# Computational analysis of the equine nasal and gut microbiome and their influence on the colonization of MDR bacteria

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium
(Dr. rer. nat.)

Submitted to the
Department of Biology, Chemistry, Pharmacy
of the Freie Universität Berlin

by

*Silver Anthony Wolf*

2023

The presented work was conducted between August 2019 and August 2023 under the guidance of Prof. Dr. Dr. h.c. mult. Lothar H. Wieler at the Robert Koch Institute, Berlin, Germany.

$1^{st}$ Reviewer: Prof. Dr. Dr. h.c. mult. Lothar H. Wieler

$2^{nd}$ Reviewer: Prof. Dr. Haike Antelmann

Date of defence: $27^{th}$ of November 2023

# Acknowledgement

First and foremost, I would like to thank Dr. Torsten Semmler and Dr. Birgit Walther for their supervision and endless support throughout my doctoral thesis. I am extremely grateful for all the opportunities they have provided for me, allowing me to learn and conduct my research with much appreciated freedom while also allocating extensive time for shared discussions, providing comprehensive feedback in regards to my progress.

I would further like to sincerely thank Prof. Dr. Dr. h.c. mult. Lothar H. Wieler for enabling this work to be conducted at the Robert Koch Institute in Berlin and for his excellent guidance throughout such a difficult period of time. In particular, I want to thank Prof. Dr. Dr. h.c. mult. Lothar H. Wieler and Prof. Dr. Haike Antelmann for both sharing their vast interdisciplinary expertises within the fields of microbiology and public health in regards to this project.

Furthermore, I would like to thank my colleagues at the Robert Koch Institute, in particular within the Genome Competence Centre (MF1) and the doctoral college (RoKoDoKo), for all the insightful discussions and exciting collaborations we developed over the recent years. I especially want to thank all collaborators who contributed towards this project, but most of all Anne Kauter for her extensive patience with handling the samples and the great discussions we shared on analysing complex microbiome data sets.

Next, I would like to thank my former colleagues Dr. Lennard Epping and Dr. Sumeet K. Tiwari for the great teas and laughs we shared while tackling new bioinformatical methods together. My special thanks goes to Felix Hartkopf and Mustafa Helal for their continuous support of our shared projects and for always having my back in times of need. I also would like to extend my thanks to Dr. Inga Eichhorn for her dedication towards proofreading this thesis.

Finally, I want to thank all of my family and friends for supporting me throughout the recent years. I particularly want to thank my father Dr. Siegmund S. Wolf and my brother Alexander T. Wolf for all of their much appreciated encouragement and of course, Lisa-Marie Epp for her unconditional love and support.

# Statement of authorship

I hereby declare that this thesis and all of the accompanying materials (code, images, tables) have been composed by myself, and describe my own work, unless otherwise acknowledged throughout the text. All sources of information are explicitly listed within the thesis. Verbatim extracts are represented through quotation marks. This work has not been accepted in any previous application for a degree.

# Table of Contents

# List of Figures

# List of Tables

# Abstract

Antimicrobial resistance (AMR) is an emerging global One Health issue, affecting both human and veterinary medicine, as well as the environment. Previous research revealed that clinics providing healthcare for companion animals, such as horses, in fact represent 'hot spots' for the local spread of multidrug-resistant (MDR) pathogens. Equines receiving perioperative antibiotic prophylaxis (PAP), consisting of gentamicin and penicillin, were hereby found to be frequently colonized by MDR bacteria. Since the duration of antibiotic therapy influences the local selective pressure and recovery time of the microbiome, two distinct PAP regimens were comparatively investigated, particularly in regards to the emergence of genes conferring resistance to antibiotics, biocides and metals (ARGs) within the enteral and nasal microbiota.

Hospitalized horses subjected to colic surgery were randomized into two groups receiving PAP, either as a single dosage prior to surgery (SSG) or across 5 consecutive days post-surgery, including the initial dosage (5DG). Faecal and nasal samples were collected on days 0 (hospital admission), 3 and 10 (post-surgery). Samples were subsequently 16S rRNA ($n = 78$) and metagenome shotgun ($n = 64$) sequenced. Two novel computational workflows were developed for analysis of the respective microbiome and resistome data sets (Meta16s and MetaGEN). Results were then tested for statistical correlation with study groups, sampling time points and taxonomic diversities.

The results display unique metagenomes associated with each of the equine patients. Beyond the impact of hospital stay and surgery, PAP caused microbiome perturbations in both nasal and gut environments, resulting in the accumulation of ARGs throughout the study period. Genes conferring resistance to beta-lactamases and aminoglycosides increased significantly in relative abundance over time within the 5DG, while the latter stagnated within the SSG. Taxonomic alpha diversity was found to be negatively correlated with the reconstructed ARGs across the sample set. Metagenomes with low taxonomic diversity hereby illustrated elevated abundances of resistance genes. While ARGs were found to be primarily associated with the family of *Enterobacteriaceae*, increased taxonomic plasticity was observed towards day 10. Draft genomes of abundant bacteria

were reconstructed, revealing the presence of genotypically MDR Gram-negative pathogens, including *Acinetobacter* spp., *Escherichia coli* and *Klebsiella* spp., among the patient set.

The study has enabled insights on the multifaceted effects of both hospitalization and PAP on the equine microbiome. A significant increase in ARGs conferring resistance to aminoglycosides and beta-lactamases, particularly within metagenomes representing the 5DG, provides evidence for the selective effect of prolonged PAP in hospitalized horses subjected to colic surgery. The presence of MDR pathogens stresses the need to further strengthen antibiotic stewardship throughout veterinary hospital environments. The presented novel computational workflows, as well as the described data sets, will enable further studies within the field of microbiome research.

# Zusammenfassung

Antibiotikaresistenzen (AMR) stellen eine zunehmende globale One Health-Herausforderung dar, welche sowohl die Human- und Veterinärmedizin als auch die Umwelt betrifft. Vorherige Untersuchungen haben ergeben, dass Kliniken, in denen Haustiere (wie z.B. Pferde) behandelt werden, 'Hotspots' für das Auftreten und die lokale Verbreitung von multiresistenten (MDR) Infektionserregern darstellen. Pferde, welche eine perioperative Antibiotikaprophylaxe (PAP) bestehend aus Gentamicin und Penicillin erhielten, wurden dabei besonders häufig mit MDR-Bakterien besiedelt. Da die Dauer der Antibiotikatherapie den lokalen Selektionsdruck und die Erholungszeit des Mikrobioms beeinflusst, wurden zwei unterschiedliche PAP-Schemata vergleichend untersucht, insbesondere im Hinblick auf das Auftreten von Genen innerhalb der enteralen und nasalen Mikrobiota, welche Resistenzen gegen Antibiotika, Biozide und Metalle (ARGs) vermitteln.

Hospitalisierte Pferde, die einer Kolikoperation unterzogen wurden, sind randomisiert in zwei Gruppen eingeteilt worden und erhielten entweder eine PAP bestehend aus nur einer einmalig Dosis vor der Operation (SSG) oder an 5 aufeinanderfolgenden Tagen, einschließlich der initialen Dosis (5DG). Kot- und Nasenproben wurden dabei an den Tagen 0 (Klinikaufnahme), 3 und 10 (nach der Operation) entnommen. Die Proben wurden anschließend 16S rRNA ($n = 78$) und Metagenom-Shotgun ($n = 64$) sequenziert. Zwei neue bioinformatische Programme wurden für die Analyse der jeweiligen Mikrobiom- und Resistom-Datensätze entwickelt (Meta16s und MetaGEN). Deren Ergebnisse wurden dann auf statistische Korrelation mit den Studiengruppen, dem Zeitpunkt der Probenentnahme und der taxonomischen Diversität geprüft.

Die Ergebnisse zeigen individuell zusammengesetzte Metagenome der Pferdekohorte. Neben dem Einfluss des Klinikaufenthaltes und der Operation als solches beeinflusste die PAP das Mikrobiom sowohl des Nasen- als auch des Darmmilieus, welches mit der Akkumulation von ARGs über den gesamten Studienzeitraum einherging. Gene, welche Resistenzen gegen Beta-Laktamasen und Aminoglykoside vermitteln, nahmen in ihrer relativen Häufigkeit im Laufe der Zeit signifikant in der 5DG zu, während letztere für die SSG stagnierten. Zudem wurde eine negative Korrelation zwischen der taxonomischen Alpha-Diversität und den rekonstruierten

ARGs über den gesamten Probensatz festgestellt. Insgesamt wurde verdeutlicht, dass Metagenome mit einer geringeren taxonomischen Vielfalt ein vergleichweise erhöhtes Vorkommen von Resistenzgenen aufwiesen. ARGs wurden darüber hinaus primär mit der Familie der *Enterobacteriaceae* assoziiert, wobei eine erhöhte taxonomische Plastizität gegen Tag 10 beschrieben wurde. Rekonstruierte Ganzgenome konnten das Vorkommen von zahlreichen genotypisch MDR Gram-negativen Erregern, einschließlich *Acinetobacter* spp., *Escherichia coli* und *Klebsiella* spp., in der Patientengruppe aufzeigten.

Diese Studie trägt dazu bei, ein Verständnis für die multifaktoriellen Auswirkungen von Hospitalisierung und PAP auf das Mikrobiom von Pferden zu entwickeln. Eine signifikante Zunahme von ARGs welche Resistenzen gegen Aminoglykoside und Beta-Laktamasen verleihen, primär in Metagenomen der 5DG, liefert Hinweise auf die selektive Wirkung der PAP-Regime bei Kolik-operierten und hospitalisierten Pferden. Das Vorkommen von MDR Erregern unterstreicht die Notwendigkeit, den Umgang mit Antibiotika im Veterinärklinikumfeld weiter zu verbessern. Die präsentierten neuen bioinformatischen Programme, sowie die beschriebenen Datensätze, werden hierbei weitere Studien im Bereich der Mikrobiomforschung ermöglichen.

# Abbreviations

**AMR** ..... Antimicrobial Resistance

**ARG** ..... Antibiotic Resistance Gene

**bp** ..... Base-Pair

**CDC** ..... Centers for Disease Control and Prevention

**DNA** ..... Deoxyribonucleic Acid

**ECDC** ..... European Centre for Disease Prevention and Control

**ESBL** ..... Extended-Spectrum Beta-Lactamase

**Gbp** ..... Giga-Base-Pair

**HGT** ..... Horizontal Gene Transfer

**Kbp** ..... Kilo-Base-Pair

**MAG** ..... Metagenome-Assembled Genome

**Mbp** ..... Mega-Base-Pair

**MDR** ..... Multiple Drug Resistance

**MGE** ..... Mobile Genetic Element

**MIC** ..... Minimum Inhibitory Concentration

**MRSA** ..... Methicillin-Resistant *Staphylococcus aureus*

**NCBI** ..... National Center for Biotechnology Information

**NGS** ..... Next Generation Sequencing

**OTU** ..... Operational Taxonomic Unit

**PAP** ..... Perioperative Antibiotic Prophylaxis

**PCA** ..... Principal Component Analysis

**PCR** ..... Polymerase Chain Reaction

**QC** ..... Quality Control

**rRNA** ..... Ribosomal Ribonucleic Acid

**SNP** ..... Single-Nucleotide Polymorphism

**SRA** ..... Sequence Read Archive

**SSG** ..... Single-Shot Group

**WGS** ..... Whole Genome Sequencing

**WHO** ..... World Health Organization

**5DG** ..... 5-Day Group

# Glossary

**Alpha diversity:** "[The] diversity measured within a particular ecosystem or sample. It is commonly characterized by [...] richness, evenness and phylogenetic diversity." [1]

**Antimicrobial resistance gene:** "[...] a gene that confers resistance to antibiotics when it is present or increases susceptibility [...] when it is absent." [2]

**Beta diversity:** "[Diversity] comparison between particular ecosystems or samples. It is commonly analysed through ecological and phylogenetic distances estimated from the sample composition." [1]

**Commensals:** "Organisms (for example, microbes) that are involved in a form of symbiosis in which one organism derives a benefit while the other is unaffected." [3]

**Dysbiosis:** "A condition in which the normal microbiome population structure is disturbed, often through external pressures such as disease states or medications." [3]

**Diversity index:** "A mathematical expression that combines species richness and evenness as a measure of diversity." [4]

**Evenness:** "A measure of the homogeneity of abundances in a sample or a community." [4]

***K-mers*:** "All possible sequences of length $k$ from a read obtained through DNA sequencing." [5]

**Metadata:** "[All] the variables and data relevant for the study providing information about the samples included." [1]

**Microbiota:** "The community of microorganisms, including bacteria, viruses and fungi, which are found within a specific environment (for example, the human gut)." [6]

**Microbiome:** "The totality of microbes, their genetic information and the milieu in which they interact. Microbiomes typically consist of environmental or biological niches containing complex communities of microbes." [3]

**Operational taxonomic unit:** "A group of closely related individuals or sequences (often 97% sequence similarity threshold)." [5]

**Probiotic:** "Living microorganisms that are thought to confer a benefit to the host." [3]

**Relative abundance:** "The quantitative pattern of rarity and commonness among species in a sample or a community." [4]

**Reservoir:** "[...] refers to a bacterium that harbours resistance or virulence genes [...] and serves as a source from which other bacteria may acquire these genes." [7]

**Resilience:** "A term in ecology indicating the capacity of a system to absorb disturbance and to reorganize itself while undergoing change, so as to retain essentially the same function, structure and identity." [3]

**Resistome:** "[...] the ensemble of genes encoding antimicrobial resistance in a given microbiome [...]" [8]

**Richness:** "The number of species in a community, in a landscape or marinescape, or in a region." [4]

**Zoonoses:** "[...] diseases and infections that are naturally transmitted between vertebrate animals and man. Emerging zoonoses are [...] diseases caused either by [...] new agents or by previously known microorganisms appearing in places or in species in which the disease was previously unknown." [9]

# 1 Background & Introduction

## 1.1 Antimicrobial Resistance

The accidental discovery of the antibiotic penicillin in 1928 significantly altered the world we live in today [10, 11]. Many advancements of modern medicine, including cancer chemotherapy and invasive interventions such as heart surgery, joint replacements or organ transplants require usage of antibiotics, such as beta-lactams, but are threatened by the emerging crisis of antibiotic resistances [12, 13]. Antibiotics are pharmaceutical drugs used to perturb essential biochemical processes or attack molecular structures within bacteria, resulting in their death or stagnation of growth [13–15]. The umbrella-term 'antimicrobial' hereby generally refers to drugs used to treat infections caused by a wider range of microorganisms, not solely bacteria, but also protozoan parasites and fungi [14]. Antibiotics and their accompanying resistances have ancient origins [16, 17], as they are naturally produced by fungi [14] and have been found across all known environments, from frozen tundras and caves to isolated populations of humans [15, 18–20].

These natural compounds have since been used to synthesize an array of pharmaceutical drugs over the last 100 years after their initial discovery [10, 14]. From the 1940s to the 1960s, a total of 20 classes of antibiotics were described [14], but the spread of newly resistant bacterial strains closely followed any large-scale application of these drugs [14, 19]. For example, resistance to penicillin began to emerge in the 1940s, less than 20 years after its initial discovery [21, 22]. During this time, antibiotic usage was found to be widespread for treating injured soldiers on the battlefields of World War 2, as well as the introduction of these drugs in form of growth promoters for livestock [23]. For newer classes of antibiotics, the emergence of bacterial resistances was reported up to five years after their widespread distribution [14]. This proliferation of resistant strains has since begun to turn diseases which used to be easily curable deadly once more [13]. The situation has steadily been worsening due to the ongoing failure to discover novel antimicrobial drugs throughout the $21^{st}$ century [13, 23, 24]. Over the last 20 years, only two new classes of antibiotics (lipopeptides and oxazolidinones) have been approved by international agencies [24]. In fact, the most recent drug class to be effective against a board spectrum of pathogenic Gram-negative bacteria are

quinolones, which were first described in the 1960s [24]. These ongoing challenges within antibiotic research and discovery are already resulting in large implications for public health [24].

The emergence of antimicrobial resistance (AMR) is therefore an urgent global health issue and represents one of the greatest hazards to healthcare systems worldwide [23, 25, 26]. With some bacterial strains already so extensively resistant that they are unable to be treated with antibiotics at all [27], the 'resistance epidemic' [25] threatens many achievements of modern medicine [28]. AMR has become an issue amongst almost all pathogenic bacteria associated with infectious diseases, encompassing both Gram-positives and negatives [23, 26]. In 2019 alone, an estimated 1.27 million deaths worldwide were directly attributed to AMR [29]. Of these, over 929,000 deaths were associated with a small range of six pathogens, lead by the species of *Escherichia coli* [29]. Clinical studies outline that a majority ($> 50\%$) of invasive infections of *Escherichia coli*, *Klebsiella pneumoniae* and *Staphylococcus aureus* already illustrate some level of resistance to a range of commonly used antibiotics [14]. In the US, multiple drug resistant (MDR) bacteria, such as methicillin-resistant *Staphylococcus aureus* (MRSA), have already been estimated to cause more deaths than HIV [30]. MDR bacteria, including *Escherichia coli* and *Staphylococcus aureus*, also remain ongoing challenges within veterinary medicine, due to limited treatment options and the potential for zoonotic transmission [31]. Species of *Staphylococci* hereby represent frequent agents of infection across livestock animals [32]. Gram-negative pathogens are of particular concern, since some strains have evolved resistance against nearly all classes of antibiotics available for treatment [14].

The World Health Organization (WHO) recently performed a multicriteria decision analysis in order to allow prioritization of research and funding towards new agents against a series of highly multidrug-resistant pathogens [24]. This list included *Mycobacterium tuberculosis* as the highest extensively-resistant pathogen and classified other priority bacteria into: critical (*Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacteriaceae*), high (*Enterococcus faecium*, *Staphylococcus aureus*, *Helicobacter pylori*, *Campylobacter* spp., *Salmonella* spp., *Neisseria gonorrhoeae*) and medium (*Streptococcus pneumoniae*, *Haemophilus influenzae*, *Shigella* spp.) [24]. Similar outlines have also been proposed by national

health agencies, including the Centers for Disease Control and Prevention (CDC), which declared *Clostridium difficile*, carbapenem-resistant or extended-spectrum beta-lactamase (ESBL)-producing *Enterobacteriaceae*, *Campylobacter* spp., non-typhoidal *Salmonella* and *Shigella* as particular threats for public health [26]. MDR bacteria, including those specifically addressed by the aforementioned agencies, cause increased mortality rates, lengths of hospitalization and medical costs when compared to infections resulting from antibiotic-susceptible strains [27, 33, 34]. MRSA infections, for example, are commonly associated with doubled death rates and hospital expenses compared to non-resistant variants [35]. With the continuing trend of increasing resistance rates, the beginning of a post-antibiotic age has been proposed, a time in which even minor infections may become untreatable once more [13, 23, 28].

### 1.1.1 Mechanisms of Antimicrobial Resistance

Antibiotics commonly target essential functions of the bacterial cell, such as cell wall maintenance, protein biosynthesis, DNA replication or repair mechanisms [15]. Throughout history, microbes have developed a range of unique strategies to alleviate these effects, including degradation of the drug through enzymes, alteration of the antibiotic target molecule, differential regulation of metabolic pathways to circumvent the effect or mitigation of substance accumulation through efflux pumps [15, 22, 26, 28, 34]. The development and spread of AMR are direct results of the multifaceted genetic processes that microbes can utilize in order to rapidly adapt to new environmental stressors [22]. Bacteria in particular possess a high level of genetic plasticity compared to humans [36]. *Escherichia coli* (*E. coli*), for example, have been described to range in genomic sizes from 4.2 to 6.0 Mbp, with only a few hundred genes shared across the entire species [37]. In contrast, the human genome is comparatively static in its size of 3 Gbp (haploid), encompassing approximately 20,000 protein-encoding genes, many of which are highly conserved [38, 39]. Both organisms differ slightly in their average mutation rates, with gut-associated *E. coli* described to develop $6.9 \times 10^{-7}$ SNPs per nucleotide site per year [40] and humans evolving within the order of $0.5 \times 10^{-9}$ per bp per year [41]. However, given that *E. coli* can double their population size in 33 min under laboratory conditions, while humans span generational times of at least 29 years [42], SNPs associated with increased fitness can rapidly emerge and spread across microbial organisms. Drug resistance in particular can result

from a variety of changes accumulating within the bacterial genome. For example, polymorphisms in the chromosomal DNA can induce the upregulation of intrinsic resistance mechanisms or render specific antibiotic components ineffective against a bacterial strain [13]. In fact, many antibiotic resistance genes (ARGs) are believed to have evolved naturally over time in environmental microbes and were later transferred to other species [43]. Individuals possessing such resistance mechanisms have the ability to survive under extreme conditions, including, but not limited to, antibiotic exposure, and may hereby spread their corresponding genes through vertical as well as horizontal genetic recombination (also known as horizontal or lateral gene transfer, HGT) to other members of the population [44, 45]. This unique ability allows the bidirectional acquisition and exchange of additional genetic material between non-related commensal and pathogenic individuals [26, 28, 46, 47], promoting the development of new zoonotic strains [48].

Mobile genetic elements (MGEs), such as conjugative plasmids, transposons, integrons, insertion sequences or prophages [18, 26, 49, 50], frequently encode ARGs or virulence genes [17, 50, 51]. While HGT, particularly the transfer of mobile ARGs, is primarily constrained by phylogenetic distance, environmental conditions and ecological barriers [28], the mechanisms of conjugation, transduction and transformation are not limited to a single species [26, 49] and can enable the exchange of genetic material even across members of different genera [22]. This allows for competitive advantages of even distantly related bacteria throughout times of antibiotic exposure, enabling the evolution of new MDR strains [51]. Mono-resistant strains are also likely predisposed to accumulate additional resistances through the careful balancing of the increased fitness costs associated with MDR and the positive selection of SNPs within metabolism-associated genes [27]. Clonal spread [33] and cross-selection [16, 46] in particular have been shown to significantly affect the dissemination of resistances across the environment. In addition, resistance genes to biocides, which play an important role in disinfectants, and antimicrobial metals, such as zinc, are frequently co-selected and co-localized together with AMR genes [8, 16, 20, 46]. The entirety of ARGs within an environment is referred to as the 'resistome' [28, 52, 53], which includes many MGEs not constrained to single individual species [22]. ARGs can be divided into two classes: (1) taxa-specific, chromosomally encoded genes and (2) through mutational processes or HGT acquired resistances [28, 54]. Intrinsic resistances (1) are

generally common within environmental bacteria and represent genes not directly associated with anthropogenic antimicrobial exposure [54]. While they may be mobilized, they commonly remain confined to specific taxonomic groups [54]. In contrast, acquired resistances (2) have the potential to disseminate across multiple species and environments [54]. Although many ARGs have evolved naturally and pre-date modern medicine by several millennia [13], their occurrence has drastically increased over the last 50 years due to the excessive selective pressure of antibiotics that bacteria of the $21^{st}$ century must endure [17, 43].

### 1.1.2 Surveillance of Antimicrobial Resistance

Today, large-scale usage of antibiotics is common practice [48, 55]. Over the past 60 years, millions of tons of antibiotics have been produced and distributed across various industries and sectors [14]. Combined with the continuing mis- and overuse of therapeutic and prophylactic dosages, especially within the livestock and healthcare sectors [23, 26, 56], resistance rates have continued to increase [48, 55]. The CDC estimated that in the US approximately 50% of prescribed antibiotics are not optimally utilized or needed at all [57]. Another major application field for antibiotics lies within veterinary medicine, more specifically livestock farming [14], which in some countries outweighs usage in human medicine [19, 51]. In the US, surveys revealed that only 52% of the general public were aware of antibiotic application in livestock production and aquaculture [14], while almost 80% of the nation's antibiotics were being used as growth promoters, infection prophylaxes and direct treatments within the veterinary field [28], although this has been turning since the ban on growth promoters within the EU (in 2006) and the US (in 2017) [58].

However, while both humans and animals consume antibiotics, they do not metabolize these substances completely, resulting in the release of remnants back into the environment through faeces and urine [15, 50, 51, 59, 60]. Due to half-lives ranging up to several years [15], remains of antibiotics can easily disseminate into sewer systems and wastewater plants [51] or are directly applied to soil as contaminants of fertilizers [59], inducing the risk of selection for resistant bacteria within these environments [26, 60]. On a positive note, this issue has started to gain increasing international attention over the last decade [45], illustrated by the multiple call-to-arms action plans against AMR [23, 24, 53], resulting in antibiotics

now being widely regarded as pollutants [16, 19, 45]. AMR has further been associated with several socio-economic and environmental factors, such as sanitation, antimicrobial usage and bacterial abundance [60]. National efforts have therefore begun to increase monitoring of AMR in order to inform policy makers about the emergence and spread of resistances [53, 61]. In Europe, multiple initiatives are currently working on developing integrated genomic surveillance systems for pathogens of public health importance, including the European Centre for Disease Prevention and Control (ECDC) [62] and the Robert Koch Institute [63]. Internationally, recent adoption of the Muscat Manifesto[1] by over 40 nations to reduce antibiotic usage through improved stewardship highlights an important step in the battle against AMR.

While additional research is needed with respect to the development of novel antibiotic agents, this represents an expensive and slow progress [24]. As such, additional public health interventions must be upheld in order to address the current healthcare burden of MDR [24]. The CDC recently outlined several strategies to slow the emergence of resistances: preventing infections, improving antibiotic stewardship, developing new drugs and diagnostic tests as well as increasing surveillance efforts [14]. In order to slow the spread of antibiotic resistance, overall usage of antibiotics must decrease, accompanied by prudent stewardship and infection control [14, 64]. This is especially true for veterinary healthcare, where alternatives to the use of antibiotics should be considered whenever possible [33].

It is important to note that while public health is of primary concern, other sectors outside of human and veterinary medicine are also heavily affected by AMR, including agriculture, livestock farming and aquaculture [22]. Within these sectors, antibiotic usage remains at an all-time high, elevating rates of resistances across a wide range of diverse environments [22]. Thus, AMR is a multifaceted, interdisciplinary issue affecting humans, animals, the environment and food chains [24] and therefore falls under the concept of One Health. The One Health framework was initially introduced in 2004 and represents a concept which interconnects public health with animal health and environmental reservoirs [19, 23]. By spanning across multiple sectors and industries, AMR is required to be evaluated under

[1] https://www.amrconference2022.om/MuscatManifesto.html

such a shared framework in order to establish coordinated surveillance programs for pathogens and the emergence of resistances [53, 65]. These enable the quantitative assessment of AMR, including underlying factors such as the actual usage of antimicrobials, as well as monitoring the development and spread of resistant pathogens [66]. Combining both genomic and epidemiologic data within an interconnected One Health approach allows to evaluate the real healthcare burden of harmful bacteria and their resistances [24]. Such systems remain essential in order to limit the global AMR burden of disease [19, 60].

With increasing rates of resistances, new methods are urgently needed in order to enhance our understanding on the biochemical mechanisms behind AMR [48] and to monitor these within a given community [44]. Phenotypic antimicrobial susceptibility tests, such as minimum inhibitory concentration (MIC) assessment [53, 55], disk diffusion assays [49, 64] and broth microdilution [67], as well as gene targeted quantitative polymerase chain reaction (PCR) [25, 64] represent gold standards for the *in vitro* assessment of bacterial resistances. While these methods have dominated AMR surveillance over the last decades [12], new technologies are currently emerging to tackle the expanding issue of global resistances, with one of these represented by the field of metagenomics [25]. Metagenomics enables culture-free, rapid characterization of clinical samples without prior knowledge of the involved pathogens or resistances [53, 68]. It allows the identification of ARGs disseminated across a microbial community *in silico*, without the need to isolate and culture the bacteria [22, 53]. Instead, samples are taken directly from an environmental source, DNA extraction and sequencing are performed, and the resulting genomic data are screened for resistance profiles, which in turn enable the identification of potential hot spots, transmission pathways [44] or natural reservoirs of AMR [69]. This enables the characterization of resistance reservoirs across different environments or microbial hosts, including humans and animals [25]. With the rapid spread of resistant pathogens posing a global health threat, methods such as metagenomics can be utilized to establish and support integrated One Health surveillance systems for monitoring the emergence and spread of AMR [19, 24, 53].

## 1.2 Metagenomics

Metagenomics represents a comparatively new discipline, primarily driven by recent advancements in DNA sequencing technologies [70]. It hereby lies within the intersection of multiple interdisciplinary research fields, including data analysis, microbial ecology and molecular genomics [71]. The prefix 'meta' originates from the Greek language and roughly translates to 'after' or 'beyond', essentially describing metagenomics as research 'beyond a single genome' [70]. The term 'metagenomics' was first coined by Handelsman et al. in 1998 [36] to describe a culture-independent, unbiased, molecular survey of genetic material present within an environmental sample [72]. This bypasses the need to isolate and produce pure cultures [22, 73] through bulk sequencing of samples taken directly from an environmental source [74]. These samples can be highly diverse in their origin, ranging from soil [59, 75, 76] to water [49, 50, 77] and air samples [69], even including anthropogenic environments, such as food sources [78], hospital surfaces [79] or faeces [52, 80, 81]. The representation of the total genomic content of a sample is commonly referred to as the 'metagenome' [5, 71], encompassing many individual organisms across their specific ecological niches and habitats [3]. While metagenomic analyses have historically focused on studying bacterial ecosystems, microbial communities in fact comprise of a mixture of bacteria, archaea, microbial eukaryotes and viruses all co-inhabiting within a shared environment [73, 82]. With increasing sequencing depths, recent studies have therefore begun to expand their focus towards microbial eukaryotes [5, 82] and viruses [83, 84]. Early large-scale metagenomic projects include the Human Microbiome Project [85] and the Global Ocean Sampling Expedition [86], with more recent initiatives studying the spread of AMR across subway systems [87, 88].

The cutting-edge nature of metagenomics has enabled new insights across a wide range of research fields throughout the last decade, including healthcare, microbial ecology and infectious diseases [72, 73, 89]. Metagenomics has been described as the method for exploring the microbial 'dark matter' [13, 83, 90] and continues to play a key role in unravelling novel species that cannot be obtained through culture [22, 73, 91, 92]. However, while strain culturing represents the gold standard to this day for performing in-depth studies on a small proportion of earth's microbes [36], a large majority of the planet's microbial biodiversity cannot be readily

cultured. This essentially limits the application of traditional genomic analyses, which require the isolation and growth of a species in culture before further DNA extraction can be performed [93]. Bacteria in particular represent a highly diverse kingdom of life, encompassing exceptional genetic variability, such as extremely divergent gene contents between members of even the same species [94]. Current studies estimate that 99% of all environmental bacteria remain unculturable [15, 43, 86, 90], which results in the absence of these microorganisms from genomic reference databases [95, 96], as well as being unable to study and monitor their resistomes through traditional methods [15]. For example, early research on the soil microbiome estimated that less than 0.1% of microbes can be readily cultured, with a single gram of soil containing up to 10,000 undescribed prokaryotic species [36]. The remaining 99.9% of soil-colonizing bacteria have since revealed a world of high genetic diversity [76], spanning many new taxonomic classifications, including several novel phyla [36]. However, even anthropogenic environments remain highly uncharacterized, as a study of the New York City public transport system concluded that 48% of the sampled DNA could not be allocated to known reference genomes, indicating the presence of a wide range of undescribed microbial species [97]. This holds particularly true for human microbiomes, which represent a major frontier for molecular analyses and genetics to this day [3]. Estimations suggest that only 30% of gut microbiota can be readily cultivated [98] and over 50% still lack corresponding reference genomes within current databases [96], stressing the need of culture-free sequencing methods for further studying these environments.

Over the recent decade, extensive research within the field of metagenomics has enabled the expanse of the known microbial diversity across several branches of life [92], including those widely unexplored such as the virosphere [83, 84, 99]. Metagenomics is also being utilized to characterize the interaction of a wide range of medical conditions, including inflammatory bowel disease, obesity and diabetes, with human microbiome compositions [100]. While the idea of isolating all microbes and trying to predict their pathogenicity is impossible *in vitro*, identifying and characterizing potential pathogens *in silico* through unbiased metagenomic surveys might represent a useful tool for future molecular healthcare [84]. With the continually decreasing costs of sequencing, metagenomics may therefore eventually be applied throughout routine microbiological diagnostics in the future [84],

for example to monitor changes within the gut microbiota [101, 102] or as a surveillance tool for assessing the spread of AMR [103]. Since the human genome was first published in 2001, sequencing technologies have in fact become over a million times cheaper and faster [104]. With this rapid increase in throughput, large quantities of metagenomic data can be easily produced, but bioinformatic analyses remain a common bottleneck for many studies [105]. For example, numerous taxonomic profiling tools have been developed for the analysis of metagenomic data [106] and while this highlights the current interest in the subject at hand [107], additional research is required to define clear gold standards, including broadly accepted methods, best practices and scientific guidelines [108].

Two common strategies have hereby been established to enable the culture-free assessment of metagenomes: either the amplification of highly conserved gene markers (such as 16S rRNA sequencing, described in section 1.2.1) or by performing untargeted sequencing of the entire DNA content (known as shotgun metagenome sequencing, described in section 1.2.2) [92]. Metagenomic analyses commonly dwarf isolate-based sequencing projects in size [109] and may even rival eukaryotic genomic analyses in terms of complexity [36]. Through sequencing, millions of short nucleotide strands are generated and further utilized to characterize the underlying microbial population by determining the taxonomic and functional structure (composition and distribution) of a sample [72, 110, 111]. This enables, beyond other, to study complex relationships between a habitat/environment (i.e. a host) and individual species through the quantification of co-habiting taxa present within a bacterial community at the time of sequencing [73, 112]. While both 16S rRNA and shotgun metagenome sequencing capture relative abundances of microbiota and are commonly utilized for microbiome characterization [1, 113, 114], they strongly differ in their respective data analysis workflows and outputs due to variations within the underlying methodologies [115]. As such, the following sections will characterize both methods in greater detail. Potential advantages and disadvantages will be compared between these approaches, particularly in regards to downstream analyses and technological limitations.

### 1.2.1 16S rRNA Sequencing

Historically, researchers have utilized the 16S ribosomal RNA (16S rRNA) since the 1970s for molecular identification and quantification of unknown prokaryotic

communities [92, 93, 112, 116, 117]. In both bacteria and archaea, the 16S rRNA gene encodes a central component for the small 30S RNA subunit of the 70S ribosomes [3, 114, 115, 118]. The 16S rRNA hereby plays an important role within the translational processes of the prokaryotic ribosomes [119]. As ribosomal translation represents an essential pathway for life, its components are highly conserved across the phylogenetic tree [111, 114] and thus represent unique genetic markers for microbial ecology [115, 120, 121]. An overview of the 16S rRNA gene is illustrated in figure 1.1. The 16S rRNA gene is approximately 1,600 bp in length [119] (1,542 bp in *E. coli* K12 [102, 114]) and can be segmented into nine variable regions (V1-V9) which are flanked by conserved sequences (C) [31, 114–118]. While the conserved regions can be found ubiquitously across the evolutionary scale [85], variable regions allow the distinction of 16S rRNA gene variants between different clades of microbial taxa [111, 116, 118].

Throughout the recent decades, short-read amplicon sequencing was primarily utilized to assess these variable regions [114, 117, 118]. Primers for 16S rRNA gene sequencing are taxonomically universal [111] and target specific pairs of conserved DNA surrounding a variable region of interest [5, 118, 122]. Although the selection of primer pairs and their targeting regions have been described to vary in taxonomic resolution and sequencing efficiency [102, 119, 123], commonly employed regions include V1-V3, V3-V4, V5-V6 or solely V4 (see figure 1.1) [114, 118]. Once primers bind to a particular conserved sequence, PCR is utilized for further amplification of the enclosed variable regions [112, 118]. This occurs simultaneously across thousands of potentially unique 16S rRNA genes and taxa, resulting in millions of copies from the targeted sequence [118].

