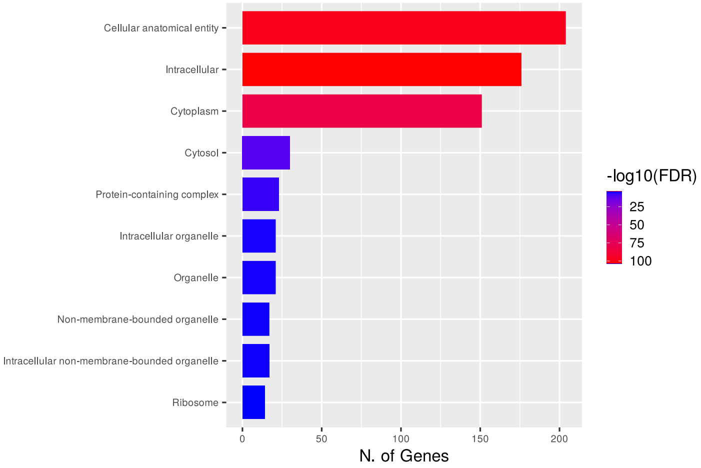


**Figure 3.30:** Top cellular processes in *A. muciniphila* in NR

Conversely, in R, *A. muciniphila* showed enrichment primarily in the Intracellular (FDR = 2.4614e$^{-101}$, nGenes = 174, fold enrichment = 4.79), Cytoplasm (FDR = 1.9039e$^{-79}$, nGenes = 148, fold enrichment = 4.86), and Cytosol (FDR = 8.3765e$^{-15}$, nGenes = 34, fold enrichment = 5.10) illustrated in Figure 3.31.

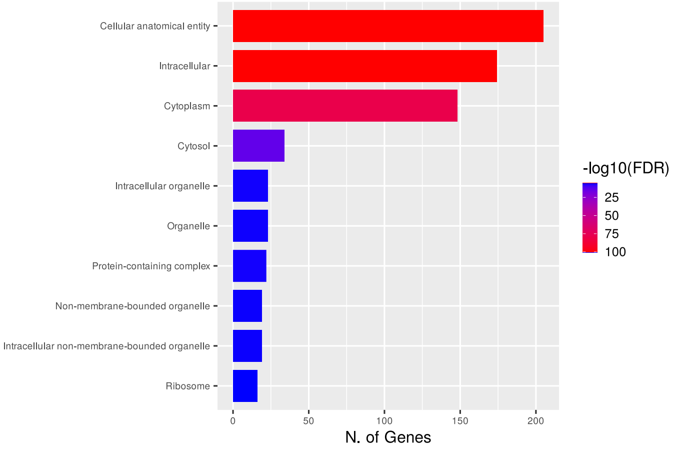


**Figure 3.31:** Top cellular processes in *A. muciniphila* in R

For *B. dorei* present in NR, significant enrichment was observed in the Intracellular (FDR = 7.0860e$^{-56}$, nGenes = 162, fold enrichment = 2.99), Cytoplasm (FDR = 6.3656e$^{-50}$, nGenes = 148, fold enrichment = 3.09), and Cytosol (FDR = 6.0304e$^{-10}$, nGenes = 37, fold enrichment = 3.31) shown in Figure 3.32.

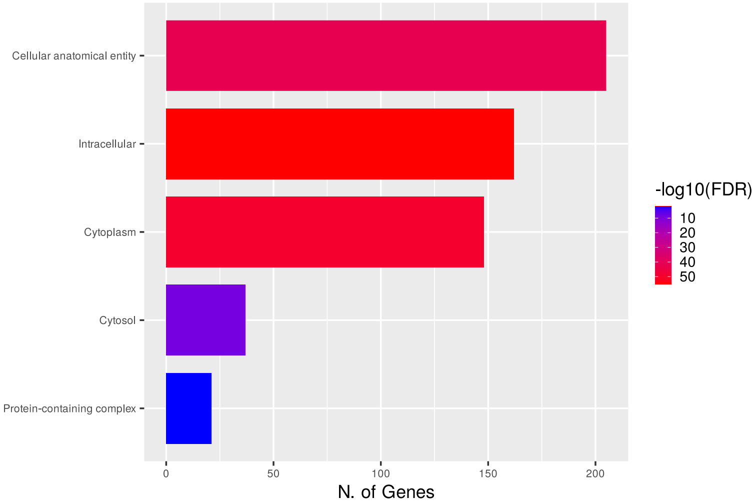


**Figure 3.32:** Top cellular processes in *B. dorei* in NR

In contrast, responders exhibited enrichment primarily in the Intracellular (FDR = 7.3963e$^{-59}$, nGenes = 178, fold enrichment = 2.92), Cytoplasm (FDR = 5.5915e$^{-57}$, nGenes = 167, fold enrichment = 3.10), and Cytosol (FDR = 1.9181e$^{-13}$, nGenes = 45, fold enrichment = 3.57) demonstrated in Figure 3.33.

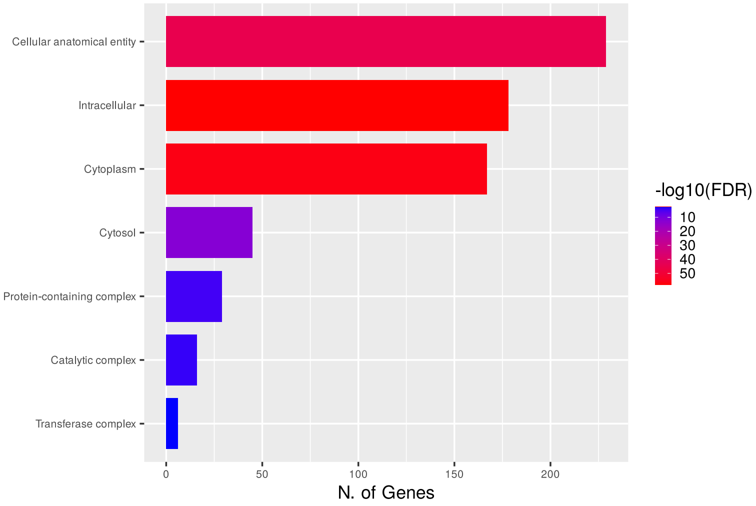


**Figure 3.33:** Top cellular processes in *B. dorei* in R

In NR, *B. stercoris* showed significant enrichment in the Intracellular (FDR = 5.3199e$^{-87}$, nGenes = 187, fold enrichment = 3.85), Cytoplasm (FDR = 5.5518e$^{-76}$, nGenes = 169, fold enrichment = 3.96), and Cytosol (FDR = 1.3326e$^{-15}$, nGenes = 43, fold enrichment = 4.23) shown in Figure 3.34.

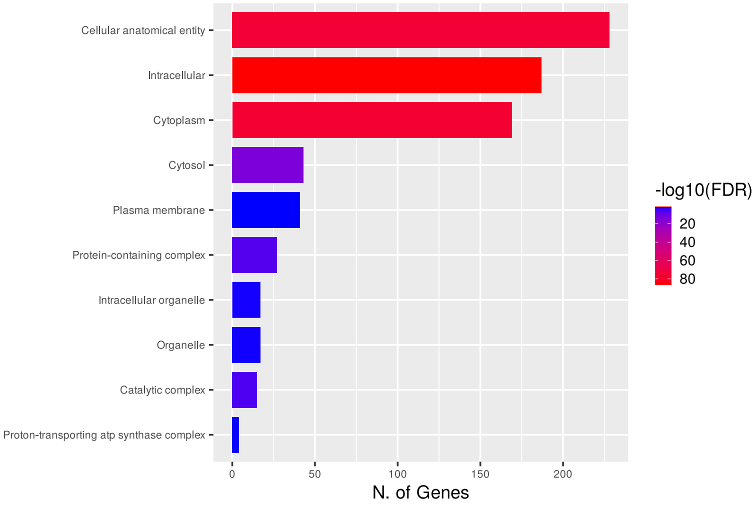


**Figure 3.34:** Top cellular processes in *B. stercoris* in NR

Conversely, R displayed enrichment primarily in the Intracellular (FDR = 4.1460e$^{-73}$, nGenes = 168, fold enrichment = 3.71), Cytoplasm (FDR = 1.4633e$^{-64}$, nGenes = 152, fold enrichment = 3.82), and Cytosol (FDR = 1.2369e$^{-17}$, nGenes = 44, fold enrichment = 4.65) illustrated in Figure 3.35.

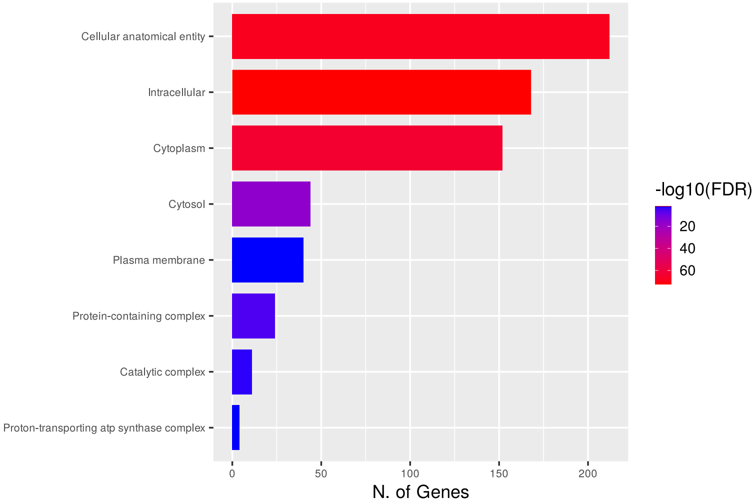


**Figure 3.35:** Top cellular processes in *B. stercoris* in R

In NR, B. uniformis showed significant enrichment in the Proton-transporting ATP synthase complex (FDR = 0.005247, nGenes = 6, fold enrichment = 4.82), Catalytic complex (FDR = 1.9876e$^{-07}$, nGenes = 29, fold enrichment = 3.03), and Cytosol (FDR = 1.5124e$^{-23}$, nGenes = 99, fold enrichment = 2.85) shown in Figure 3.36.

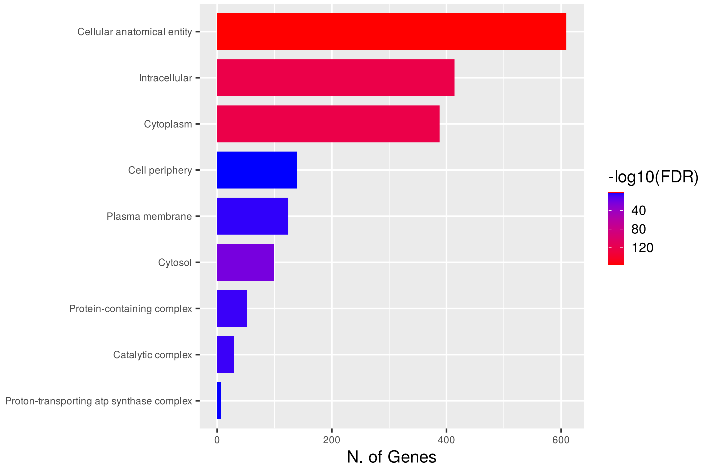


**Figure 3.36:** Top cellular processes in *B. uniformis* in NR

Conversely, R exhibited enrichment primarily in the Proton-transporting ATP synthase complex, catalytic core F(1) (FDR = 0.001341, nGenes = 5, fold enrichment = 6.98), Proton-transporting two-sector ATPase complex, catalytic domain (FDR = 0.008143, nGenes = 5, fold enrichment = 5.23), and Proton-transporting ATP synthase complex (FDR = 0.003583, nGenes = 6, fold enrichment = 5.03) demonstrated in Figure 3.37.

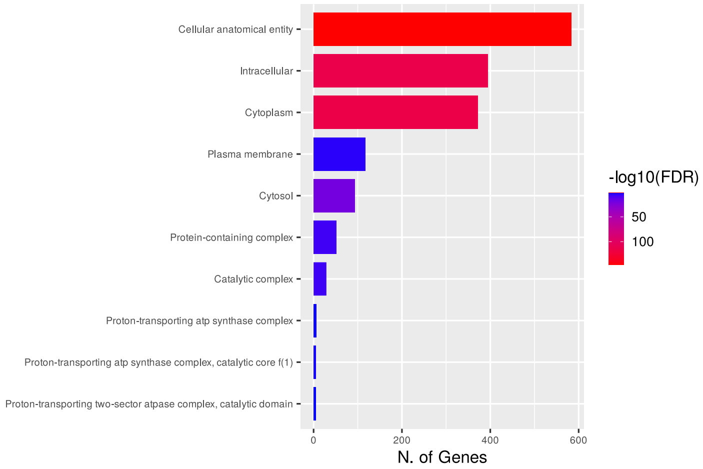


**Figure 3.37:** Top cellular processes in *B. uniformis* in R

In NR, B. vulgatus showed significant enrichment in the DNA repair complex (FDR = 0.004182, nGenes = 5, fold enrichment = 7.51), Cytosol (FDR = 4.3346e$^{-11}$, nGenes = 40, fold enrichment = 3.40), and Cytoplasm (FDR = 2.2362e$^{-62}$, nGenes = 170, fold enrichment = 3.17) shown in Figure 3.38.

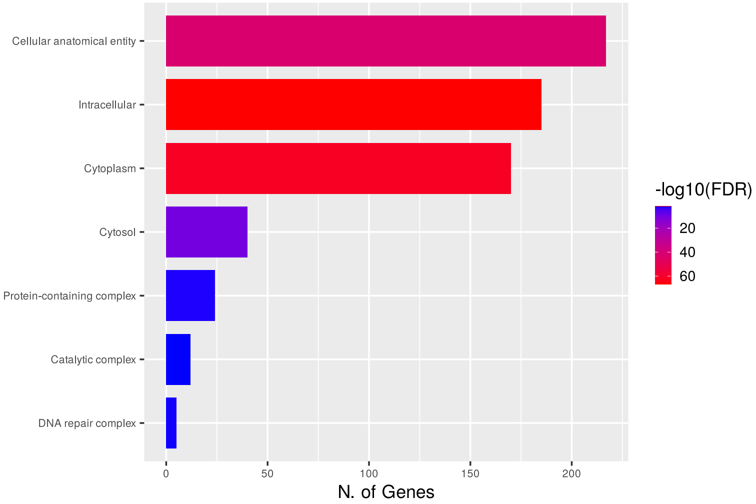


**Figure 3.38:** Top cellular processes in *B. vulgatus* in NR

Conversely, R exhibited enrichment primarily in the DNA repair complex (FDR = 0.004483, nGenes = 5, fold enrichment = 7.65), Cytosol (FDR = 2.2813e$^{-11}$, nGenes = 40, fold enrichment = 3.46), and Cytoplasm (FDR = 4.3766e$^{-57}$, nGenes = 163, fold enrichment = 3.09) illustrated in Figure 3.39.

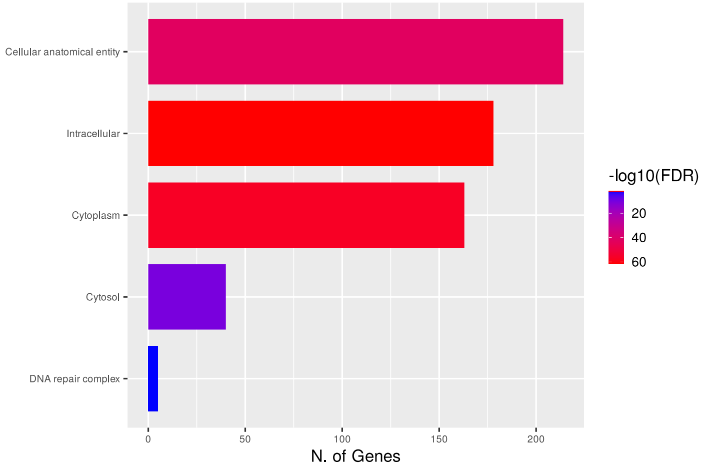


**Figure 3.39:** Top cellular processes in *B. vulgatus* in R

For *F. prausnitzii* in NR, significant enrichment was noted in the Proton-transporting two-sector ATPase complex, catalytic domain (FDR = 7.7376e$^{-06}$, nGenes = 4, fold enrichment = 35.84), Proton-transporting two-sector ATPase complex (FDR = 3.7194e$^{-06}$, nGenes = 5, fold enrichment = 25.20), and Proton-transporting ATP synthase complex (FDR = 3.7194e$^{-06}$, nGenes = 5, fold enrichment = 25.20) illustrated in Figure 3.40.

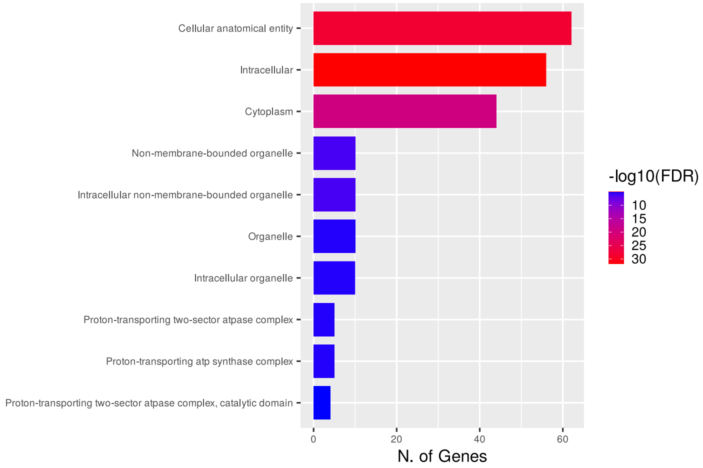


**Figure 3.40:** Top cellular processes in *F. prausnitzii* in NR

Conversely, R exhibited enrichment primarily in the Proton-transporting two-sector ATPase complex (FDR = 2.5088e$^{-05}$, nGenes = 5, fold enrichment = 17.96), Proton-transporting ATP synthase complex (FDR = 2.5088e$^{-05}$, nGenes = 5, fold enrichment = 17.96), and Ribosome (FDR = 1.868e$^{-06}$, nGenes = 9, fold enrichment = 9.40) shown in Figure 3.41.

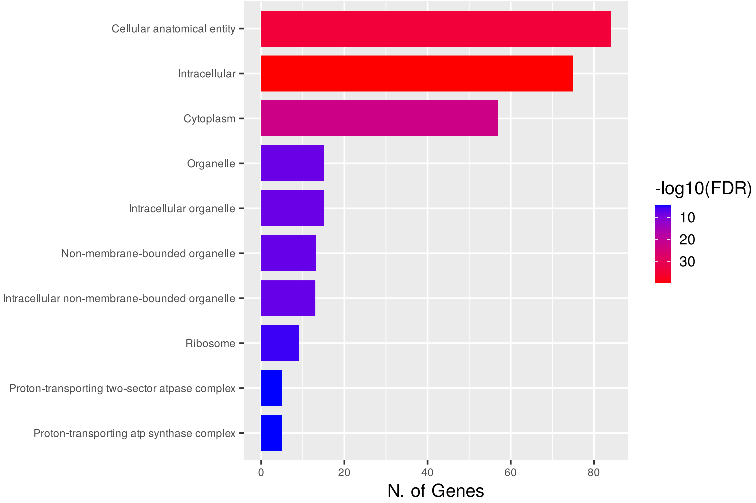


**Figure 3.41:** Top cellular processes in *F. prausnitzii* in R

Lastly, in NR, *P. distasonis* displayed significant enrichment in the Endodeoxyribonuclease complex (FDR = 0.00126, nGenes = 4, fold enrichment = 13.23), Endonuclease complex (FDR = 0.00179, nGenes = 4, fold enrichment = 11.90), and DNA repair complex (FDR = 0.00501, nGenes = 4, fold enrichment = 9.16) demonstrated in Figure 3.42.

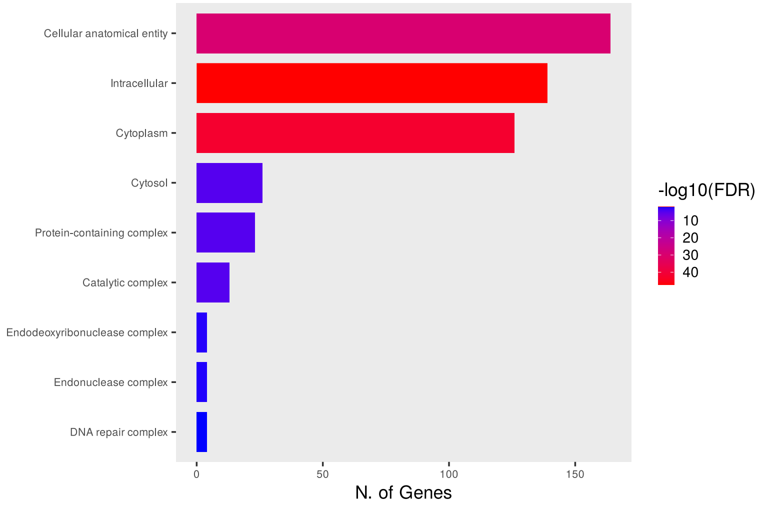


**Figure 3.42:** Top cellular processes in *P. distasonis* in NR

This section underscored the enrichment of genes in the distinct cellular component associated with different microbial species and their potential implications in treatment response.

### 3.5.3 *Molecular Functions*

In NR, *A. muciniphila* demonstrated significant enrichment in key molecular binding and catalytic activity functions. The top three most significant functions identified were Purine ribonucleotide binding (FDR = 5.6117e$^{-42}$, nGenes = 78, fold enrichment = 6.04), Nucleotide-binding (FDR = 4.8138e$^{-46}$, nGenes = 89, fold enrichment = 5.64), and Nucleoside phosphate binding (FDR = 4.8138e$^{-46}$, nGenes = 89, fold enrichment = 5.64) shown in Figure 3.43.

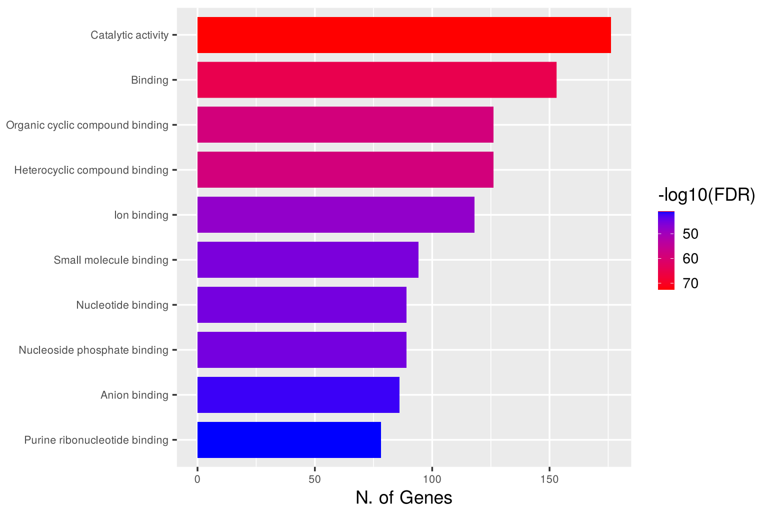


**Figure 3.43:** Top molecular functions in *A. muciniphila* in NR

In R, the most significant functions were Purine ribonucleotide binding (FDR = 7.2878e$^{-40}$, nGenes = 76, fold enrichment = 5.89), Nucleotide binding (FDR = 6.0487e$^{-44}$, nGenes = 87, fold enrichment = 5.51), and Nucleoside phosphate binding (FDR = 6.0487e$^{-44}$, nGenes = 87, fold enrichment = 5.51) illustrated in Figure 3.44.