Amplified products are then sequenced using high-throughput Next Generation Sequencing (NGS) platforms in order to obtain short nucleotide fragments, known as reads, corresponding to the variable 16S rRNA gene regions of choice [5, 118]. Metagenomic studies have commonly been performed on the widely adapted Illumina platform [1], which utilizes sequencing by synthesis [26, 124] in order to generate billions of short reads at a comparatively low cost [107, 114]. Adapter-ligated DNA fragments are captured by a surface of complementary sequences and are multiplied through the process of bridge amplification [114]. Sequencing is finally performed via the base-by-base introduction of fluorescently labelled terminator

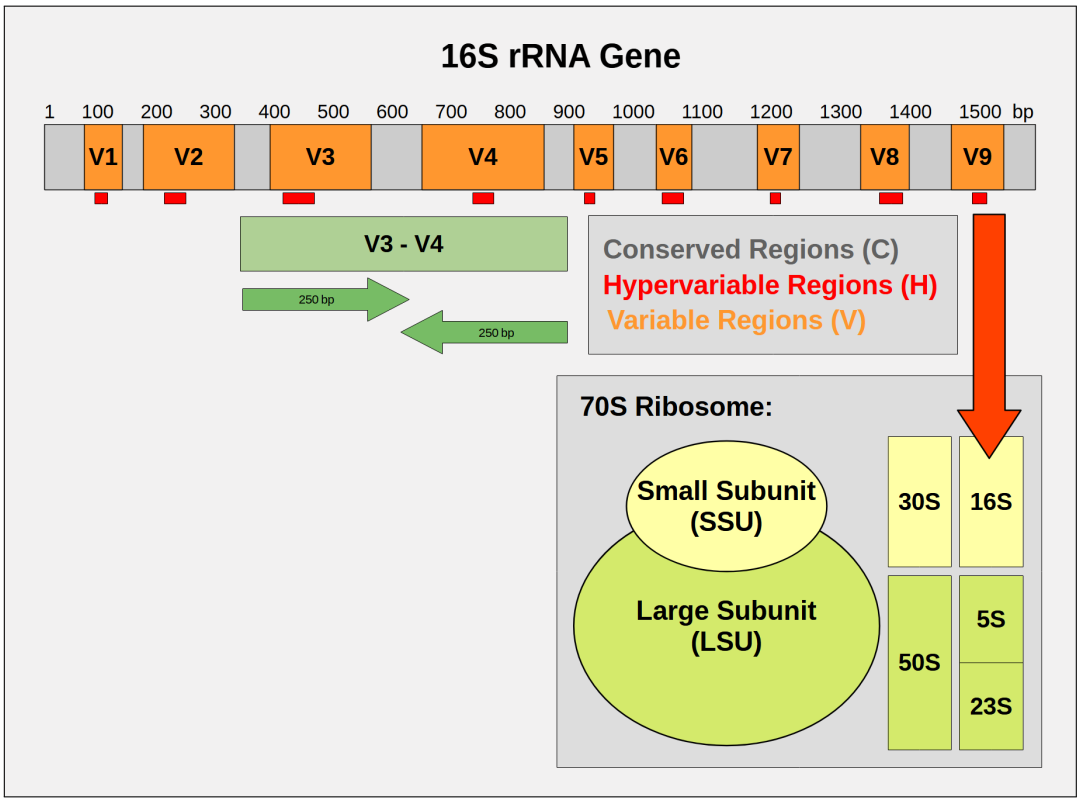


**Figure 1.1: 16S rRNA Sequencing.** Conceptional visualization of the process for 16S rRNA amplicon sequencing. Bacterial 16S rRNA genes are approximately 1,600 bp in length (shown on top) and encode a central component for the 30S RNA subunits of the 70S ribosomes (illustrated on the right). The 70S ribosomes consist of both small (30S) and large subunits (50S), which comprise of multiple rRNAs and proteins (16S, 23S, 5S). The 16S rRNA genes can further be segmented into nine variable regions (V1-V9 in orange), flanked by conserved sequences (in grey). V1-V9 each contain small segments of high taxonomic-variability (hypervariable regions in red), which represent targets for the amplicon sequencing. Primers (in dark green) hereby bind to the flanking conserved regions, followed by subsequent amplification of the enclosed variable regions (shown as an example for V3-V4).

sequences [114]. The resulting reads are typically 50-300 bp in length [111, 125], with an average error rate of 0.1% per nucleotide [5, 126, 127]. Once the amplicon regions have been sequenced, reads can be allocated to specific taxa during downstream analysis, commonly referred to as operational taxonomic units (OTUs) [92, 112]. As the abundance of individual OTUs infers which microbial taxa were present in the original sample at the time of sequencing [128], tables of OTU counts provide the basis for all downstream microbiome analyses [114, 129].

The major advantage of 16S rRNA sequencing is its high-throughput speed and cost efficiency [92, 117], enabling the rapid diagnostic of potentially infectious species [98, 116]. Analogous approaches to have also been developed for microbial eukaryotes (18S rRNA sequencing) [1, 93, 114] and fungi (ITS sequencing) [5, 92, 111, 128]. However, these methods possess several drawbacks, including biases induced by variations in primer binding affinities, region selection, amplicon size and the amount of utilized PCR cycles [5, 96, 102, 130]. Since 16S rRNA sequencing focuses solely on a single genomic region for taxonomic profiling, it has been described to be effective for microbial community characterization from phylum- to genus-level, however lacks accuracy on the species-level [112]. This illustrates another important limitation in particular: 16S rRNA gene multiplicity, which occurs frequently across bacterial clades [96, 102, 117, 120]. These heterogeneous copies may differ in their respective sequences, leading to the existence of multiple gene variants for a single organism [120], which can bias abundance estimation towards taxa with high copy numbers [1, 117]. The average bacterial genome possesses approximately 4 copies of the 16S rRNA gene [120], although some taxa may carry up to 15 or more [1, 31, 73]. For some prokaryotes the intragenomic heterogeneity of these genes are especially high (e.g. $> 5\%$ for *Halobacteriales*), which may introduce further biases regarding overall phylogenetic assessment [120]. Several hypotheses exist as to why these bacteria carry multiple rRNA genes within their genomes, including that these can provide a multiplication effect for the process of translation, which in turn increases adaptability to novel environments through differential expression [120]. However, due to the nature of high-throughput sequencing, the bias induced by multiple 16S rRNA gene copies remains comparatively small [131]. While methods exist to correct these copy number counts, they suffer from low accuracy and are therefore not commonly utilized within microbiome studies [1].

### 1.2.2 Shotgun Metagenome Sequencing

With the emergence of high-throughput sequencing platforms, the field of metagenomics evolved from focusing on a few selected marker genes to being able to characterize the entire taxonomic and functional content of a sample [132]. Compared to earlier amplification methods, including the aforementioned 16S rRNA sequencing, modern technologies, such as shotgun sequencing, are not bound to predefined regions of gene markers and allow the characterization of random frag-

ments of DNA [118]. Deep sequencing of metagenomic samples (known as shotgun metagenome sequencing) results in billions of random short reads from a pool of microbial communities [91]. These reads are then utilized downstream in order to estimate the taxonomic profile of the original sample by assigning each fragment to a microbial lineage based on similarity to available reference sequences [91].

When compared to 16S rRNA sequencing, shotgun metagenomics is not limited to a single phylogenetic group and as such can be used to characterize bacteria, eukaryotes and viruses present within a microbial community [112]. This enables the detection of microbes that otherwise would have been missed through amplicon sequencing [111] and even allows the characterization of entirely novel sequences [72, 84, 112]. Its independence from gene markers enables profiling at increased taxonomic resolutions (up to species-level), resulting in improved downstream analyses, including diversity estimations [112, 117, 121]. Shotgun metagenomics furthermore allows the characterization of individual genes (such as ARGs) across samples, enabling insights into the functional potential of the metagenome [112]. This method is also less prone to PCR biases [76, 99, 110, 112, 121], although it has been described to suffer from decreased sensitivity [53] and underrepresentation of taxa-specific sequences, such as fungi [130], when compared to targeted approaches [52]. However, the largest downside of shotgun metagenomics is that it requires extensive data analysis and is overall more cost- and labour-intensive when compared to 16S rRNA amplicon sequencing [112]. Shotgun metagenomics is also much more dependent on the selection of sequencing parameters, such as read lengths and depths, which are known to affect downstream analyses, including taxonomic and resistance gene profiling [133, 134]. Sequencing depth describes the total number of fragments obtained from a sample through the process of NGS [128]. Different levels of depth can hereby influence the ability to capture and quantify the enclosed microbial communities [135], with increasing values enabling the detection of minority species [128]. Since metagenomes are typically 'open', i.e. possess no fixed point defining that they are complete, various approximations exist to estimate the essential parameter of sequencing depth [73], which will be presented in more detail throughout chapter 2. Regarding read lengths, fragments are recommended to encompass approximately 300 bps in order to be utilized for accurate taxonomic assignment, although this can commonly be achieved through partial overlapping of paired-end reads post-sequencing [130].

As illustrated by the example of sequencing parameters, there exists a wide range of standards yet to be established throughout the field of metagenomics. With the continuous technological advancements in NGS, thousands of new species are currently being discovered, leading to the ongoing issue of taxonomically classifying potentially novel sequences [136]. This is particularly challenging due to intrinsic properties of metagenomic samples, enabling the characterization of high- and low-abundance species, while also accounting for the increased levels of sequence variation across bacterial strains [110, 136]. Another fundamental issue of metagenomics is handling shared, homologous regions of DNA and correctly assigning reads containing these to specific taxa [110]. These challenges have previously limited our insights into the microbial world surrounding us [72], including bacterial 'dark matter' [13, 83] which contains a majority of cryptic environmental species that cannot be easily cultivated and are therefore of increasing interest, especially in regards to the emergence of AMR [26].

### 1.2.3 Targeted Analysis of Metagenomes

In addition to describing the taxonomic content of a sample, metagenomic shotgun sequences can be utilized for characterization of the underlying microbial functional potential, based on the abundance of well-defined gene clusters. This commonly encompasses resistance (section 1.2.3.1), virulence (section 1.2.3.2) and mobile genetic elements (section 1.2.3.3).

#### 1.2.3.1 The Resistome

One collection of particular interest is represented by the 'resistome', which describes the broad spectrum of ARGs present within an environment or habitat [52, 137]. Patterns of resistance genes are hereby categorized and compared across sample sets in order to gain insights into the underlying structure of the microbiome, including active selective pressures [20]. In wet-lab-based microbiology, AMR can be assessed through various phenotypic tests, such as strain cultivation and the determination of antibiotic-specific MIC values [64, 66]. However, these approaches are less feasible within large-scale metagenome studies, as they are limited to characterizing antibiotic resistances across a few selected species [53, 60], which represents an issue for gut microbiome studies focusing on heterogeneous populations of uncultureable bacteria. Phenotypic tests are therefore often limited

to detecting AMR in specific indicator organisms, for instance *Escherichia coli* or *Enterococcus* spp., which represent members of the natural gut microbiota that are easy to cultivate and known to potentially harbour ARGs [53, 66]. Generating metagenomic shotgun reads can circumvent this limitation and allow an unbiased assessment of the resistome within a microbiome of interest [47]. Resistomes are known to vary in gene composition, associated microbiota and dissemination dynamics across different environments and conditions [20]. For example, the healthy human gut microbiome harbours a wide range of ARGs, spanning several antibiotic classes [20, 26, 43, 53]. ARGs are not necessarily bound to a single substance and may even induce tolerance to multiple different compounds, as frequently seen for biocide and heavy metals resistance genes [8]. However, a majority of ARGs hosted by intestinal bacteria are only distantly related to medically relevant resistance genes and as such primarily represent immobile intrinsic features of commensal microbes [2]. While these resistance genes should nonetheless be monitored, their existence does not resemble an immediate threat for public health due to the fact that only a small subset of commensal gut bacteria actually interact with foreign pathogens [47]. Resistance genes of clinical significance include *amp*C, *erm*B, *ndm*-1, *tem*1 and *van*A [47]. *Staphylococci*, for example, utilize active efflux pumps (such as *tet*K or *tet*L) to achieve tetracycline resistance [7]. In comparison, aminoglycoside and beta-lactamase resistances are based on enzymatic inactivation through operons such as *bla*, which are frequently mobile and can encoded chromosomally or plasmidial [7]. Beta-lactamase resistance can alternatively be achieved through the expression of additional penicillin-binding proteins, such as *mec*A [7].

#### 1.2.3.2 The Virulome

Pathogenic bacteria have evolved their genomic repertoire in order to invade, colonize and damage a host, often effectively bypassing the immune system [17]. This, combined with limited treatment options due to AMR, stresses the impact of these bacteria on public health [17]. Virulence hereby consists of a complex regulatory cross-talk across multiple genes and pathways [138]. Virulence factors play an important role for the pathogenicity of bacterial strains by increasing infectivity and counteracting human immune defences [17]. They often contribute to pro-inflammatory environments, promoting a state of dysbiosis within the microbiome [17]. The totality of virulence genes within a microbiome is known as the

'virulome' [77]. These factors encompass a wide range of genetic elements encoding, for example, additional secretion pathways such as *gsp* [139], or cell-surface structures like fimbriae (*fim*) which enable microbes to bind to their environment [140]. Other virulence factors are involved in the production of bacterial toxins, siderophores or regulator genes [17]. Bacteria of the genus *Staphylococcus*, for instance, may additionally possess genetic components involved in biofilm formation, as well as genes related to toxin production [7].

#### 1.2.3.3 The Mobilome

The entirety of sequences involved in HGT within an environment are known as the 'mobilome' [26, 53, 141]. MGEs play a key role in the adaptation of microorganisms to new environmental conditions [70, 142]. The majority of these consist of plasmids and phages [143], but may also include integrative conjugative elements, transposons and inverted repeats [141]. These microbial sequences are commonly underrepresented in public databases, particularly when compared to the wealth of available chromosomal data [143]. Plasmids represent a widely abundant type of MGEs [142], with over half of all marine bacterial species carrying at least one plasmid [143]. They resemble fragments of additional, extrachromosomal DNA with the ability to replicate autonomously and are frequently exchanged through HGT [142, 144]. Plasmids primarily exist in circularized form, varying in size from small (2-10 Kbp) to large (> 50 Kbp) [142, 145]. Their sequences include genetic markers such as relaxases, a group of genes involved in DNA mobilization and conjugative transfer, which have been associated with AMR conferring MGEs [8]. They also commonly carry 'cargo genes', including additional copies of ARGs conferring co-resistances [87, 146].

## 1.3 A Brief Overview of the Mammalian Gut Microbiome

Humans and animals are host to multiple complex and diverse microbial ecosystems, known as microbiomes [147]. These communities of microorganisms have dynamically co-evolved with their hosts over millions of years [92, 135, 138, 148, 149], involving microbes in almost every facet of health [3, 115, 147, 150]. They are commonly referred to as microbiota and evolve continuously over a lifetime [3], varying in composition and abundance across different anatomical sites [3, 96, 135, 151]. A single human can hereby harbour $3.9 \times 10^{13}$ bacteria, which

represents a 1:1 ratio in comparison to host cells [152]. Recent advancements in high-throughput metagenomic sequencing have enabled to study these diverse communities and sparked interest in various host-associated environments, including the gut and nasal microbiomes of equines [31].

Mammalian intestines contain a vast population of microbes, including bacteria, archaea, phages, fungi and protozoa [134, 153, 154]. In humans, gut microorganisms are estimated to range between 10-100 trillion individuals [85, 155], spanning thousands of different species [98] with a combined gene pool over 150 times the size of the human genome [80]. These microbes consist primarily of commensals and mutualists, which symbiotically aid their hosts in essential nutrient metabolism and immune responses [17, 31, 135, 150, 153, 155–158]. They hereby closely interact with their surrounding environment [148], occupying a wide range of unique functional niches within their respective hosts [85].

### 1.3.1 The Human Gut Microbiome

One microbiome of particular relevancy to health is represented by the microbiota colonizing the gastrointestinal tract, commonly referred to as the gut microbiome [155, 156, 159]. Since the human gut represents one of the most well-studied host-associated microbiomes [135], research has focused on describing the composition and health impact of the enclosed microbiota. For example, increased age has been linked with a decrease in microbial diversity [3, 122, 153, 156] and even the circadian rhythm has been shown to influence the functional profile of the gut microbiome [160]. In contrast, instability of the microbiome has been associated with the development of various complex diseases, including diabetes, depression and cancer [149, 156, 160, 161]. Tetracyclines and beta-lactamases in particular represent two frequently prescribed drug classes which are utilized to treat a variety of infections, including gastrointestinal pathogens [52, 123, 162]. Their corresponding ARGs are therefore abundant within the gut microbiota of healthy adults [43, 52, 123, 162]. However, extensive antimicrobial usage can impact the microbiota for months or even years following treatment [163, 164], with recovery times being highly variable across individuals [5].

The healthy human gut microbiome has been described to be composed of few highly abundant bacterial species and many low-abundance taxa [112]. These

predominantly consist of the phyla *Bacillota*, *Bacteroidota*, *Actinomycetota* and *Verrucomicrobiota*, with their abundance greatly varying across different parts of the intestines [150, 156]. However, characterization on more specific taxonomic levels has been challenging, as up to 50% of the species inhabiting the gut still lack corresponding genomic reference sequences within current databases [133]. Although only 1% of the human microbiota composes of the *Pseudomonadota* phylum [98], this taxonomic group includes multiple potential pathogens, such as *Escherichia*, *Klebsiella* and *Pseudomonas*. Their abundance however remains comparatively low in healthy microbiomes [150]. The conserved, intestinal microbiome of other mammals also contains a variety of microbial taxa, ranging up to more than 400 unique bacterial species within pigs [66].

### 1.3.2 The Equine Gut Microbiome

Domesticated horses (*Equus caballus*) [165] are mammals with distinct physiology, behaviour and diseases which occupy a unique niche through their close interactions with humans [166]. They are both physically (as companion animals) and culturally (as livestock) closely linked to humans [123] and have played an important role in the development of society, particularly as working animals, with domestication starting approximately 4,000 years ago [166] and their subsequent diversification into various breeds [167]. Encompassing a worldwide population of 60 million, the horse remains a core component of many societies to this day, especially as companion animals for sports, transportation and agriculture [167].

*Equidae* represent a family of non-ruminant herbivores, carrying both large and complex gastrointestinal systems [157, 168], encompassing up to $10^{15}$ bacterial cells [31]. Horses are foraging herbivores that feed on fibrous, cellulose-rich grasses [167]. As hindgut fermenters with a particularly diverse gut microbiome [135, 168], their microbiota plays an important role in energy production, digestion and overall health [157, 169]. A majority of their daily energy requirements are provided by short-chain fatty acids, including acetate, butyrate and propionate, which are produced through anaerobic fermentation of fibre within the large intestine [31, 153, 157, 168]. Cellulosic forage is hereby metabolized by fibrolytic microbes within the gut [31, 168]. The resulting fatty acids provide a unique energy source for the equine host and its microbiome [157]. The gut microbiota has furthermore been described to symbiotically aid its host by neutralizing toxins, protecting

against pathogens, regulating gene expression within the epithelial cells and has been associated with the overall immune response [157]. This complex ecosystem maintains a sensitive homeostasis through a combination of genetics, environment and lifestyle.

The gut environment represents a dynamic microbial community [156], exhibiting high individual variation in regards to taxonomic composition and metabolic potential [31, 115, 122, 149, 159]. Even genetically identical mice differ in their microbiota due to minor variations regarding early life exposures and environmental parameters [5]. Diet and age, for example, represent two factors known to actively affect the microbiome [156, 170]. Diet in particular represents a major driver for gut diversity [153, 154, 171], with food intake known to impact taxonomic composition [3, 163]. For equines, transitioning between distinct types of feed has been shown to have profound influences on dominant phyla within the gut environment [153, 157]. Age is another factor known to affect the structure of the gut microbiome [31, 98, 167]. Increased age has also been associated with a decrease in microbial diversity in equines, with adults (5-12 years of age) displaying an enriched diversity compared to elders (19-28 years) [31, 153, 157]. Contrary to initial findings, foals are actually born with a complex intestinal microbiota, which continues to develop until 50-60 days postpartum [31, 157, 164]. The early gut microbiome is hereby colonized by parental (milk and birth canal), as well as environmental microbes immediately after birth [31, 155, 164].

However, while individual variations exist, maintaining a stable microbiome is crucial for the overall health of the host species [157]. Current research outlines that the gut microbiome is inhabited by a species-rich community of microorganisms [150] and that a sensitive homeostasis of these highly diverse microbial populations acts as a colonization resistance against foreign bacteria [3, 31]. While the introduction or extinction of specific microbes can induce variation in the functional profile of the microbiota [3], natural microbiomes, such as the healthy gut, share high levels of functional redundancy across community members [85]. Multiple species hereby share the ability to engage within the same metabolic processes, enabling the microbiome to remain comparatively resilient against taxonomic fluctuations induced by environmental changes [15]. While this allows the microbiome to persist many environmental influences, it remains susceptible to

persistent stressors such as dietary changes, lifestyle and travelling [3, 31, 154, 157]. This may explain how seemingly minor variations can have profound impacts on gut composition and host health [159].

The term 'core microbiome' is therefore introduced to represent the set of microorganisms present in all, or a vast majority, of samples within a habitat [3, 31, 85]. This typically includes the most stable, consistent populations of bacterial species known to be involved in key functions of their habitat, exercising crucial roles within their ecosystem [31, 157]. The mammalian core microbiome contains bacteria such as *Clostridiales* and aerobic *Lachnospiraceae*, which are involved in butyrate production and thus play an important role in providing a protective function for the gut environment [31]. The core gut community of equines appears to comprise of a smaller set of bacterial species, which may indicate why horses are increasingly susceptible to disruption of the microbiome equilibrium, frequently resulting in the development of gastrointestinal diseases [31, 153]. Their intestinal microbiome consists of four major phyla: *Bacillota*, followed in relative abundance by *Bacteroidota*, *Pseudomonadota* and *Actinomycetota* [157, 167, 168]. This small group of core microbes contains taxa present across equines of different ages, diets and treatments [153, 157]. While the species-level composition of the horse microbiome still remains largely undescribed, microbial profiling revealed that *Staphylococcus* spp. and *E. coli* represent two common commensals within the equine gut and the mucosal membrane [33]. Additionally, species of the phyla *Actinomycetota*, *Bacteroidota* and *Pseudomonadota* are frequently found in high abundance across plant-associated microbiomes [172]. Microbiota are often classified as beneficial or harmful for the health of their host [150] and while this distinction is more complex than the simple presence-absence of particular taxa, the primary phyla of the healthy equine hindgut have been described to include *Bacillota*, *Bacteroidota* and *Verrucomicrobiota* [31, 168]. Methanogenic archaea also represent abundant inhabitants of the equine colon [31].

Dysbiosis describes a rapidly occurring imbalance within the sensitive microbiome equilibrium, resulting in altered microbiota composition and overall reduced species diversity [31, 150, 164]. In equines, dysbiosis can result in inflammation of the gut, heavily impacting the essential fermentation processes of the microbiome [31, 164]. This increases susceptibility to foreign infections and promotes the de-

velopment of various metabolic disorders [31, 100]. Causes for these drastic shifts in microbiome composition include sudden waves of bacterial deaths induced by extreme stressors such as antibiotic administration or diseases [157]. While some of these stimuli may only persist for short periods of time, they can have profound effects on the overall structure of the host microbiome [156]. Multiple hypotheses exist as to how and when pathogen overgrowth occurs. On the one hand, impaired species diversity has been associated with increased vulnerability to foreign infections and diseases [15]. On the other side, loss of individual key taxa can result in a cascading effect causing additional extinction events [3]. The overrepresentation of specific phyla, such as *Bacillota* and *Pseudomonadota*, for example, has been linked to a disruption of the gut environment [81]. Gut health has also diagnostically been described through the computation of several metrics, including the *Bacteroidota/Bacillota* and *Bacillota/Pseudomonadota* ratios [131, 169, 173]. The abundance of *Bacteroidota* commonly outweighs *Bacillota* within a healthy gut environment, while the ratio is found to be impaired throughout dysbiosis [81, 156].

#### 1.3.2.1 Diseases Affecting the Gut Microbiota

Diseases affecting the gastrointestinal tract represent the leading cause of mortality in horses and include equine colic [168]. Large knowledge gaps hinder effective prevention and treatment options for this condition [31]. Colic is a broad syndrome of prolonged abdominal pain caused by disruptions within the gut microbiome [157, 169, 174]. It represents one of the major causes of morbidity and mortality in equines [157] and remains an ongoing issue throughout veterinary medicine, with annual incidences of 4.2% in the US alone [174]. Even with immediate medical attention, including operative surgery, only 63% of diagnosed horses are able to recover from this serious health condition [31, 169, 174]. While the development of colic remains poorly understood, it is influenced by multiple risk factors, such as age, diet and environmental conditions [169]. In addition, changes within the faecal microbiota can be observed throughout the onset of colic [169]. Elevated abundances of *Pseudomonadota*, including Gram-negative bacteria such as *Enterobacteriales* and *Pseudomonadales*, some of which are known to be involved in intestinal nitrogen fixation, have been associated with episodes of colic [31, 169]. The enteral diversity of *Pseudomonadota* is hereby mainly driven through environmental intake [31].

#### 1.3.2.2 Impact of Antibiotics on the Gut Microbiota

Antimicrobials have had an incredible impact on human and veterinary medicine through their ability to efficiently disable pathogens, representing one of the most frequently prescribed drugs for the treatment and prophylaxis of clinical infections [157]. Infectious complications following gastrointestinal surgery remain an ongoing issue throughout human and equine medicine and are associated with high rates of mortality and morbidity [150], alongside increased recovery times and hospitalization lengths [174]. To minimize the likelihood of post-surgical site complications, short-term perioperative antibiotic prophylaxis (PAP) are routinely administered in order to prevent wound infections following open surgeries [155, 163, 174, 175].

In equine medicine, a common PAP for colic surgery consists of a mixture of penicillin and gentamicin, which is administered up until 5 days after the intervention [175]. Antibiotic administration has been shown to induce various degrees of post-operative changes within the microbiome, depending on the individual substance and its potential for intestinal penetration [3, 163, 176]. While PAP are typically lower in dosage compared to treatments of acute infections, they can cause effects comparable to those of antibiotic therapies [56]. Even a single perioperative dosage is able to disrupt the gut microbiota composition, often outweighing the effects of the surgical procedure [163]. For human surgery, a single dosage of PAP is deemed sufficient for patients lacking significant risk factors, with current guidelines establishing that any usage past the 24-hour mark yields no additional health benefit and may even be associated with detrimental side effects, such as increased chances of MDR bacterial infections [177, 178]. Although antibiotic administration beyond 24 hours post-intervention has been described as being therapeutic and lacks significant health benefits compared to single dosage PAP, perioperative prophylaxis in equine medicine routinely spans multiple days post-surgery [174]. This gives reason to reconsider current standardized protocols for PAP usage within equine medicine [174].

Antibiotics hereby affect all susceptible microbiota, driving changes within the gut microbiome on high taxonomic levels [168], while also providing new ecological niches for resistant strains to emerge [156]. Although specifics may differ depend-

ing on the drug, host and administration route [123], prolonged treatment with antimicrobials has generally been shown to decrease diversity within the gut environment [5, 156], perturbing community composition and resulting in dysbiosis [179]. Additional side effects include long-term influences on the host microbiota, such as the loss of key species and the development of AMR [31, 157], both of which have been associated with adverse clinical outcomes and increased rates of pathogenic infections [135, 176, 180]. Even prophylactic dosages induce positive selection in regards to the development of resistant strains [49]. ARG abundance spikes sharply following successive antibiotic treatment, correlating with decreasing microbiome diversity and richness [33, 59, 123]. Microbial shifts induced by antibiotics are detectable a few days after administration, however it may take weeks to restore the original diversity of the gut [157]. For equines, one study reported the largest impact on the diversity and richness of equine faecal samples at 5 days post-administration, although the original diversity was not fully restored by day 25-30 post-treatment [181], illustrating that oral antimicrobials can have prolonged effects on the gut microbiome and should be used with caution [31, 157]. Specific ARGs may even remain in the microbiome for extensive periods of time ($>$ 34 days) [122, 123], acting as potential reservoirs long after administration has concluded [49, 56].

#### 1.3.2.3 MDR within the Gut Microbiota

Strains associated with MDR have quadrupled over the recent two decades [182], with both the isolation and local spread of MDR bacterial strains being well reported within equine healthcare [33, 175]. These pathogens are of serious concern, since treatment options remain limited and they harbour the potential to zoonotically infect staff, resulting in further transmission across equine patients and the environment [175]. Veterinary healthcare clinics hereby represent hotspots for the local spread of MDR bacteria, including *Acinetobacter* spp., ESBL-producing *Enterobacterales* and MRSA, which remaining ongoing challenges within these facilities [175]. Further studies identified *E. coli* and *S. aureus* as dominant nosocomial pathogens throughout facilities providing healthcare for horses [33], which relates to the increasing rates of carriage and faecal shedding of ESBL-producing *E. coli* reported across equine clinics [175]. The gut microbiome in particular can act as a reservoir for multiple ARGs and pathogens [167], including the aforementioned bacteria, with both hospital stay and antibiotic exposure representing

crucial risk factors for the colonization of ESBL-producing *E. coli* in equines [175]. Frequent contact between humans and horses, e.g. through riding schools or clinics, might furthermore be linked to an enhanced risk for the zoonotic acquisition of ESBL-producing *E. coli* [123].

## 1.4 A Brief Overview of the Mammalian Nasal Microbiome

The oral and gut microbiome of humans have been found to be closely interlinked, indicating that the oral bacterial population contributes to seeding of the gut environment [122]. Combined, the oral and intestinal microbiota are estimated to contain over 45 million genes, effectively encompassing 2000-fold more genetic material than the human genome [183]. However, in contrast to the gut, the nasal microbiome still remains largely unexplored. Nasal swabs are hereby particularly challenging to analyse, due to the low amount of bacterial DNA they contain [184]. Contamination of these low-biomass samples with foreign sequences is therefore an ongoing issue for nasal microbiome analyses, requiring dedicated methodology and careful evaluation [5, 128, 184].

### 1.4.1 The Human Nasal Microbiome

In humans the most abundant taxa of the nasal microbiota encompass *Corynebacterium*, *Neisseria*, *Staphylococcus* and *Streptococcus* [151]. A healthy microbiome is maintained through close interactions of these bacterial groups, inhibiting the growth of pathogens and modulating the immune system, similar to the gut equilibrium [184]. Pulmonary diseases, such as asthma or cystic fibrosis, have also shown to distinctively influence the microbiota in both composition, diversity and metabolic potential [184, 185]. One particularly common microbe of the nasal microbiome is represented by *S. aureus*, which asymptomatically colonizes over a third of the human population [138]. The oral microbiota have further been described to contain *S. mitis*, *S. oralis*, *S. pneumoniae*, *S. salivarius* and *Veillonella* spp., but large interindividual differences exist [186]. Phages are also highly abundant, with estimates of over 35 times more phages than bacteria within the oral cavity of humans [143]. Opportunistic pathogens with the ability to inhabit the nasal mucosa include highly adaptable species such as *Acinetobacter baumannii* [34, 182, 187], which colonize poultry populations [188] and occur seasonally within cattle [189]. Several species of *Acinetobacter* represent pathogens associ-

ated with high rates of AMR [146]. *A. baumannii* in particular have been shown to co-inhabit both the nasal and gut microbiomes in cases of acute infection [189].

### 1.4.2 The Equine Nasal Microbiome

*Equidae* are nostril breathers with high ventilation throughputs of over 60 l/min at rest [184]. Their nasal microbiomes represent part of a series of distinct microbial communities inhabiting the airways [184], with their composition reflecting the general health status of the host [151]. Studies focussing on the lung microbiomes of equines revealed that the three major phyla include *Pseudomonadota*, *Bacteroidota* and *Bacillota* [184]. The nasal microbiota is known to closely interact with the environment of its host [151] and is as such easily affected by abiotic factors, including seasonal variations [3]. The encompassed bacterial community is further highly susceptible to environmental influences, such as dust particles, which have been described to represent a potential reservoir for the transmission of ESBL-producing *E. coli* due to their absorption by the nasal microbiome [54, 175].

# 2 Materials & Methods

## 2.1 Study Design

Over the recent decade, metagenomic research has begun to unravel the link between microbiome composition and host health [122, 157]. Technological advancements in NGS have enabled the characterization of individual taxa present within the microbiota and assessment of environmental influences (such as PAP) on the microbiome structure [157]. This work is part of a multiomics One Health project to study the effect of different antibiotic prophylaxes on the nasal and gut microbiota of hospitalized horses. This chapter will first describe the general study outline (section 2.1), as well as the corresponding sample sets which were subsequently analysed (sections 2.1.1 and 2.1.2), before establishing some computational concepts regarding 16S rRNA and metagenome shotgun analyses (sections 2.2 and 2.3) and defining the overarching research goals of this thesis (section 2.4).

The presented work is part of the interdisciplinary research network #1Health-PREVENT[2], funded by the German Federal Ministry of Education and Research (grants 01KI1727F and 01KI1727; 01KI2009D and 01KI2009F). It was further supported by the German Research Foundation, through the NGS-CC research infrastructure (project 407495230) within the NGS competence network (project 423957469). The funding parties did not influence data interpretation or writing of this thesis. All sequencing was performed at the Competence Centre for Genomic Analysis in Kiel, Germany. Within this #1Health-PREVENT project, the effect of two different PAPs on the gut and nasal microbiomes of hospitalized equines were comparatively investigated through a randomized, longitudinal pilot study. Since the equine microbiome remains poorly understood, additional research is required to determine the effects of PAP and the development of AMR within these environments [31, 157]. Previous multiomics studies focused on describing the clinical impact of the different PAP regimes [174], the local dynamics of ESBL-producing *E. coli* within the clinic environment [175] and the effect of PAP on the gut microbiome through 16S rRNA sequencing [190]. This thesis expands on the aforementioned results by computationally characterizing the im-

[2]https://www.zoonosen.net/en/forschungsnetz/verbunde-nachwuchsgruppen/1health-prevent

pact of PAP on the gut and nasal microbiome of hospitalized equines, as well as the effect on the resistome, through a combination of metagenome shotgun and 16S rRNA sequencing.

This cross-sectional study was conducted at a horse clinic in Germany between January 2018 and February 2020. According to German law for research on animal subjects, comparison of two PAP regimens does not require regulatory approval (Landesamt für Gesundheit und Soziales, Berlin, 18.04.2017). Written consent for study participation was acquired from the respective horse owners at hospital admission. Equine patients subjected to colic surgery were enrolled based on strict recruitment criteria. Horses were required to be older than 1 year, with no clinical signs of infectious diseases or antibiotic treatment prior to the intervention and had to successfully recover from anaesthesia after surgery [174, 175]. In addition, patients were excluded from participation if their hospital stay ended prematurely or there was any deviation from the antibiotic regime [175]. Hospitalized horses undergoing laparotomy received PAP either as a conventional, continuous 5-day prophylaxis (5DG) or as a single-shot (SSG). Equine patients were assigned their respective study group through a sealed envelope drawn prior to surgery [175]. Both groups were intravenously administered a PAP containing a mixture of penicillin and gentamicin, commonly utilized for equine surgery [174, 175]. While the SSG received only a single dosage of perioperative prophylaxes, administration of antibiotics was continued for the 5DG until day 5 after the intervention. In order to monitor changes within the microbiota, faecal and nostril samples were collected longitudinally across three time points: directly upon hospital admission ($t_0$), 3 days ($t_1$) and 10 days ($t_2$) post-surgery [175]. In total, 63 faecal and 79 nasal swabs were collected from 31 hospitalized equine patients. A comprehensive overview of all samples analysed within this study is illustrated in appendix table A.2. This includes dedicated IDs for each equine patient (ranging from 1 to 98), as well as identifiers for specific samples of the gut and nasal microbiomes.