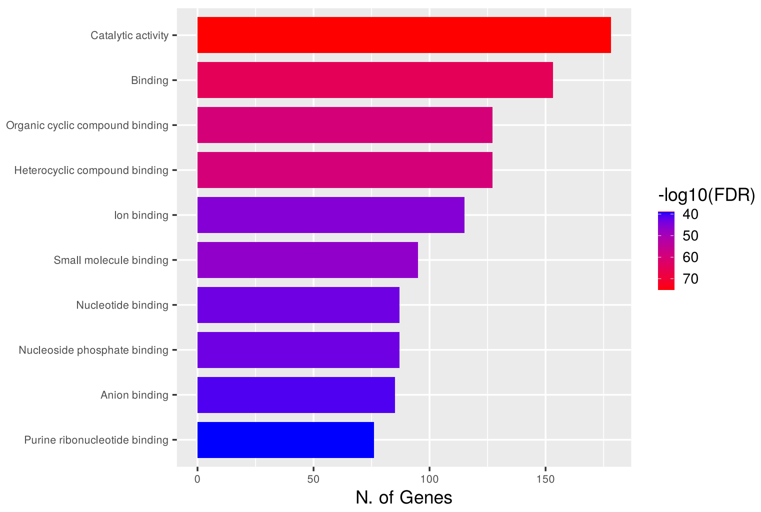


**Figure 3.44:** Top molecular functions in *A. muciniphila* in R

For NR, *B. dorei* was found in functions related to Ribonucleotide binding (FDR = 2.3051e$^{-21}$, nGenes = 74, fold enrichment = 3.34), Nucleotide binding (FDR = 7.0620e$^{-25}$, nGenes = 85, fold enrichment = 3.32), and Nucleoside phosphate binding (FDR = 7.0620e$^{-25}$, nGenes = 85, fold enrichment = 3.32) shown in Figure 3.45.

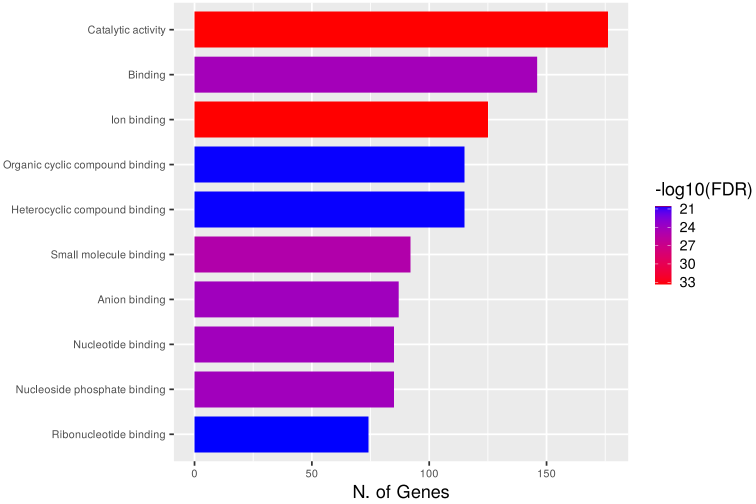


**Figure 3.45:** Top molecular functions in *B. dorei* in NR

In R, the significant functions included Nucleotide binding (FDR = 7.6106e$^{-27}$, nGenes = 94, fold enrichment = 3.26), Nucleoside phosphate binding (FDR = 7.6106e$^{-27}$, nGenes = 94, fold enrichment = 3.26), and Ribonucleotide binding (FDR = 1.9716e$^{-22}$, nGenes = 81, fold enrichment = 3.25) demonstrated in Figure 3.46.

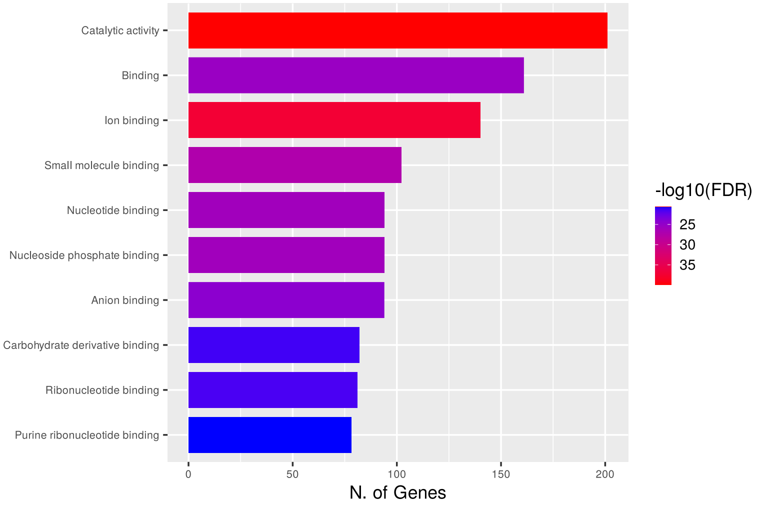


**Figure 3.46:** Top molecular functions in *B. dorei* in R

In R, *B. stercoris* showed enrichment in Nucleotide binding (FDR = 2.6162e$^{-27}$, nGenes = 82, fold enrichment = 3.74), Nucleoside phosphate binding (FDR = 2.6162e$^{-27}$, nGenes = 82, fold enrichment = 3.74), and small molecule binding (FDR = 8.4471e$^{-31}$, nGenes = 91, fold enrichment = 3.73) shown in Figure 3.47.

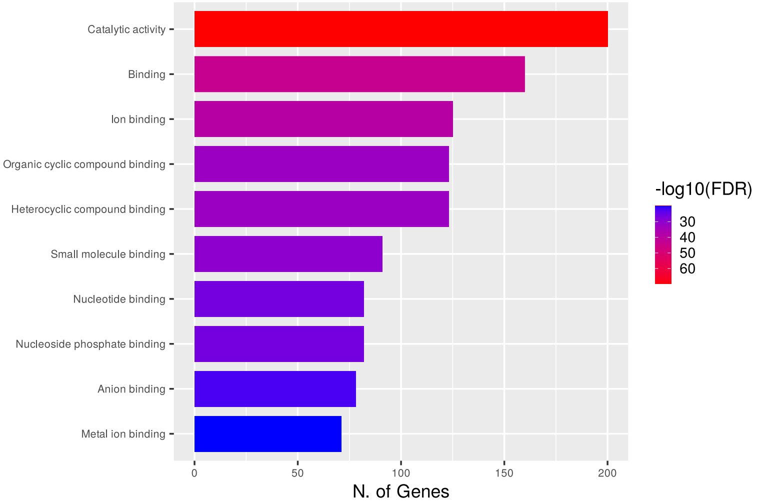


**Figure 3.47:** Top molecular functions in *B. stercoris* in NR

R exhibited significant pathways in small molecule binding (FDR = 1.177e$^{-27}$, nGenes = 83, fold enrichment = 3.71), Metal ion binding (FDR = 1.719e$^{-22}$, nGenes = 70, fold enrichment = 3.68), and Nucleotide binding (FDR = 6.358e$^{-24}$, nGenes = 74, fold enrichment = 3.67) illustrated in Figure 3.48.

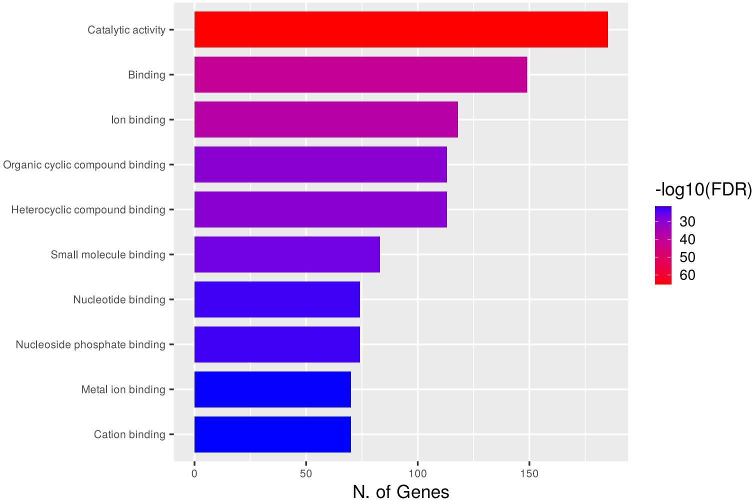


**Figure 3.48:** Top molecular functions in *B. stercoris* in R

In NR, *B. uniformis* exhibited significant functions in Cation binding (FDR = 6.581e$^{-39}$, nGenes = 171, fold enrichment = 2.70), Small molecule binding (FDR = 2.555e$^{-44}$, nGenes = 203, fold enrichment = 2.59), and Anion binding (FDR = 1.568e$^{-38}$, nGenes = 182, fold enrichment = 2.58) demonstrated in Figure 3.49.

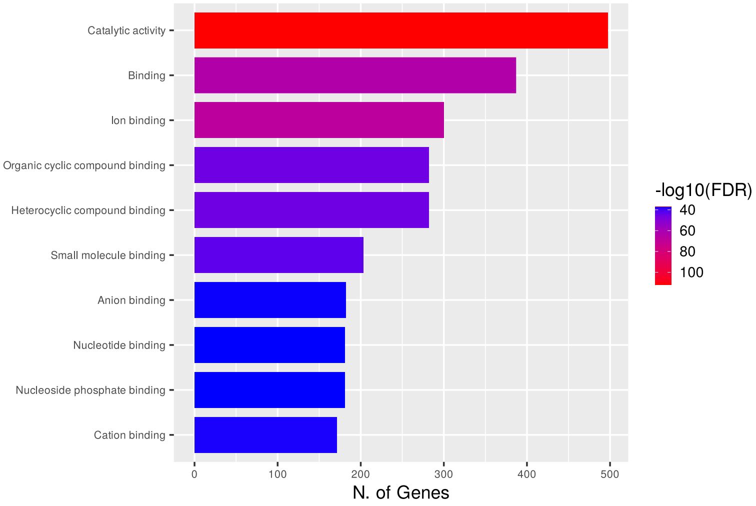


**Figure 3.49:** Top molecular functions in *B. uniformis* in NR

In R, the significant functions were Cation binding (FDR = 4.409e$^{-36}$, nGenes = 163, fold enrichment = 2.67), Metal ion binding (FDR = 7.205e$^{-36}$, nGenes = 162, fold enrichment = 2.67), and Anion binding (FDR = 7.599e$^{-35}$, nGenes = 172, fold enrichment = 2.53) shown in Figure 3.50.

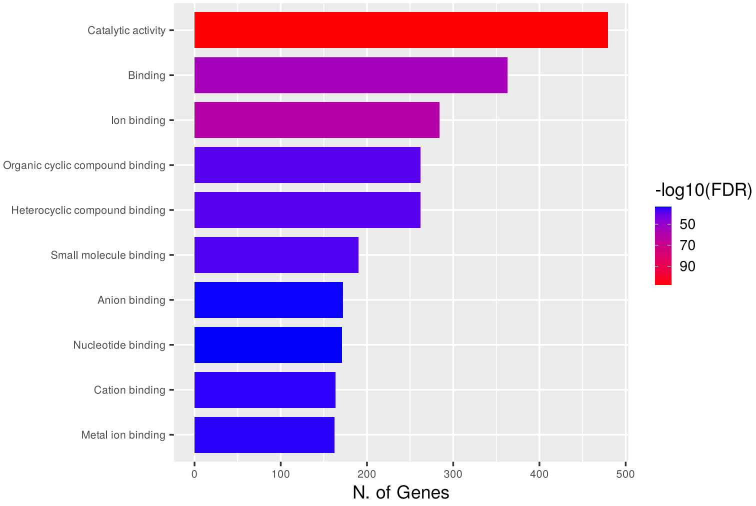


**Figure 3.50:** Top molecular functions in *B. uniformis* in R

For NR, *B. vulgatus* had the most significant functions in Purine nucleotide binding (FDR = $1.103e^{-29,}$ nGenes = 90, fold enrichment = 3.64), Purine ribonucleotide binding (FDR = $3.780e^{-29}$, nGenes = 89, fold enrichment = 3.61), and Purine ribonucleoside triphosphate binding (FDR = $3.780e^{-29}$, nGenes = 89, fold enrichment = 3.61), with notable enrichment in Nucleotide binding, Anion binding, and small molecule binding pathways as shown in Figure 3.51.

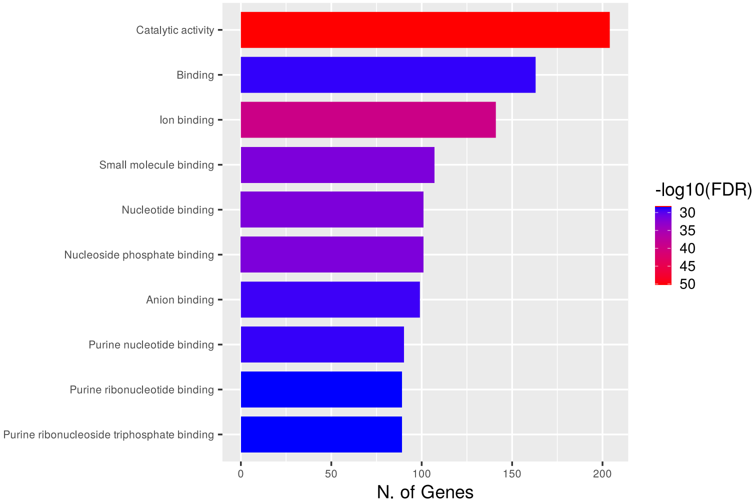


**Figure 3.51**: Top molecular functions in *B. vulgatus* in NR

R showed significant pathways in Purine nucleotide binding (FDR = $3.744e^{-24}$, nGenes = 82, fold enrichment = 3.36), Purine ribonucleotide binding (FDR = $1.451e^{-23}$, nGenes = 81, fold enrichment = 3.33), and Carbohydrate derivative binding (FDR = $3.162e^{-24}$, nGenes = 84, fold enrichment = 3.30), with substantial enrichment in Nucleotide-binding, Anion binding, and small molecule binding pathways as illustrated in Figure 3.52.

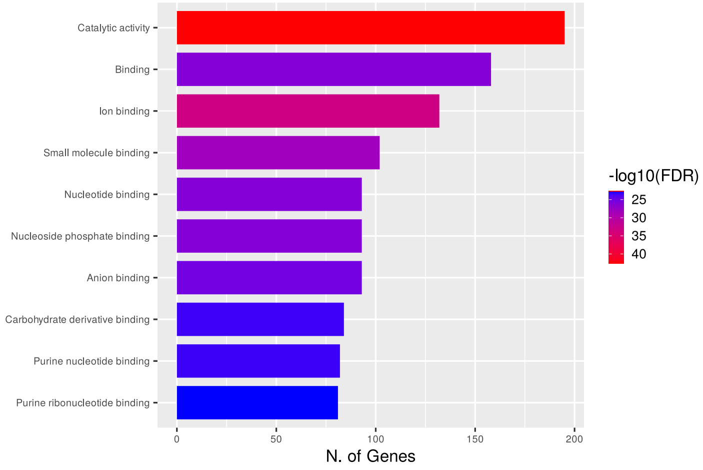


**Figure 3.52:** Top molecular functions in *B. vulgatus* in R

For NR, *F. prausnitzii* showed significant pathways in Purine nucleotide binding (FDR = 1.062e$^{-10}$, nGenes = 24, fold enrichment = 5.07), Purine ribonucleotide binding (FDR = 6.470e$^{-10}$, nGenes = 23, fold enrichment = 4.87), and Nucleotide binding (FDR = 7.661e$^{-11}$, nGenes = 26, fold enrichment = 4.71), with notable enrichment in Small molecule binding, Ion binding, and pathways associated with organic cyclic and heterocyclic compound binding, shown in Figure 3.53.

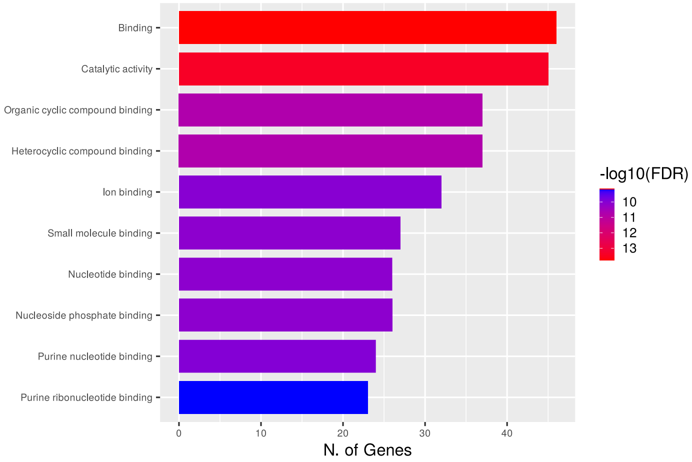


**Figure 3.53:** Top molecular functions in *F. prausnitzii* in NR

R had significant functions in Purine nucleotide binding (FDR = 3.711e$^{-14}$, nGenes = 33, fold enrichment = 4.93), Carbohydrate derivative binding (FDR = 2.960e$^{-14}$, nGenes = 34, fold enrichment = 4.81), and Purine ribonucleotide binding (FDR = 1.996e$^{-13}$, nGenes = 32, fold enrichment = 4.79), with enrichment in Nucleotide binding and pathways associated with organic cyclic and heterocyclic compound binding as shown in Figure 3.54.

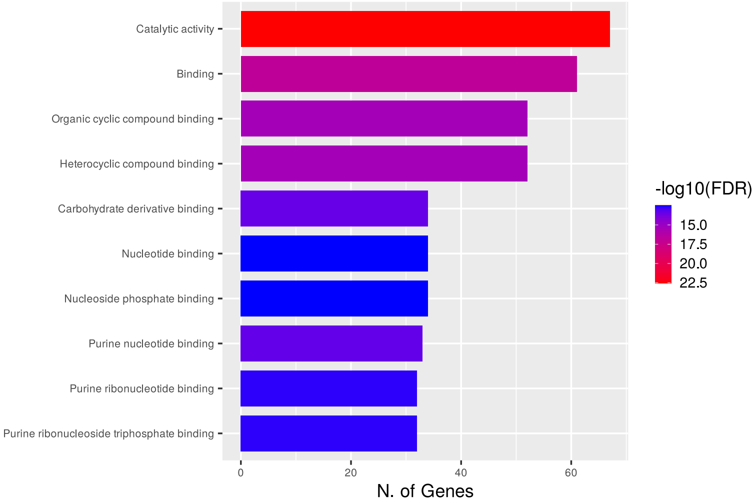


**Figure 3.54:** Top molecular functions in *F. prausnitzii* in R

For NR, *P. distasonis* exhibited significant enrichment in Carbohydrate derivative binding (FDR = 1.230e$^{-15}$, nGenes = 60, fold enrichment = 3.12), Nucleotide binding (FDR = 7.759e$^{-17}$, nGenes = 66, fold enrichment = 3.06), and small molecule binding (FDR = 1.883e$^{-17}$, nGenes = 71, fold enrichment = 2.96), with significant enrichment in pathways associated with anion binding and nucleoside phosphate binding, illustrated in Figure 3.55.

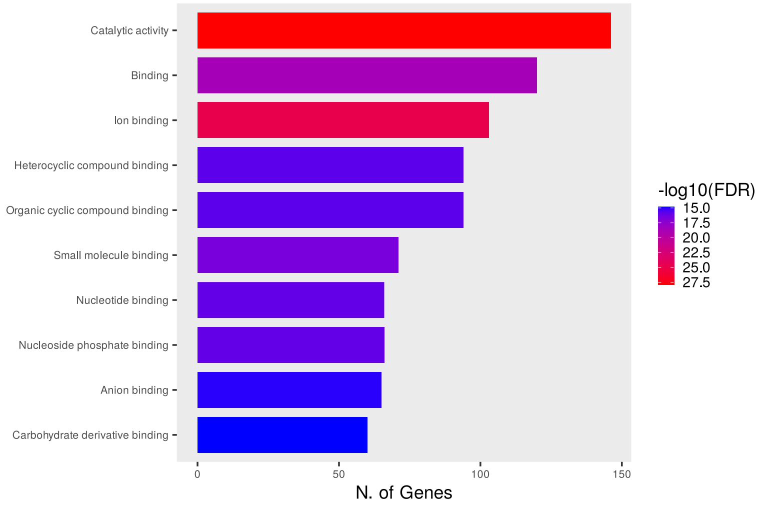


**Figure 3.55:** Top molecular functions in *P. distasonis* in NR

Across both NR and R, various bacterial species exhibited significant enrichment in molecular binding and catalytic activity functions. Key functions such as Purine ribonucleotide binding, Nucleotide-binding, and small molecule binding were consistently significant. Differences in the fold enrichment and the number of genes involved highlight the variations between the NR and R.

## 3.6 Statistical Analysis

### 3.6.1 *Alpha Diversity of Species, Genes Counts & Genetic Variants*

The α diversity of species was analyzed with response (R and NR) across three different time points (T0, T1, and T2). R demonstrated an increase in α diversity from T0 (3.386) to a peak at T1 (3.864), followed by a slight decrease at T2 (3.718). In contrast, NR exhibited more stable alpha diversity, with a slight increase at T1 (3.564) from T0 (3.347) and a subsequent decline at T2 (3.270). Overall, R had higher α diversity compared to NR across the observed time points, with mean values of 3.578 for R and 3.417 for NR. Figure 3.56 illustrates the α diversity in R and NR.

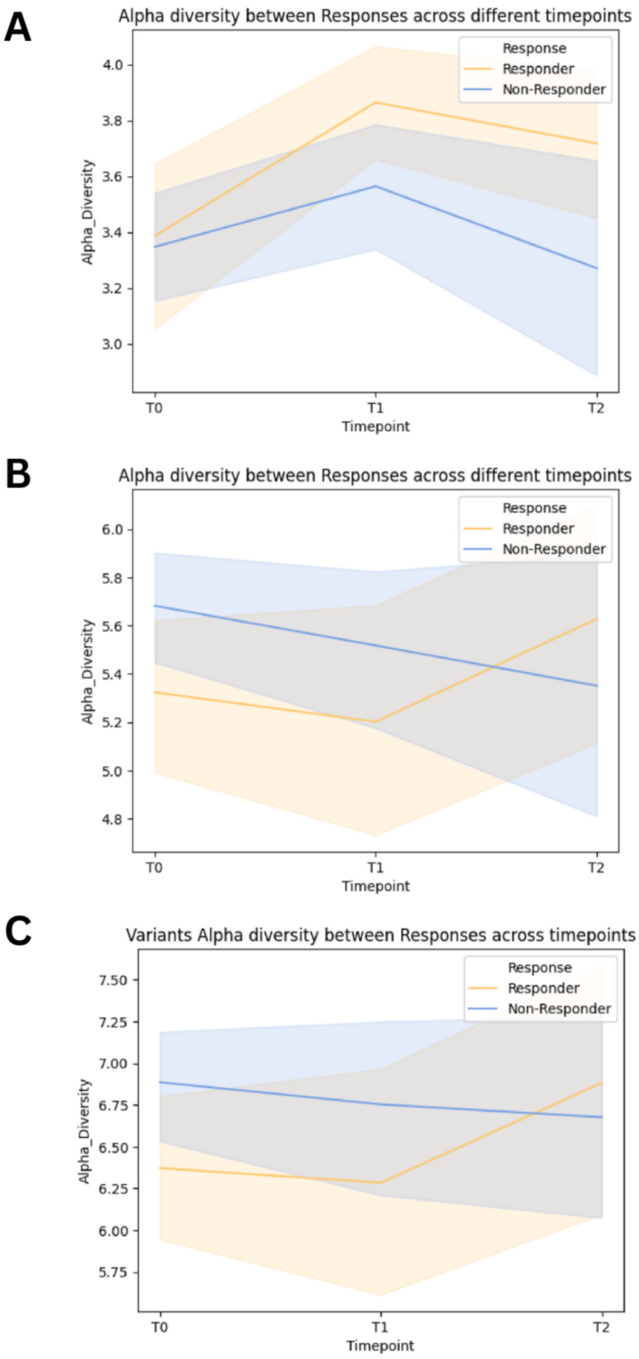


**Figure 3.56:** α diversity of **A.** microbial species; **B.** gene counts; and **C.** genetic variants in R vs. NR

The α diversity of gene counts was assessed with response status across different time points. R showed a decrease in α diversity from T0 (5.324) to T1 (5.202), followed by an increase at T2 (5.628). Conversely, NR exhibited a steady decline in α diversity from T0 (5.683) to T1 (5.518) and further to T2 (5.351). Across all time points, NR had a higher mean α diversity (5.585) compared to R (5.347).