### 2.1.1 Faecal Samples

Faeces are frequently selected as a proxy for the analysis of the gut microbiome due to limitations in obtaining intestinal samples [134]. Recent studies have demonstrated that the microbial population structure of faeces do not strongly differ

from samples taken directly from the large intestine of equines [31, 154, 168, 169]. Faecal samples therefore represent adequate, non-invasive proxies for the study of changes occurring within the gut microbiome [167, 168]. They encode inter-individual differences [85], while being comparatively stable against factors such as age or equine breeds [31].

Faecal samples were directly transferred into tubes and remained cooled until shipment on dry ice to the sequencing facility [175]. DNA extraction and sequencing was performed by collaborators at the Competence Centre for Genomic Analysis in Kiel, Germany. A total of 63 faecal samples were selected for metagenome shotgun sequencing based on the previously established criteria of the study. Genomic DNA was extracted from the sample collection using a commercially available extraction kit (QIAamp Fast DNA Stool Mini Kit, Qiagen, Germany), according to the manufacturers protocol. Paired-end libraries ($2 \times 150$ bp) were created and subsequently sequenced on an Illumina MiSeq platform. Sequencing parameters, such as read depth, are known to affect the ability to detect rare taxa and AMR genes [18, 66]. Read lengths of $2 \times 150$ bp were selected to enable comparison with contemporaneous studies of the microbiome [191, 192]. For metagenomic shotgun sequencing, a depth of 1 million read pairs has been demonstrated to sufficiently capture a majority of the gut microbiome diversity in humans [65, 115], with sequencing beyond 60 million read pairs not demonstrating any further improvement regarding the taxonomic classification of faecal samples [134]. Samples were therefore deeply sequenced to a sufficient depth between these two extremes, resulting in an average of 31 million read pairs per sample. This generated a total of 3 billion reads, encompassing over 600 billion bp of genomic gut microbiome data for the entire sample set. Raw data has been deposited in the sequence read archive (SRA) of the National Center for Biotechnology Information (NCBI) under the BioProject ID PRJNA998844.

To additionally expand upon this collection, further literature research was performed in order to identify similar, publicly available data sets. Since technical variability from experimental methods is known to influence downstream analyses [5, 135], samples were selected to be comparable in sequencing depth, geographic location and methodology to the established collection. A set of five additional equine faecal samples were identified and retrieved from the SRA of the NCBI

using the search terms 'equine' AND 'metagenome' AND 'gut' OR 'faeces', analogous to prior research [16]. The selected samples had previously been utilized to reconstruct a large amount of novel metagenomic draft genomes originating from the equine gut [167], enabling enhanced classification resolution for the presented study throughout downstream analysis. Corresponding raw data were downloaded from BioProject PRJNA590977 in .fastq format through SRA Explorer[3].

### 2.1.2 Nostril Swabs

In addition to studying the effects of PAP on the aforementioned gut metagenome, the nasal microbiome was also investigated within the context of this thesis. Nostrils of the equine patients were hereby sampled using swabs and subsequently sequenced. Since nasal swabs represent low-biomass samples, they are particularly susceptible to foreign contamination. A majority of the samples ($n = 78$) were therefore chosen to be 16S rRNA sequenced, which effectively voids the presence of any host DNA fragments due to the absence of this gene in eukaryotes. Metagenome shotgun sequencing was additionally performed on one sample (J32522), representing the prolonged 5DG.

DNA extraction was hereby performed analogously to the faecal samples by collaborations at the Competence Centre for Genomic Analysis in Kiel, Germany. For 16S rRNA sequencing, the selection of the variable region defines the ability to characterize the bacterial composition of a sample [122]. An overview of the 16S rRNA gene is visualized in figure 1.1. The V2 and V4 regions are both generally considered to enable highly consistent estimations for taxonomic profiling [115]. Furthermore, V1-V2 and V3-V4 illustrated favourable classification results when compared to other fragments of the 16S rRNA gene [116]. The V1-V3 region in particular has been preferred for oral microbiome studies in humans [122]. Read lengths above 100 bp also remain essential for 16S rRNA profiling [102], with longer sequences resulting in increased sensitivity throughout downstream analysis [45]. Nasal swabs were subsequently V3-V4 $2 \times 300$ bp amplicon sequenced on an Illumina MiSeq platform using a commercially available primer set. Sequences were received in the adapter-trimmed, demultiplexed .fastq format. Raw data has been deposited in the SRA under the BioProject ID PRJNA998844.

[3] https://sra-explorer.info/

## 2.2 16S rRNA Analysis

16S rRNA analyses, sometimes also referred to as metataxonomics, represent methods which utilize data preprocessing (section 2.2.1), taxonomic assignment (section 2.2.2), normalization (section 2.2.3) and diversity estimation (sections 2.2.4 and 2.2.5) to gain insight into the microbial composition of a sample by characterizing fragments of the 16S rRNA gene [111, 118]. 16S amplicon and metagenomic shotgun sequencing hereby share a variety of overlapping methods and data types [92, 118]. They both heavily utilize taxonomic counts of abundance tables, representing the amount of reads matching specific taxa, in order to compute microbial compositions and diversities [118]. The following sections will therefore describe the computational concepts utilized for 16S rRNA (section 2.2) and metagenomic sequencing (section 2.3), respectively.

### 2.2.1 Data Preprocessing

Microbiome sequences generally represent a highly fragmented and noisy source of data [43, 118]. Additional preprocessing is therefore strictly required before any further analyses can be conducted in order to distinguish between real biological signals and errors arising from imperfect sequencing [1, 5, 114, 131]. Quality control (QC) is performed through a combination of previously established thresholds, designed to remove technical noise. Initial QC mainly consists of primer removal and demultiplexing [183]. Next, paired-end reads are joined into longer fragments of DNA in order to improve lengths by overlapping sequences [1, 130, 151, 193]. Finally, filtering of short sequences, reads containing ambiguous bases and chimeric fragments is performed on the data set [118, 193].

### 2.2.2 OTU Clustering

Following initial QC, the next step for preprocessing 16S rRNA amplicon data is summarizing high-quality reads into sets of representative sequences, known as operational taxonomic units (OTUs) [1, 5, 92, 96, 127, 183]. This resolves sequencing errors through the merging of DNA fragments at a predefined similarity threshold, computationally known as clustering [114, 193]. Amplicons are hereby commonly clustered on an identity threshold of 97% [135, 151, 179, 183], meaning that reads with minor errors are assigned to a shared consensus sequence repre-

sented by the OTU [5, 118]. OTUs resemble groups of closely related bacterial species [194], defined by curated 16S rRNA databases such as GreenGenes [195] and RDP [196]. However, these databases vary in composition and nomenclature, resulting in differences when utilized for taxonomic profiling [111, 116, 197].

Next, taxa abundances are summarized as large matrices containing the observed number of reads per OTU across a sample set [5, 118]. Known as abundance, feature or OTU tables, columns represent samples and rows list counts of individual taxa [183]. OTUs with low counts are commonly removed from abundance tables. Known as singletons, these counts likely represent false-positive sequencing artefacts or spurious low-level contaminants and are therefore discarded [98, 119, 121, 153, 154]. Removing low-abundance features has further been shown to improve comparative analysis, in particular statistical methods, later downstream [198]. Various count thresholds have been proposed for this filtering [98, 121, 153, 198].

### 2.2.3 Rarefaction Analysis

However, summarized OTU tables do not reflect absolute abundances of bacterial taxa, but instead represent count data resulting from a process of random sampling [113, 183], limited by the amount of reads produced from the sequencing platform [118]. OTU tables are of extremely high dimensionality, commonly consisting of thousands of taxa [5], with individual counts varying greatly across sample sets [118]. Since a majority of OTUs are only observed within a few samples [118], their distribution is often skewed [92, 199], extremely sparse (with up 90% of these matrices being zero) [5, 71, 183] and overdispersed [113, 198]. Their raw abundances and sums reflect only a fraction of the original composition within the environment [71, 199] and can therefore not be directly compared across different samples [118, 183]. Variations in sequencing depth can also heavily affect the count data, confounding downstream analyses such as the assessment of taxonomic diversity [5, 128]. In order to enable comparisons across samples, further normalization of the data remains essential [71, 118, 200].

Rarefying is a normalization method commonly used for comparing sample sets with uneven sequencing depth [4, 153, 200]. It was initially introduced as a method to deal with the influence of rare taxa on diversity estimation [183] and has since been established as a standard for normalizing metagenomic data [87].

First, a minimum sequencing depth is assessed through rarefaction analysis [112, 183]. A measure for diversity (such as the total amount of taxa) is hereby successively plotted against varying sizes of subsampled sequencing depths [1, 4, 183]. Mathematically, successive random subsampling of the data is performed and the enclosed number of taxa are visualized [173]. At first, rarefaction curves rise exponentially due to the inclusion of highly abundant species [4]. With increasing sequencing depth, many new taxa are added to these subsets [173, 201]. After these common species have been recorded, rare species are discovered at a much slower rate [4, 173]. With the rate of new taxa steadily decreasing, the curve effectively levels off into a plateau [4], eventually reaching an asymptote at which no further species are recorded [169, 173]. As the slope of this curve approaches zero it levels out, indicating that a majority of the biological diversity has been covered [1, 128, 168]. Once this threshold has been established, reads are randomly drawn from the OTU table without replacement until each sample reaches a uniform sequencing depth [183, 200]. The remaining fragments are then discarded from further analyses [200]. For 16S rRNA analyses of the equine gut microbiome, a depth of 10,000 reads has been described to sufficiently capture the general microbiota composition [154].

The rarefied OTU tables can then be utilized for further downstream analyses, including the assessment of differential abundance. Differential analysis can be performed by using parametric tools such as the edgeR package [202]. Originally designed for transcriptomic analyses, these methods have since been established within the field of metagenomics [8, 17, 71, 118, 186, 198]. In edgeR, count matrices are normalized by comparing all samples to a selected reference [200]. A scaling factor is then computed by utilizing the weighted trimmed mean of log abundance ratios after trimming the upper and lower 30% of the data [183, 200]. Observed abundances are then modelled using a negative binomial distribution [118, 154, 183], which represents an extension of the Poisson distribution more suitable for the comparison of count data, as it allows increased variability [183]. Under the assumption that a majority of the count data does not differ between conditions, differential abundance is finally assessed for each taxon [71].

### 2.2.4 Diversity Estimation

Diversity represents a core concept in ecology which is utilized to characterize habitats [4], quantify their community structure [111] and assess overall ecosystem stability [135]. It can be divided into several components: the taxonomic composition per sample (alpha diversity), the species variation across a collection of a data sets (beta diversity) and the sum of both for an entire habitat (gamma diversity) [201, 203, 204]. Due to the nondeterministic process of sampling through sequencing, it is impossible to directly assess these values for complex microbiomes. Instead, the aforementioned diversity measurements need to be mathematically estimated from the data through the computation of ecological indices [205]. Diversity indices represent statistics computed directly from a sample set [206], such as abundance matrices of 16S rRNA or shotgun metagenome data. They encode varying degrees of diversity based on the classification of individuals within a population into pre-selected groups (e.g. taxonomic levels) [206]. Using a combination of multiple diversity indices enables in-depth insights into the underlying ecological structure of a sample set and can reveal different facets of the data throughout downstream analysis. The following subsections will therefore describe both, the concepts of alpha (section 2.2.4) and beta diversity (section 2.2.5) in further detail. In addition, various forms of statistical representations will be assessed for these measurements. An overview of the concepts of diversity is illustrated in figure 2.1.

Alpha diversity is commonly referred to as the diversity within an individual sample, such as a microbial population [5, 110, 118, 128]. It is calculated samplewise in order to compare means across different groups of individuals [168, 201]. While multiple indices exist for assessing these values, the most common measures include richness ($R$) [205], evenness ($J$) [112] and the Shannon diversity ($H$) [207]. Which measure to use will depend on the aspect of diversity an analysis will focus on: richness, evenness or a combination of both [4]. The most fundamental form of diversity is represented by richness ($R$), which plainly describes the total amount of unique taxonomic groups (for example microbial species) per sample [118, 130, 135, 173, 184, 205], with no regard to their abundances [208]. In contrast, evenness ($J$), assesses the homogeneity of the distribution of

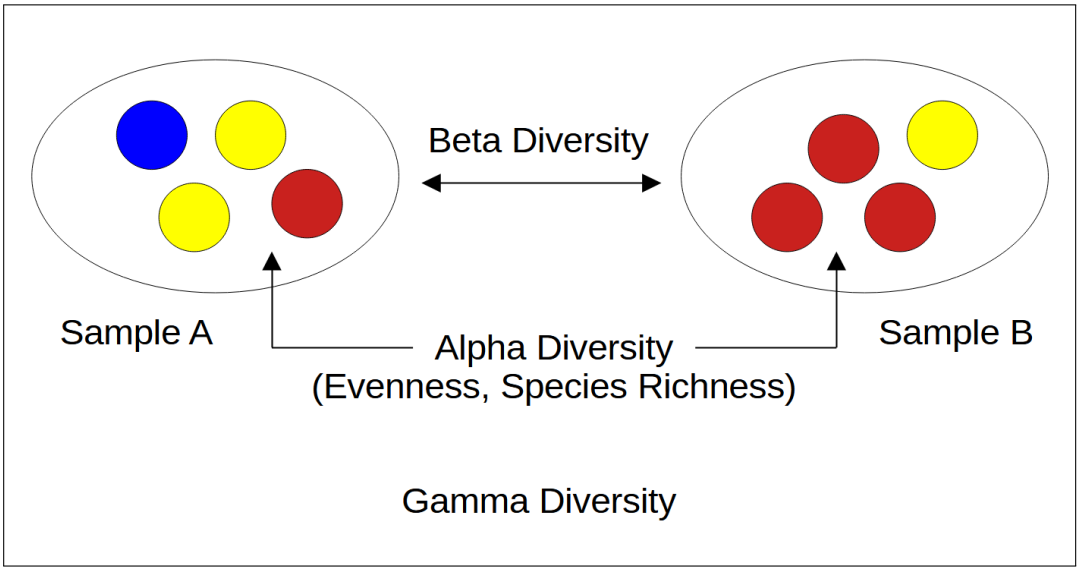


**Figure 2.1: Measures of Diversity.** Metagenomic samples (A and B) consist of individual microbes (circles), spanning a variety of species (colours). Alpha diversity refers to sample-specific indices which depend on the taxonomic richness and evenness of a community. Beta diversity, in contrast, is utilized to compare species profiles across multiple different samples. The totality of taxonomic variance within a large habitat is summarized as gamma diversity, which encompasses values of both, alpha and beta diversity.

taxa abundances within a sample [112, 118]. It is mathematically defined as:

$$J = H/\log(R) \qquad [134]$$

with low values illustrating an uneven composition and large values reflecting taxonomic equality [203]. The most commonly utilized measure for alpha diversity is represented by the Shannon index (also known as the Shannon-Wiener index), which combines information on both the richness and evenness of a community [4, 134, 184, 205]. It weights the amount of species by their relative evenness through a mathematical function [73, 128, 135, 208]. The formula is based on early work in the field of communication theory, originating from the question on how to predict the next letters of an incoming message [205, 207]. This definition of uncertainty is used for ecological diversity assessment to this day and is commonly represented by the Shannon function ($H$):

$$H = -\sum_{i=1}^{n} p_i \ln p_i \qquad [134, 205, 208]$$

in which for each taxon ($i$) within a community of species ($n$), its proportion ($p_i$) is assessed and multiplied by its natural logarithm. The total of all abundances are then summed and sign inversed. Shannon indices directly depend on the richness and evenness of a sample, but are, in contrast to other ecological measures, comparably less sensitive to $R$ and instead more directed by $J$ [4]. A population with many species of similar abundance is hereby more diverse when compared to a community of few species with high variability [73]. As $H$ starts at 0, larger values are associated with increased diversity [128].

### 2.2.5 Dissimilarity Computation

As measures for alpha diversity are based on species counts and their distribution within a community, they may oversimplify the true taxonomic profile of a sample [119]. Different samples may even share identical values for alpha diversity, although their communities vary greatly in composition [208]. Additional measures are therefore needed for directly comparing the species profiles across multiple ecological communities, including the commonly utilized concept of beta diversity [204, 208]. Beta diversity is often referred to as a measure for reflecting taxonomic distances between samples of a data set [5, 118, 129, 170, 201]. Beta diversity is calculated between the abundance profiles of each sample within a set [128]. Mathematically, pairwise distances are computed between points of high dimensionality and are summarized within a square matrix [5, 128, 129, 208]. Historically, this has been utilized to compare samples across different habitats or environmental conditions [4].

Comparable to alpha diversity, there exists a wide range of ecological distance and dissimilarity measures for quantifying beta diversity between microbial profiles, with limited agreement on best practices [4, 118, 208]. The specific choice of distance metric can hereby impact the obtained outcome and should be selected with the research question in mind [5, 129]. Pairwise distances can either be based on the presence-absence of individual taxa or quantitative abundance data [129]. In addition, they can be supplemented with phylogenetic information or be applied without these [128]. One common non-phylogenetic measurement for the beta diversity is represented by the Bray-Curtis dissimilarity index ($BC$) [118, 128, 209]. The $BC$ distance metric is based on the abundance of individual taxa

[5] and is mathematically defined as:

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j} \qquad [116]$$

between samples $i$ and $j$. $C_{ij}$ represents the sum of lesser counts for shared species, while $S_i$ and $S_j$ illustrate the total amount of counts per sample [116]. The resulting values range between 0 and 1, with larger values representing increased differences across samples, while small values indicate high similarity [116, 128]. These can be utilized for the visualization of dissimilarity and to compute mean distances across groups of samples [201].

Microbiome tables typically span thousands of entries, resulting in extremely high dimensionality, which is challenging to handle throughout downstream statistical analyses [113]. Methods for the reduction of dimensionality, such as principal component analyses (PCAs), are therefore commonly applied for metagenomic data sets [5, 210]. Mathematically, these methods compress high-dimensional data (such as OTU tables) into a reduced set of axes (principal components), which try to retain the main features and distances between the samples [5, 118, 199]. Known as principal components, they encompass much of the original variance, while reducing overall complexity of the target data [211]. The first components aim to explain a majority of the variability, allowing direct comparisons of individual samples and groups based on a selected few values [211]. These are typically utilized for visualization in 2D plots, enabling the detection of sample clusters. Groups within PCA plots can further be coloured according to the available metadata, such as treatment groups, in order to assess the influence of multiple parameters on sample orientation [128]. For metagenomics, PCA is commonly performed to visualize distances of beta diversities [5, 118]. Hereby, the most informative features are projected onto ordination plots [208], enabling clustering of the individual data points.

## 2.3 Metagenome Shotgun Analysis

Metagenomic shotgun sequencing generates a rich array of data, which requires dedicated computational methods in order to gain biological insights into the samples at hand [73, 212]. In some aspects, analysis of metagenome shotgun data is comparable to 16S rRNA sequencing, in that NGS data preprocessing (section

2.3.1) and taxonomic assignment (section 2.3.2) represent core components for all downstream analyses by utilizing random short reads instead of gene marker fragments. Metagenome shotgun analyses further expand on these with the additional ability to perform functional assessment (section 2.3.3), (co-)assembly (section 2.3.4) and full-length resistance gene profiling (section 2.3.5).

### 2.3.1 Data Preprocessing

As metagenomic shotgun sequences consist of a mixture of genomes [105], they are subjected to the same QC metrics as traditional isolate data would be. *In silico* QC remains an essential step for NGS [124] and includes the removal of sequencing artefacts, including low-quality or contaminated reads [73, 105]. Removing these greatly improves the accuracy of downstream metagenome analyses, such as taxonomic quantification, microbial diversity assessment and assembly [105, 111, 117]. For this, various QC metrics are evaluated based on the raw sequencing data, including adapter abundances, base contents, duplication rates, *k-mer* distributions, overrepresented sequences and quality scores [124, 213]. NGS reads are commonly stored in the .fastq format [132], which encodes nucleotide sequences as well as individual per-base quality values. Since quality assessment generally depends upon the technology with little respect to the sample source [213], tools used for the QC of NGS can commonly be applied to both, isolate and metagenomic data. However, since metagenomic reads are generated through random shotgun sequencing, they also commonly include off-target DNA, such as host fragments or other contaminants [92, 111]. Raw reads therefore need to be mapped to the host genome in order to remove these sequences prior to further analysis [53, 56, 112, 115, 170, 214].

### 2.3.2 Taxonomic Profiling

One central component of metagenomic analyses is to perform *in silico* taxonomic classification of the sequenced DNA fragments using a combination of statistical and empirical models [61]. A wide range of profiling methods have been proposed for the assessment of the taxonomic origin of random shotgun sequences [72]. As amplicon analyses are based on partial sequencing of the 16S rRNA gene, they are limited in resolution to the genus-level [5, 92, 115, 122, 155], due to variable regions being near indistinguishable between individual species within a genus [96,

102, 194]. While full-length 16S rRNA genes can provide insights into the species-level [31, 115] and technologies exist to capture, sequence and analyse these [117, 215], they are not yet widely utilized and are still subject to ongoing research and extensive benchmarking. In contrast, high-throughput metagenomic shotgun sequencing has been established for over a decade, allowing the development of a multiplicity of tools for abundance estimation at the species-level [110].

Taxonomic profiling generally assigns DNA fragments of a sample to nodes within a phylogenetic tree based on their respective sequence similarity to a corresponding reference database [110, 114, 216]. With BLAST-based approaches being computationally infeasible for large environmental data sets, alternative rapid lookup algorithms have been developed for the taxonomic classification of short metagenomic reads [106, 111, 114, 121, 217]. For example, the Kraken2 software can assign taxonomic labels to random shotgun reads without the need to utilize pre-defined phylogenetic markers for profiling [104, 114, 134, 217, 218]. Using exact matches of short subsequences, each read is associated with multiple nodes within a phylogenetic tree, which are back-traced until they converge at a shared taxonomic label, known as the lowest common ancestor [96, 97, 111, 114, 218]. This requires the construction of a specialized database from genomic reference sequences and corresponding taxon IDs prior to analysis [116, 218]. First, low-complexity regions of a set of input genomes are masked from further analysis [218]. Sequences are divided into short fragments (known as *k-mers*) based on a user-specified length ($k$), which are labelled with the taxonomic classification of their genomic origin [46, 116]. The resulting database structure then contains all *k-mers* present within the reference genome set, as well as their taxon IDs [111]. Next, a query DNA sequence is also divided into *k-mer* subsequences, which are compared to the pre-computed database in order to determine their taxonomic origin [46]. Finally, the classifications of all *k-mers* across a read are aggregated into a single prediction of the sequence using a minimizer-based approach [217, 218]. Ambiguous *k-mers* mapping to multiple species are assigned the lowest common ancestor of those taxa [111, 116, 121, 218]. By default, Kraken2 utilizes $k = 31$, which represents a fine-tuned trade-off between prediction specificity and sensitivity [104, 111]. Through this process, Kraken2 can assign taxonomic labels to a majority of the sequencing fragments, independent of their origin or functional association, which is of particular interest for resistance, virulence and

MGE characterization (section 2.3.5).

However, assignments made by Kraken2 do not directly represent taxonomic abundances and may instead underestimate certain species since every read is assigned to its lowest common ancestor [116]. Taxa sharing high genomic similarities, such as members of *Bacillus*, *Pseudomonas* and *Shigella*, or those prone to genera-spanning HGT, are particularly difficult to classify with Kraken2 at species-level [104, 111, 121, 218]. In order to address these issues, estimations provided by Kraken2 can be extended with the Bracken software [104], which represents a post-processing step aiming to improve classification accuracy [106, 111, 218]. Taxonomic assignments of Kraken2 are hereby re-estimated based on the overall *k-mer* distribution and taxa abundance within the classification database [104, 218]. For this, Bracken utilizes a Bayesian algorithm to probabilistically redistribute sequences assigned to higher nodes of the taxonomy down to species-level counts [106, 111, 116, 218]. Reads are re-assigned to the taxon with the highest similarity or remain unclassified [104]. Since this stochastic approach is known to induce some technical noise, a low-abundance threshold is introduced in order to discard counts below a specific value, 10 reads by default [104, 106, 167, 192, 215].

### 2.3.3 Functional Analysis

The gut microbiome contains a wide range of species that inhabit unique ecological niches, fulfilling distinct functional roles within this complex ecosystem [4, 71]. The abundance of microbial genes known to be involved in metabolic pathways has been linked to the development of various diseases in humans and animals [85]. Through measuring the gene content of metagenomic samples, the overall biochemical potential of an environment can be assessed [36]. Metagenomes can generally be treated as single ecological units [36], with the presence-absence of specific pathways assumed to induce beneficial functions or have detrimental effects on the survival of the enclosed microbial populations [72, 86]. Instead of focusing on individual strains, the fundamental goal of these approaches is to assess the gene pool of the entire bacterial community and to characterize their genetic traits and local dynamics [148]. Functional profiling can generate lists of thousands of genes that might be difficult to investigate by themselves [219]. Further aggregation of genes into higher-level categories (such as pathways) is therefore commonly performed in order to gain insights into biochemical differences [159].

Various approaches exist to determine the metabolic potential of metagenomes [71, 220]. One comparatively straightforward method utilizes the abundances of short *k-mer* sequences to identify differences in both taxonomical and functional profiles across sample groups [87, 147]. The computed *k-mers* are hereby assigned to biochemical categories, which are then assessed for differential abundance using statistical tests, such as the Wilcoxon rank-sum [147]. Through the presence-absence of individual fragments, co-occurrences of taxa and genes can be determined, enabling insights into the metabolic capacity of a metagenome set [159]. Another method for assessing functional differences is through computational workflows such as HUMAnN3, which utilizes metagenomic reads to profile the genetic potential and relative abundance of microbial pathways within a sample [5, 80, 114, 148, 185]. HUMAnN3 first performs gene-marker-based screening to identify the species-content of a sample in order to select an appropriate set of reference genomes [114, 148]. These are then utilized to characterize concatenated input reads through mapping against the established gene catalogue, including annotations for gene families (UniRef90), biochemical functions (KEGG) and associated pathways (MetaCyc) [100, 115, 185, 221, 222].

### 2.3.4 Assembly & Binning

Metagenomic data can further be utilized to reconstruct draft genomes of environmental species through the process of reference-free *de novo* assembly [72, 92, 111, 191]. This is especially advantageous for poorly explored microbiomes, which are challenging to analyse due to the absence of reference sequences within current databases [71, 95, 111]. Additionally, since short reads frequently map to homologous regions shared across multiple genes, they may not be discriminative enough to fully characterize the functional content of a sample [217]. *De novo* assembly represents an effective method for the reconstruction of longer genetic features, including regions of HGT and full-length resistance genes [22, 132], enabling additional insights into the microbiome structure [8]. The general concept of genome reconstruction can be applied to metagenomics as well: through the process of assembly overlapping reads are iteratively merged into longer contiguous sequences (known as contigs), which are subsequently ordered into scaffolds [73, 90, 114, 223]. However, not all assumptions for isolate assembly hold true for metagenomic samples, as these possess unique traits which need to be considered accordingly [93].

In contrast to genomic data, complex metagenomes encompass the entire genetic heterogeneity present in a population of species at the time they were sequenced [72], spanning thousands of unique taxa [93, 132] with different abundances [111], consisting of dozens of related strains across billions of short fragments [109, 224]. Furthermore, metagenomic sequences vary greatly in abundance depending on the dominance of the enclosed species, resulting in highly non-uniform read coverages across taxa [111, 225], with some being much lower than typically required for isolate reconstruction [109]. In addition, multiple species may share conserved genomic regions, complicating the process of assembly even further [109, 225]. As each metagenome represents a unique snapshot of microbial communities at the time a sample was taken, any contigs resulting from assembly do not represent a single individual or even strain [172]. With the overall quality depending on the underlying composition of the microbiome [226], the resulting contigs are commonly highly fragmented [221], decreased in accuracy [71] or cannot be reconstructed at all [73, 217], compared to isolate assembly. Complex and highly diverse microbiomes are hereby especially challenging to reconstruct [86, 217]. As such, metagenomic assembly remains a challenging task [109] with high computational requirements [225], especially memory usage [217]. However, various strategies have been developed in order to improve the continuity of metagenomes [227], including co-assembly of related samples [108, 176, 228]. Pooling read data of several associated samples into a single assembly greatly increases sequencing depth and completeness of the metagenome [95, 217, 228]. This enables the reconstruction of particularly low-abundance species [172, 191, 229], which are commonly missed during individual reconstruction [217]. On the other hand, co-assembly also drastically increases the data set size and complexity [224, 230] and may result in artefacts when merging samples from different sources [192]. In contrast, independent reconstruction tends to generate less, but often more higher-quality, draft genomes compared to co-assembly, primarily due to the reduced complexity of the sequencing data set [95, 224]. It is therefore common to perform both, individual and co-assembly throughout metagenomic analyses [185, 191, 224, 228].

Assembly methods used for the reconstruction of cultured isolates commonly require high and uniform read coverages, resulting in fragmented and error-prone assemblies when applied for metagenomic data sets [109]. This is why specialized assemblers have been developed to account for the computational requirements of

(co-)assembling billions of short metagenomic reads [93, 109, 111, 114, 176, 225]. MEGAHIT [225] represents a frequently utilized metagenomic assembler which illustrates good benchmarking performance [5, 90, 217, 231]. It uses De Bruijn graphs for metagenomic reconstruction [109, 114], which decompose reads into *k-mers* represented by interconnected nodes [93, 223]. The process of assembly is then represented by traversing through the edges of this graph in order to form long contiguous sequences from a selected path [90, 223]. Choosing an appropriate *k-mer* step size is of particular importance as it affects the overall assembly quality [93, 225, 232]. Increasing the value of $k$ will generally result in longer sequences across regions with high depth, but will break low-coverage contigs [223]. As a result, the recovery of low-abundance genomes increases when utilizing small *k-mer* sizes, while abundant taxa are more easily recovered through large *k-mers* [232]. Since the selection of a specific value for $k$ is non-trivial, MEGAHIT iteratively utilizes multiple *k-mers* through the construction of graphs from small *k-mers* before moving to larger sizes in order to refine the assembly connectivity [90, 93, 111, 114, 225, 232]. Despite this, metagenomic reconstruction remains a challenging process. Assemblies are often highly fragmented, with average contig lengths of 1-2 Kbp [95, 142] due to strain variation and differential coverage [43, 111]. The fraction of reads that can be assembled therefore typically remains low across microbiome studies ($< 20\%$) [228], requiring additional read-based analyses to be performed. As downstream analyses are known to be affected by these values [233, 234], QC also remains a crucial step throughout metagenomic reconstruction [111, 229]. Tools such as MetaQUAST can be used to validate the quality of the metagenomic assemblies [109, 231, 232]. By mapping contigs to a set of genomic references, MetaQUAST can evaluate various quality statistics, including fragmentation, gene content and misassembly rate [93, 109, 111, 231].

Metagenomic reconstruction results in sets of contigs as an output [93]. The next step focuses on the automated clustering of these sequences into groups of similarity that represent closely related taxa [93, 111, 221, 233, 234]. Known as binning [73], this process utilizes statistical properties of the contigs in order to cluster these into taxonomic groups [93, 111, 114]. Binning commonly results in thousands of genomic groups, each representing a different bacterial species with a unique role within the ecosystem of interest [221]. Large, complete bins are referred to as metagenome-assembled genomes (MAGs) [95] and represent draft-

level genomes of taxa present within the data set, many of which cannot be easily cultured *in vitro* [22, 229, 234]. However, binning is unable to distinguish genetic material from closely related organisms, as the reconstructed assemblies represent population genomes [111, 229]. Taxonomic lineages often share similarity in sequence composition characteristics, such as nucleotide frequencies, the presence of specific gene markers and differential coverage, all of which have been utilized as classification criteria for binning [5, 109, 111, 148, 229, 234]. Tetra-nucleotide frequencies, for example, represent a fundamental characteristic of genetic DNA [142, 235]. Contigs with similar abundances of 4-*mer* base combinations are hereby more likely to be derived from the same taxon [235]. In addition, fragment statistics can be assessed across multiple samples through the mapping of reads back to the assembly and evaluation of the resulting co-abundance patterns in order to assist throughout binning [95, 111, 224, 228]. Various methods have been established for the classification of metagenomic sequences into bins, including MetaBAT2 [234]. MetaBAT2 further utilizes probabilistic species-distances derived from available reference sequences [111] to group contigs into bins with which they correlate [234]. This process has been described to be especially advantageous when utilizing high amounts of metagenomic samples [234].

The resulting MAGs are subjected to additional QC in order to determine their suitability for further downstream analyses [191, 216, 229]. Dereplication can be performed through the dREP tool [224], which identifies identical MAGs from a large data set and removes these, while retaining a high-quality representative genome for each taxon. First, the completeness (how many conserved genes were identified) and contamination (how many of these were single-copy) of each bin are verified [5, 111, 228, 229]. If known genetic markers are absent from a genome, it must be incomplete, while in contrast, if single-copy genes are identified in multiplicity within a MAG, the bin is likely contaminated [111]. Distances between high-quality MAGs ($> 75\%$ completeness, $< 25\%$ contamination) are then assessed, before the remaining assemblies are clustered and dereplicated [94, 224].

Taxonomic classification of high-quality bins can be achieved by comparing these to a suitable reference database [95, 172]. One tool for this task is represented by the GTDB-Tk [236], which determines the taxonomic rank of MAGs using catalogues of gene markers and the assessment of similarity values to reference

sequences [87, 237]. First, gene calling is performed on the set of MAGs by predicting open reading frames [236]. Identified genes are then scanned for sequence similarity to a pre-established database of bacterial and archaeal gene markers [221]. MAGs are then assigned to a taxonomic level based on the proportion of the identified marker genes and assembly similarity to the closest reference genome [96]. Taxon-specific markers are hereby multiple-sequence aligned and stored within a multi-fasta file [236]. This can be utilized to assess phylogenetic relationships across the MAG set [108, 148, 192].

### 2.3.5 Resistome, Virulome & Mobilome Profiling

The process of assembly recovers large genomic fragments which are inherently more informative than individual reads, as they allow the characterization of full-length gene sequences [93]. Resulting contigs can further be contextualized within their genetic environment, essentially linking functional predictions with taxonomic assignments [228]. Taxonomic assignment of contigs facilitates an improved understanding of resistance mechanisms and can reveal the structural relationships surrounding these [22]. The main advantage of metagenomics is that there is little knowledge required *a priori* regarding which bacteria or genes are present within a given sample [25]. This enables the usage of wide-range databases for broad-spectrum screening of genotypic elements inducing pathogenicity, such as ARGs, virulence-associated genes or MGEs [17, 50]. While the genotypic presence of these elements does not necessarily induce true pathogenicity or phenotypic resistance [50, 54, 79], their profiles can be utilized as early warning systems for the development of these since they allow the detection of silent genes which carry the potential to become active in the future [67]. Metagenomic contigs can be annotated through a variety of different computational tools [112], primarily based on comparisons to specialized reference databases [22, 44, 212].