The α diversity of genetic variants was evaluated concerning response status over time points. R showed a decrease in α diversity from T0 (6.372) to T1 (6.284), followed by an increase at T2 (6.883). In contrast, NR exhibited a slight decline in α diversity from T0 (6.886) to T1 (6.753) and further to T2 (6.677). Overall, NR maintained a higher mean α diversity (6.814) compared to R (6.436) throughout the timepoints.

### 3.6.2 *Median Comparison of Species, Genes Count & Genetic Variants Using Wilcoxon Test*

The α diversity data of species was analysed by response status (R and NR) across time points (T0, T1, and T2). R showed an increase in α diversity from T0 to T1, followed by a slight decrease at T2, while NR maintained a more stable pattern. Despite these trends, the Wilcoxon test results indicated that the differences in α diversity between R and NR were not statistically significant, with p-values of 0.4082, 0.0833, and 0.181 for the respective time points. Therefore, although there are observable differences in trends, they do not reach statistical significance as shown in Figure 3.57.

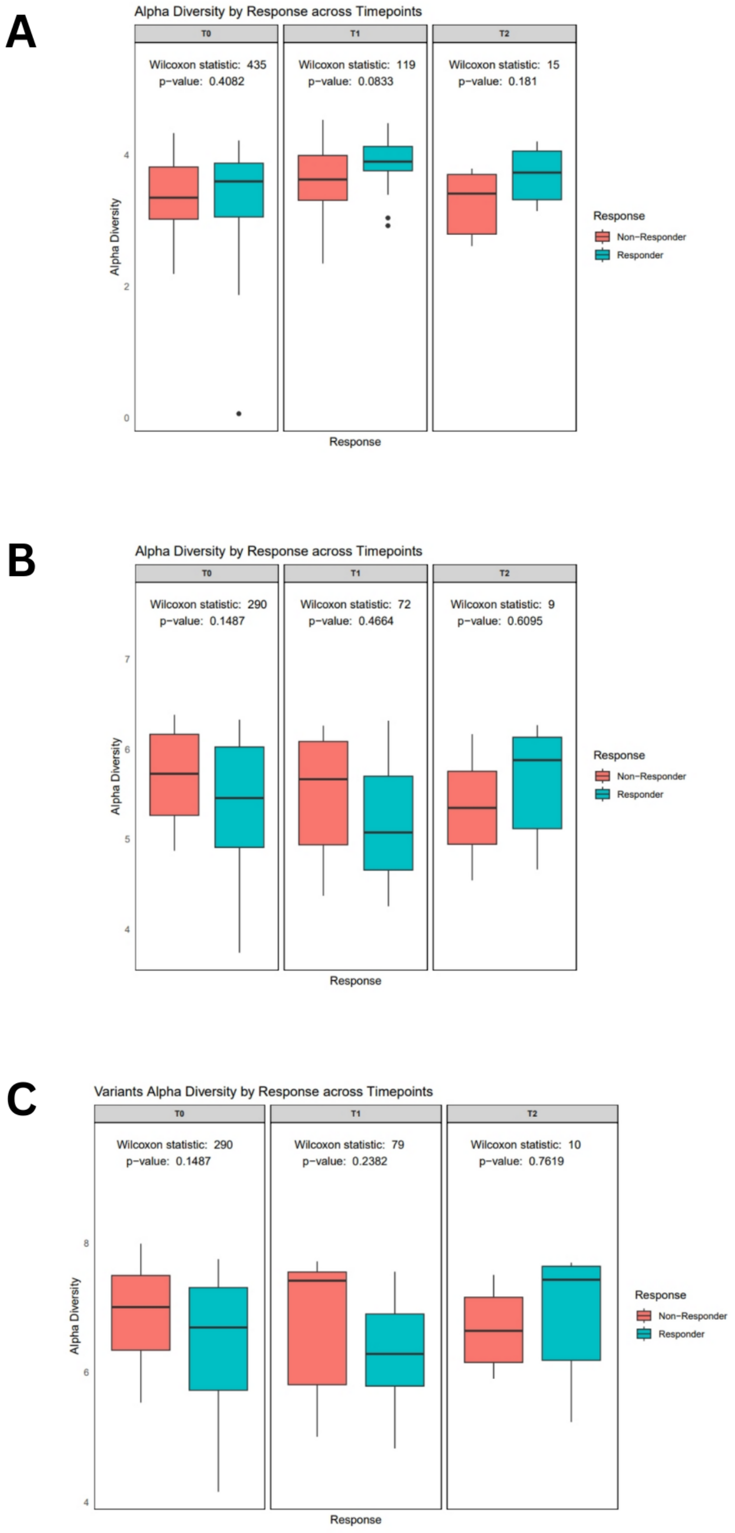


**Figure 3.57:** Wilcoxon test on α diversity of **A.** Species B. gene counts **C.** genetic variants

The α diversity of gene counts was examined by response across time points. The Wilcoxon test showed that R experienced a decrease in α diversity from T0 to T1, followed by an increase at T2, while NR exhibited a steady decline. However, the differences in α diversity between R and NR were not statistically significant, with p-values of 0.1487, 0.4664, and 0.6095 for the respective time points. These findings suggest that the trends in α diversity, though observable, do not reach statistical significance.

The α diversity of genetic variants was analysed by response across time points. R showed a decrease in α diversity from T0 to T1, followed by an increase at T2. In contrast, NR exhibited a slight decline from T0 to T1, with a further decrease at T2. The Wilcoxon test results indicated that the differences in α diversity between R and NR were not statistically significant, with p-values of 0.1487, 0.2382, and 0.7619 for the respective time points. Thus, while there are noticeable trends, they do not reach statistical significance.

### 3.6.3 *Associations Using MaAsLin2*

The analysis of gene associations with response (R vs. NR) revealed several trends in gene expression differences. In R, the expression of *lpdA* was significantly lower compared to NR, with a coefficient of -0.967 and an FDR of 0.2335. Similarly, *nadB* exhibited reduced expression in R, with a coefficient of -0.591 and an FDR of 0.2335. The genes *sufD* and *uxaC* also showed decreased expression in R, with coefficients of -1.14 and -0.952 respectively, both having an FDR of 0.2335. Additionally, *ftsA*, *obgE*, *rhaT*, and *xylE* also showed lower expression in R, with coefficients of -0.586, -0.363, -0.521, and -0.654, and FDR values around 0.2354. Hence, responders generally exhibit lower expression levels of these genes compared to NR. Figure 3.58 shows the association of genes with treatment response.

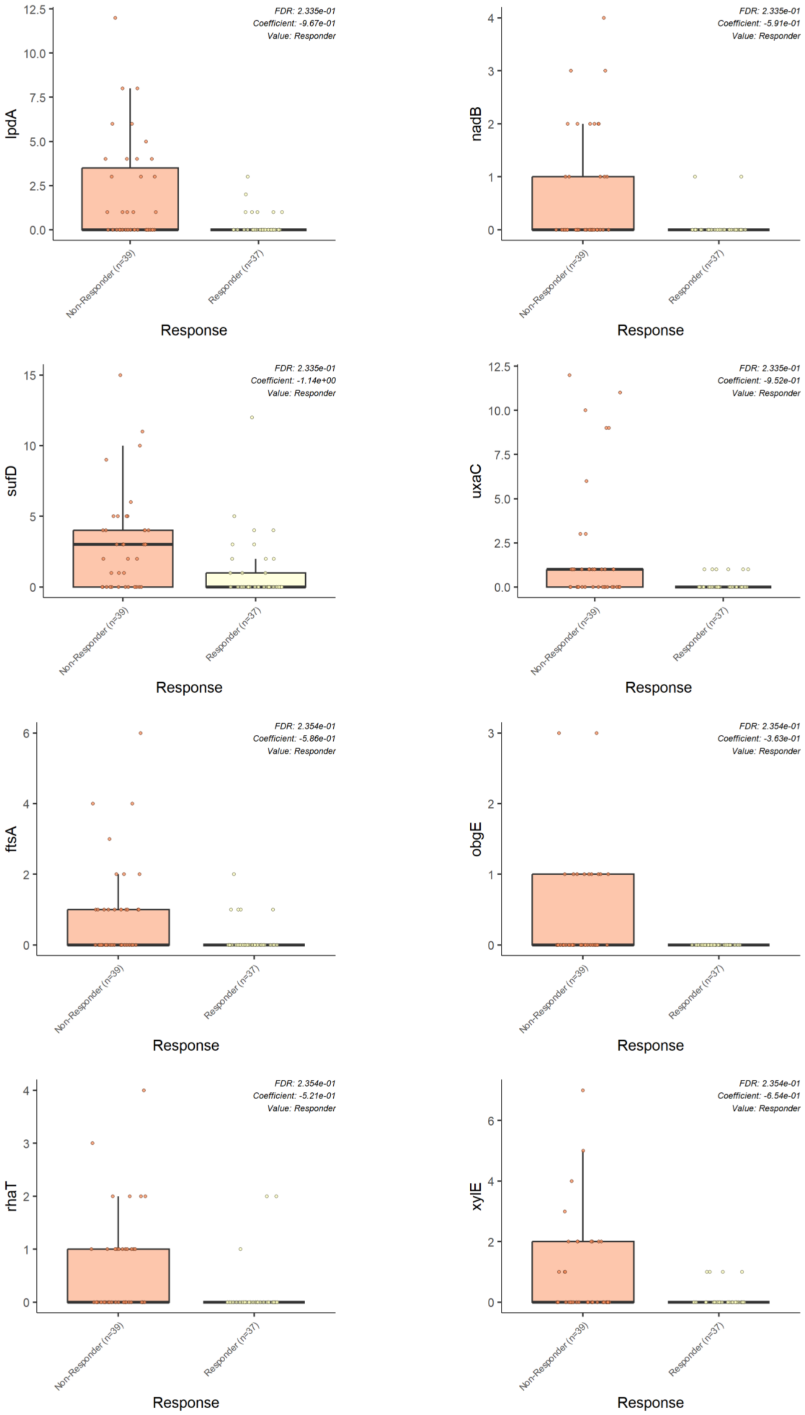


**Figure 3.58:** Genes identified by MaAsLin2 associated with treatment response

## 3.7 Association via Machine Learning Models

### *3.7.1 Evaluation Through ML Models*

The RF classifier showed robust performance with an accuracy of 76%, precision of 76%, recall of 79%, and an F1-score of 77%, effectively distinguishing between R and NR. The model's ROC AUC of 0.83 further underscores its excellent discriminative capability. Noteworthy genes identified include *sufD*, *uxaC*, and *nadB*, indicating their significant roles in response mechanisms. Additionally, critical variants such as missense variant c.817T>C p.Cys273Arg and missense variant c.697T>C p.Tyr233His were highlighted. Figure 3.59 shows the performance of the RF classifier via the ROC curve.

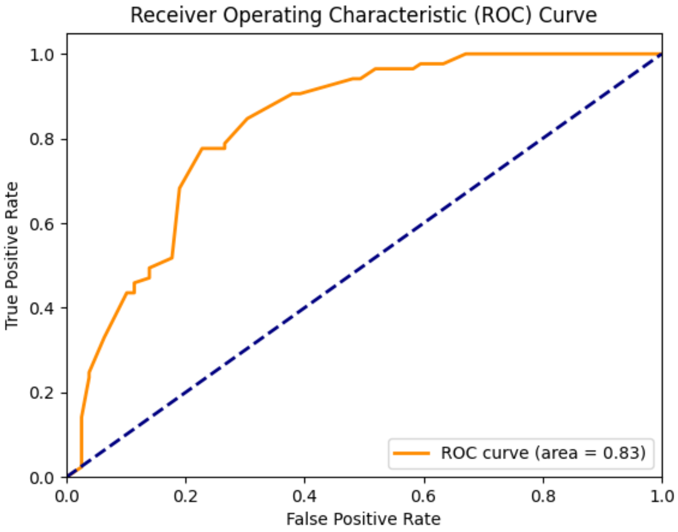


**Figure 3.59:** RF classifier model performance

The Logistic Regression model demonstrated good performance in classifying R and NR, with an accuracy of 68%, precision of 65%, recall of 84%, and an F1-score of 73%. The model's ROC AUC of 0.75 indicates a high level of discriminative capability.

Important variants identified include missense variant c.817T>C p.Cys273Arg, missense variant c.697T>C p.Tyr233His, and missense variant c.931G>A p.Ala311Thr. Figure 3.60 shows the performance of the logistic regression model via ROC curve. The identification of these variants provides valuable insights for understanding the genetic basis of treatment response and could inform targeted therapeutic strategies.

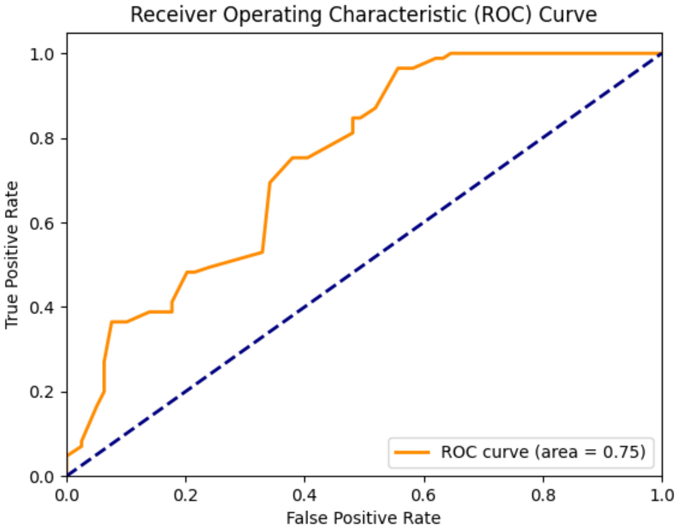


**Figure 3.60:** Logistic regression model performance

The SVM model demonstrated strong performance with an accuracy of 71%, precision of 69%, recall of 79%, and an F1-score of 74%. The ROC AUC of 0.76 indicates good discriminative ability. Significant genes identified include *sufD*, underscoring its relevance in treatment response. Key variants such as missense variant c.931G>A p.Ala311Thr and missense variant c.698A>G p.His233Arg were also highlighted, suggesting their crucial roles in distinguishing between R and NR. Figure 3.61 shows the performance of the SVM model via the ROC curve.

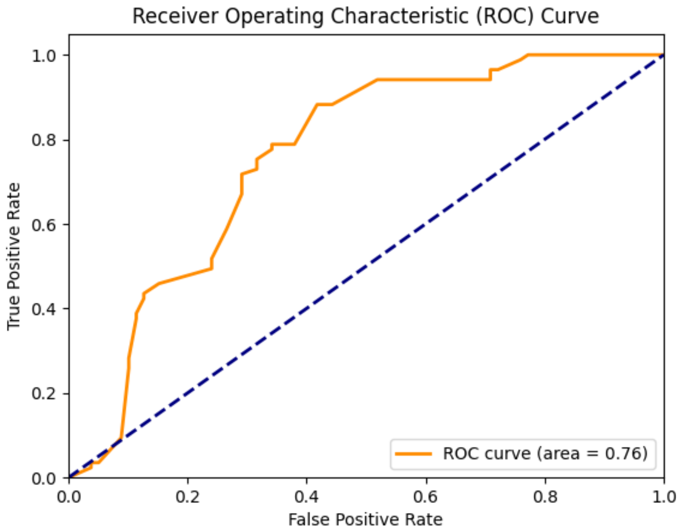


**Figure 3.61:** SVM model performance

The XGBoost model demonstrated outstanding performance with an accuracy of 78%, precision of 76%, recall of 84%, and an F1-score of 80%. The ROC AUC of 0.83 indicates exceptional discriminative ability. Significant genes identified include *obgE*, *nadB*, and *rhaT*, highlighting their relevance in predicting treatment response. Key variants such as missense variant c.698A>G p.His233Arg, missense variant c.931G>A p.Ala311Thr, and missense variant c.817T>C p.Cys273Arg were also identified, underscoring their significant roles in classification. Figure 3.62 illustrates the performance of the XGBoost model via the ROC curve.

The Decision Tree model performed well, achieving an accuracy of 84%, precision of 77%, recall of 85%, and an F1-score of 80%. The ROC AUC of 0.84 indicates strong discriminative ability. Significant genes identified include *sufD*, rhaT, and *xylE*, highlighting their relevance in treatment response prediction.

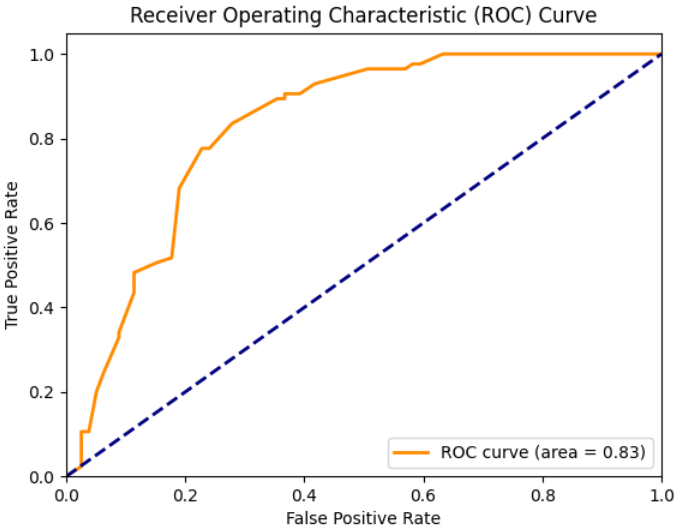


**Figure 3.62:** XGBoost model performance

Key variants such as missense variant c.931G>A p.Ala311Thr, missense variant c.989T>C p.Leu330Ser, and missense variant c.503C>T p.Ala168Val were also identified. Figure 3.63 demonstrates the performance of the Decision tree model via the ROC curve.

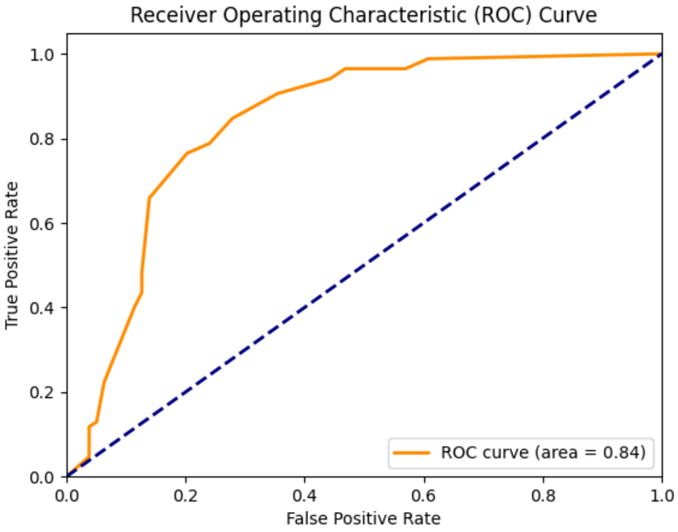


**Figure 3.63:** Decision tree model performance

The GBM model demonstrated strong performance with an accuracy of 79%, precision of 77%, recall of 84%, and an F1-score of 80%. The ROC AUC of 0.85 indicates excellent discriminative capability. Key genes identified include *sufD*, *rhaT*, and *xylE*, highlighting their relevance in predicting treatment response. Significant variants such as missense variant c.503C>T p.Ala168Val, missense variant c.989T>C p.Leu330Ser, and missense variant c.931G>A p.Ala311Thr were also identified. Figure 3.64 shows the performance of the GBM model via ROC curve.

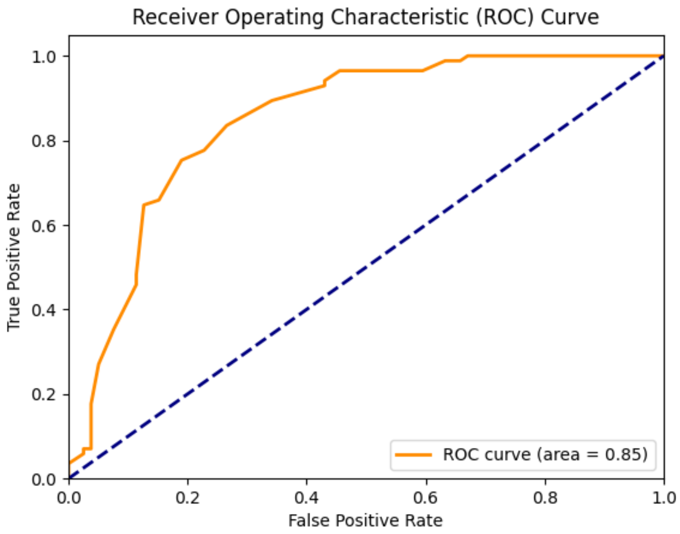


**Figure 3.64:** GBM model performance

GBM and DT emerged as the top-performing models with ROC AUC values of 0.85 and 0.84, respectively. Key genes consistently identified across multiple models include *sufD*, *rhaT*, and *xylE*. Significant variants such as missense variant c.503C>T p.Ala168Val, missense variant c.931G>A p.Ala311Thr, and missense variant c.697T>C p.Tyr233His were pivotal in distinguishing between R and NR. These genetic markers provide critical insights for developing targeted therapeutic strategies and advancing our understanding of treatment response mechanisms.

**Table 3.5:** Performance matrix of ML models

| | Non-Responders | | | Responders | | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Precision** | **Recall** | **F1 - Score** | **Precision** | **Recall** | **F1 - Score** | **Accuracy** | **AUC** |
| Random Forest | 0.76 | 0.73 | 0.75 | 0.76 | 0.77 | 0.76 | 0.76 | 0.83 |
| Logistic regression | 0.75 | 0.52 | 0.61 | 0.65 | 0.84 | 0.73 | 0.68 | 0.75 |
| SVM | 0.73 | 0.62 | 0.67 | 0.69 | 0.79 | 0.74 | 0.71 | 0.76 |
| XGBoost | 0.80 | 0.72 | 0.76 | 0.76 | 0.84 | 0.80 | 0.78 | 0.83 |
| Decision tree | 0.81 | 0.72 | 0.77 | 0.77 | 0.85 | 0.80 | 0.79 | 0.84 |
| GBM | 0.81 | 0.73 | 0.77 | 0.77 | 0.84 | 0.80 | 0.79 | 0.85 |

### 3.8 Functional Annotation of Genes

The gene *lpdA* encodes a protein involved in the oxidative decarboxylation of pyruvate and other alpha-keto acids. It plays a critical role in cellular respiration and energy production. In the context of cancer, alterations in metabolic pathways, including those involving *lpdA*, can contribute to tumor growth and survival. This gene is involved in the pyruvate metabolism pathway and is present in various bacterial species, highlighting its role in fundamental metabolic processes across different organisms. *nadB* gene is responsible for the biosynthesis of NAD (nicotinamide adenine dinucleotide), a crucial coenzyme in redox reactions. NAD is essential for cellular metabolism and energy production. In cancer, NAD metabolism is often dysregulated, leading to changes in cellular energy homeostasis and redox balance, which can promote cancer cell proliferation and survival. The *nadB* gene is part of the NAD biosynthesis pathway and is found in several bacterial species, indicating its conserved role in essential cellular functions.