There exists a wide range of curated databases for *in silico* resistome profiling [46, 48, 67, 238]. However, one major caveat of most resistance databases is that they primarily focus on ARGs identified within culturable, human pathogens, practically disregarding a large fraction of environmental genes [25]. This issue is addressed by MEGARes [46], an ARG database specifically designed for the analysis of environmental data which has since been utilized for various metagenomic analyses around the globe [87]. It comprises of a comparatively large gene catalogue,

encompassing reference sequences for antibiotic, biocide and metal resistances [46]. While extensive databases, such as MEGARes, enable the identification of ARGs, they are inherently limited to profiling known resistance genes [44, 67]. Their predictions can therefore be extended through additional machine learning-based tools, which have the ability to categorize novel sequences, effectively addressing this database-related issue [43, 44]. Trained on multiple resistome databases, DeepARG utilizes a multi-layered model of interconnected nodes in order to map complex sequence data to resistance gene potentials [44]. All of the aforementioned methods actively identify ARGs based on their presence-absence within the metagenome. However, metagenomic assembly techniques also have their own drawbacks, including the inability to reconstruct low-abundance ARGs [25, 162]. These must therefore be assessed independently, through the process of read mapping [25, 46, 159]. As identical parts of homologous genes frequently result in unspecific read assignments [162], aggregation to higher-level classifications remain essential [66, 198].

The public health risk of an individual ARG depends on multiple factors, including the resistance it confers, its prevalence amongst human pathogens and its potential for mobility [18, 53]. One major drawback of using metagenomic read data for assessing resistances is its inability to distinguish between zoonotic and immobile environmental genes [25]. This however can be addressed through local contextualization of genes, in particular by assessing their genomic surroundings via assembly and subsequent binning of the sequences [25, 50]. MGEs play an essential role in the transmission of resistances. ARGs and virulence factors are frequently transmitted through MGEs, such as plasmids, enabling them to be exchanged across a wide range of bacteria. Differentiating between sequences originating from chromosomal DNA and those representing plasmids is a non-trivial task, especially within the context of metagenomics [239]. However, plasmids generally share a selection of features that allow them to be distinguished from chromosomal sequences, including differing nucleotide compositions, lack of primary metabolic genes and the presence of elements improving host fitness (such as virulence or resistance genes) [142]. Specialized software, such as PlasClass [239], can utilize these subtle changes in sequences in order to predict the source of assembly-level contigs. PlasClass is based on a selection of classifiers trained on *k-mer* frequency vectors of established plasmidial and chromosomal sequences

[239]. These are utilized to assign a given sequence to its genomic origin under a specified probability [142, 239].

## 2.4 Scientific Aims of this Thesis

The primary goals of this thesis were (1) the development of two novel and innovative bioinformatic workflows for the extensive, computational analysis of 16S rRNA and metagenome shotgun sequencing data, (2) to perform in-depth characterization of the equine nasal and gut microbiota, particularly in regards to taxonomic and resistance gene profiling, in order to (3) investigate and illustrate the effect of different PAPs on the host microbiome structure.

Metagenomics illustrates a novel method for the *in silico* assessment of taxonomic abundances, genotypic resistances and virulence components across microbial ecosystems (2) [66]. Particular challenges however include the technology of metagenomics still being in its infancy, with limited gold standards currently available [92, 217, 229, 232], especially in regards to computational resistome profiling [65]. Due to the high complexity of these 'big data' sets, analyses are often time intensive [198, 221, 240], requiring substantial expertise in bioinformatics in order to assess the most appropriate methodology, software and parameters [133]. Turning raw sequencing data into lists of microbial populations and gene pathways represents a non-trivial task [214, 228, 241], challenged by large biological, technical and inter-individual variation [92, 113]. While few workflows exist for metagenomic data processing [87, 108, 132, 212, 214, 240], they have not been specifically designed to handle high abundances of uncharacterized equine microbiome data and as such do not encompass the full scope of analyses required throughout this study (2, 3). It therefore remained essential to design an appropriate workflow dependent on the available data, biological characteristics and research questions at hand and to validate different methodologies in order to enable precise taxonomic and functional quantifications [70, 118, 211, 217, 241].

In order to meet the requirements of this work, development of customized software solutions were necessary (1). Two completely novel software workflows were established within this thesis: a pipeline for the characterization of 16S rRNA sequences (Meta16s) and a workflow for analysing shotgun metagenome data (Meta-

GEN). Both pipelines will be presented in detail within the context of the current literature across sections 3.1.1 and 3.1.2. Bioinformatical workflows commonly combine multiple tools in order to obtain a comprehensive overview of a data set [194, 221]. A variety of computational methods for analysing high dimensional metagenomic data sets were therefore evaluated throughout this work and tested in regards to reproducibility and robustness [242]. The Meta16s and MetaGEN pipelines are both open-source and utilize sets of peer-reviewed, state-of-the-art tools for individual analysis substeps. Particular focus was set on the few available guidelines for sustainable software development regarding microbiome analyses, including tracking of the utilized commands and software versions through a version control system [5, 241, 243]. All code was run using the Python3 (v3.9.13) programming language with the Snakemake (v6.7.0) framework. Statistical analyses were performed through R (v4.3.1), with multiple testing corrected significance levels of $p \leq 0.05$ applied whenever appropriate. Computations were performed on an AMD EPYC 7H12 @ 3.30 GHz workstation, containing 128 cores, 256 threads and 504 GB of memory, running Ubuntu 20.04.4 LTS.

# 3 Results

## 3.1 Development of Metagenomic Analysis Pipelines

Within the context of this thesis, two novel bioinformatical pipelines were developed specifically for the analysis of complex microbiome data sets. Termed Meta16s, for utilization with 16S rRNA sequences, and MetaGEN, for the analysis of metagenome shotgun data, both were applied to the nasal and gut equine samples collected across this study. Their full workflows will be described throughout the following sections for Meta16s (section 3.1.1) and MetaGEN (section 3.1.2).

### 3.1.1 Meta16s Workflow

16S rRNA workflows are generally divided into several interconnected analysis steps, including data preprocessing (section 2.2.1), taxonomic profiling (section 2.2.2) and downstream analyses, such as normalization (section 2.2.3), diversity estimation (section 2.2.4), and visualization (section 2.2.5) [111, 118]. Meta16s represents a command-line tool written in the programming languages Python3 and R, which implements several state-of-the-art methods for streamlining 16S rRNA data analyses, based on previous research [154]. Meta16s is open-source and available on GitHub[4] under the GPLv3 licence. A full overview of the workflow is illustrated in figure 3.1. The following section will provide a detailed description of the individual analysis steps applied by Meta16s.

Amplicon fragments of the 16S rRNA gene are initially merged using the fastq-join function of the QIIME pipeline (v1.9.1) [244] with a default, minimum overlap of 6 bp, maximum of 60 bp and a maximum difference of 8% [154]. Next, reads containing ambiguous bases, short sequences and low-quality fragments are removed from the data set [118, 130, 193]. Fragments that are subsequently too short or long after merging (e.g. $< 300$ bp) are removed using the Seqkit framework (v0.16.1) [245]. Quality filtering is performed through the split_libraries function of QIIME [244] with a Phred quality score of 20, allowing no ambiguous (N) bases [154]. The remaining reads are then exported and utilized for taxonomic assignment.

[4] https://github.com/SiWolf/Meta16s

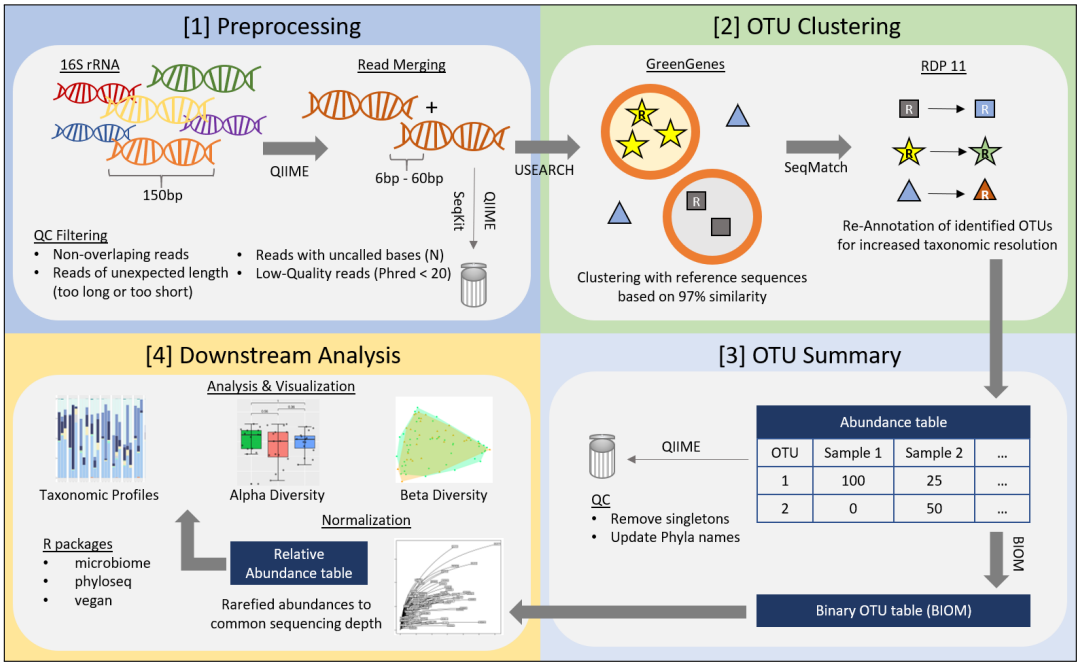


**Figure 3.1: Meta16s Workflow.** The workflow of Meta16s encompasses several interconnected analysis steps. First, paired-end reads are merged into longer sequences of DNA, which are processed based on multiple quality parameters (1). Next, the remaining fragments are compared against two reference databases for improved taxonomic annotation (2). Assigned reads are then summarized in tabular form, low-abundance hits are removed and taxonomic labels are revised once more according to current nomenclature (3). Finally, the computed table can be utilized for further downstream analyses, including normalization, diversity estimation and visualization (4). OTU: Operational Taxonomic Unit, QC: Quality Control.

The resulting, quality-controlled sequences are then subjected to OTU clustering and subsequent taxonomic annotation. Meta16s utilizes USEARCH (v11.0.667) [246] for clustering of the reads with the current GreenGenes reference sequences (gg_13_8_otus) [195] based on the established 97% similarity threshold. However, due to the limited sequence specificity of short 16S rRNA gene fragments, it is common that many OTUs cannot be sufficiently annotated, especially on lower hierarchical levels [118]. In order to increase taxonomic resolution throughout profiling, the Meta16s workflow employs an additional annotation step [154]. Generated OTUs are hereby taxonomically assessed through the SequenceMatch pipeline of RDPTools[5] (v2.0.3), which assigns clusters to their closest represen-

[5]https://github.com/rdpstaff/RDPTools

tative sequences within the more-recently updated RDP 11 collection [196]. Sequences not matching the GreenGenes database [195] are dereplicated before further classification through SequenceMatch is conducted.

In order to ensure consistency of the Meta16s results with current research, taxonomic labels assigned by RDP [196] are updated once more according to the newest nomenclature. This primarily concerns the recent renaming of multiple bacterial phyla [247], which still need to be adapted into current 16S rRNA databases. In addition, Meta16s utilizes QIIME [244] to remove any low-abundance noise with less than 4 counts across all samples [153]. This results in a revised and filtered OTU table, which is finally exported into the BIOM format, a binary data type that is easily accessible for downstream analysis [111].

The BIOM-formatted OTU table is then imported into a standardized data object through the phyloseq R package (v1.43.0) [248]. Next, the rarefy_even_depth function of phyloseq is utilized with a fixed seed ($r = 1$) to initiate downsampling of the abundance matrix [96]. In order to retain as much data as possible, subsampling is performed to the level of the highest shared sequencing depth, as suggested previously [60, 153, 160, 169, 173]. Normalized subsampled data are then utilized for further analyses, including diversity estimation and ordination [168, 198].

While there exists a wide range of computational tools for performing alpha diversity estimation [249], Meta16s utilizes the microbiome package[6] (v1.21.0) throughout its workflow. For the assessment of beta diversities, the distance function of phyloseq [248] is used to compute Bray-Curtis distances between the taxonomic profiles of each sample. These are then converted into 2D plots through dimensionality reduction of the ordinate function. Finally, Meta16s utilizes edgeR (v3.41.8) [202] for differential abundance estimation on the family-level.

### 3.1.2 MetaGEN Workflow

Metagenome shotgun sequences encode a particularly rich array of biological information, requiring large-scale computing and dedicated data management in order

[6]https://github.com/microbiome/microbiome

to gain insight into the ecological structures contained within [73, 212]. MetaGEN represents a command-line tool implemented in the programming languages Python3 and R, that employs state-of-the-art methods to preprocess (section 2.3.1), taxonomically profile (section 2.3.2), functionally analyse (section 2.3.3), assemble (section 2.3.4) and assess the pathogenicity (section 2.3.5) of complex metagenomic samples. 16S rRNA amplicon and metagenome shotgun sequencing share a variety of data types, mathematical properties and analysis requirements [92, 118]. MetaGEN therefore includes many of the conceptional methods described within the Meta16s workflow (section 3.1.1), but extends these with multiple shotgun-specific tools, which will be described in detail throughout this section.

Workflows for shotgun metagenome analyses commonly require high-performance clusters [221] and must therefore be scaleable to available computational resources. MetaGEN utilizes the workflow management system Snakemake, which enables automatic scaling of computational tasks to a set of given resources [243]. Each analysis step of MetaGEN is implemented as a rule within the Snakemake framework, strictly defining the data input, output, tools and versions to be utilized. In order to facilitate reproducible analyses across local and remote systems, Snakemake supports the usage of conda environments, which enable automatic installation of any required dependencies [243]. Through the utilization of this modular system, individual components of MetaGEN can also be easily adapted and extended with novel software in the future [221]. Being aware of the potential terabytes of data that metagenomic analyses generate, MetaGEN was designed with emphasis on efficient storage usage, keeping as much relevant information as possible, while retaining as little redundancy as needed. Individual tools used within MetaGEN were selected on a variety of criteria, including applicability for the analysis at hand, recommendations within the current literature [93, 114, 214], benchmarking results [232], comparison to other metagenomic pipelines [1, 87, 95], as well as overall compatibility and stability [221]. Particular focus was also set on utilizing publicly available open-source software to enable full transparency of the methods and programming code behind MetaGEN [134]. As parameter choices are known to impact the results of metagenomic analyses, they have been carefully evaluated based on current literature and benchmarkings [232]. Meta-

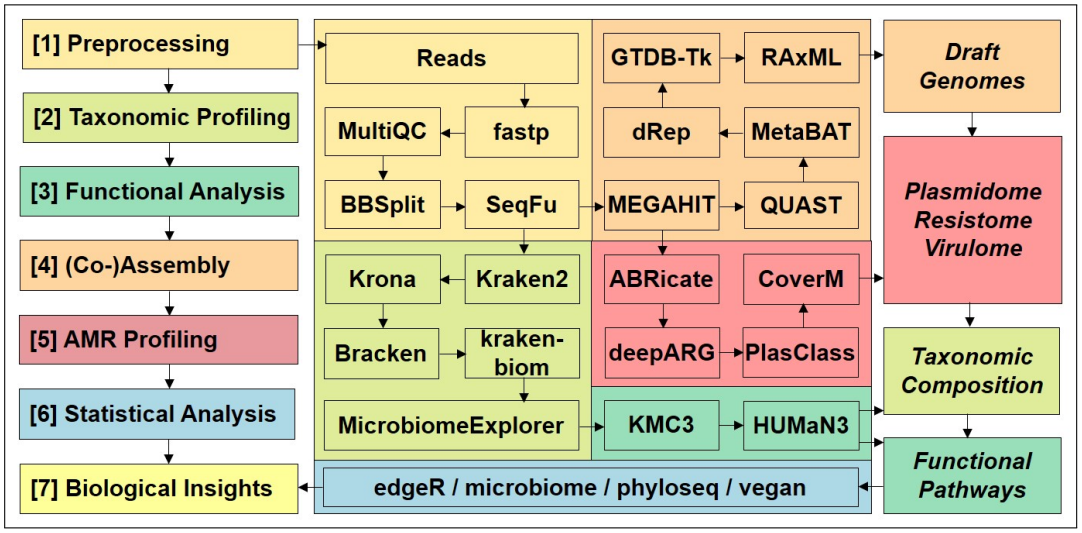


**Figure 3.2: MetaGEN Workflow.** MetaGEN requires raw metagenome shotgun reads as an input and performs several comprehensive analyses based on these (colours). First, reads are preprocessed by removing adapter sequences and host DNA (1). Next, quality-controlled reads are taxonomically profiled (2), before functional analysis is conducted (3). In parallel, metagenomic reads are assembled and binned to reconstruct draft genomes (4), which are further utilized for resistome, virulome and mobilome profiling (5). Finally, results generated by MetaGEN are statistically compared between sample groups (6) in order to gain biological insight into the data set (7). AMR: Antimicrobial Resistance.

GEN is open-source and freely available on GitHub[7] under the GPLv3 licence. An overview of its workflow is illustrated in figure 3.2.

MetaGEN requires a folder of .fastq files as an input, containing raw, paired-end sequences of metagenomic samples. Next, the input data are initially preprocessed using the fastp software [124]. This tool aggregates a variety of functions for NGS data processing, such as automatic adapter detection, quality trimming and reporting, and has subsequently been utilized throughout metagenomic studies [214]. In MetaGEN, fastp (v0.22.0) is used to trim adapter sequences as well as any low-quality read tails. Low quality reads commonly result from the imperfect process of sequencing [105]. As such, reads were trimmed to a recommended Phred quality score of $> 20$ using a sliding window approach [25, 47, 60, 214], corresponding to a per-base-accuracy of 99% [97] or an error-rate of 1% [73].

[7] https://github.com/SiWolf/MetaGEN

Quality statistics on the reads prior to and after processing are also exported into .html and .json formats [124, 214]. These are utilized to generate summarized quality reports across all samples through the MultiQC (v1.11) software [250]. In addition, BBSplit (v38.93) is used to identify and remove any host contamination [108]. Components of the BBTools suite[8] have been widely applied for the removal of specific sequences from genomic queries [229], with BBSplit in particular enabling the mapping of reads to multiple reference genomes. MetaGEN performs contamination removal through mapping of the metagenomic reads to a set of reference sequences, including the current *Equus caballus* (RefSeq accession: GCF_002863925.1) and *Homo sapiens* (telomere-to-telomere assembly, GenBank accession: GCA_009914755.4) genomes, the latter of which can result from imperfect handling of the samples. Finally, quality-controlled, contaminant-free reads are once more statistically assessed using the SeqFu (v1.6.0) package [251], before further downstream analyses are conducted.

A variety of methods were assessed for phylogenetic classification within MetaGEN. Since retrieving taxonomic labels for each metagenomic read was required throughout further downstream analysis, Kraken2 (v2.1.2) [218] was selected as the primary software for this task. While a majority of taxonomic classifiers generate comparable results [106], Kraken2 has been described as especially suited for gut microbiome analyses [217]. The default Kraken2 database was built using reference genomes of RefSeq for archaea, bacteria, viruses, plasmids, humans and vector sequences (dated 10.08.2021), encompassing 51 GB in size. Kraken2 classification was then performed on read-level using a confidence threshold of 0.0 [115]. Taxonomic reports were visualized using the Krona (v2.8) software [252], which generates interactive pie chart-like plots of the microbial community structure. Species re-estimation was enabled through Bracken (v2.6.1) [104], using a defined read length of 150 bp [96, 115]. While actual read lengths may differ as a result of the trimming process (section 2.3.1), a majority of fragments remain at this length, minimizing the impact of this parameter [104]. In order to reduce low-abundance false positives (those resulting from few, falsely mapped reads), hits containing less than 10 counts were discarded by default [104, 216]. The re-estimated abundances were then converted into the accessible BIOM format using

[8] https://sourceforge.net/projects/bbmap/

the kraken-biom package[9] (v1.0.1), which can be utilized for further visualization through software such as MicrobiomeExplorer (v1.9.0) [249].

Next, KMC3 (v3.1.2rc1) [253] is utilized by MetaGEN for the sample-wise computation of *k-mer* distributions across sequencing reads. Setting a specific *k-mer* length remains crucial for these analyses, as short *k-mers* have an increased chance to occur randomly, resulting in decreased biological relevancy [147]. Long *k-mers*, on the other hand, can might miss features of interest and are computationally challenging to analyse [147]. MetaGEN therefore utilizes a default threshold of $k = 25$, in order to provide sufficient mappability for taxonomic and functional profiling. Group-specific *k-mers* are identified and assigned to the Swiss-Prot database[10] (dated 30.05.2023) using a best-hit approach [12, 132, 228] with BLASTX (v2.12.0) [254]. In addition, MetaGEN utilizes HUMAnN3 (v3.6) [220] for in-depth pathway and gene abundance estimation. Its corresponding databases (ChocoPhlAn, MetaPhlAn and UniRef90) were downloaded and setup on the 11.04.2022. Subsequently, lists of associated pathways were statistically compared between sample groups.

The MEGAHIT software (v1.2.9) [225] was chosen as the genomic reconstruction tool for MetaGEN due to its ability to perform large-scale co-assemblies [114], based on its decreased memory requirements [1, 93, 111]. It generated viable results when benchmarked against other assembly tools [90, 232] and has since been utilized across many metagenomic frameworks [214]. Since all samples originate from a common microbiome, their sequences share significant overlaps, allowing comprehensive co-assembly to be performed [230]. Metagenomic reads were both individually and co-assembled using MEGAHIT. For individual reconstruction, the default *k-mer* list of (21, 29, 39, 59, 79, 99, 119, 141) was utilized, while for the more computationally exhaustive co-assembly a customized range of (27, 37, 47, 57, 67, 77, 87) was used for MEGAHIT, as recommended throughout current literature [191]. As binning methods require longer contigs [111], assembled fragments $< 1$ Kbp were discarded since they provide little gene-centric information [87, 90, 172]. The resulting contigs were re-named with sample-specific identifiers

[9] https://github.com/smdabdoub/kraken-biom

[10] https://www.uniprot.org/help/downloads

using BBRename (v38.93) of the BBTools suite[11]. MetaQUAST (v5.2.0) [231] was then utilized for assessing the overall assembly quality, limiting reference-based comparisons to 5 taxa per sample. Next, additional taxonomic profiling was performed on the remaining contigs through Kraken2 (v2.1.2) [218] with the previously established database. In addition, open reading frames were identified and extracted using Prodigal (v2.6.3) [255]. Prior to binning, Bowtie2 (v2.4.4) [256] was used to map the metagenomic reads to their corresponding assemblies, followed by indexing the mapping file through SAMtools (v1.11) [257]. Bowtie2 was utilized due to its high accuracy and speed [214]. As this study included over 60 deeply sequenced metagenome samples, MetaBAT2 (v2.12.1) [234] was selected as the tool for defining MAGs, since it specializes on differential abundance binning, a process which leverages coverage information from similar data sets to improve accuracy [221]. It performed comparatively well throughout benchmarking [1], with its default settings being optimized for a wide range of data sets and additional tweaking only marginally benefitting the results [234]. Binning was subsequently performed on both, individual and co-assemblies using only slightly modified settings (--minContigLength 1000 and --minContigDepth 2) [191]. The resulting bins were then taxonomically assessed through GTDB-Tk (v2.2.6) [236]. They were further dereplicated into a set of high-quality MAGs using dRep (v3.4.0) [224] for both bacteria and archaea. Finally, conserved gene-marker alignments of dRep were utilized for phylogenetic reconstruction of the dereplicated sets through RAxML-NG (v1.1.0) [258] and visualized within the iTOL platform [259].

Since the length of Illumina sequencing reads limits genomic contextualization [18], assembly-level profiling of ARG and virulence genes is initially performed by MetaGEN. ARGs are hereby characterized, assessed for mobility and quantified [18]. For this, MetaGEN utilizes ABRicate[12] (v1.0.1) in combination with the MEGARes [46] and VFDB [260] databases, both dated 2021-Mar-27 respectively. Scans were performed on all MAGs using stringent 90% identity and coverage thresholds in order to limit false positives and to ensure biological relevancy of the detected hits [28, 65, 141, 175]. ABRicate reports were generated through the included --summary function [176]. ARG and virulence results were then sum-

[11] https://sourceforge.net/projects/bbmap/

[12] https://github.com/tseemann/abricate

marized within count matrices and normalized by the total nucleotide count for each metagenomic assembly [51]. Using the previously identified coding sequences of Prodigal (v2.6.3) [255], novel ARG profiling was performed through DeepARG (v1.0.2) [44]. Metagenomic contigs were further evaluated for mobility using PlasClass (v0.1.1) [239]. A sequence was described to originate from a plasmid if its probability score was $> 0.5$ [239]. In order to characterize low-abundance ARGs, which would not have been reconstructed through assembly, additional profiling was performed via metagenomic read mapping. For this, the MEGARes database [46] was first downloaded (dated 05.05.2023) and subsequently preprocessed. As confirmation of SNPs on metagenomic data remains limited [46], genes requiring polymorphisms for conferring resistance were removed from the collection [87, 141]. In addition, ARGs were grouped into representative sequences in order to prevent unspecific mappings by limiting the redundancy of the database [51, 78, 261]. The remaining genes were clustered at 90% identity and coverage into a non-redundant catalogue using MMseqs2 (v13.45111) [262]. Representative sequences were then utilized for read-level abundance estimation through CoverM[13] (v0.6.1) [47]. Reads were required to map in proper pairs with an identity and coverage of 90% towards the processed database [51, 60].

## 3.2 Computational Analysis of the Gut Microbiome

MetaGEN (v0.7.8) was then utilized for the computational analysis of the equine gut microbiome data. Hospitalized horses subjected to colic surgery were hereby treated with either a conventional, 5-day prophylaxis (5DG) or a single-shot dosage (SSG) of gentamicin and penicillin [175]. Samples were collected across three time points: on hospital admission ($t_0$), 3 days ($t_1$) and 10 days ($t_2$) post-surgery. An overview of the sample set is illustrated in appendix table A.2.

### 3.2.1 Data Assessment

For analysis of the equine gut microbiome, a total of 63 faecal samples were collected and subsequently $2 \times 150$ bp metagenome shotgun sequenced to an average depth of 31 million read pairs. An additional 5 compatible samples were selected from NCBI to be included as reference sequences. The final data set consisted of

[13] https://github.com/wwood/CoverM

68 samples, encompassing over 4 billion reads containing 680 billion bp of genomic data. Before any bioinformatical analyses were performed, extensive QC of the shotgun metagenome data were conducted through MetaGEN. Adapter clipping, quality trimming and host DNA removal (both human and equine) was subsequently performed using MetaGEN as described prior (section 2.3.1). A total of 4 billion reads, composed of 650 billion bp (over 95% of the original data), remained after preprocessing and were utilized for further taxonomic study (section 2.3.2). Initial abundance estimation at the phylum-level revealed highly uneven sequencing depths (ranging from 200,000 to 80 million), as well as increased percentages of unclassified fragments (up to 91%) using current databases (appendix figure A.1). Additional, database-independent methods were therefore essential throughout later downstream analyses (sections 3.2.4 and 3.2.5). The overall sequencing depth was further normalized and evaluated using rarefaction analysis. After taxonomic redistribution, profiles were randomly subsampled in order to assess the ecological saturation of the sample set (figure 3.3). Re-estimated species counts obtained from the data remained highly uneven, ranging from 200,000 to 31 million, with an average of 3.6 million across the sample set. To enable cross-group comparisons, samples were therefore normalized to a shared minimum depth of 200,000 read counts.

### 3.2.2 Taxonomic Profiles

After read normalization, focus was shifted onto assessing the bacterial diversity of the equine gut microbiome in more detail. Throughout the following section, the microbial landscape of the equine gut will be described across the sample set, focussing on the various time points and distinct study groups. All samples were subjected to extensive taxonomic profiling using MetaGEN as described previously (section 2.3.2). For now, unclassified reads were excluded from further analysis (appendix figure A.1). Raw bacterial compositions were initially assessed through sample-wise comparisons using Krona plots (see figure 3.4 for sample D61010). Dominant taxa were identified across multiple taxonomic levels, including phylum (*Pseudomonadota*, *Bacteroidota*, *Bacillota*, *Actinomycetota* and *Cyanobacteria*), genus (*Bacteroides*, *Acinetobacter*, *Pseudomonas*, *Prevotella*) and species (*Bacteroides fragilis*, *Escherichia coli*, *Bacteroides heparinolyticus*, *Bacteroides thetaiotaomicron*, *Pseudomonas putida*).

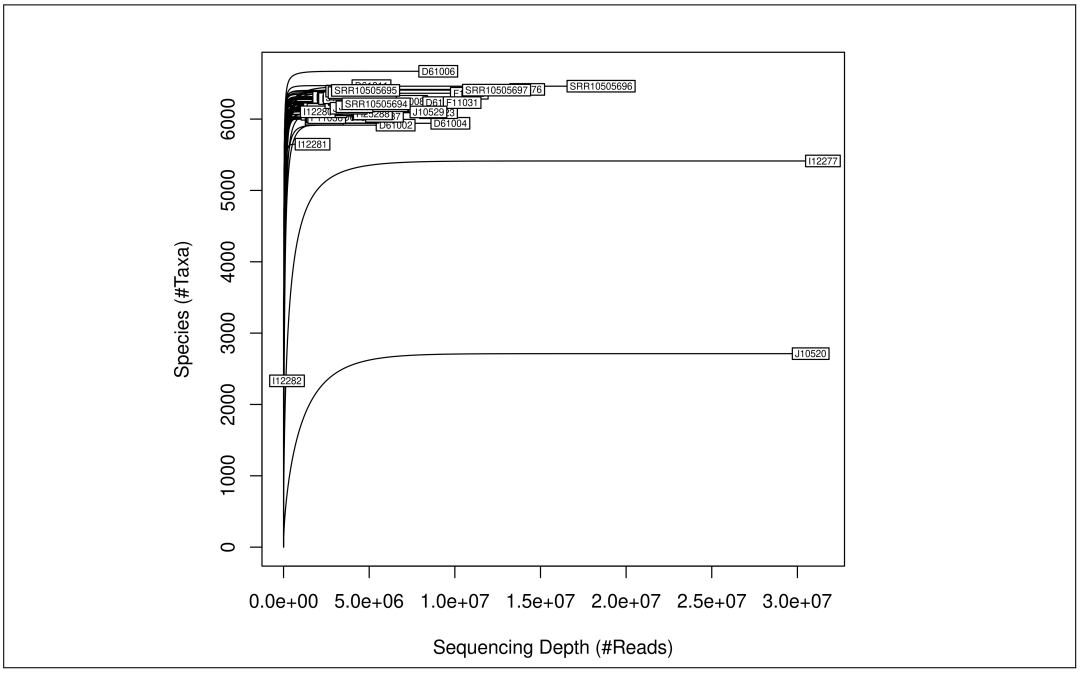


**Figure 3.3: Rarefaction Analysis (Gut Microbiome).** Rarefaction curve of the equine metagenome shotgun data. Illustrated are random subsets of differing depth for the sequencing data and the amount of taxa classifiable throughout these. Each line represents a different sample of the collection ($n = 68$). All graphs rise quickly and then stagnate within a plateau-phase, indicating that the total diversity of the microbiome can be observed without utilizing the full sequencing depth.

Additional aggregation of the data was required in order to compare the different prophylaxis regimens (SSG and 5DG) across the selected time points ($t_0$, $t_1$, $t_2$) and against the additional sample set obtained from NCBI. Rarefied taxonomic counts were therefore summarized at phylum-level and averaged across the aforementioned groups to characterize differences and similarities between these. The results were visualized in heatmaps (figure 3.5 and appendix figure A.2). Phyla identified within the gut environment were represented in the order from highest to lowest relative abundance by *Pseudomonadota*, *Bacteroidota*, *Bacillota*, *Actinomycetota*, *Chordata*, *Spirochaetota*, *Euryarchaeota*, *Cyanobacteria* and *Verrucomicrobiota*. Averaged abundances of these groups varied across the time points, with *Pseudomonadota* and *Bacteroidota* expanding in particular from $t_0$ to $t_1$ across both prophylaxis groups, before rebounding at $t_2$. In contrast, the phyla *Bacillota*, *Actinomycetota*, *Chordata*, *Spirochaetota*, *Euryarchaeota*, *Cyanobacteria* and *Verrucomicrobiota* decreased at $t_1$, before returning to their original abundances

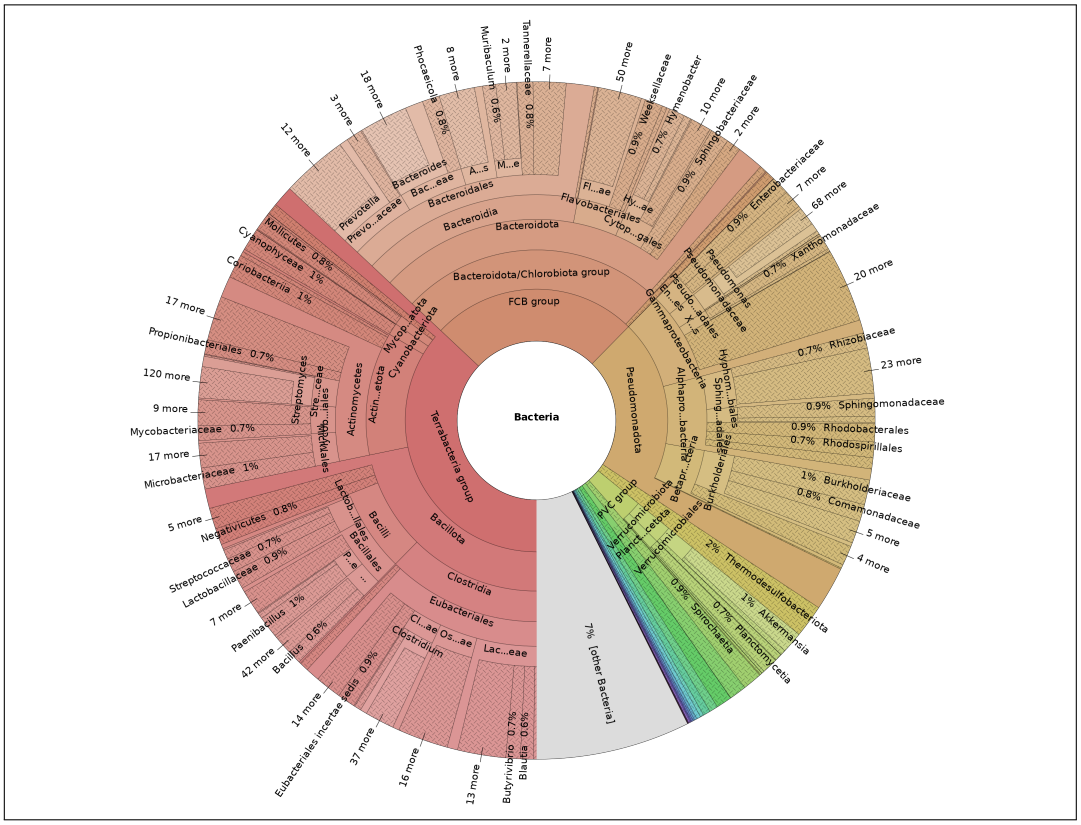


**Figure 3.4: Krona Plot (Gut Microbiome).** Pie chart-like visualization of the bacterial content of sample D61010 using the Krona software. The presented overview illustrates a breakdown of the most abundant phyla (inner circle) and the composition of these on up to family-level (nested sections). Taxonomic labels are visually clustered by colour according to their phylum, with species-rich groups encompassing larger sections of the plot. Listed percentages represent the abundance of a respective bacterial group within D61010.

towards $t_2$. Hierarchical clustering (appendix figure A.2) revealed similarities between the profiles of the 5DG and SSG at $t_0$, as well as $t_1$, with major differences identified throughout $t_2$. However, averaged abundances of these high-level taxonomic groups do not enable sufficient resolution to further characterize the underlying structural compositions.