The *sufD* gene is involved in the assembly of iron-sulfur clusters, which are vital cofactors for numerous enzymes. Iron-sulfur clusters play a significant role in various cellular processes, including DNA repair, respiration, and metabolic pathways. Dysregulation of iron-sulfur cluster assembly can impact mitochondrial function and genomic stability, contributing to cancer development. The *sufD* gene is part of the iron-sulfur cluster assembly pathway and is present in a range of bacterial species. The *uxaC* gene encodes an enzyme involved in the degradation of uronic acids, which are components of complex carbohydrates. This metabolic pathway is crucial for the utilization of plant-derived polysaccharides. While *uxaC* itself is not directly implicated in cancer, alterations in carbohydrate metabolism and the tumor microenvironment can influence cancer progression. The *uxaC* gene is found in bacterial species that metabolize plant-derived sugars.

The *ftsA* gene is a key component of the bacterial cell division machinery. It plays a critical role in the formation of the divisome, a protein complex essential for bacterial cytokinesis. While *ftsA* is not directly related to human cancer, studying bacterial cell division genes can provide insights into fundamental biological processes and potential antimicrobial targets. The *ftsA* gene is present in various bacterial species involved in cell division. The *obgE* gene encodes a GTP-binding protein involved in various cellular processes, including ribosome assembly, stress response, and cell cycle regulation. In cancer, dysregulation of cell cycle and stress response pathways can contribute to tumor growth and resistance to therapy. The *obgE* gene is part of stress response pathways and is found in numerous bacterial species, indicating its role in fundamental cellular processes.

The *rhaT* gene is involved in the transport and metabolism of rhamnose, a sugar found in plant cell walls. While *rhaT* itself is not directly linked to cancer, changes in sugar metabolism and the availability of nutrients in the tumor microenvironment can influence cancer progression. The *rhaT* gene is present in bacterial species that metabolize plant-derived sugars. The *xylE* gene encodes an enzyme involved in the degradation of xylose, a sugar found in hemicellulose from plant biomass. Alterations in sugar metabolism pathways can affect the tumor microenvironment and cancer cell metabolism. The *xylE* gene is part of the xylose degradation pathway and is found in bacterial species that utilize plant-derived sugars.

Addressing the treatment response, the genes *lpdA*, *nadB*, *sufD*, *obgE*, and *xylE* show significant associations with R and NR categories. Alterations in these genes' metabolic pathways might be indicative of their roles in influencing treatment outcomes, potentially contributing to cancer cell proliferation, survival, and response to therapy.

# CHAPTER 4: CONCLUSIONS AND FUTURE RECOMMENDATION

The comprehensive analysis of microbial species, strains, genetic variants, and genes in R and NR to treatment has provided valuable insights into the complex interactions within the gut microbiome and their potential impact on treatment efficacy.

In R, the most abundant microbial species included *P. vulgatus*, *B. uniformis*, *F. prausnitzii*, *P. dorei*, and *A. muciniphila*. These species are known for their beneficial roles in gut health and immune modulation, which might contribute to a favorable response to treatment. In contrast, NR showed higher abundances of *P. distasonis* and *P. dorei*, species that may be associated with a less favorable treatment response. Strain diversity profiling revealed that certain strains were consistently present across all time points in both R and NR, such as *P. dorei*, *B. uniformis*, and *F. prausnitzii*, indicating their stable role in the microbiome. Unique strains to R included *L. eligens*, *E. coli*, and *P. merdae*, whereas *P. distasonis* and *B. fragilis* were specific to NR. These findings suggest potential microbial markers for predicting treatment efficacy.

Among the seven most abundant common species across all time points; *A. muciniphila*, *B. dorei*, *B. stercoris*, *B. uniformis*, *F. prausnitzii*, *P. distasonis*, and *P. vulgatus*, NR exhibited a higher number of genetic variations with 47,969 compared to R containing 35,615. This indicates a possible link between genetic diversity and treatment response. Specifically, genes such as *ftsA*, *lpdA*, *nadB*, *obgE*, *rhaT*, *sufD*, *uxaC*, and *xylE* were significantly associated with treatment response, showing distinct patterns of variation between the two groups. These genes are involved in critical biological processes, including cell division, metabolic pathways, and stress responses, which could influence treatment outcomes. Machine learning models, particularly GBM with AUC of 85% and DT having AUC of 84%, effectively identified key genes and variants associated with treatment response, providing robust predictive power with high accuracy, precision, and recall. Important genes such as *sufD*, *rhaT*, and *xylE*, along with significant variants like missense variant c.817T>C p.Cys273Arg, missense variant c.503C>T p.Ala168Val, missense variant c.698A>G p.His233Arg, missense variant c.680C>T p.Thr227Ile, missense variant c.697T>C p.Tyr233His, and missense variant cc.931G>A p.Ala311Thr, were consistently highlighted, underscoring their relevance in distinguishing between R and NR.

Future research should focus on the insights gained from the identified microbial species, strains, and genetic variants to enhance treatment outcomes for NSCLC patients. Longitudinal studies with larger cohorts are essential to validate these microbial and genetic markers, enabling a deeper understanding of their dynamic changes over time and their influence on treatment response. Experimental validation of the functional roles of identified genes and microbial strains, particularly those consistently present across all time points such as *P. dorei*, *B. uniformis*, and *F. prausnitzii*, will provide crucial insights into their mechanisms of action and potential therapeutic targets.

Microbiome engineering approaches, such as probiotics or faecal microbiota transplantation, should be explored to modulate the gut microbiome composition in favor of beneficial species like *A. muciniphila*, which could enhance treatment efficacy in NR. By altering the gut microbiota to support beneficial strains, it may be possible to shift the microbial balance toward a state that supports better treatment outcomes. Integrating microbial and genetic profiling into clinical practice can pave the way for personalized treatment strategies, allowing for tailored interventions based on an individual's microbiome and genetic makeup. For instance, profiling the presence of specific genetic variants such as missense variant c.817T>C p.Cys273Arg or missense variant c.503C>T p.Ala168Val could guide the customization of therapeutic approaches to improve patient outcomes.

Employing multi-omics approaches, which combine genomics, transcriptomics, proteomics, and metabolomics, will offer a comprehensive view of the biological processes involved in treatment response and help identify additional therapeutic targets. This holistic approach can reveal how different layers of biological information interact and contribute to treatment response, providing a more complete picture of the underlying mechanisms. By focusing on these areas, future research can build on the current findings to develop more effective treatments and improve outcomes for patients undergoing immunotherapy and other cancer treatments in NSCLC.

# REFERENCES

Abdelsalam, N. A., Hegazy, S. M., & Aziz, R. K. (n.d.-a). Elud. *Gut Microbes*, *15*(2), 2249152. https://doi.org/10.1080/19490976.2023.2249152

Abdelsalam, N. A., Hegazy, S. M., & Aziz, R. K. (n.d.-b). The curious case of Prevotella copri. *Gut Microbes*, *15*(2), 2249152. https://doi.org/10.1080/19490976.2023.2249152

Ağagündüz, D., Cocozza, E., Cemali, Ö., Bayazıt, A. D., Nanì, M. F., Cerqua, I., Morgillo, F., Saygılı, S. K., Berni Canani, R., Amero, P., & Capasso, R. (2023). Understanding the role of the gut microbiome in gastrointestinal cancer: A review. *Frontiers in Pharmacology*, *14*, 1130562. https://doi.org/10.3389/fphar.2023.1130562

Bai, J., Gao, Z., Li, X., Dong, L., Han, W., & Nie, J. (2017). Regulation of PD-1/PD-L1 pathway and resistance to PD-1/PD-L1 blockade. *Oncotarget*, *8*(66), 110693–110707. https://doi.org/10.18632/oncotarget.22690

Brunner-Weinzierl, M. C., & Rudd, C. E. (2018). CTLA-4 and PD-1 Control of T-Cell Motility and Migration: Implications for Tumor Immunotherapy. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.02737

Chamoto, K., Hatae, R., & Honjo, T. (2020). Current issues and perspectives in PD-1 blockade cancer immunotherapy. *International Journal of Clinical Oncology*, *25*(5), 790–800. https://doi.org/10.1007/s10147-019-01588-7

Cheng, W. Y., Wu, C.-Y., & Yu, J. (2020). The role of gut microbiota in cancer treatment: Friend or foe? *Gut*, *69*(10), 1867–1876. https://doi.org/10.1136/gutjnl-2020-321153

Cheng, X., Wang, J., Gong, L., Dong, Y., Shou, J., Pan, H., Yu, Z., & Fang, Y. (2022). Composition of the Gut Microbiota Associated with the Response to Immunotherapy in Advanced Cancer Patients: A Chinese Real-World Pilot Study. *Journal of Clinical Medicine*, *11*(18), 5479. https://doi.org/10.3390/jcm11185479

Ehudin, M. A., Golla, U., Trivedi, D., Potlakayala, S. D., Rudrabhatla, S. V., Desai, D., Dovat, S., Claxton, D., & Sharma, A. (2022). Therapeutic Benefits of Selenium in Hematological Malignancies. *International Journal of Molecular Sciences*, *23*(14), 7972. https://doi.org/10.3390/ijms23147972

Escors, D., Gato-Cañas, M., Zuazo, M., Arasanz, H., García-Granda, M. J., Vera, R., & Kochan, G. (2018). The intracellular signalosome of PD-L1 in cancer cells. *Signal Transduction and Targeted Therapy*, *3*(1), 1–9. https://doi.org/10.1038/s41392-018-0022-9

Garg, S., Sharma, N., Bharmjeet, & Das, A. (2023). Unraveling the intricate relationship: Influence of microbiome on the host immune system in carcinogenesis. *Cancer Reports*, *6*(11), e1892. https://doi.org/10.1002/cnr2.1892

Gong, J., Chehrazi-Raffle, A., Reddi, S., & Salgia, R. (2018). Development of PD-1 and PD-L1 inhibitors as a form of cancer immunotherapy: A comprehensive review of registration trials and future considerations. *Journal for ImmunoTherapy of Cancer*, *6*(1), 8. https://doi.org/10.1186/s40425-018-0316-z

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*, *359*(6371), 97–103. https://doi.org/10.1126/science.aan4236

Hakozaki, T., Richard, C., Elkrief, A., Hosomi, Y., Benlaïfaoui, M., Mimpen, I., Terrisse, S., Derosa, L., Zitvogel, L., Routy, B., & Okuma, Y. (2020). The Gut Microbiome Associates with Immune Checkpoint Inhibition Outcomes in Patients with Advanced Non-Small Cell Lung Cancer. *Cancer Immunology Research*, *8*(10), 1243–1250. https://doi.org/10.1158/2326-6066.CIR-20-0196

Ishida, Y., Agata, Y., Shibahara, K., & Honjo, T. (1992). Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *The EMBO Journal*, *11*(11), 3887–3895. https://doi.org/10.1002/j.1460-2075.1992.tb05481.x

Jelinek, T., Paiva, B., & Hajek, R. (2018). Update on PD-1/PD-L1 Inhibitors in Multiple Myeloma. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.02431