Taxonomic profiles were therefore further investigated using stacked bar charts. The composition of each sample was hereby evaluated (figure 3.6), allowing the assessment of individual variability, as well as the identification of potential outliers, illustrated by samples obtained from equine patient 33. The data were also averaged and visualized as summarized bar charts samples of each study group

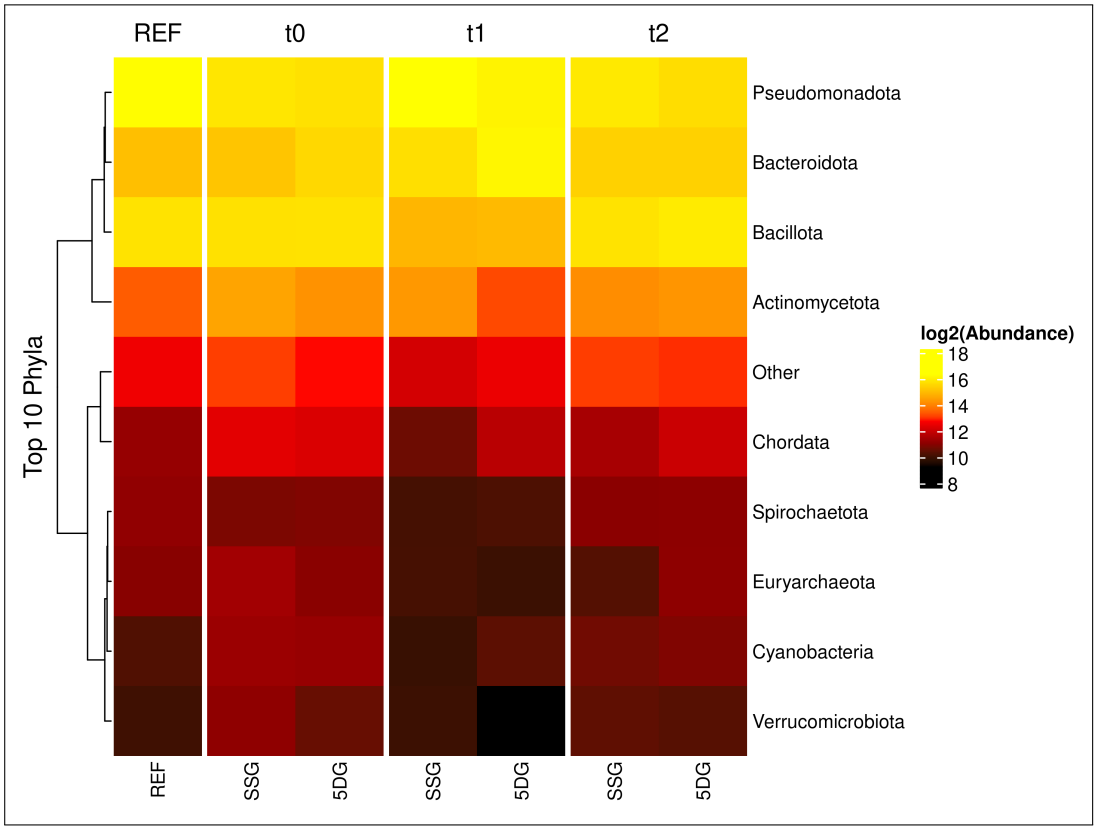


**Figure 3.5: Heatmap of Phylum-Level Abundances (Gut Microbiome).** Taxonomic profiles were summarized at phylum-level and averaged across the sample groups. The heatmap visualizes the abundance ($\log_2$ of normalized counts) from low (dark red) to high (yellow) across the data set. Columns represent the different study groups at specific time points. Rows illustrate the 9 most abundant phyla, with minor ranks being grouped into 'other'. Phyla are clustered by their similarity to each other in the represented abundance profiles (dendrogram on the left). REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

(appendix figure A.3). These underlined the previously characterized fluctuations on phylum-level, with *Pseudomonadota* and *Bacteroidota* expanding towards $t_1$, before rebounding at $t_2$. In addition, differences between the study group samples and NCBI references were detected. The NCBI group illustrated an increased abundance of *Pseudomonadota*, as well as decreased rates of *Actinomycetota*, *Chordata* and other (minor) phyla compared to the study group. In contrast, samples from the hospitalized patients illustrated enhanced abundances of *Bacteroidota*, independent of the sampling time point.

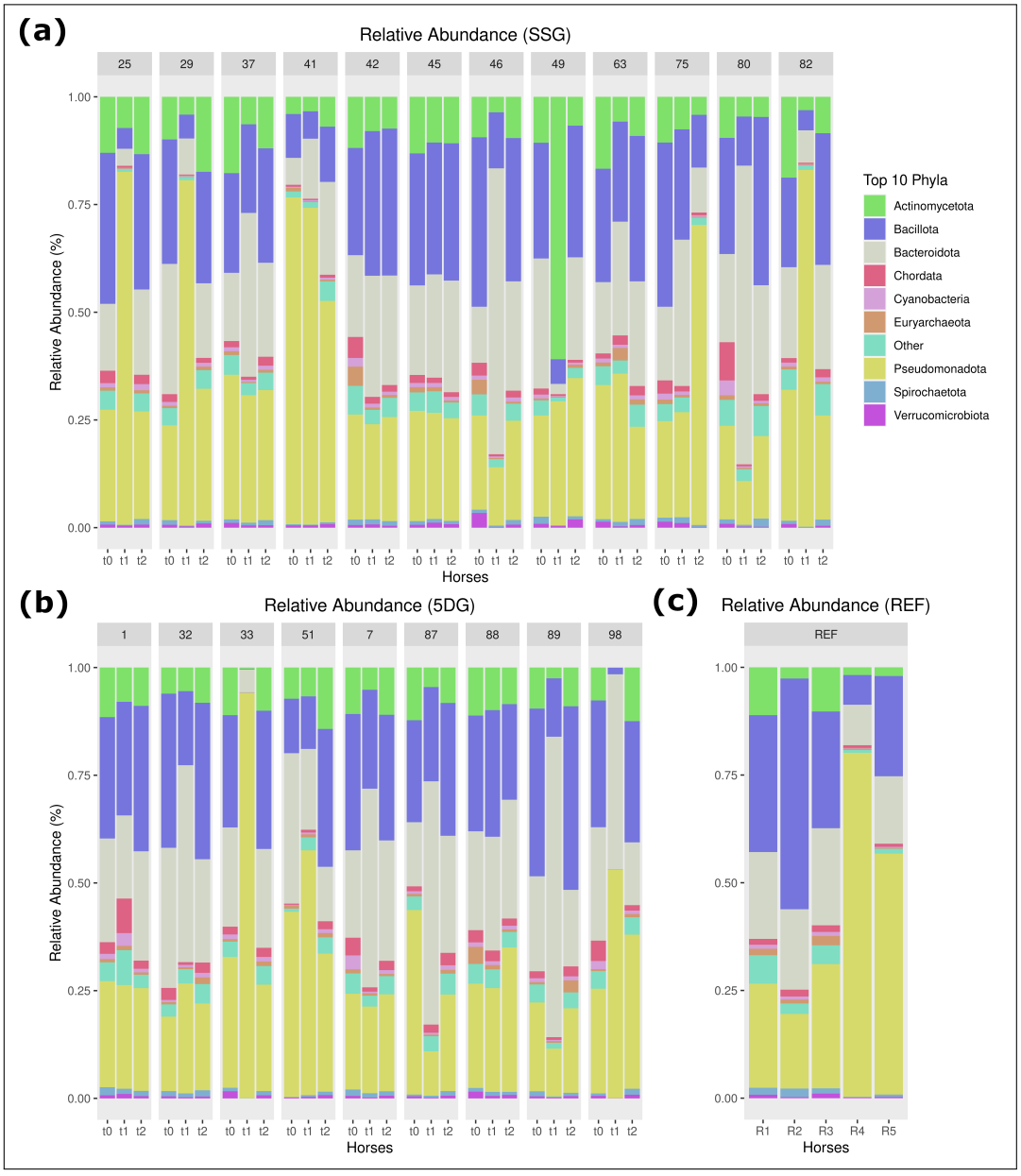


**Figure 3.6: Bar Charts of Phylum-Level Compositions (Gut Microbiome).** Stacked bar charts representing the composition of individual samples on the phylum-level for the 9 most abundant taxonomic groups, with minor ranks being grouped into 'other' (colours). Individual horses ($n = 26$) are clustered into the three prophylaxis groups (a) SSG, (b) 5DG and (c) NCBI references. Samples were collected for each equine patient (top row IDs) across three time points ($t_0$, $t_1$ and $t_2$). Visualized are the percentages of phyla per group over time. REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

Finally, the conserved set of microbes shared across all 68 samples was evaluated, representing the 'core' microbiota of the analysed sample collection. To characterize these, low-abundance taxa were first removed, based on a prevalence threshold of 95% across the sample set. A total of 3,115 taxa were identified as the conserved microbiota of the equine gut samples. This is in stark contrast to the preserved microbiome computed by utilizing samples of the SSG (increased to 5,358 taxa) and 5DG (restricted to 1,429 taxa) independently. Next, temporal relationships were investigated by assessing the conserved microbiota set per time point for both the SSG and 5DG. Venn diagrams were utilized for visualization of the corresponding taxonomic overlaps. An overview of the preserved microbiota is provided in figure 3.7. The overlap in conserved taxa visibly shifted between the SSG and 5DG over time, originating with 2,364 taxa at $t_0$, increasing to 5,494 at $t_1$ and returning to 2,304 towards $t_2$. A majority of these differences were associated with the total increased amount of conserved taxa within the SSG compared to the 5DG. This characterization also enabled insights into unique species at more specific taxonomic levels across the time points and sample groups, including $t_0$ (SSG: *Acinetobacter equi*, *Citrobacter farmeri*, *Lelliotta amnigena*; 5DG: *Streptococcus* sp. KS 6, *Streptococcus* sp. NPS 308), $t_1$ (SSG: *Bifidobacterium bifidum*, *Citrobacter cronae*, *Ruminococcus bicirculans*; 5DG: *Citrobacter* sp., *Serratia quinivorans*, *Lysobacter* sp.) and $t_2$ (SSG: *Streptomyces* sp.; 5DG: *Lelliottia nimipressuralis*, *Moraxella ovis*).

Next, relative abundances were further investigated at a more specific taxonomic level (family). Differential analysis was performed using MetaGEN to identify changes occurring within the described microbiome structure for bacterial families of interest. An overview of statistically significant differences (p-value $\leq 0.05$) is illustrated in appendix table A.3. For samples representing the 5DG, significant changes in abundance were observed for 47 ($t_0$ to $t_1$), 55 ($t_1$ to $t_2$) and 5 ($t_0$ to $t_2$) microbial families. The microbiome of the SSG, in contrast, was significantly altered for 45 ($t_0$ to $t_1$), 20 ($t_1$ to $t_2$) and 9 ($t_0$ to $t_2$) taxonomic groups. At $t_0$ (hospital admission) variation between the SSG and 5DG was limited to 7 families. This peaked at $t_1$ with 66 taxa, before decreasing towards $t_2$ with 0 differential families. A majority of these microbial groups reduced in abundance from $t_0$ to $t_1$, followed by recovery towards $t_2$. Abundances of the *Enterobacteriaceae* family heavily fluctuated throughout the study period. This group was subsequently

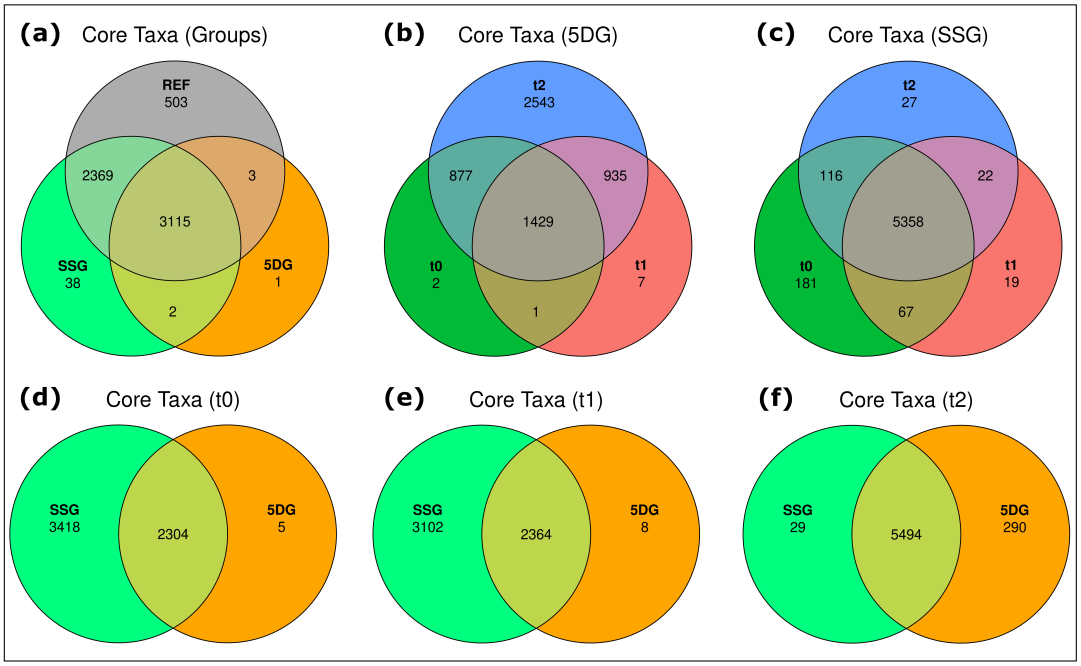


**Figure 3.7: Conserved Taxa (Gut Microbiome).** Venn diagrams illustrating the conserved ('core') microbial taxa shared between the different prophylaxis groups and time points. Taxa with prevalence $> 95\%$ were compared across the corresponding data sets. (a) Illustrates the conserved taxa shared across all samples, while (b) and (c) represent subsets of the 5DG and SSG respectively. (d)-(f) visualize the temporal component of the study, with changes occurring in the microbiome overlap between the SSG and 5DG over time. Illustrated taxa were analysed based on their presence-absence and not on their abundance across groups. REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

compared across the sample set in more detail (figure 3.8). Normalized abundances of *Enterobacteriaceae* increased significantly from $t_0$ (mean SSG: 6,472; mean 5DG: 20,613) to $t_1$ (SSG: +27,468; 5DG: +8,140) and decreased towards $t_2$ (SSG: −14,150; 5DG: −22,921) for both prophylaxis groups.

### 3.2.3 Diversity Analysis

Furthermore, similarity and diversity within (alpha diversity) and between (beta diversity) the metagenomic samples was comparatively investigated (see section 2.2.4 for more details). First, alpha diversity indices were assessed for each sample of the collection and subsequently averaged for further comparison (see appendix table A.1). This primarily focused on indices regarding richness, evenness and

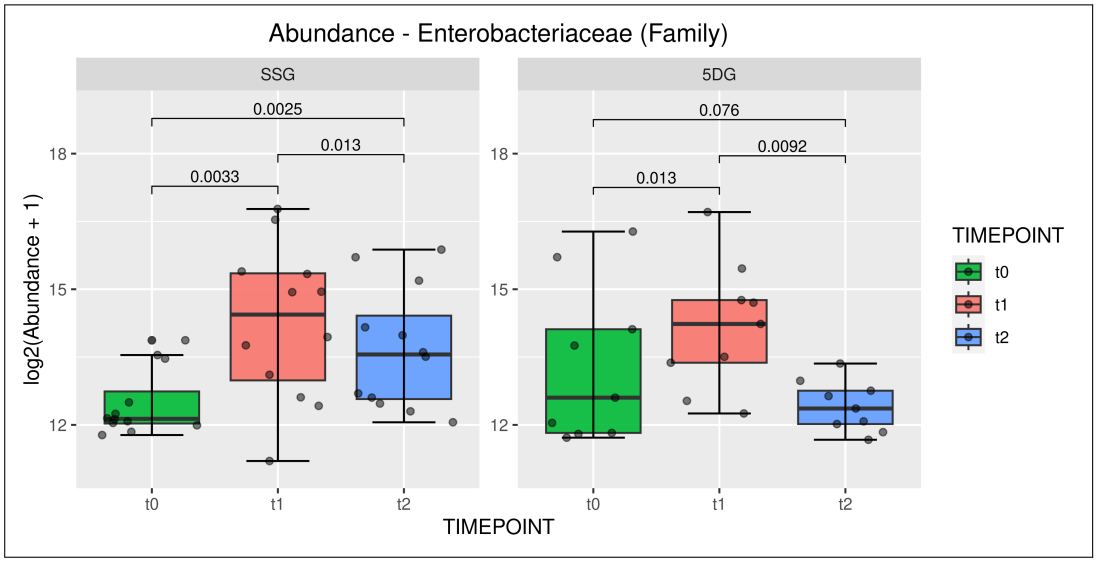


**Figure 3.8: Abundance of *Enterobacteriaceae* (Gut Microbiome).** The abundance of the family of *Enterobacteriaceae* was assessed and comparatively investigated across the sample set. Individual abundances were summarized as box plots for each prophylaxis group (SSG: $n = 12$, 5DG: $n = 9$) and subsequently compared over time (colours) using Wilcoxon rank-sum tests. A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

diversity of each sample. On average, the gut metagenome samples contained 5,979 taxa, an evenness of 0.01 and a Shannon diversity of 6.34. Taxa ranged from 1,545 to 6,544 across the entire set. Evenness was also highly variable, with values ranging between $7 \times 10^{-4}$ and $4 \times 10^{-2}$. The computed Shannon indices encompassed samples associated with values from 1.70 to 7.53. The lowest diversity values (1.70 and 3.47, respectively) were primarily associated with samples of the 5DG, while the highest values (7.49 and 7.53) mapped to the SSG. Significant differences in alpha diversity were determined between prophylaxis groups and time points through Wilcoxon rank-sum tests, visualized as box plots in figure 3.9. Compared to the diversity at $t_0$ (mean SSG: 7.12; mean 5DG: 6.54), values were found to be impaired at $t_1$ (SSG: $-1.71$; 5DG: $-1.44$), with recovery following shortly towards $t_2$ (SSG: $+1.27$; 5DG: $+2.09$) for both the SSG and 5DG. At $t_2$, diversity had been re-established for both groups compared to hospital admission (SSG: 6.68; 5DG: 7.20). While evenness at $t_0$ (mean SSG: 0.02; mean 5DG: 0.01) was also found to be impaired throughout $t_1$ (SSG: $-0.01$; 5DG: $-0.01$) and increased towards $t_2$ (SSG: $+0.01$; 5DG: $+0.02$), differences were detected in

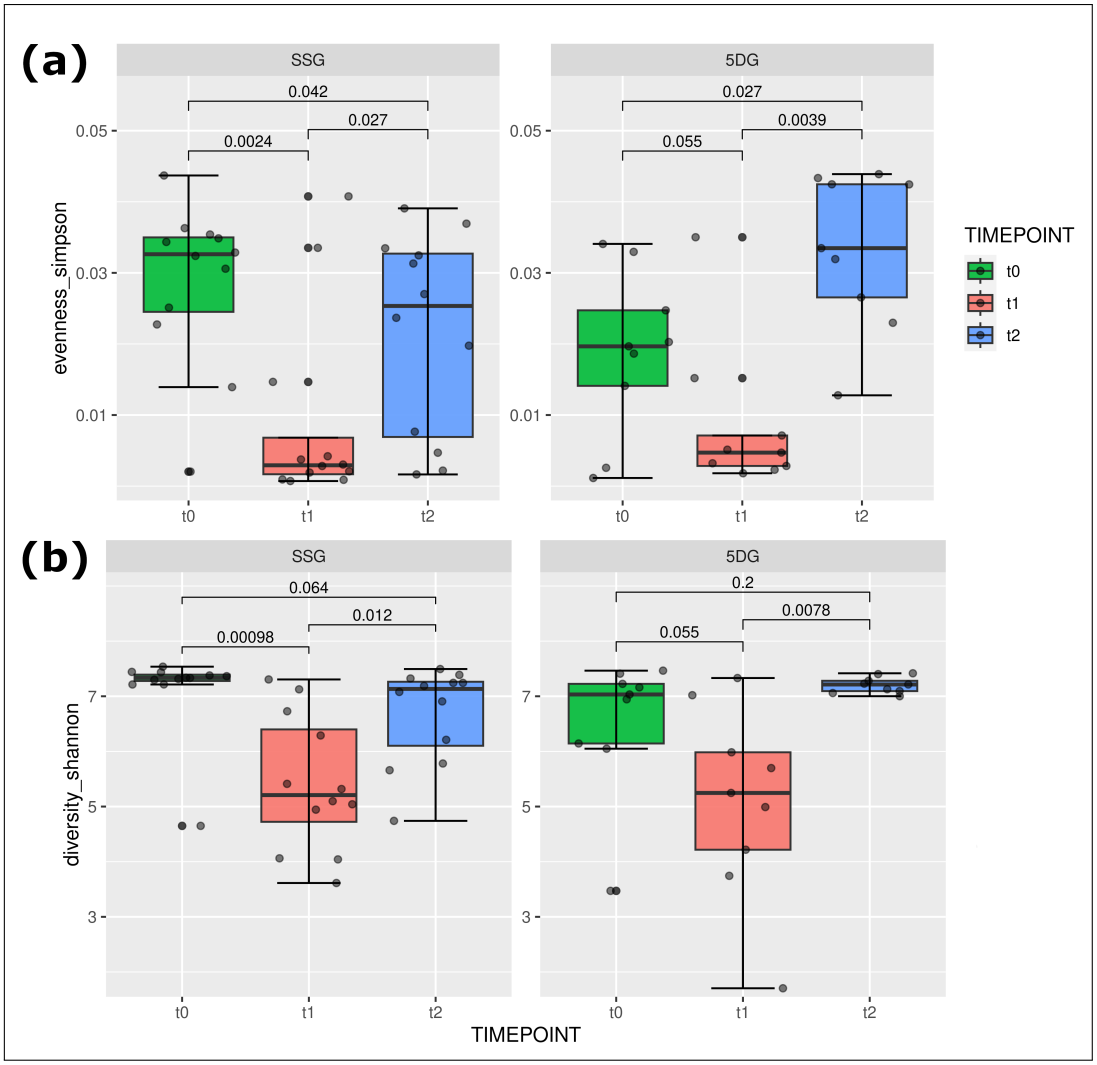


**Figure 3.9: Alpha Diversity (Gut Microbiome).** Comparison of diversity indices across the sample groups (SSG: $n = 12$, 5DG: $n = 9$) using box plots. (a) Illustrates the evenness, while (b) represents the Shannon diversity over time (colours). Both values decrease from $t_0$ to $t_1$ and recover towards $t_2$. Wilcoxon rank-sum tests were performed to infer significant differences. A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

comparison to the initial findings at hospital admission. Compared to $t_0$, the SSG illustrated decreased values of evenness at $t_2$, while this measure was found to be elevated for samples of the 5DG (SSG: 0.02; 5DG: 0.03).

Beta diversity was assessed between the samples through computation of Bray-Curtis distances from the normalized abundance tables at the species-level. This was followed by subsequent principal component analysis (PCA), which was used to characterize how the microbial community structure was influenced across time and antibiotic prophylaxis. Individual samples were subsequently visualized within PCA plots, labelled by groups and time points. Principal Component 1 (PC1) hereby accounted for 34% and Principal Component 2 (PC2) for 12.7% of the total variance observed. An overview is illustrated in figure 3.10. PCA resulted in multiple clusters of the samples, indicating major differences in taxonomic profiles between groups. A general shift in microbiome composition is detected over time for both study groups, dominated by particularly high variability at $t_1$. In contrast, samples of $t_0$ and $t_2$ possess similarities in their taxonomic abundances and cluster comparatively close together, although this process is also dominated by high individuality, especially visible throughout $t_2$.

### 3.2.4 Metagenome-Assembled Genomes

Metagenomics is based on the concept of comparing environmental DNA fragments to characterized databases in order to deduct the underlying microbial community of a sample through sequence similarity. While this represents a sensible approach, it remains intrinsically challenged by currently available genomic references, which often span only a minor fraction of the entire tree of life, resulting in limited classification accuracy throughout downstream analysis. Another method was therefore utilized to counteract the aforementioned issue: reconstruction and analysis of microbial draft genomes via metagenomic assembly and subsequent binning. Described in full detail throughout section 2.3.4, sample-wise short-read reconstruction was performed in order to generate metagenomic assemblies.

Quality-controlled sequences produced assemblies ranging in size from 11 Mbp to 1.1 Gbp, with an average of 336 Mbp per sample. Minor fragments of length $< 1$ Kbp were hereby discarded. The resulting assemblies remained highly fragmented with an average N50 value of only 3.9 Kbp. In order to characterize low-abundance microbes, additional co-assembly of the entire sample set was performed, utilizing a total of 328 GB of equine gut microbiome data originating from the 68 samples. The co-assembly encompassed 11 Gbp of microbial sequences and was combined with the individual assemblies for subsequent binning.

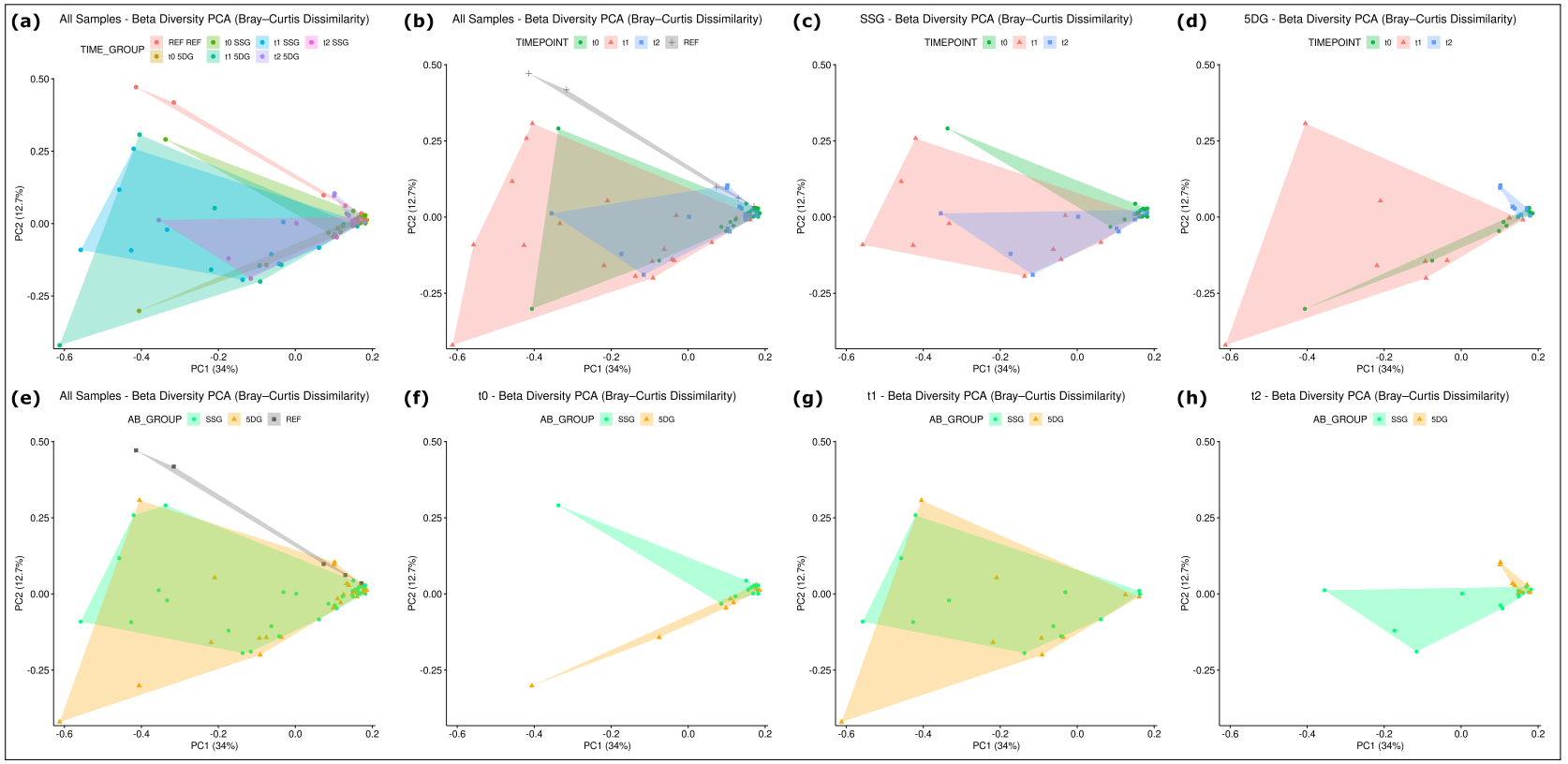


**Figure 3.10: Beta Diversity (Gut Microbiome).** PCA based on Bray-Curtis distances of the gut metagenome samples. Visualized is the complete sample collection across different subsets (colours). (a) Illustrates the full data set, coloured by prophylaxis group and time point. (b) and (e) represent identical subsets, but are coloured by group and temporal metadata, respectively. (c) and (d) represent subsets of the SSG and 5DG, labelled by time point. (f)-(h) illustrate subsets for each time point, distinguished by prophylaxis group. PCA: Principal Component Analysis, REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

Binning generated a total of 10,613 genomes of varying quality, which were utilized for additional QC and visualization. As the taxonomic classification of MAGs are affected by values of completeness and contamination (based on the presence of conserved, single-copy genes), a subset of high-quality draft genomes was first extracted prior to further analysis. Quality filtering and dereplication into non-redundant representative sequences generated 1,381 unique MAGs of high-quality (completeness $\geq$ 75%, contamination $\leq$ 25%), consisting of 1,360 bacterial and 21 archaeal draft genomes. Together, these MAGs spanned 22 phyla, 26 classes, 56 orders, 110 families, 322 genera and 179 species, although not all could be classified at this level. Representative genomes were then utilized for phylogenetic reconstruction, visualized in figure 3.11 for bacteria and 3.12 for archaea, respectively. The bacterial tree was primarily dominated by *Bacteroidota* (38% of MAGs) and *Bacillota* (42% of MAGs), while the archaeal phylogeny consisted of the *Methanobacteriaceae*, *Methanomethylophilaceae* and *Methanocorpusculaceae* families. Bacterial MAGs also included full draft genomes for the species *Bacteroides fragilis*, *Escherichia coli* and *Streptococcus equinus*.

### 3.2.5 Resistome, Virulome & Mobilome Profiles

Particular focus was set on characterization of the equine resistome, virulome and mobilome in order to investigate the impact of antibiotic prophylaxis regimens and hospitalization on the emergence of pathogenic bacteria, especially in regards to MDR. Metagenomic assemblies were therefore profiled for the genotypic presence of ARGs using the MEGARes database. A total of 2,215 genetic elements conferring resistance to antibiotics, biocides and heavy metals were identified across the sample set. These encompassed 17 different drug classes, ranging from components inducing resistance towards aminoglycosides to trimethoprim. An overview of the results is visualized in figure 3.13. Equines arrived at the clinic with unique microbiomes, medical histories and resistomes ($t_0$). The abundance of resistance components strongly increased throughout hospitalization and prophylaxis ($t_1$), with the highest resistance gene counts ($\geq$ 200) identified within samples of the 5DG. A majority of these ARGs were associated with beta-lactamases (*bla*EC), metal (*mdt*A) and biocide (*emr*A) resistances. Towards $t_2$, resistance genes diminished in abundance, with specific components remaining in few individuals for longer periods of time.

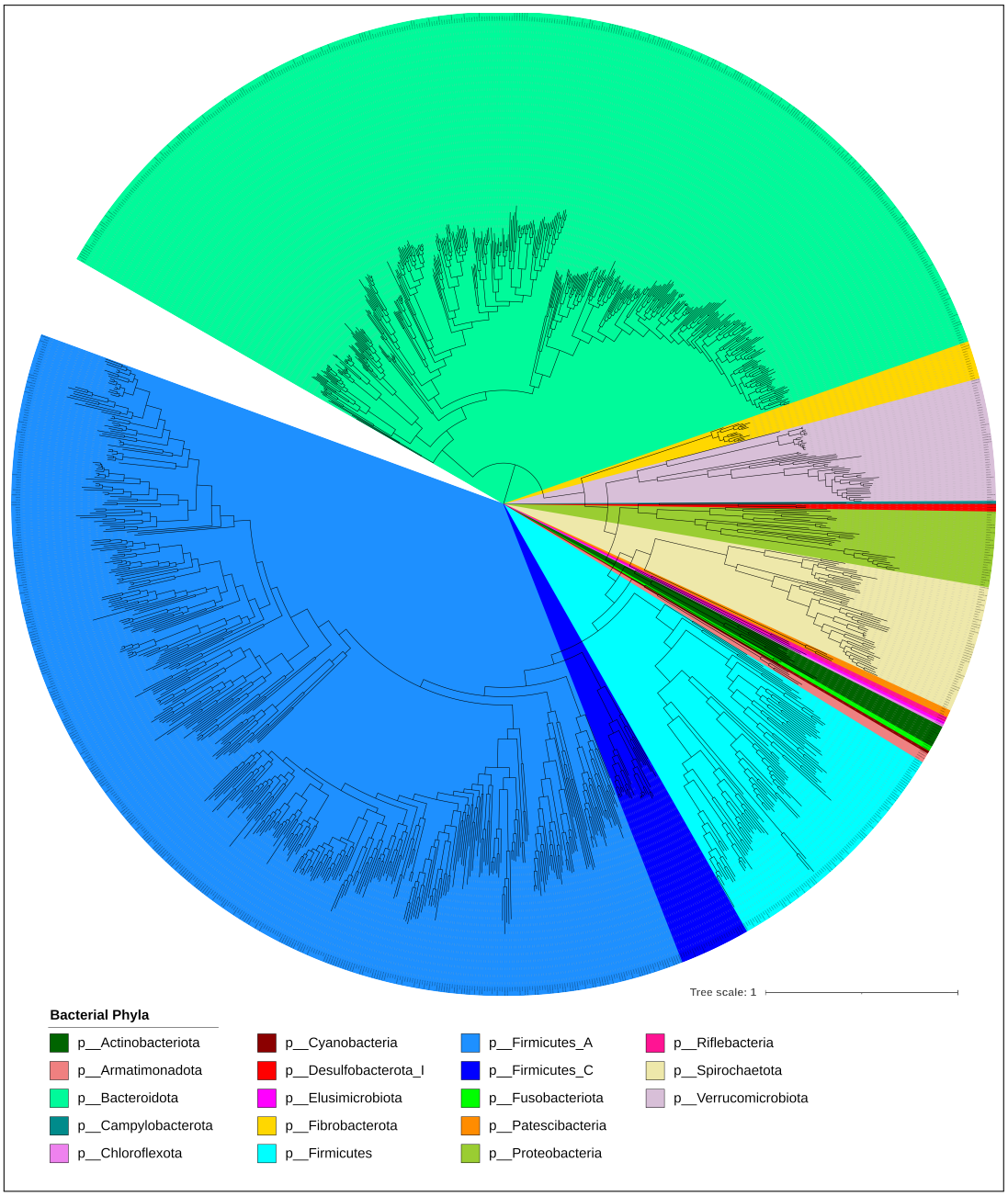


**Figure 3.11: Bacterial MAGs (Gut Microbiome).** Circular representation of a maximum likelihood phylogeny of bacterial MAGs visualized in iTOL. 1,360 high-quality MAGs were reconstructed from the equine gut metagenome collection. Sequences of conserved housekeeping genes were further utilized for the inference of the phylogenetic relationships. Bacterial MAGs were coloured according to their taxonomic classification on phylum-level.