Jin, Y., Dong, H., Xia, L., Yang, Y., Zhu, Y., Shen, Y., Zheng, H., Yao, C., Wang, Y., & Lu, S. (2019). The Diversity of Gut Microbiome is Associated With Favorable

Responses to Anti–Programmed Death 1 Immunotherapy in Chinese Patients With NSCLC. *Journal of Thoracic Oncology*, *14*(8), 1378–1389. https://doi.org/10.1016/j.jtho.2019.04.007

Katayama, Y., Yamada, T., Shimamoto, T., Iwasaku, M., Kaneko, Y., Uchino, J., & Takayama, K. (2019). The role of the gut microbiome on the efficacy of immune checkpoint inhibitors in Japanese responder patients with advanced non-small cell lung cancer. *Translational Lung Cancer Research*, *8*(6), 847–853. https://doi.org/10.21037/tlcr.2019.10.23

Kim, C. G., Kim, K. H., Pyo, K.-H., Xin, C.-F., Hong, M. H., Ahn, B.-C., Kim, Y., Choi, S. J., Yoon, H. I., Lee, J. G., Lee, C. Y., Park, S. Y., Park, S.-H., Cho, B. C., Shim, H. S., Shin, E.-C., & Kim, H. R. (2019). Hyperprogressive disease during PD-1/PD-L1 blockade in patients with non-small-cell lung cancer. *Annals of Oncology*, *30*(7), 1104–1113. https://doi.org/10.1093/annonc/mdz123

Kim, Y., Lee, D., Kim, D., Cho, J., Yang, J., Chung, M., Kim, K., & Ha, N. (2008). Inhibition of proliferation in colon cancer cell lines and harmful enzyme activity of colon bacteria by Bifidobacterium adolescentis SPM0212. *Archives of Pharmacal Research*, *31*(4), 468–473. https://doi.org/10.1007/s12272-001-1180-y

Lee, K. A., Luong, M. K., Shaw, H., Nathan, P., Bataille, V., & Spector, T. D. (2021). The gut microbiome: What the oncologist ought to know. *British Journal of Cancer*, *125*(9), 1197–1209. https://doi.org/10.1038/s41416-021-01467-x

Lee, S.-H., Cho, S.-Y., Yoon, Y., Park, C., Sohn, J., Jeong, J.-J., Jeon, B.-N., Jang, M., An, C., Lee, S., Kim, Y. Y., Kim, G., Kim, S., Kim, Y., Lee, G. B., Lee, E. J., Kim, S. G., Kim, H. S., Kim, Y., … Park, H. (2021). Bifidobacterium bifidum strains synergize with immune checkpoint inhibitors to reduce tumour burden in mice. *Nature Microbiology*, *6*(3), 277–288. https://doi.org/10.1038/s41564-020-00831-6

Li, S., Zhu, S., & Yu, J. (2023). The role of gut microbiota and metabolites in cancer chemotherapy. *Journal of Advanced Research*. https://doi.org/10.1016/j.jare.2023.11.027

Limeta, A., Ji, B., Levin, M., Gatto, F., & Nielsen, J. (n.d.). Meta-analysis of the gut microbiota in predicting response to cancer immunotherapy in metastatic melanoma. *JCI Insight*, *5*(23), e140940. https://doi.org/10.1172/jci.insight.140940

Liu, Y., Zugazagoitia, J., Ahmed, F. S., Henick, B. S., Gettinger, S. N., Herbst, R. S., Schalper, K. A., & Rimm, D. L. (2020). Immune Cell PD-L1 Colocalizes with Macrophages and Is Associated with Outcome in PD-1 Pathway Blockade Therapy. *Clinical Cancer Research*, *26*(4), 970–977. https://doi.org/10.1158/1078-0432.CCR-19-1040

Longhi, G., van Sinderen, D., Ventura, M., & Turroni, F. (2020). Microbiota and Cancer: The Emerging Beneficial Role of Bifidobacteria in Cancer Immunotherapy. *Frontiers in Microbiology*, *11*, 575072. https://doi.org/10.3389/fmicb.2020.575072

Machicote, A., Belén, S., Baz, P., Billordo, L. A., & Fainboim, L. (2018). Human CD8+HLA-DR+ Regulatory T Cells, Similarly to Classical CD4+Foxp3+ Cells, Suppress Immune Responses via PD-1/PD-L1 Axis. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.02788

Maia, M. C., Poroyko, V., Won, H., Almeida, L., Bergerot, P. G., Dizman, N., Hsu, J., Jones, J., Salgia, R., & Pal, S. K. (2018). Association of microbiome and plasma cytokine dynamics to nivolumab response in metastatic renal cell carcinoma (mRCC). *Journal of Clinical Oncology*, *36*(6_suppl), 656–656. https://doi.org/10.1200/JCO.2018.36.6_suppl.656

Makuku, R., Khalili, N., Razi, S., Keshavarz-Fathi, M., & Rezaei, N. (2021). Current and Future Perspectives of PD-1/PDL-1 Blockade in Cancer Immunotherapy. *Journal of Immunology Research*, *2021*, e6661406. https://doi.org/10.1155/2021/6661406

Mao, J., Wang, D., Long, J., Yang, X., Lin, J., Song, Y., Xie, F., Xun, Z., Wang, Y., Wang, Y., Li, Y., Sun, H., Xue, J., Song, Y., Zuo, B., Zhang, J., Bian, J., Zhang, T., Yang, X., … Zhao, H. (2021). Gut microbiome is associated with the clinical response to anti-PD-1 based immunotherapy in hepatobiliary cancers. *Journal for ImmunoTherapy of Cancer*, *9*(12), e003334. https://doi.org/10.1136/jitc-2021-003334

Noguchi, T., Ward, J. P., Gubin, M. M., Arthur, C. D., Lee, S. H., Hundal, J., Selby, M. J., Graziano, R. F., Mardis, E. R., Korman, A. J., & Schreiber, R. D. (2017). Temporally Distinct PD-L1 Expression by Tumor and Host Cells Contributes to Immune Escape. *Cancer Immunology Research*, *5*(2), 106–117. https://doi.org/10.1158/2326-6066.CIR-16-0391

Procaccianti, G., Roggiani, S., Conti, G., Brigidi, P., Turroni, S., & D'Amico, F. (2023). Bifidobacterium in anticancer immunochemotherapy: Friend or foe? *Microbiome Research Reports*, *2*(3), 24. https://doi.org/10.20517/mrr.2023.23

Routy, B., Le Chatelier, E., Derosa, L., Duong, C. P. M., Alou, M. T., Daillère, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M. P., Fidelle, M., Flament, C., Poirier-Colame, V., Opolon, P., Klein, C., Iribarren, K., Mondragón, L., Jacquelot, N., Qu, B., … Zitvogel, L. (2018). Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. *Science*, *359*(6371), 91–97. https://doi.org/10.1126/science.aan3706

Simin, J., Tamimi, R. M., Engstrand, L., Callens, S., & Brusselaers, N. (2020). Antibiotic use and the risk of breast cancer: A systematic review and dose-response meta-analysis. *Pharmacological Research*, *160*, 105072. https://doi.org/10.1016/j.phrs.2020.105072

Sivan, A., Corrales, L., Hubert, N., Williams, J. B., Aquino-Michaels, K., Earley, Z. M., Benyamin, F. W., Man Lei, Y., Jabri, B., Alegre, M.-L., Chang, E. B., & Gajewski, T. F. (2015). Commensal Bifidobacterium promotes antitumor immunity and facilitates anti–PD-L1 efficacy. *Science*, *350*(6264), 1084–1089. https://doi.org/10.1126/science.aac4255

Song, P., Yang, D., Wang, H., Cui, X., Si, X., Zhang, X., & Zhang, L. (2020). Relationship between intestinal flora structure and metabolite analysis and immunotherapy efficacy in Chinese NSCLC patients. *Thoracic Cancer*, *11*(6), 1621–1632. https://doi.org/10.1111/1759-7714.13442

Tanoue, T., Morita, S., Plichta, D. R., Skelly, A. N., Suda, W., Sugiura, Y., Narushima, S., Vlamakis, H., Motoo, I., Sugita, K., Shiota, A., Takeshita, K., Yasuma-Mitobe, K., Riethmacher, D., Kaisho, T., Norman, J. M., Mucida, D., Suematsu, M., Yaguchi, T., … Honda, K. (2019). A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature*, *565*(7741), 600–605. https://doi.org/10.1038/s41586-019-0878-z

Ting, N. L.-N., Lau, H. C.-H., & Yu, J. (2022). Cancer pharmacomicrobiomics: Targeting microbiota to optimise cancer therapy outcomes. *Gut*, *71*(7), 1412–1425. https://doi.org/10.1136/gutjnl-2021-326264

Trefny, M. P., Kaiser, M., Stanczak, M. A., Herzig, P., Savic, S., Wiese, M., Lardinois, D., Läubli, H., Uhlenbrock, F., & Zippelius, A. (2020). PD-1+ natural killer cells in human non-small cell lung cancer can be activated by PD-1/PD-L1 blockade. *Cancer Immunology, Immunotherapy*, *69*(8), 1505–1517. https://doi.org/10.1007/s00262-020-02558-z

Xia, C., Su, J., Liu, C., Mai, Z., Yin, S., Yang, C., & Fu, L. (2023). Human microbiomes in cancer development and therapy. *MedComm*, *4*(2), e221. https://doi.org/10.1002/mco2.221

Xu, X., Lv, J., Guo, F., Li, J., Jia, Y., Jiang, D., Wang, N., Zhang, C., Kong, L., Liu, Y., Zhang, Y., Lv, J., & Li, Z. (2020). Gut Microbiome Influences the Efficacy of PD-1 Antibody Immunotherapy on MSS-Type Colorectal Cancer via Metabolic Pathway. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.00814

Yang, K., Li, J., Sun, Z., Zhao, L., & Bai, C. (2020). Retreatment with immune checkpoint inhibitors in solid tumors: A systematic review. *Therapeutic Advances in Medical Oncology, 12*, 1758835920975353. https://doi.org/10.1177/1758835920975353

Yang, Q., Wang, B., Zheng, Q., Li, H., Meng, X., Zhou, F., & Zhang, L. (2023). A Review of Gut Microbiota-Derived Metabolites in Tumor Progression and Cancer Therapy. *Advanced Science*, *10*(15), 2207366. https://doi.org/10.1002/advs.202207366

Ye, L., Hou, Y., Hu, W., Wang, H., Yang, R., Zhang, Q., Feng, Q., Zheng, X., Yao, G., & Hao, H. (2023). Repressed Blautia-acetate immunological axis underlies breast cancer progression promoted by chronic stress. *Nature Communications*, *14*(1), 6160. https://doi.org/10.1038/s41467-023-41817-2

Zeng, W., Wang, Y., Wang, Z., Yu, M., Liu, K., Zhao, C., Pan, Y., & Ma, S. (2023). Veillonella parvula promotes the proliferation of lung adenocarcinoma through the nucleotide oligomerization domain 2/cellular communication network factor 4/nuclear factor kappa B pathway. *Discover. Oncology*, *14*, 129. https://doi.org/10.1007/s12672-023-00748-6

Zeriouh, M., Raskov, H., Kvich, L., Gögenur, I., & Bennedsen, A. L. B. (2023). Checkpoint inhibitor responses can be regulated by the gut microbiota – A systematic review. *Neoplasia*, *43*, 100923. https://doi.org/10.1016/j.neo.2023.100923

Zhan, Z., Liu, W., Pan, L., Bao, Y., Yan, Z., & Hong, L. (2022). Overabundance of Veillonella parvula promotes intestinal inflammation by activating macrophages via LPS-TLR4 pathway. *Cell Death Discovery*, *8*(1), 1–12. https://doi.org/10.1038/s41420-022-01015-3

*BBMap*. (2023, November 23). SourceForge. https://sourceforge.net/projects/bbmap/