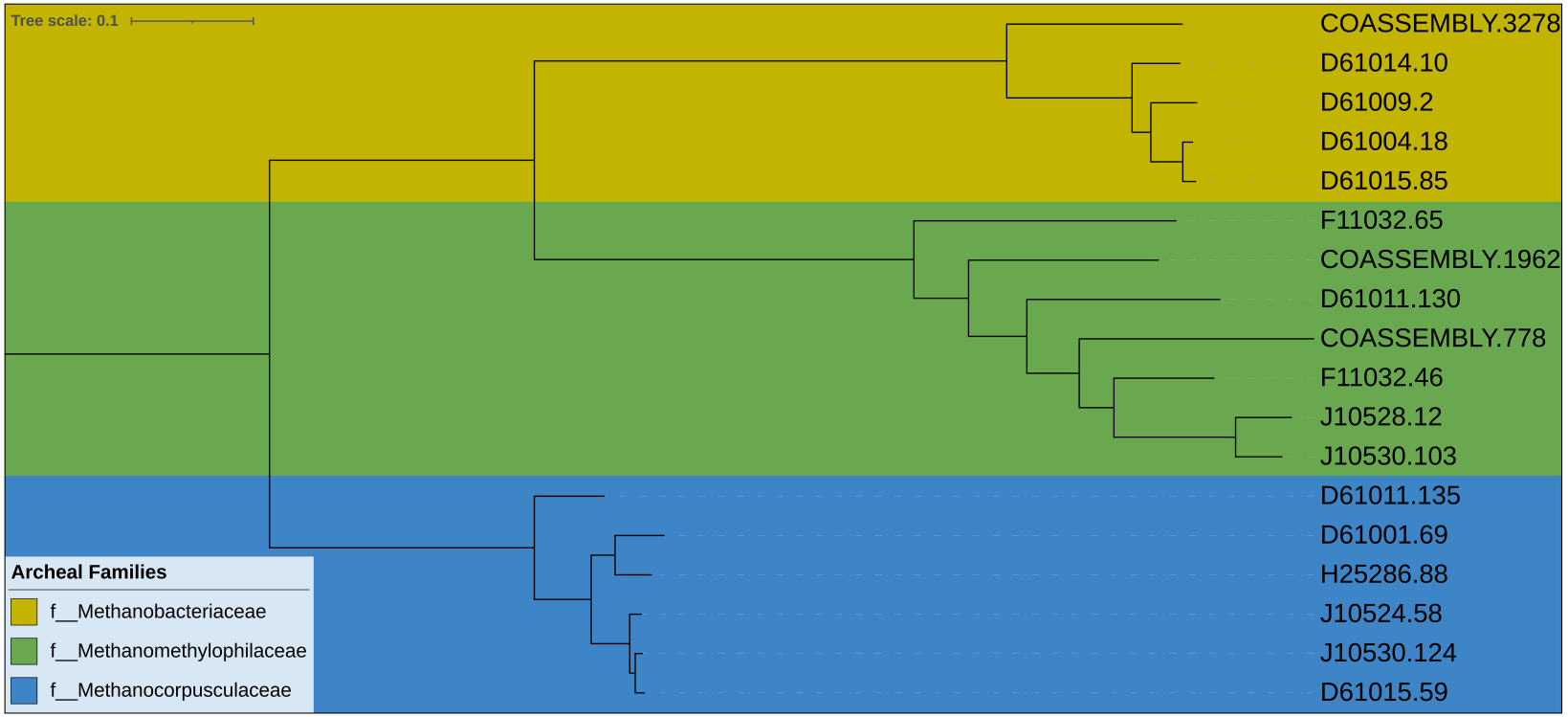


**Figure 3.12: Archaeal MAGs (Gut Microbiome).** Linear representation of a maximum likelihood phylogeny of archaeal MAGs visualized in iTOL. 21 high-quality MAGs were reconstructed from the equine gut metagenome collection. Sequences of conserved housekeeping genes were further utilized for the inference of the phylogenetic relationships. Archaeal MAGs were coloured according to their taxonomic classification on family-level.

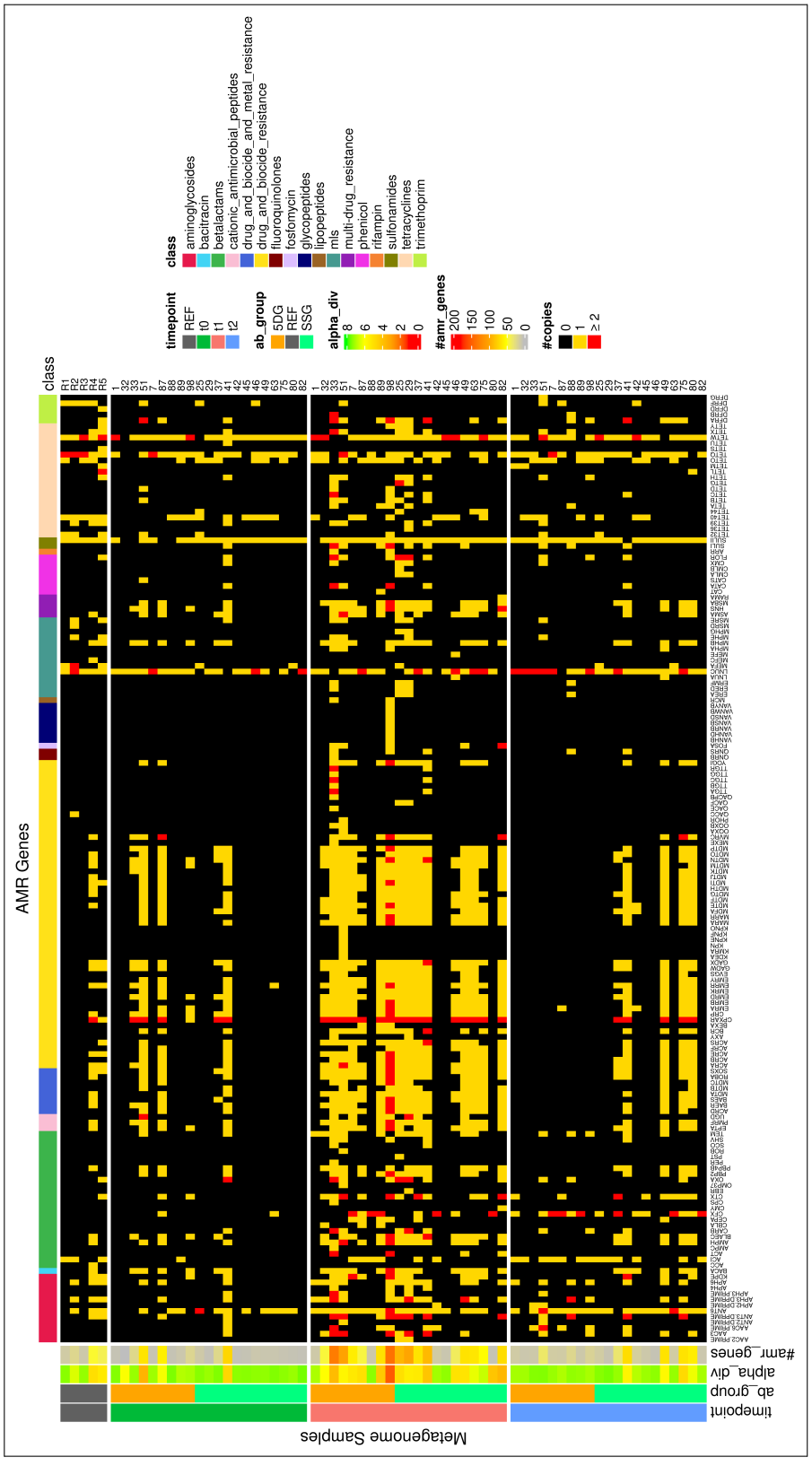


**Figure 3.13: Resistance Components (Gut Microbiome).** Visualized are the abundances of genetic resistance components across the sample set. These include ARGs, as well as genes conferring resistance against biocides and heavy metals. Samples (right) are divided horizontally by groups and time points (left). Columns represent individual resistance elements (bottom) across different classes of components (top). The heatmap colours indicate absence (black) or abundance (low to high, yellow to red) of specific elements. Additional metadata, including the total abundance of ARGs, as well as alpha diversity and IDs of the equine patients are plotted alongside the heatmap (left and right). AMR: Antimicrobial Resistance, REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

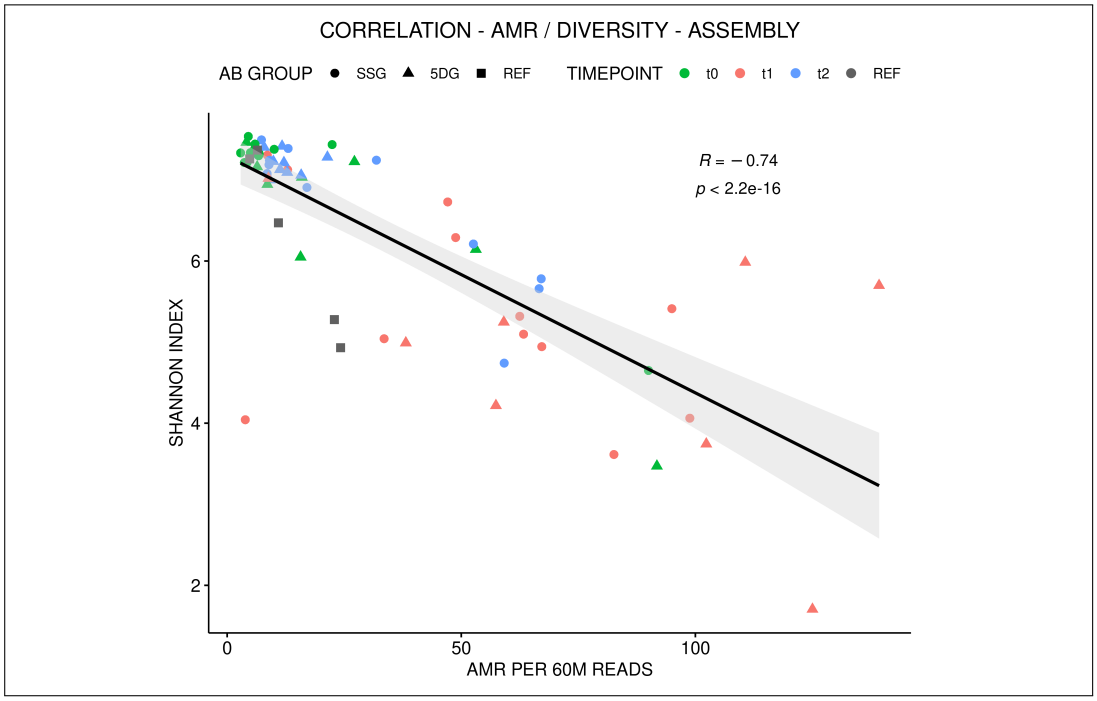


**Figure 3.14: Correlation of Resistance & Diversity (Gut Microbiome).** Spearman correlation of normalized resistance components and microbiome alpha diversity across the sample set. Individual samples were coloured according to their time points, with shapes illustrating the distinct study groups. Samples with low alpha diversity were more likely to carry high amounts of resistance elements. Most individuals with high resistance gene counts were associated with the 5DG. AMR: Antimicrobial Resistance, REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

Resistance gene counts were further normalized according to sequencing depth to enable group-wise comparisons. Interestingly, equines with decreased values of alpha diversity possessed more resistance genes compared to patients with high gut microbiome diversity. Counts of resistance components were therefore tested for correlation with alpha diversity values. A significant, negative Spearman correlation ($R = -0.74$, $p < 2.2e-16$) was identified between resistance genes and alpha diversity across the sample set, independent of study group or time point. Visualized in figure 3.14, samples with diminished microbiome diversity were more likely to possess elevated abundances of resistance genes compared to samples with high diversity. While many samples of $t_0$ and $t_2$ carried few ARGs, resistance components were highly abundant within $t_1$ samples and particularly associated with samples of the prolonged prophylaxis group (5DG).

In order to account for low-abundance resistance genes, additional mapping-based methods were utilized for resistome profiling. These elements can generally not be reconstructed through the assembly process and must therefore be analysed based on the assignment of metagenomic reads. Preprocessed fragments were hereby mapped to the dereplicated MEGARes database with stringent thresholds (see section 2.3.5 for more details) in order to further characterize the abundance of ARGs. Reads mapped to 239 unique genetic components, encompassing resistances to a total of 35 drug classes. Hierarchical clustering resulted in three general branches of resistance components: highly abundant (aminoglycosides, beta-lactamases, tetracyclines), common (biocides and metals) and rare hits (bacitracin and phenicols) identified within the gut environment. An overview of resistance genes (appendix figure A.4) and drug classes (appendix A.5) was illustrated using heatmaps. Next, comparisons were performed on class-level counts. Count abundances of ARGs were subsequently summarized to drug classes and assessed across time points for the SSG and 5DG. The results were visualized as box plots for the SSG (appendix figure A.6) and 5DG (appendix figure A.7). Findings include the significant increase of ARGs against phenicols (5DG: $p = 0.013$, SSG: $p = 0.092$) and trimethoprim (5DG: $p = 0.036$, SSG: $p = 1.0$) within the 5DG. In addition, ARGs associated with resistance to aminoglycosides (5DG: $p = 0.058$, SSG: $p = 0.84$) and beta-lactamases (5DG: $p = 0.0092$, SSG: $p = 0.0.0054$) expanded over time, primarily within the 5DG. Further aggregation of resistance elements revealed insights into the individuality of ARG accumulation (appendix figure A.8).

Genetic resistance determinants of the reconstructed draft genomes were further assessed in regards to the characterization of MDR. Processed MAGs of both prophylaxis groups contained up to 50 unique resistance elements, including genes conferring resistances to biocides and metals. The latter were excluded from further evaluation, as they do not directly induce AMR. A majority of the metagenomic bins (97%) were found to be genotypically fully susceptible to antibiotics, with few MAGs (3%) illustrating varying levels of AMR. Of all MAGs encompassing ARGs, a total of 40 contained genes conferring resistance to more than two antibiotic classes, indicating likely genotypic MDR for the associated bacterial groups. These taxonomic bins were primarily represented by *E. coli*, but also illustrated draft genomes classified as *Acinetobacter*, *Flavobacterium*,

| Classification | # of MAGs | % of MAGs |
|---|---|---|
| Total MAGs | 10,613 | 100% |
| Susceptible (No ARGs) | 10,346 | 97.5% |
| Resistant (ARGs of 1-2 antibiotic classes) | 227 | 2.1% |
| MDR (ARGs of > 2 antibiotic classes) | 40 | 0.4% |

**Table 3.1: MDR MAGs (Gut Microbiome).** Draft genomes were profiled for ARGs and classified based on the amount of resistance determinants towards different drug classes. A total of 10,613 MAGs were reconstructed through binning, of which a majority (10,346, 97.5%) were genotypically susceptible to all classes of antibiotics. 227 MAGs (2.1%) contained genes inferring resistance towards 1-2 antibiotic classes. 40 MAGs (0.4%) encompassed resistance determinants to more than 2 classes of antibiotics, representing taxonomic groups with genotypic MDR. ARG: Antibiotic Resistance Gene, MAG: Metagenome-Assembled Genome, MDR: Multiple Drug Resistance.

*Klebsiella* and *Rahnella*, all of which contained enrichments of resistance genes against multiple drug classes. They commonly included ARGs conferring resistance towards aminoglycosides, antimicrobial peptides, beta-lactamases, bacitracin and MLS. An overview of the results is illustrated in table 3.1.

Metagenomic assemblies were further profiled for virulence-associated factors. A total of 1,343 virulence-associated genes were characterized across the sample set. Analogously to AMR profiling, an overview heatmap was generated (appendix figure A.9). Patterns similar to those illustrated by the ARG profiles were detected, whereby a majority of virulence genes were abundant throughout $t_1$, while few were individually present at $t_0$ and $t_2$. The highest counts of virulence factors ($\geq$ 200) were identified within samples of the 5DG at $t_2$. Abundant virulence genes were involved in secretion mechanisms (*esp*L1-Y4, *gsp*C-M) and fimbrial proteins (*fae*D-J, *fim*A-I, *foc*H, *pap*B-K, *sfa*B-Y). Additional full-length genes were reconstructed for factors associated with the production of exotoxins (*ast*A, *cdt*A-C), bacterial motility (*flg*B, *fli*C), capsule formation (*kps*D-T), adhesion (*afa*A-F, *omp*A) and transporter systems (*fep*A-G, *iro*B-N). Further statistical analyses were performed and concluded that virulence abundance also significantly correlated with taxonomic alpha diversity ($R = -0.75$, $p = 3e - 13$) throughout the sample collection (appendix figure A.10).

Taxonomic association and mobility of ARGs are of particular interest for assessing the overall transferability of the aforementioned genes. One challenge is represented by the process of accurately identifying the phylogenetic origin of a gene, especially considering high genomic plasticity and HGT. By re-running the metagenomic classifier Kraken2 on the assembled contigs containing ARGs, it was possible to associate specific resistances with taxonomic groups. Kraken2 represents a method commonly utilized throughout microbiome research, which assigns taxa based on the majority of species-specific *k-mers* per contig. Finally, the putative mobility of resistance and virulence determinants were evaluated. Contigs containing these elements were analysed for their likelihood to be carried by a plasmid. Predictions with accuracy $> 0.5$ were labelled as potentially mobile-associated. These mobile contigs were taxonomically characterized and statistically compared across time points and study groups (figure 3.15).

Plasmid-associated ARGs increased significantly in abundance from $t_0$ to $t_2$ within the 5DG ($p = 0.014$, Wilcoxon rank-sum test). While mobile virulence genes increased towards $t_1$, abundances remained non-significant between $t_0$ and $t_2$ for both of the sample groups (5DG: $p = 0.1$, SSG: $p = 0.2$). Taxonomic allocation of these fragments (figure 3.16) revealed that a majority of the mobile ARGs were associated with the bacterial family of *Enterobacteriaceae* (28%), followed by *Moraxellaceae* (20%), *Bacteroidaceae* (8%) and *Lachnospiraceae* (6%). However, a large proportion of mobile ARGs (38%) were allocated to minor families or could not be linked to a corresponding taxonomic label at all. Plasmid-associated resistance genes commonly consisted of *aac*, *tet* and *sul* ARGs. Mobile virulence factors were also primarily linked to the family of *Enterobacteriaceae*. A comprehensive overview of bacterial families associated with mobile ARGs is illustrated in appendix figure A.11.

## 3.3 Computational Analysis of the Nasal Microbiome

Next, Meta16s (v0.1.7) was utilized for the analysis of the equine nasal microbiome data. This data set consisted primarily of 78 nostril swabs subjected to 16S rRNA sequencing, supplemented with an additional metagenome shotgun sample (J32522). The latter of which was analysed using MetaGEN (v0.7.8).

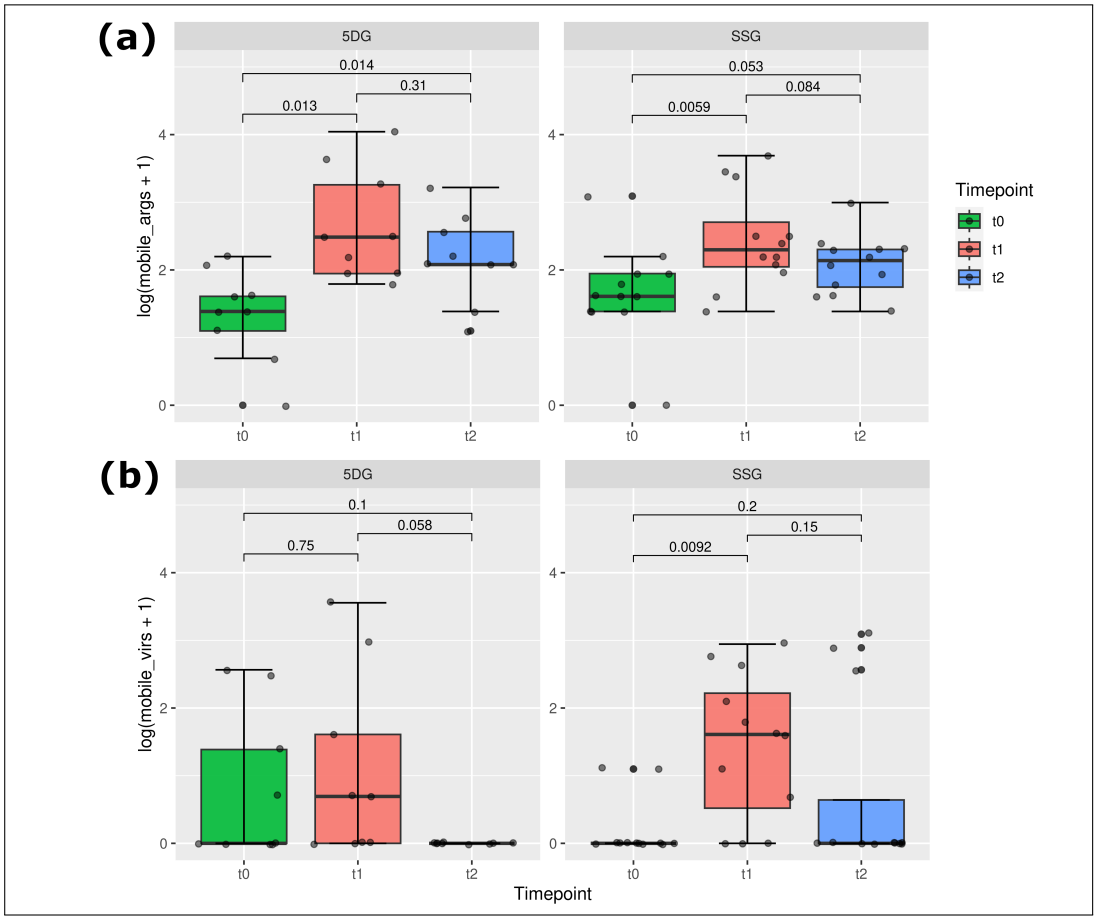


**Figure 3.15: Plasmidial Resistance & Virulence Abundances (Gut Microbiome).** Abundances of plasmid-associated resistance (a) and virulence (b) determinants were summarized per prophylaxis group and subsequently visualized as box plots (SSG: $n = 12$, 5DG: $n = 9$). Total counts were statistically compared across prophylaxis groups and time points using Wilcoxon rank-sum tests. A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

### 3.3.1 Data Assessment

16S rRNA sequencing data was processed according to section 2.2. The collection of 78 nasal samples were hereby utilized for analysis using the Meta16s pipeline. Raw data encompassed 424 MB of V3-V4 amplicon sequencing data. Clustering of this data set with the GreenGenes database resulted in a total of 6,539 high-quality OTU sequences. Samples contained on average 7,808 OTU counts, but sequencing depth varied greatly across the data set (ranging from 987 to 30,749). Rarefaction analysis was subsequently performed to evaluate the ecological saturation of these

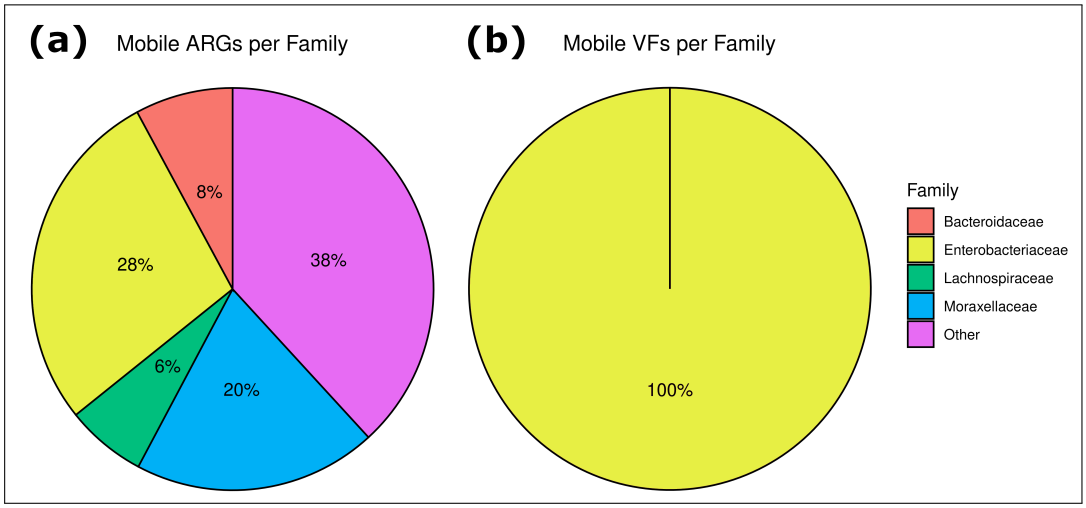


**Figure 3.16: Taxonomic Association of Plasmidial Contigs (Gut Microbiome).** Contigs of identified resistance (a) and virulence (b) determinants were characterized for plasmid-association. Predictions were then summarized in the form of pie charts for the entire data collection, based on their respective taxonomic annotations. Colours represent the 4 major bacterial families associated with these genetic elements. Minor taxonomic ranks are grouped into 'other'. ARGs: Antibiotic Resistance Genes, VFs: Virulence Factors.

sequences. An overview is illustrated in figure 3.17. As OTU counts were highly uneven, downsampling was applied to enable cross-group comparisons. Through rarefaction, samples were normalized to a shared, minimum depth of 987 OTU counts.

### 3.3.2 Taxonomic Profiles

Next, taxonomic profiles were generated through annotation of the established OTUs with the RDP database. Abundances were subsequently evaluated on multiple hierarchical levels. OTU counts were further investigated through heatmaps and bar charts. At the phylum-level, the most abundant taxonomic groups comprised of *Pseudomonadota*, *Actinomycetota*, *Bacillota*, *Bacteroidota*, *Cyanobacteria*, *Fusobacteriota*, *Chloroflexota*, *Acidobacteriota* and *Candidatus Saccharibacteria*. Comparative analyses revealed that the presence of these groups fluctuated across time. While the major phyla of *Pseudomonadota*, *Actinomycetota*, *Bacillota* and *Bacteroidota* remained comparatively stable, minor taxa fluctuated strongly throughout time. Counts associated with *Cyanobacteria* and *Fusobacteriota* increased until $t_2$, while the phyla of *Chloroflexota*, *Acidobacteriota* and *Candida-*

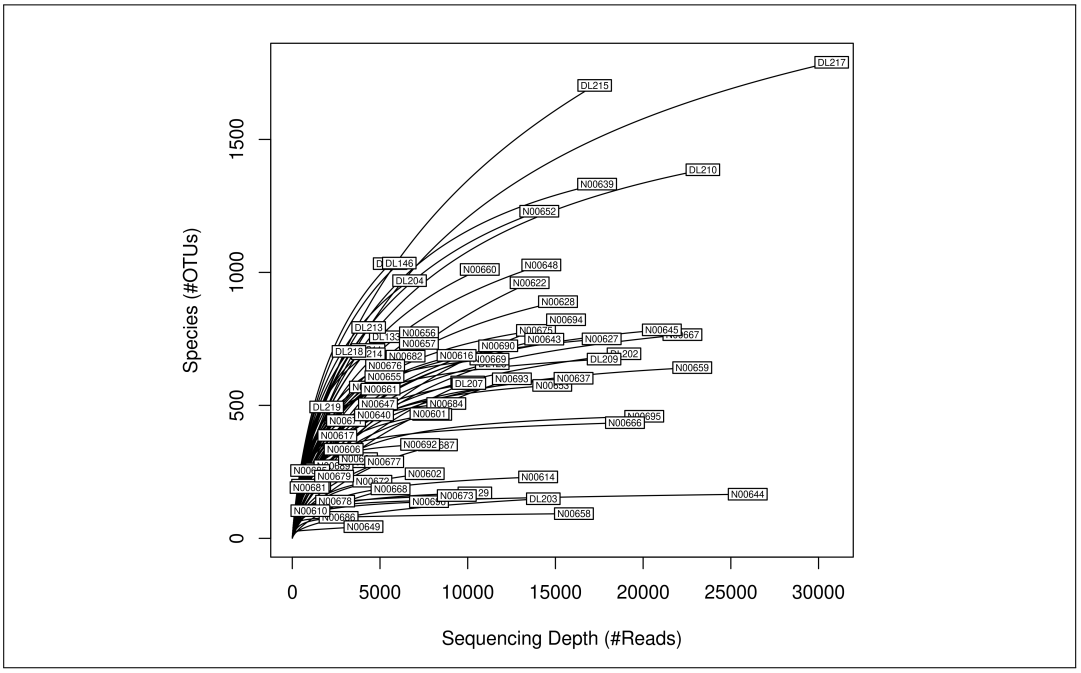


**Figure 3.17: Rarefaction Analysis (Nasal Microbiome).** Rarefaction curve of the equine 16S rRNA data. Illustrated are random subsets of differing depth for the sequencing data and the amount of OTUs classifiable throughout these. Each line represents a different sample of the collection ($n = 78$). All graphs rise quickly and then stagnate within a plateau-phase, indicating that the total diversity of the microbiome can be observed without utilizing the full sequencing depth. OTUs: Operational Taxonomic Units.

*tus Saccharibacteria* slowly diminished across both sample groups. Overviews of the phylum-level are visualized in figure 3.18, with more comprehensive insights into the individual samples provided in appendix figures A.14 and A.15.

Since phylum represents a high-level taxonomic rank, additional insights into more specific phylogenetic classifications were required. The family-level illustrates the highest resolution possible to achieve through OTU clustering of V3-V4 16S rRNA fragments. Data were therefore also aggregated onto the family-level. An overview is represented in figure 3.19 and appendix figure A.13. Insights into the individual compositions per sample are visualized in appendix figure A.16. The most abundant families of the nasal microbiome comprised of *Sphingomonadaceae*, *Micrococcaceae*, *Streptococcaceae*, *Corynebacteriaceae*, *Flavobacteriaceae*, *Microbacteriaceae*, *Moraxellaceae*, *Pasteurellaceae* and *Chloroplast*. Their abundance also varied, with a majority of these groups increasing in counts over time. Finally,

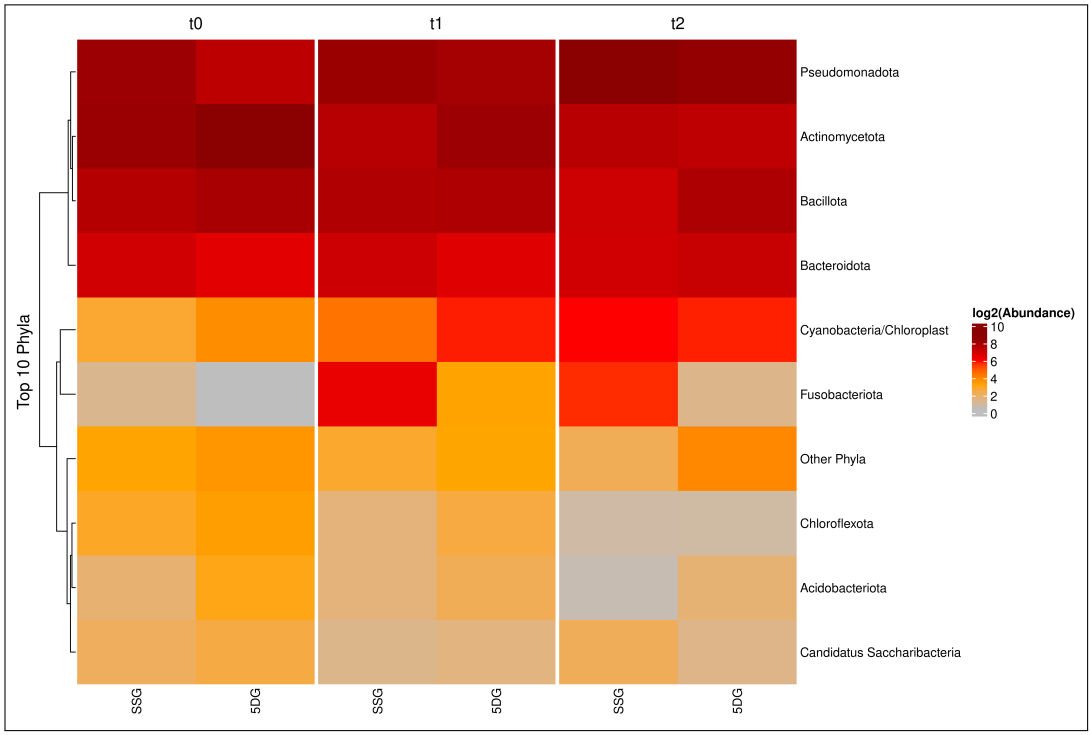


**Figure 3.18: Heatmap of Phylum-Level Abundances (Nasal Microbiome).** Taxonomic profiles were summarized at phylum-level and averaged across the sample groups. The heatmap visualizes the abundance ($\log_2$ of normalized counts) from low (yellow) to high (dark red) across the data set. Columns represent the different study groups at specific time points. Rows illustrate the 9 most abundant phyla, with minor ranks being grouped into 'other'. Phyla are clustered by their similarity to each other in the represented abundance profiles (dendrogram on the left). SSG: Single-Shot Group, 5DG: 5-Day Group.

differential abundance was statistically assessed on family-level. Group-wise comparisons were performed for multiple families of interest. The family of *Enterobacteriaceae* increased significantly over time within both prophylaxis groups (figure 3.20). Similar trends were also revealed for the family of *Fusobacteriaceae* (appendix figure A.17). The abundance of *Ruminoccocaceae* only increased significantly within the SSG (appendix figure A.18). In contrast, abundances of *Staphylococcaceae* decreased significantly over time for both groups (appendix figure A.19). *Streptococcaceae* illustrated minor variances within the SSG, with a significant increase at $t_1$ and decrease towards $t_2$, which were not identified for samples of the 5DG (appendix figure A.20). An overview of all differential families is listed in appendix table A.3.

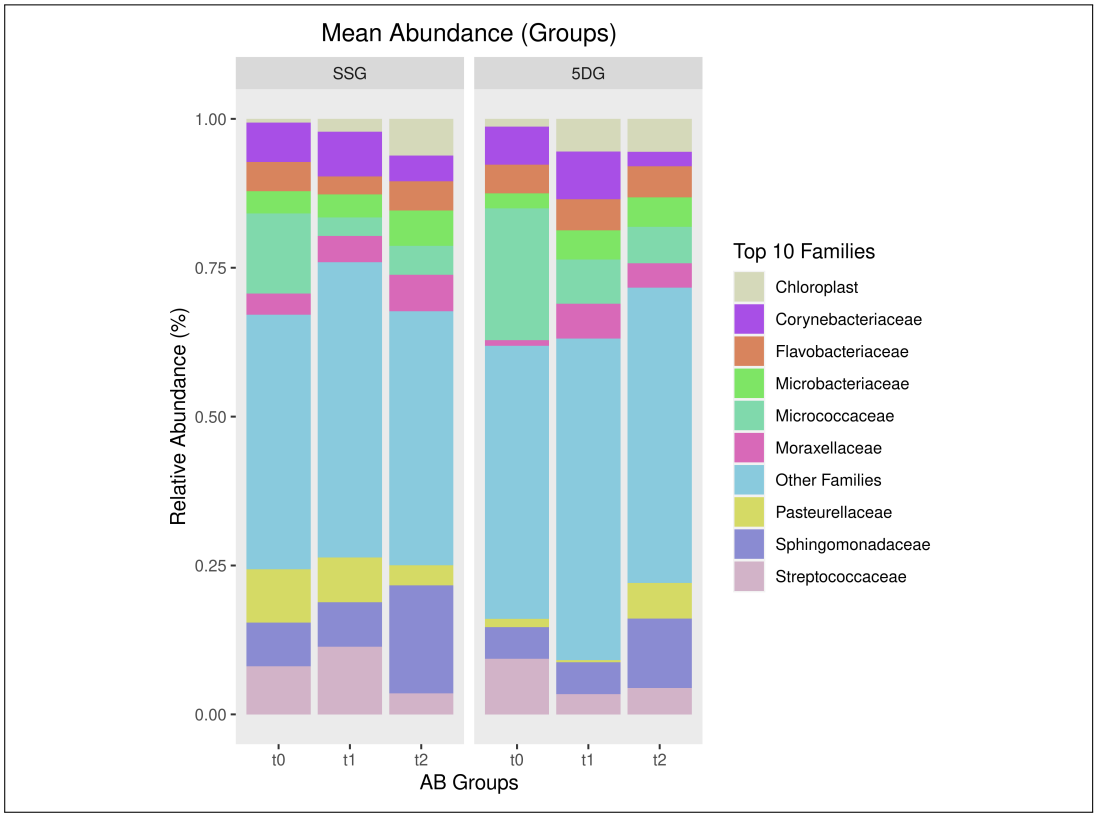


**Figure 3.19: Summarized Bar Charts at Family-Level (Nasal Microbiome).** Stacked bar charts representing the averaged composition of each sample group on the family-level. The 9 most abundant families are visualized as percentages of each bar, with minor ranks being grouped into 'other'. SSG: Single-Shot Group, 5DG: 5-Day Group.

### 3.3.3 Diversity Analysis

Ecological indices regarding richness, evenness and diversity were assessed for the nasal microbiome environment based on the established OTU tables (see appendix table A.1). On average, nostril swabs contained a total of 439 distinct OTUs (richness), with an evenness of 0.14 and an overall alpha diversity of 4.10. Shannon indices ranged from 1.08 to 5.52. Variations in alpha diversity were primarily driven by differences in richness and evenness. Taxonomic richness ranged from 33 to 954 OTUs for the sample collection. With values between 0.01 to 0.64, evenness also varied greatly across the data set. The comparison of prophylaxis groups across time was visualized in figure 3.21 and appendix figure A.21. Differences were assessed using Wilcoxon rank-sum tests. While evenness and alpha diversity differed between both groups at $t_1$, variations remained mostly non-significant.

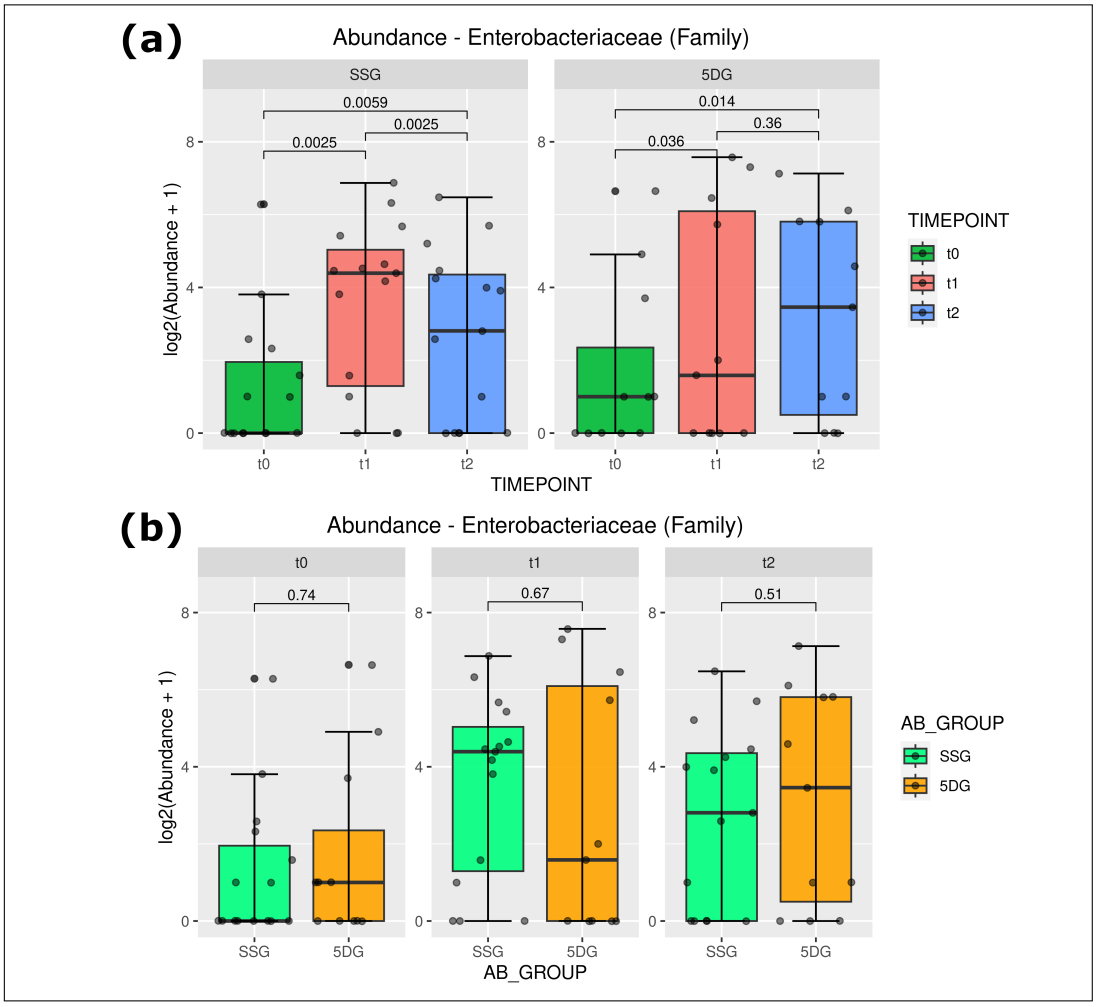


**Figure 3.20: Abundance of *Enterobacteriaceae* (Nasal Microbiome).** The abundance of the family of *Enterobacteriaceae* was assessed and comparatively investigated across the sample set. Individual abundances were summarized as box plots for each prophylaxis group (a) and subsequently compared over time (b) using Wilcoxon rank-sum tests (SSG: $n = 15$, 5DG: $n = 11$). A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

In addition, beta-diversity was assessed through PCA of computed Bray-Curtis distances. Principal Component 1 (PC1) accounted for 11.1%, while Principal Component 2 (PC2) encompassed 6.5% of the total variance observed. Visualized in figure 3.22, clear differences between the prophylaxis groups can be described. A general shift in microbiome structure is visible over time, whereby samples of the SSG appear much more variable in their taxonomic composition compared to the 5DG.

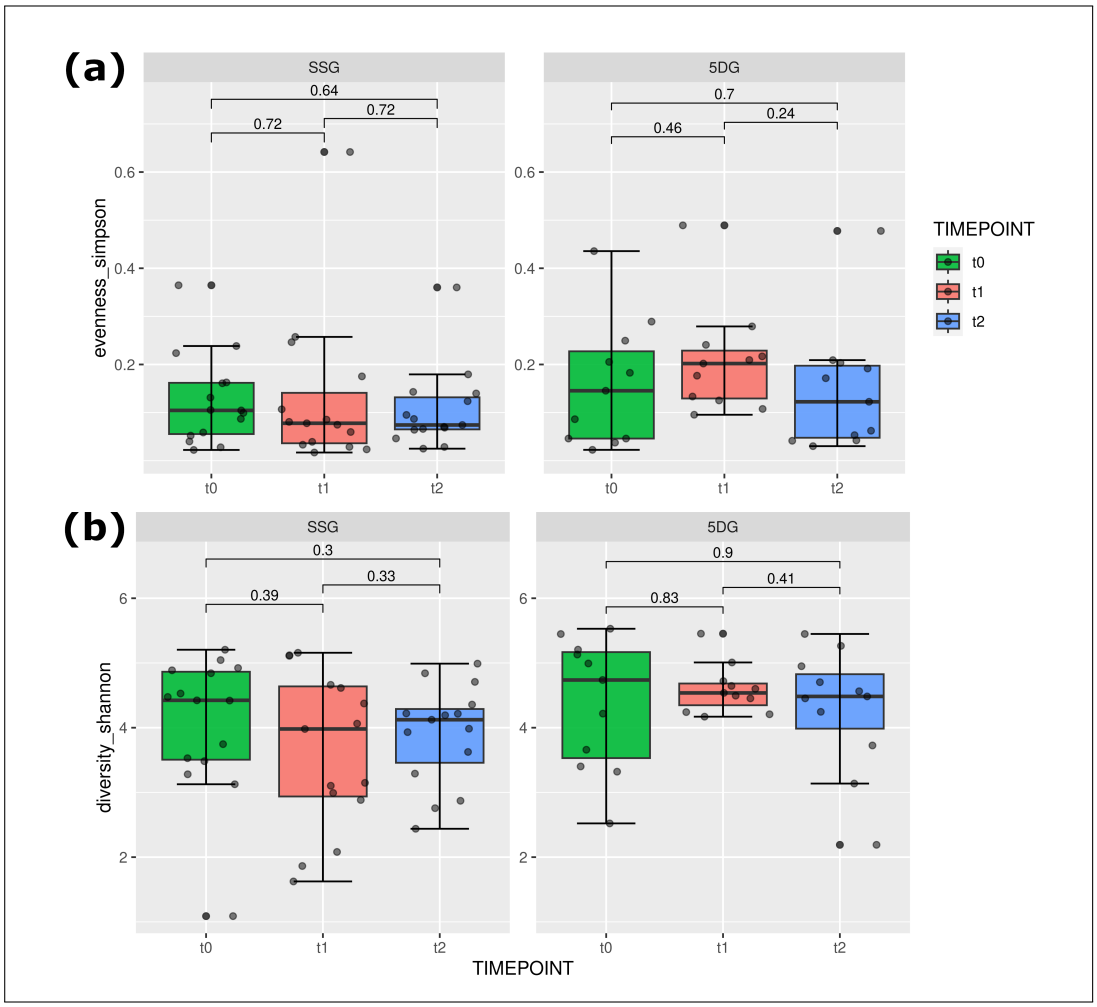


**Figure 3.21: Alpha Diversity (Nasal Microbiome).** Comparison of diversity indices across the sample groups (SSG: $n = 15$, 5DG: $n = 11$) using box plots. (a) Illustrates the evenness, while (b) represents the Shannon diversity over time (colours). Both values decrease from $t_0$ to $t_1$ and recover towards $t_2$. Wilcoxon rank-sum tests were performed to infer significant differences. A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

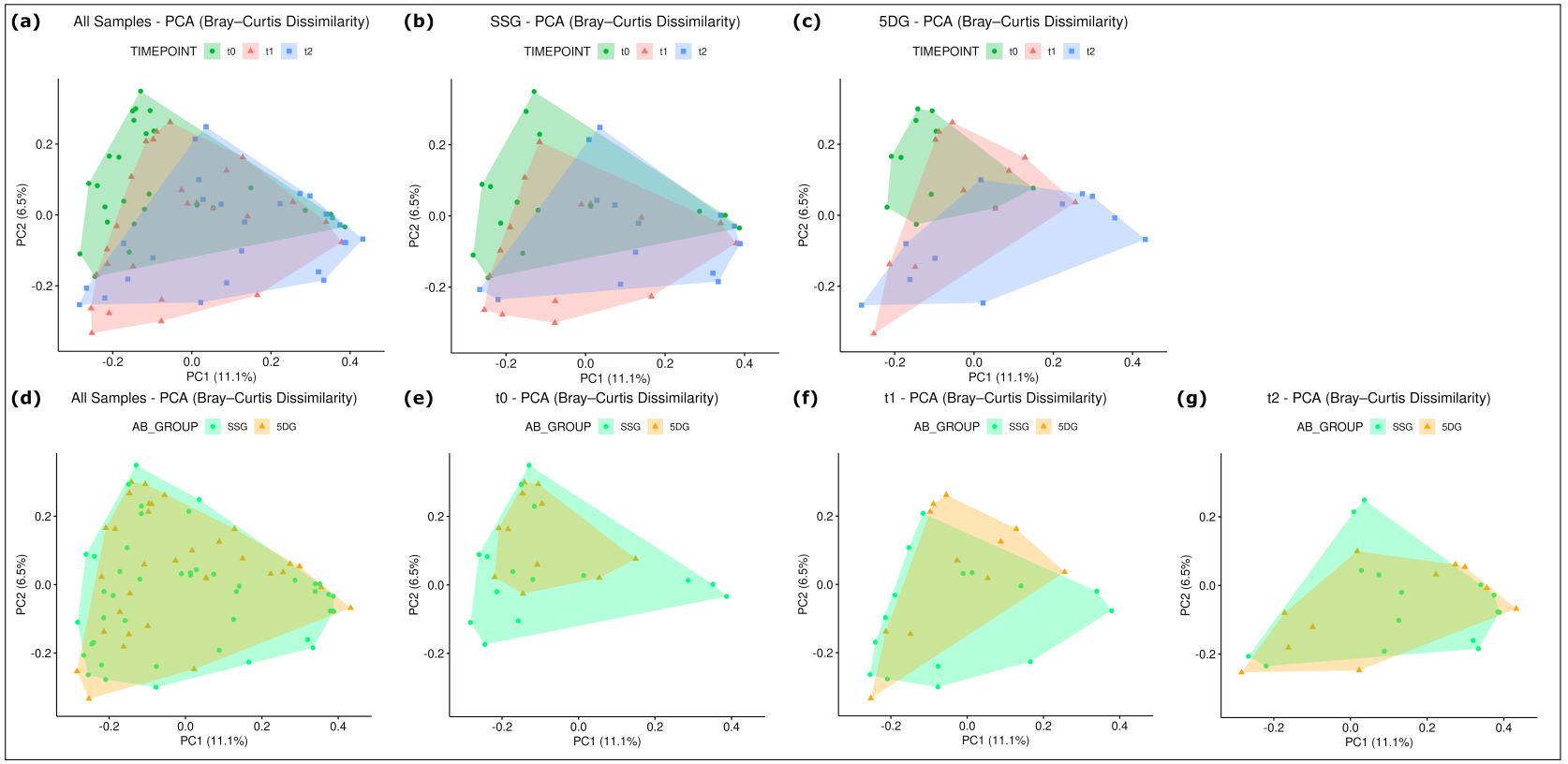


**Figure 3.22: Beta Diversity (Nasal Microbiome).** PCA based on Bray-Curtis distances of the nasal microbiome samples. Visualized is the complete sample collection across different subsets (colours). (a) Illustrates the full data set, coloured by time point. (b) and (c) represent subsets of the corresponding prophylaxis groups, coloured by time points. (d) represents the full data set once more, with samples labelled by group. (e)-(g) illustrate subsets for each time point, distinguished by prophylaxis group. PCA: Principal Component Analysis, SSG: Single-Shot Group, 5DG: 5-Day Group.

### 3.3.4 Resistome, Virulome & Mobilome Profiles

Finally, a representative sample of the 5DG (J32522) was metagenome shotgun sequenced in order to gain insights into the resistome of the nasal environment. Analysis of this sample was conducted using MetaGEN. Taxonomic profiles on family-level differed slightly from the aforementioned results (see appendix figure A.12), whereby groups of *Hymenobacteraceae* and *Methylobacteriaceae* were identified as highly abundant in addition to those characterized through the previous 16S rRNA sequencing. Metagenome sequences encompassed multiple resistance genes, including *aac*3, *carb*B, *erm*C and *sul*II, many of which have been associated with plasmids. *De novo* resistome profiling revealed additional ARGs conferring reduced susceptibility towards aminoglycosides, diaminopyrimidines, fluoroquinolones, fosfomycins, glycopeptides, MLS and tetracyclines. An overview of the identified resistance determinants are listed in appendix table A.4. Further taxonomic annotation revealed these to be hosted by multiple bacterial species, including *Acinetobacter nosocomialis*, *Escherichia coli* and *Staphylococcus aureus*.

# 4 Discussion

Microbiomes represent complex, dynamic ecosystems influenced by environmental conditions and host regulation [92]. The enclosed microbiota sense and respond to changes within their habitats, reflected through adaptations in gene content and expression patterns [185]. However, the complex processes behind the adaptability and dissemination of bacteria remain poorly understood, especially in regards to colonization resilience and the emergence of MDR within microbiomes dominated by species which cannot be cultured under current laboratory conditions [22, 92]. High-throughput metagenomics enables collective insights into these processes at a community level [185], furthering our understanding on how external factors can influence the host microbiome composition.

## 4.1 Novel Pipelines for Metagenomic Analyses

With the rapid progression of metagenomics over the recent decade, microbiome studies in particular have represented an ongoing 'hot topic' within genomics and science [31, 135, 154]. As technologies continue to evolve, novel methods are essential for gaining comprehensive insights into these complex NGS data. Metagenomic sequencing generates enormous amounts of 'big data', which require precise computational analyses in order to gain biological insights [26]. Due to the rapid advancements in this active field of research, there currently exists little consensus on gold standards regarding study designs and computational analyses [92, 183, 217, 232]. However, the quality of metagenomic analyses are known to be strongly influenced by experimental, biological, environmental and computational factors [92], in particular sample handling, sequencing methodology and bioinformatics workflow [53, 66]. Different bioinformatic approaches can hereby have major impacts on the filtering, clustering, taxonomic assignment and binning processes of microbiome data [92]. Few computational workflows exist for the analysis of metagenomic sequences [87, 108, 132, 212, 214]. However, these pipeline solutions have not been specifically designed to handle high abundances of uncharacterized equine microbiome data and do not encompass the full scope of analyses required throughout this study (described in section 2.1), particularly large-scale co-assembly (section 2.3.4), as well as high- and low abundance resistome profiling (section 2.3.5). It therefore remained essential to develop appropriate analysis

workflows based on the presented equine patient sets, in order to gain biological insights into the data at hand. Due to this, two novel pipelines were designed and implemented within the context of this thesis, termed MetaGEN and Meta16s, enabling comprehensive analyses of metagenomic shotgun and 16S rRNA sequencing data.

### 4.1.1 Meta16s for 16S rRNA Data

Meta16s utilizes the well-established approach of clustering 16S rRNA gene reads into sets of OTUs, representing DNA fragments of high similarity [1, 5, 92, 96, 183]. While there exist alternative methods to group 16S rRNA fragments based on their similarity [127, 193], OTU assignment was selected as the primary method for Meta16s in order to remain comparable to previous studies of the equine microbiome [154]. Due to the high technical noise of amplicon sequencing, including sequencing artefacts and low-quality variants, rigorous data preprocessing was further conducted to ensure true presence of the 16S rRNA fragments. Thresholds were hereby selected based on established research, including the overlapping of read pairs and removal of any ambiguous reads [154]. Quality-controlled fragments were then utilized for OTU clustering. Each OTU is assigned a taxonomic label, based on the utilized reference database. Historically, the GreenGenes database [195] represents one of the most frequently utilized databases for 16S OTU clustering, however it remains limited in taxonomic resolution due to its deprecated updates. This represents the primary reason why another source, the RDP database [196], was chosen for re-annotation of the identified OTU clusters. First, clustering was performed with the GreenGenes database, followed by additional annotation using RDP 11. This enables increased taxonomic resolution across the sample set, through the improved nomenclature of the RDP. The resulting abundance table is then cleaned for singletons, rarefied and utilized throughout downstream analysis, including taxonomic profiling and diversity estimation. Meta16s hereby enables the automized transformation of raw 16S rRNA reads into cleaned, applicable abundance tables, which represent the foundation for further microbiome analyses. These are then utilized for in-depth diversity estimation and visualization.

### 4.1.2 MetaGEN for Metagenome Shotgun Data

MetaGEN extends the capabilities of Meta16s through additional analysis steps specific to shotgun metagenome data. Particular focus was set on expanding the limitations of 16S amplicon sequencing through functional profiling, metagenomic assembly, and resistome annotation, all of which require shotgun data as an input. This enables characterization of metagenome-assembled draft genomes and full-length resistance determinants alongside taxonomic and functional results. Since a majority of the equine gut microbiome remains unclassified by current databases, additional reference-free approaches are utilized by MetaGEN. This includes large-scale metagenomic binning for the reconstruction of high-quality draft genomes, which expanded the known diversity of the microbiomes with thousands of annotated MAGs. As the collected dataset represents the largest equine gut metagenome sample set currently available, computationally expensive co-assembly was performed to utilize the full genomic capacity. This enabled the reconstruction of additional, low-abundance species which otherwise would have been missed by other pipelines. The resulting MAGs were further scanned for resistance and virulence determinants in order to gain deep insights into the distribution of these pathogenic elements across the taxonomic profiles.

Both, the Meta16s and MetaGEN pipelines represent standardized workflows for reproducible in-depth analyses of complex microbial communities, extending the capabilities of currently available software tools. Given an appropriate set of genetic references, such as equines and humans within the context of this study, these tools enable the flexibility to analyse a variety of host-associated microbiomes. MetaGEN[14] and Meta16s[15] are both open-source and freely available on GitHub under the GPLv3 licence. While individual (sub)tools might become obsolete in the future, the general workflow on how to analyse complex microbiome 'big data' and the types of biological results which can be generated from these represent essential resources for future studies within the field of metagenomics. Throughout this thesis, MetaGEN and Meta16s were both applied to a range of microbiome data sets in order to investigate the influence of hospitalization and antibiotic prophylaxes on the gut resistome of colic horses.

[14]https://github.com/SiWolf/MetaGEN

[15]https://github.com/SiWolf/Meta16s

## 4.2 Characterization of the Equine Microbiome

The microbiome represents an important health-associated indicator across all mammalian species. The findings outlined within this thesis illustrate how the multifaceted processes of surgery, hospitalization and perioperative prophylaxis affect the gut and nasal microbiota of equines. Throughout surgery and hospital care, equine patients were subjected to multiple environmental stimuli and internal changes (i.e. pain) known to influence their microbiomes. It hereby remained essential to distinguish between factors which affected all study participants (the process of hospitalization, surgery and PAP) and group-specific differences (induced by the differing lengths of prophylaxes). Within the following chapter, the established results will be discussed in light of the aforementioned aspects. This provides new insights into how both hospitalization and PAP affected the equine microbiome, particularly in regards to the accumulation of MDR.

### 4.2.1 At Hospital Admission

Equines arrived at the clinic with individual, unique microbiota compositions, resistance and virulence gene profiles ($t_0$), indicating that their microbiomes were primarily shaped through the individual backgrounds of the patient set. Variations in geographic origins, feed or stable conditions, as well as previous diseases and medical-histories represent a selection of discrete factors contributing to these observations [171]. PCA revealed insights into the diverse microbiomes for both the nasal (figure 3.22) and gut (figure 3.10) samples at the time of hospitalization. However, while microbiome compositions of the equine patients were driven by high individuality, they also possessed a multiplicity of similarities. A majority of samples hereby clustered together tightly with minor distances in beta diversity. This reveals the existence of a baseline microbiome composition, which underlines the suitability of the presented methods for further comparisons across multiple taxonomic levels over time. However, although the gut microbiome was represented comparatively well through PCA (with a total of 46.7% of taxonomic variability visualized using PC1 and PC2), the nasal microbiome could not be described as clearly through dimensionality reduction (only representing 17.1% of the variance). These results indicated that the equine nasal microbiome data is particularly multidimensional and cannot be fully captured using this approach [211]. Additional analyses were therefore required in order to gain insight into

the underlying taxonomic shifts occurring throughout the microbiome. While 3 of the 5 samples from NCBI (R1-R3, see appendix table A.2) appeared similar to the gut microbiomes of the equine patients, R4 and R5 represented outliers with vastly different taxonomic compositions compared to the collection. Given that these sequences were matched on a best-hit basis, external factors such as geographic distance or other environmental parameters, including differences in living conditions (stables vs free-range), likely contributed to the observed distances, as discussed previously [171]. However, this is particularly interesting, since the NCBI sequences represented healthy horses with no recent history of antibiotic usage [167], indicating that, while major taxonomic groups remain comparatively stable across the equine gut microbiome, variance in abundances and the presence-absence of minor taxa play a key role in overall host health and the onset of disease.

Comparing nasal and gut communities of the same individuals revealed insights into the fundamental differences in structural composition between these microbiomes. At hospital admission, the gut microbiome consisted of 2,304 conserved taxa shared by the entire equine patient collection (figure 3.7). While this was primarily dominated by *Pseudomonadota*, *Bacteroidota* and *Bacillota* (appendix figure A.2), high individuality was present even on phylum-level (figure 3.6). The nasal microbiomes encompass unique microbiota, clearly differing from those associated with the gut, as illustrated by the only few taxa shared between both environments. While similarities exist between these microbiomes on high taxonomic levels, the nasal microbiome was additionally inhabited by *Actinomycetota*, a phylum encompassing genera such as *Corynebacterium*, which represent common colonizers of the mucosa of humans and animals [151]. In addition, taxonomic composition of the nasal environment fluctuated even more strongly across individuals compared to the gut microbiota (appendix figure A.14), with biodiversity persistently increased within the faecal samples compared to the nasal swabs (average of 6.34 vs 4.10), primarily due to elevated values of richness (5,979 vs 439). Furthermore, high levels of evenness were detected within the nasal samples (0.14 vs 0.01).

Equines possess large and complex digestive tracts, for which a wide range of microbes play essential roles in food processing and energy production [157, 168]. They therefore host extremely diverse bacterial communities [135, 168], which

encompass high levels of functional redundancies in order to remain stable in metabolic potential over time [85]. While healthy gut environments are comparatively resilient against pathogen overgrowth, they can still be heavily affected by external stimuli, such as antibiotic prophylaxis [15]. In contrast, the nasal microbiome illustrates a much more transient environment, dominated by comparatively few taxonomic groups (5,979 vs 439). It thereby hosts a less diverse, but more evenly populated microbiota, heavily fluctuating through the direct surroundings of the equine patients.

A limited amount of resistance and virulence genes were reconstructed at $t_0$ (figure 3.13 and appendix figure A.9) for the equine patient data set. While a majority of these could be associated with the family of *Enterobacteriaceae* (appendix figure A.11), and the presence of this group was also verified independently within the gut (figure 3.8) and nasal (figure 3.20) environments, fragments containing these genes were primarily not associated with plasmidial resistance (figure 3.15), indicating passive carriage with limited HGT across these pathogens. ARG profiles at $t_0$ were also highly similar to the gut samples collected from NCBI, revealing comparable base-levels of individuality within the equine resistomes and virulomes. The reconstructed ARGs therefore represent non-mobile resistance components of the normal microbiota, which are present even in healthy microbiomes [20, 26, 53]. Additional insights were gained through low-abundance ARG profiling (appendix figure A.4), which revealed the ubiquitous presence of sequences associated with sulfonamide (*sul*I) and tetracycline (*tet*Q, *tet*W) resistances. Stringent thresholds and customization of the resistome database were utilized to ensure the true presence of these fragments within the metagenomic data set (section 2.3.5). However, since a majority of these could not be fully reconstructed through the assembly process, fragments representing these genes are present throughout the gut environment at a barely detectable, low-abundance level.

### 4.2.2 3 Days Post-Surgery

After 3 days of hospitalization and post-surgical prophylaxis ($t_1$), clear differences were detected in taxonomic profiles and diversity within the equine microbiomes compared to their initial compositions at hospital admission ($t_0$). Surgery, hospitalization and antibiotic prophylaxis induced large structural changes within the microbiome of the equine gut. This resulted in sets of highly diverse and unbal-

anced interindividual gut microbiota [153]. PCA of the gut (figure 3.10) and nasal (figure 3.22) environments revealed increased variability across microbial profiles throughout $t_1$. As Bray-Curtis distances strongly increased, individuals appeared shifted from their original ordinations within the PCA, indicating strong rearrangements within the underlying taxonomic compositions. High intra-individual variability in microbiome composition clearly illustrate a form of temporal instability [103], induced by the aforementioned processes of surgery, hospitalization and PAP. Previous findings of the accumulation of MDR bacteria within the equine patients further strengthen this observation [175].

By contextualizing the computed diversity values, additional insights into the complex relationships between community members were gained and described as follows. Throughout $t_1$, diversity dropped below the level of the NCBI comparison group. Both Shannon and evenness values decreased from $t_0$ to $t_1$ (figures 3.9 and 3.21), representing a rapid depletion of microbial diversity within the equine microbiome, resulting in a state of dysbiosis. However, this was primarily observed within the gut microbiome and was less profound for samples of the nasal environment. The reasons for this are likely multifactorial. Firstly, the gut microbiome may have been generally more susceptible to foreign influences throughout hospitalization due to the previous colic surgery, antibiotic prophylaxis or both. While the nasal microbiome is also visibly subject to compositional changes, it is intrinsically less diverse compared to the gut environment and through its transient nature less prone to be affected by the aforementioned factors. Secondly, as high variability continues to play a major role throughout the process of microbiome stabilization, it is clear that the presented results were also influenced to a certain degree by individuality. The gut and nasal microbiomes of equine patient 33, for example, visibly diverged from the remaining data set through an immediate expansion of *Pseudomonadota* (figure 3.6b and appendix figure A.14b). This patient was diagnosed throughout $t_1$ as suffering from colitis, a medical condition likely associated with the observed microbiome shifts of this individual. The *Pseudomonadota* phylum consists of a major pool for mobile ARGs within the gut resistome, as it includes many clinically relevant pathogens [28].

Antibiotic-induced depletion of the gut microbiome has been shown to favour the colonization of pathogens and the development of various gastrointestinal diseases,

also in equines [181, 263, 264]. Exponential accumulation of genetic elements associated with resistances and virulence was detected throughout $t_1$ (figure 3.13 and appendix figure A.9). Specifically, full-length genes conferring resistance against aminoglycosides, beta-lactamases, biocides and metals all proliferated strongly across the entire sample collection at $t_1$. However, the effect of individuality also remained clearly visible within the resistome profiles, with each sample encompassing a unique set of ARGs. These findings were further supported using independent read mapping against the MEGARes database (appendix figure A.4). High abundances of virulence-associated genes were also characterized across the study period, specifically at $t_1$. A majority of these consisted of fimbrial adhesion proteins, such as *fim*A-I [140], and secretion elements, including *gsp*C-M [139], which are common across Gram-negative bacteria. However, full-length genes were also reconstructed for genes involved in exotoxins (*ast*A, *cdt*A-C), cell motility (*flg*B, *fli*C), capsule formation (*kps*D-T) and overall adhesion (*afa*A-F, *omp*A). Exotoxins and adhesion factors in particular increase virulence of the associated strains, enabling enhanced connectivity towards host cells and the potential to induce cell death [17, 187]. Additional genes involved in cell motility and capsule formation, on the other hand, increase adaptability of the strains within changing environmental conditions, such as those imposed throughout surgery, hospitalization and PAP.

Co-occurring with the accumulation of resistance and virulence genes was the expansion of *Enterobacteriaceae*, increasing significantly in abundance from $t_0$ to $t_1$ across samples of both, the nasal and gut microbiome (figures 3.8 and 3.20). The family of *Enterobacteriaceae* spans more than 80 different genera, including bacterial species such as *E. coli*, which are common colonizers of the human and animal gut habitat, but are also abundant across environmental reservoirs [65]. *Enterobacteriaceae* represented the primary microbial group associated with ARGs at $t_1$ (appendix figure A.11), a majority of these genes were also characterized as being potentially plasmid-bound (figure 3.15). This underlines previous results describing the abundance and local dynamics of ESBL-producing *E. coli* within equine clinics [33, 175]. While both, putative plasmid-associated resistance and virulence genes, increased in normalized counts over time, mobile virulence factors continued to remain too scarce to be statistically significant for samples of

the 5DG ($p = 0.75$). However, a significant increase was noted within the SSG ($p = 0.0092$), likely due to the enlarged sample size (SSG: $n = 12$, 5DG: $n = 9$).

### 4.2.3 10 Days Post-Surgery

After 10 days of hospitalization ($t_2$), the equine microbiota shifted once more in taxonomic composition. Both the microbiome and resistome showed decreased variability, marking the onset of stabilization within the microbial composition. The nasal and gut microbiota revealed clear taxonomic changes occurring throughout time, shifting the overall structure of the microbiome from one state to another. Towards day 10, their compositions diverged back towards a potentially new stable state. Visualized through PCA (figures 3.10 and 3.22), decreased variability within the equine patient set was clearly evident. In comparison to $t_1$, all microbiomes illustrated signs of stabilization and potential recovery, with increasing Shannon diversities and evenness towards a state comparable to $t_0$ (figures 3.9 and 3.21). Similar dynamics have been reported through 16S rRNA sequencing of the gut environment, revealing that PAP administration can induce fundamental shifts in the taxonomic composition of the microbiome [190]. Hierarchical clustering further revealed varying degrees of taxonomic similarities between $t_2$ and samples of the previous time points ($t_0$ and $t_1$) (appendix figure A.2), indicating that this newfound state incorporates microbiome signatures from both previous time points, but also remains shaped by high individuality [31, 115].

Towards $t_2$, further resistome accumulation halted across all groups, as was expected with the offset of PAP. Full-length resistance and virulence-associated genes reduced in abundance compared to $t_1$ (figure 3.13 and appendix figure A.9). However, genetic components, including virulence factors and resistances against biocides and metals, remained within the microbiomes at $t_2$ and likely past the study period ($>$ 10 days). ARG accumulation occurring during hospitalization can therefore have profound long-term effects on the equine resistome [31, 157, 181]. This was particularly prominent for equine patients 75 and 80, which were primarily ARG-negative at hospital admission, but illustrated excessive accumulation of resistance genes towards day 10. ARG carriage was identified across a diverse set of hosts, dominated by few, abundant intestinal taxa. Most ARGs were found to be associated with the genus *Escherichia*, in particular the aforementioned species of *E. coli*.

*E. coli* was identified as being highly abundant across the presented data set, confirming its well-established role within the equine microbiome. It represents a diverse and adaptive species [265], frequently colonizing the gut microbiome of mammals, including humans and equines [37]. *E. coli* are genetically versatile, possessing high genomic plasticity through an open pan-genome [37]. They encompass both benign as well as pathogenic strains [170], with a majority representing commensal bacteria [97]. Being part of the *Escherichia* genus they are directly involved in host metabolism and defence mechanisms of the gut, but may also act as potential reservoirs for a wide range of ARGs [59, 81]. Expansion of pathogenic *E. coli* within the gut microbiome has therefore been associated with multiple diseases and represents a potential health risk [170]. As illustrated by the high abundance of associated ARGs, resistant strains of *E. coli* play a major role in the dissemination of the aforementioned ARGs across microbiomes, including equine hospital environments as reported previously [33, 175].

## 4.3 Influence of PAP Regimens on the Equine Microbiome

Changes occurring across the sample collection were primarily driven by the processes of surgery, hospitalization and perioperative prophylaxis, as discussed above. However, different PAP regimens (SSG vs 5DG) induced additional distinct effects upon the gut and nasal microbiomes. While both groups initially illustrated comparable patterns of diversity, with a decrease towards $t_1$ and subsequent recovery at $t_2$ (figures 3.9 and 3.21), major differences were identified within the conserved taxa shared between the SSG and 5DG (figure 3.7). Most individuals of microbial communities belong to few highly abundant species [4], which has been described to hold true for the equine gut microbiome [153]. As the 'core' microbiota of the patient set represent consistent microbes identified within the studied gut samples, they are likely involved in conserved functions throughout time points and groups. A large set of persistent microbes illustrates overall high stability of the ecological habitat, with many taxa shared between non-related hosts and few individually acquired. At $t_0$, a total of 2,304 conserved taxa were identified across both prophylaxis groups (figure 3.7d). This overlap remained stable at $t_1$, encompassing 2,364 shared taxa (figure 3.7e), before increasing to 5,494 at $t_2$ (figure 3.7f). Prior to this, the SSG had retained an enhanced conserved set of taxa (figure 3.7c) compared to the 5DG (figure 3.7b), indicating overall reduced

microbial diversity within the latter. Towards $t_2$ both groups appear to converge towards a new state, independent of prophylaxis group. This directly correlates with the aforementioned results of the PCA clusterings (figures 3.10 and 3.22), which also illustrated that samples converged towards a new, shared microbiome state at $t_2$. This is likely the effect of hospitalization and other coinciding factors (such as clinical environment, feed, etc.) shaping the gut microbiome across the study period [266].

### 4.3.1 Effect on the Taxonomic Profiles

Throughout hospitalization, the major taxonomic groups remained comparatively stable, indicating that the fluctuating set of conserved taxa primarily represented low-abundance microbes. Loss of these minor, but often important, taxa can open up various ecological niches for the overgrowth of resistant and foreign bacteria to emerge, changing the overall structure of the microbiome community. A reduction of conserved taxa can increase subsequent risks for health complications and other diseases through metabolic dysfunctions [153]. Rare species (those with low abundance) are in particular much more vulnerable to extinction [4] and therefore likely could not survive PAP and hospitalization. These may contain specialist species known to hold essential roles within their ecological niches with reduced habitat ranges and greater dependence on specific environmental conditions [75, 267]. PAP can further have profound effects on slow-growing bacteria, which also likely could not handle the short exposure times to antibiotics [15].

However, low-abundance species may also partially represent transient organisms originating from the environment, such as food sources [3]. For example, high abundances of *Cyanobacteria* were detected and clearly represent transient organisms originating from feed. In addition, low levels of environmental contaminants cannot be outruled completely, as these can easily accumulate during sample collection [98]. Extensive data preprocessing was therefore performed in order to remove any host-associated DNA contamination *in silico*, described in detail throughout section 2.3.1. As this process is challenged by the high complexity of metagenomic data sets, few reads could not be removed completely, as evident by the presence of the *Chordata* phylum identified within the sample set. Fragments of this phylum primarily represented *Homo sapiens* and could likely not be dismissed via the process of mapping due to the repetitive nature of their sequences.

This issue can be circumvented in future studies through increased care during sample collection and by utilizing longer read lengths (> 150 bp), which will improve the likelihood of these fragments to be successfully mapped to a genomic reference.

Regarding microbiome diversity, limited differences were identified between the 5DG and SSG. Potentially, this was due to both groups reaching a transient, stable composition quickly after initial antibiotic administration, with minor variations in microbial compositions. This has been outlined in previous studies, during which antibiotic exposure resulted in the development of a transient state [3, 15, 268]. However, taxonomic shifts, such as those determined throughout this study (figure 3.18 and appendix table A.3), can have important health implications for the equine patients, as these have been described to contribute to the onset of diseases [157]. *Verrucomicrobiota* decreased in abundance throughout hospitalization ($t_1$), followed by a partial recovery on day 10 ($t_2$). The phylum of *Verrucomicrobiota* represents abundant endosymbionts of the environment and their absence from the gut microbiome has been associated with metabolic diseases in humans [31, 81]. Due to their importance in maintenance of the mucosal layer within the intestines [180], this group is primarily involved in gut health homeostasis [81, 123]. A decrease in *Verrucomicrobiota* has been associated with multiple diseases in equines, including colitis [180]. At $t_1$, this phylum nearly vanished completely within equine patient 33 (figure 3.6), underlying the established link with the condition of colitis. *Bacteroidota* and *Pseudomonadota* increased at $t_1$ within the gut microbiome and diminished towards $t_2$. Elevated rates of *Bacteroidota* have been associated with the general process of hospitalization [123]. Furthermore, decreased abundances of *Bacteroidota* and *Bacillota*, as well as higher occurrences of *Pseudomonadota* have been described as a state of dysbiosis across various species, including horses [169].

*Bifidobacteria* also significantly decreased in abundance over time and did not recover at $t_2$, however solely within the 5DG (appendix table A.3). In humans, *Bifidobacteria* are known to promote overall gut health and are therefore frequently utilized as probiotics [31, 150, 169]. They increase resistance against pathogen infections through modulation of pH values via the production of short-chain fatty acids [155]. Species of this family are also known to be involved in the production

of butyrate, a substance associated with a healthy colonic mucosa, which is found to be depleted in patients with inflammatory bowel disease [168]. Butyrate is an important, calorie-rich nutrient for host metabolism that inhibits the growth of facultative anaerobic enteropathogens, including *Enterobacteriaceae* [186]. Elevated abundances of *Bifidobacteria* through probiotics have been associated with decreased rates of *E. coli* colonization [150]. A reduction of butyrate-producing taxa may therefore facilitate the overgrowth of the microbiome through pathogens, including ESBL-producing *E. coli*.

However, the equine gut microbiome also contained high amounts of unclassified reads (appendix figure A.1), begging the question of which other species are present within the sample collection. Metagenomic studies of less well-described environments frequently possess high rates of unclassified sequences, commonly upwards to 68% of all data [60, 134]. MetaGEN utilizes a comprehensive Kraken2 database consisting of well-characterized sequences from RefSeq. Unclassified reads therefore originate from undefined or underrepresented microbial species within the current RefSeq database, emphasizing the absence of many uncultured microbiota from these reference sets [114, 157]. As more research is required to assess this microbial 'dark matter', future metagenome studies will be able to annotate these using revised reference collections. This is of particular importance for the gut microbiome samples, as these contained high abundances of uncharacterized microorganisms, for which an alternative method was utilized in order to enable taxonomic characterization within the scope of this study [197, 228]. The complete read set was hereby subjected to metagenomic *de novo* assembly and subsequent binning, a process which reconstructs draft genomes of unknown, yet abundant microbes. Utilizing the most comprehensive equine gut metagenome collection to date, a total of 10,613 MAGs were reconstructed through (co-)assembly. MAGs included both bacteria and archaea, representing all major phyla of the gut environment (figure 3.11). Bacterial MAGs included *E. coli* and *B. fragilis*, but also *S. equinus*, a lactic acid-producing species associated with acidosis [191]. While comparatively few archaeal MAGs were reconstructed (21 of the 1,381 high-quality draft genomes), they represented a range of methanogenic families (figure 3.12) known to be abundant inhabitants of the equine colon [31].

### 4.3.2 Accumulation of Resistance Genes

Resistome and virulome profiling revealed strong negative correlations between microbiome diversity and reconstructed full-length resistance genes ($R = -0.74$, $p < 2.2e-16$) (figure 3.14), as well as virulence components ($R = -0.75$, $p = 3e-13$) (appendix figure A.10). Increased microbiome diversity is often positively associated with host health [118], as a rich, commensal microbiota can provide protection against infectious pathogens [151]. Correlation analyses enable insights into whether ARG abundances are related to other factors, such as diversity indices [25]. While it is impossible to reveal the true relationship of resistome and microbiome through statistical association alone, they allow insights into the complex ecological processes at hand. Correlation values measure the degree of which data sets co-vary, ranging from -1 to 1, with values $> 0.7$ (such as the ones computed here) being described as particularly strong effects [128]. However, identifying these relationships is only the first step which require further follow-up studies to establish causation [5]. In this case, low microbiome diversity co-occurred with elevated abundances of resistance and virulence elements. Decreased microbiome diversity may enable foreign species to invade the microbiome and to proliferate throughout antibiotic prophylaxis [185].

While it is important to evaluate rare AMR genes within the resistome, they are notoriously challenging to identify and characterize [66]. MetaGEN utilizes a mapping-based approach to assess low-abundance AMR genes which were missed by metagenomic assembly. Through read-mapping, abundances of individual antibiotic classes were further evaluated [217]. Interestingly, summarized ARG counts illustrated unique patterns for both prophylaxis groups (appendix figure A.8). Horses of the SSG arrived at the hospital ($t_0$) with slightly elevated resistance gene hits, before decreasing in abundance at $t_1$ and recovering until $t_2$. In contrast, ARG fragments within the 5DG continuously increased throughout hospitalization, with abundances at $t_2$ surpassing levels at $t_0$. While this trend was not statistically significant, additional analyses on specific drug classes revealed clearer changes occurring throughout the resistomes. Amongst the various antibiotic classes, both aminoglycosides ($p = 0.058$ vs $p = 0.84$) and beta-lactamases ($p = 0.0092$ vs $p = 0.0.0054$) illustrated significant increases in abundance over time within the 5DG (appendix figure A.7), while increases of resistance com-

ponents against aminoglycosides remained limited across the SSG (appendix figure A.6). Both of these are directly associated with the prolonged penicillin-gentamicin prophylaxis, as penicillin represents a beta-lactamase antibiotic, while gentamicin is classified as an aminoglycoside [269]. The aforementioned increases in resistance fragments within the 5DG, and to a lesser extent the SSG, are therefore likely to mirror the effect of prolonged PAP. In addition, phenicol ($p = 0.013$ vs $p = 0.092$) and trimethoprim ($p = 0.036$ vs $p = 1.0$) significantly increased solely within the 5DG. ARGs of these classes frequently represented co-resistance genes located on shared MGEs. This was also directly associated with the observation that plasmidial ARGs significantly increased in abundance over time within the 5DG, but not within the SSG (figure 3.15).

ARGs associated with beta-lactamases resistance were hereby frequently identified throughout hospitalization, including *bla*CMY, *bla*CTX, *bla*OXA and *bla*TEM (figure 3.13). These genes are commonly located on multidrug-resistant plasmids, indicating the presence of MGEs across the data set [56]. A majority of them are directly linked with ESBL, which represents a diverse set of resistance mechanisms against beta-lactamases shared by various Gram-negative pathogens of the gut [141, 270]. Limited treatment options and worse clinical outcomes are complicated by the known mobility of ESBL-associated genes through plasmids [141, 144]. Previously, phenotypic resistance profiling was performed for the studied horses, revealing the presence of ESBL-producing *E. coli* throughout the patient set [175]. Metagenomic analyses further supported these results, illustrating an increase in ESBL colonization rates from $t_0$ to $t_1$ (figures 3.8 and 3.20). By linking reconstructed ARGs through additional taxonomic classification, it was observed that ESBL-genes were primarily associated with the family of *Enterobacteriaceae*, in particular *E. coli*, stressing the importance of this bacterium as a commensal of the gut with increased pathogenic potential.

Mobilization plays a crucial role in the adaptation of bacteria to foreign environments, through HGT and the transfer of resistances [47, 59]. Once an ARG has been is mobilized, it can rapidly disseminate across an ecosystem [47]. AMR gene content spiked shortly after antibiotic prophylaxis ($t_1$), likely since MGEs (including plasmids) were released into the microbiome, which was confirmed by assessing the abundance of plasmid-associated resistance genes. However, plasmids often

require specific environmental conditions or hosts to be able to proliferate successfully [54]. Plasmid production and HGT can hereby be triggered through various environmental factors, antibiotic exposure or host defence mechanisms [144]. While increased antibiotic stress might act as a driver for HGT to occur frequently between bacteria, it is challenging to determine whether the noted increase in AMR was solely based on selection pressure or how propagation of environmental ARGs may have influenced these observations [51]. Moreover, MGEs are commonly transferred temporally and are, due to their significant maintenance costs, lost comparatively fast after offset of the respective selection pressure [37]. However, metagenomics can only detect the presence of specific genetic material and not whether the accompanying organism is alive or dead, further stressing that the identification of ARGs within an environment does not necessarily indicate an immediate health risk [51, 53]. Assessing the pathogenic potential is further complicated within metagenomics, as no single species or strain is involved in the transmission of AMR, but instead interactions take place across a complex mixed community of microbes [26]. Further investigations are therefore needed to confirm the long-term ($>$ 10 days) preservation of these elements within the equine microbiome, as increased levels of plasmidial sequences at day 10 (figure 3.15), particularly within the 5DG, indicated expansion of resistance-associated elements through HGT [143]. The same holds true for ARGs, as their presence alone indicates that resistances were present within the environment, warranting additional research [53].

While virulence-associated elements also fluctuated throughout hospitalization, differences in abundances remained non-significant. One reason behind this, at least partially, was due to the underrepresentation of microbiome virulence components within current databases. All plasmidial virulence factors were hereby associated with the family of *Enterobacteriaceae* (figure 3.16), more specifically *E. coli*, which represents an abundant species of public health interest as described previously. However, it appears unlikely that no other bacterial family contained virulence components, such as toxins or additional adhesives. More research is therefore needed to characterize the presence of virulence factors within microbiome environments and to expand current reference databases.

### 4.3.3 Local Dynamics of MDR Bacteria

MAGs were further profiled for resistance components. AMR profiles were categorized according to guidelines for human medicine, whereby bacteria carrying resistances to three or more antimicrobial classes were classified as genotypically MDR [33, 34, 56]. While a majority of genomic bins were genetically susceptible to antibiotics (97.5%), this enabled insights into a range of ARGs containing worrisome patterns of resistances. Although archaea are known to be intrinsically resistant to a range of antibiotics [15, 176], resistance genes were solely identified within bacterial species. Of all MAGs, ARGs against 1-2 antibiotic classes were identified within 227 (2.1%) draft genomes, while 40 MAGs (0.4%) possessed resistance determinants against $> 2$ classes of drugs, making the accompanying species genotypically MDR (table 3.1). Metagenomic classification revealed that a range of microbiome colonizers have the potential to carry resistance genes against multiple drug classes. Taxonomic groups with high potential for MDR included a range of Gram-negative, public health relevant pathogens such as *Acinetobacter*, *Flavobacterium*, and *Rahnella*, but were primarily dominated by bins of *E. coli*. MDR MAGs further encompassed *Klebsiella*, which are commonly associated with nosocomial infections and represent a reservoir for a wide range of ARGs [155]. It is therefore likely, that strains and ARGs were acquired throughout the hospitalization and proliferated within the dysbiotic microbiome [155]. These taxa represent therapeutic challenges as they possess the ability to carry resistance genes against multiple classes of drugs, limiting treatment options [175].

Interestingly, comparable resistance genes were identified within the gut and nasal microbiomes. The shotgun sequenced nasal swab contained a plethora of ARGs against aminoglycosides (described throughout section 3.3.4). Summarized with the previous results, it seems plausible that the gut microbiome might be consistently seeded with resistant bacteria from the environment [54, 175]. Since the effect of hospitalization appeared stronger than the effect of the different PAP regimes, it is highly likely that the equine patients acquired resistant bacteria locally at the clinic. This would explain why highly similar patterns of ARGs proliferate across the entire sample set and remain within the microbiomes for longer periods of time even after the antibiotic prophylaxis has concluded. As the nasal

environment illustrates a brief, transitional environment before foreign pathogens are ingested, elevated ARG counts within the nasal swab underline this theory.

Healthcare facilitates are known to harbour a reservoir of transferable resistance genes, thus stressing the need to further improve hygiene concepts in order to prevent the spread of infections [19]. In human healthcare facilities, microbes may circulate for years, sustaining a consistent reservoir of AMR, while persistently causing opportunistic infections [79]. Several hypotheses exist as to how these reservoirs continue to exist, including evidence of repetitive seeding from new patients, as well as high selective pressure through hospital hygiene and cleaning interventions [79]. While resistant bacteria may revert to a susceptible state in the absence of antibiotic pressure [146], this is naturally challenged due to constant environmental exposure. Previous research illustrated that veterinary clinics also represent hotspots in regards to the accumulation and local dynamics of MDR bacteria, such as ESBL-producing *E. coli* [175]. This not only represents a health threat for patients, but also staff, as close interactions with animals poses risks towards colonization with resistant bacteria through zoonotic transmission events [28, 78]. Equine patients arriving at the clinic may already partially be colonized with ESBL-producing *Enterobacteriaceae* [174], which is also supported throughout the presented work (figure 3.13). Close spatio-temporal relationships can then result in the local spread of these pathogens [175]. Sources of local transmission include other hospitalized horses or the environment, such as contaminated stables, floors and medical equipment [175]. As equines require special needs in regards to their housing, keeping pathogenic bacteria at bay throughout these environments is particularly challenging, calling for strict hygiene management through consistent cleaning protocols [174]. Additional efforts should therefore be aimed towards improving hospital hygiene via routine resistance profiling in order to further reduce the spread and development of AMR within these clinical settings.

Antibiotic stewardship represents one pillar in the ongoing fight against the emerging MDR pandemic. Increased selection pressure caused by antibiotics favours the evolution and transfer of resistance determinants [51]. Even marginal levels of antibiotics can result in AMR, virulence and HGT within an affected bacterial population [15]. It is therefore pivotal to reduce this ecological pressure in order

to limit the risk of MDR-associated outbreaks. As such, the aim of this study was to assess the prevalence of ARGs in the equine gut microbiome as well as their dynamics throughout antibiotic prophylaxes and hospitalization. Previous results outlined that a reduced PAP duration in equine surgery did not increase the occurrence of adverse post-operative effects [174]. Based on this, it was hypothesized that the enhanced antibiotic prophylaxis of the 5DG would vastly promote the development of a resistome compared to the SSG.

While taxonomic shifts were expected to occur throughout the continuous 5-day prophylaxis group, results illustrate that even a single-shot dose disrupts the sensitive equilibrium within the gut (figure 3.9). Antibiotic prophylaxis was characterized by a moderate degree of gut microbiome dysbiosis, including a decrease in the abundance of health-promoting bacteria (*Bifidobacteria*) and an increase in various resistance and virulence-associated genes, as well as proliferation of opportunistic pathogens, such as ESBL-producing *E. coli*. Antibiotic treatment has been associated with a reduction of diversity and the depletion of commensal gut bacteria [180]. Overall, diversity decreased during hospitalization, independent of prophylaxis groups, however resistance gene counts increased predominantly within the 5DG. Reducing the usage of antibiotics will therefore also result in decreased selection pressure within this environment, enhancing the safety of the animal patients and staff members [175]. Prudent usage seems sufficient to inhibit the development of surgical site infections and represents one step in the ongoing battle against MDR amongst bacteria inhabiting skin and mucosal surfaces [174]. Thus, shortening the length of perioperative prophylaxes in equine surgery is advisable, as antibiotics are known to disrupt the gut microbiome and deplete beneficial bacteria, allowing the expansion of pathogenic strains [123]. Since total ARG abundance was decreased within the SSG compared to the 5DG, the potential benefit of reducing antibiotic prophylaxis with no clinical impairment has been demonstrated.

## 4.4 Limitations of this Study

It is important to note that this study has several limitations. For one, patient participation remains a major challenge in equine research. While the presented sample size is comparable to previous research on the equine microbiome [168],

specific methods, such as machine learning techniques, require more samples, and currently remain unfeasible with the included data set [147]. Causative conclusions and directionality therefore need to be evaluated with care and require additional validation. Increasing the amount of subjects in future studies will enhance the ability to detect variations and help to differentiate between the natural, complex microbiota and the antibiotic-states through additional, statistical power. However, although colic surgery represents a well-established intervention within equine medicine, it remains challenging to increase the amount of studied cases within a clear period of time due to the low success rate of the intervention [31, 169, 174]. The equine population included in this study therefore also spanned a broad range of different age groups, breeds, dietary habits, medical records and previous managements. While these factors likely had limited effect on the gut microbiome [154, 180], the airway microbiota is known to be heavily influenced through environmental conditions such as weather or seasonal effects [184], which cannot be outruled entirely [31, 155, 189]. These variations might have hampered the detection of otherwise statistically significant differences, as has previously been described by others [157, 168]. Upcoming surveys should therefore extend sampling to incorporate healthy control horses, in order to establish a baseline for any comparative analysis. While this was possible within the context of this thesis through the inclusion of equine gut microbiome samples from NCBI, the nasal microbiome remains largely undescribed, with limited data currently available. Furthermore, although additional follow-up between clinic discharge and recovery outside might prove challenging, this may reveal potential long-term associations induced by the hospitalization.

While this study primarily focused on the technology of metagenome shotgun data, 16S rRNA sequencing was performed for the accompanying nasal swabs. In order to identify shared resistance determinants across these environments, additional shotgun sequencing was conducted for a subset of nasal swabs. Nasal samples represent low biomass samples, known to be particularly susceptible to foreign contamination, with sequencing remaining a challenge. While 16S rRNA sequencing was able to successfully capture a wide range of taxonomic diversity within the nasal microbiome, shotgun sequences primarily consisted of host DNA ($> 99.9\%$) and were thus unusable for further microbial analysis (data not shown).

Focus was therefore set on characterizing the nasal resistome of one representative sample for the 5DG (J32522).

Regarding the identification of ARGs, different members of the microbiome may encode for the same resistance genes, inhibiting the identification of the responsible taxon [49]. In addition, as not all individuals of a species may be resistant, difficulties arise during ARGs profiling across the population [49]. Both of these issues were largely addressed via mapping-based approaches, however true phenotypic resistance can only be assessed using laboratory validation. Furthermore, it is likely that metagenomic sequencing did not unravel the full richness of AMR genes present within the samples, as these might only be recovered at $> 200$ million reads per sample [65]. This represents another reason why culture-dependent and *in silico* prediction methods must be utilized to complement each other [31].

Finally, using metagenomic data alone is limited in the assessment of the total bacterial load, as this method cannot distinguish between alive and dead material [49, 53, 76, 96, 98]. By supplementing metagenomics with gene expression or protein data, further insights can be gained into the complex symbiotic dynamics of microbiome and host [53, 92, 147]. Metagenomic RNA-Sequencing is one such method which can be applied to samples in order to evaluate the true functional activity of microbiome communities and their hosts [5, 92], as well as the transcriptomic mechanisms behind infections of identical bacteria resulting in different clinical outcomes [241].

In summary, the presented results will act as a foundation for future analyses. Based on the aforementioned findings, future studies are strongly recommended to increase sample sizes and focus on standardizing protocols, in particular sample handling and bioinformatical analyses. Additional culture-based methods can be applied to confirm the results presented throughout these studies [60, 123].

# 5 Outlook & Conclusion

As the world slowly recovers from the SARS-CoV-2 pandemic, we can reflect on the major contributions NGS has enabled in regards to the characterization of novel variants and the rapid development of vaccines [125]. The integrated genomic surveillance of microbial pathogens has never been as relevant as today [63] and metagenomics may contribute to this throughout the current decade [87]. Metagenomics represents a versatile, molecular tool that will continue to be utilized for environmental AMR surveillance and microbiome characterization, contributing to disease prevention and clinical diagnostics in the future. It can be implemented as an additional tool for the passive analysis of complex microbial ecosystems, supplementing existing whole genome surveillance systems [53]. Fuelled by interdisciplinary research, metagenome-based epidemiology enables new opportunities for outbreak detection and risk assessment [53], but also personalized medicine, particularly in regards to gut microbiome transplants and the prevention of acute dysbiosis [31, 122].

Current high-throughput sequencing technologies already enable metagenomic surveying in a timely manner, but elaborate downstream analyses are often still required, spanning a wide range of specialized software and algorithms in order to gain insights into the data at hand. However, given the pace at which this rich field is progressing, many of the ongoing issues will likely be overcome within the foreseeable future [241], particularly in regards to *in silico* ARG prediction. AMR represents a global One Health issue, not bound to single key pathogens, but instead affecting a wide range of different sectors, environments and habitats. As sequencing costs continuously fall, read depths and lengths will further increase. With the field of metagenomics continuing to evolve, new bioinformatical methods will therefore achieve even better representations of environmental ARGs. Future studies of microbiome resistomes will hereby be able to detect rarer AMR genes with greater accuracy. While computational approaches for the prediction of AMR have been in development for decades, newest research in the field artificial intelligence, including machine learning methods such as deep learning [147], have the potential to overcome existing databases issues and to predict novel resistance genes [22] through their ability to recognize subtle patterns of data with little prior knowledge [92]. Given sufficient training data sets, these

approaches have the potential to predict future states of microbiome compositions prior to any treatment [5]. Additionally, recent advancements in long-read sequencing will fuel the development of new standards for metagenomics and subsequent analysis methods [215, 230, 233]. However, all of the aforementioned *in silico* methods require further validation through established microbiological laboratory approaches, such as bacterial culturing or PCR for ARG detection, before their results can be applied within clinical contexts [49].

Throughout this work, the advantages of both 16S rRNA and metagenome shotgun sequencing have further been highlighted. While amplicon 16S rRNA sequencing is comparatively straightforward to analyse and yields deep insights into the microbiome composition, it is intrinsically limited in taxonomic resolution and cannot be utilized for functional profiling, including genotypic resistome analysis. Metagenome shotgun sequencing, on the other hand, enables insights into the entirety of the microbiome gene content, allowing characterization of resistance, virulence and plasmid-associated genes. However, analysing metagenome shotgun data is challenged by high dimensionality and requires bioinformatic expertise in order to gain biological insights. Mixing both of these methods will therefore remain essential throughout future studies. Particularly for the analysis of low-biomass samples, such as nasal swabs, 16S rRNA seems preferable compared to metagenome shotgun analyses, as it voids sequencing of eukaryotic host DNA. For functional microbiome analyses however, 16S rRNA data must be supplemented with metagenome shotgun sequencing in order to gain insights into abundant pathways and resistance determinants.

This study combined the advantages in high-throughput sequencing of 16S rRNA with the advanced resolution of shotgun metagenomics in order to describe the equine gut microbiome and to characterize its association with hospitalization, antibiotic prophylaxis and resistance accumulation in great detail. The results presented throughout this thesis illustrate a useful resource for future metagenomic analyses regarding exploration of the effect of antibiotic prophylaxes on host microbiomes. Millions of gene sequences were recovered across thousands of genomic bins, representing the largest catalogue of equine gut microbiome data available to date. In particular, the data set presented throughout this work can be utilized to generate new hypotheses and act as a starting point for further

comparative studies on the effect of antibiotic prophylaxes within equine medicine. Coupled with the open-source workflows developed for metagenomic (MetaGEN) and 16S rRNA (Meta16s) analyses, new research will enable the unravelling of the complex interactions between host microbiomes and antibiotic usage. Both MetaGEN and Meta16s were developed specifically for the reproducible analysis of metagenomic sequences and are already being utilized throughout further studies. As their frameworks are completely modular, they can be easily expanded with new bioinformatic software in the future. Combining a wide range of metagenomic tools within a customized, reusable framework was essential, as it allowed to uncover a wide range of mechanistic insights within the equine microbiome, which would have been missed using alternative workflows.

Focus was set on characterizing the impact of antibiotic prophylaxes on the composition and resistance profiles within the equine gut and nasal microbiomes. Results illustrate that hospitalization and perioperative prophylaxes significantly alter the equine microbiome. A single dosage of PAP and sustained hospitalization has already been described to induce long-lasting changes within the microbial gut community [163], including the emergence of AMR genes. ARGs conferring resistance to the classes of aminoglycosides and beta-lactamases were highly abundant across the sample set, particularly within the prolonged prophylaxis group. A range of pathogens were identified as genotypically MDR, including Gram-negative bacteria such as *Acinetobacter*, *Flavobacterium*, *Klebsiella*, *Rahnella* and *E. coli*. Resistant pathogens were further introduced throughout the hospital stay, driven by the proliferation of ESBL-producing *E. coli*. Increased resistance and virulence factors, as well as an overall diminished microbiome diversity were observed within the sample set, particularly the 5DG. While differences between the SSG and 5DG lacked frequent statistical significance, and were likely outweighed by the effect of hospitalization [175], results outline that the latter is linked with strong adverse effects on the microbiome structure and should therefore be reconsidered throughout future treatments. The presented findings demonstrate local ARG accumulation and expansion across equine patients under the selective pressure of antibiotics within a hospital environment, representing an ongoing challenge for hygiene management and work-place safety, especially in regards to the spread of MDR and zoonotic pathogens [175]. This thesis illustrates a starting point for further research on antibiotic stewardship and hospital hygiene,

providing an expansive data set for metagenome-wide association studies and the identification of additional biomarkers.

While the development of resistances is driven by extensive antibiotic usage, avoiding antimicrobials in general is not feasible as many infections can solely be treated through these drugs [156]. This is one of the reasons why attention must be shifted into the development of new strategies to counteract the negative side effects of antibiotic administration [156]. The observed results illustrate a potentially beneficial relationship between high microbiome diversity and decreased counts of AMR and virulence-associated genes. This appears plausible based on early microbiome research indicating a beneficial effect of high diversity microbiota [3, 15]. Representing a healthy gut environment, increased microbial diversity has been described as a protective mechanism against pathogen and ARG overgrowth [19]. However, the identified correlation does not necessarily equal causation. This relationship may be the result of few highly abundant taxa which proliferate strongly within the diminished microbiome and also tend to carry resistance and virulence genes (such as *E. coli*). Furthermore, metagenomic studies are limited to the assessment of relative abundances of microbial communities, as their results do not allow insights into the absolute bacterial or resistance gene loads. Additional studies are therefore required to supplement the proposed findings and to characterize the underlying effects of both decreased microbiome diversity and increased resistance gene abundance. The usage of targeted microbiome therapy through probiotics and bacterial load estimation can enable additional insights into this complex relationship. Probiotics are already being utilized to reduce unfavourable effects on the gut microbiome across humans [150, 156, 176] and equines [31]. Given further research, these may present another supplement to increase antibiotic stewardship through a reduction of excessive PAP usage.

Translating the results of microbiome studies into clinical solutions presents an ongoing effort [92]. A single dosage of perioperative antibiotics combined with hospitalization can already yield long-term impacts on the gut microbiome, including diminished diversity and elevated ARG carriage. This complements the results of previous studies on the gut microbiome [163]. Importantly, using a reduced prophylaxis (SSG) compared to the current prolonged regimen (5DG) does not have a significant impact on neither the clinical outcome [174] nor does it have a

worse effect on the microbiome composition, as illustrated within this thesis. This is in line with previous research, demonstrating that a reduced PAP timeframe did not influence the clinical outcome of equine patients [175, 190]. The results indicate that a single-shot dosage is as effective as a conventional 5-day prophylaxis, while limiting negative impacts on microbiome and resistome of the equine patients. Since prolonged antibiotic prophylaxes lacks proof of superiority regarding patient outcomes [174], current protocols should be revised accordingly. The presented results further raise important questions in regards to the local spread of ARGs within healthcare settings, their prevention and the usage of metagenomics as a molecular tool for surveillance. As additional research is needed to assess the influence of the clinical setting on the gut microbiome and to establish differences past day 10, future studies should take the presented microbiome analysis results into account accordingly.

# 6 Bibliography

[1] A. E. Pérez-Cobas, L. Gomez-Valero, and C. Buchrieser, "Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses", *Microbial Genomics*, 6(8):mgen000409, 2020.

[2] E. Ruppé et al., "Prediction of the intestinal resistome by a three-dimensional structure-based method", *Nature Microbiology*, 4(1):112–123, 2019.

[3] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease", *Nature Reviews Genetics*, 13(4):260–270, 2012.

[4] R. K. Colwell et al., "Biodiversity: concepts, patterns, and measurement", *The Princeton Guide to Ecology*, 663:257–263, 2009.

[5] R. Knight et al., "Best practices for analysing microbiomes", *Nature Reviews Microbiology*, 16(7):410–422, 2018.

[6] F. Finotello et al., "Next-generation computational tools for interrogating cancer immunity", *Nature Reviews Genetics*, 20(12):724–746, 2019.

[7] S. Nemeghaire et al., "The ecological importance of the Staphylococcus sciuri species group as a reservoir for resistance and virulence genes", *Veterinary Microbiology*, 171(3-4):342–356, 2014.

[8] V. F. Lanza et al., "In-depth resistome analysis by targeted metagenomics", *Microbiome*, 6(1):11, 2018.

[9] B. Chomel, "Zoonoses", *Encyclopedia of Microbiology*, 1:820–829, 2009.

[10] A. Fleming, "On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae", *British Journal of Experimental Pathology*, 10(3):226–236, 1929.

[11] B. L. Ligon, "Penicillin: its discovery and early development", *Seminars in Pediatric Infectious Diseases*, 15(1):52–57, 2004.

[12] A. G. McArthur and G. D. Wright, "Bioinformatics of antimicrobial resistance in the age of molecular epidemiology", *Current Opinion in Microbiology*, 27:45–50, 2015.

[13] E. D. Brown and G. D. Wright, "Antibacterial drug discovery in the resistance era", *Nature*, 529(7586):336–343, 2016.

[14] R. A. Smith, N. M. M'ikanatha, and A. F. Read, "Antibiotic resistance: a primer and call to action", *Health Communication*, 30(3):309–314, 2015.

[15] C. Roose-Amsaleg and A. M. Laverman, "Do antibiotics have environmental side-effects? Impact of synthetic antibiotics on biogeochemical processes", *Environmental Science and Pollution Research*, 23(5):4000–4012, 2016.

[16] Y. Yang et al., "Metagenomic insights into the abundance and composition of resistance genes in aquatic environments: influence of stratification and geography", *Environment International*, 127:371–380, 2019.

[17] L. de Nies et al., "PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data", *Microbiome*, 9(1):49, 2021.

[18] M. Oh et al., "MetaCompare: a computational pipeline for prioritizing environmental resistome risk", *FEMS Microbiology Ecology*, 94(7):fiy079, 2018.

[19] S. Hernando-Amado et al., "Defining and combating antibiotic resistance from One Health and Global Health perspectives", *Nature Microbiology*, 4 (9):1432–1442, 2019.

[20] C. Pal et al., "The structure and diversity of human, animal and environmental resistomes", *Microbiome*, 4(1):54, 2016.

[21] E. Abraham and E. Chain, "An enzyme from bacteria able to destroy penicillin", *Reviews of Infectious Diseases*, 3713:837, 1940.

[22] V. A. De Abreu, J. Perdigão, and S. Almeida, "Metagenomic approaches to analyze antimicrobial resistance: an overview", *Frontiers in Genetics*, 11 :575592, 2021.

[23] C. Y.-p. Lo and N. Thomas, "The macrosecuritization of antimicrobial resistance in Asia", *Australian Journal of International Affairs*, 72(6):567–583, 2018.

[24] E. Tacconelli et al., "Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis", *The Lancet Infectious Diseases*, 18(3):318–327, 2018.

[25] P. Munk et al., "A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds", *Journal of Antimicrobial Chemotherapy*, 72(2):385–392, 2017.

[26] R. De, "Metagenomics: aid to combat antimicrobial resistance in diarrhea", *Gut Pathogens*, 11:47, 2019.

[27] A. Torres Ortiz et al., "Genomic signatures of pre-resistance in Mycobacterium tuberculosis", *Nature Communications*, 12(1):7312, 2021.

[28] Y. Hu et al., "The bacterial mobile resistome transfer network connecting the animal and human microbiomes", *Applied and Environmental Microbiology*, 82(22):6672–6681, 2016.

[29] C. J. Murray et al., "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis", *The Lancet*, 399(10325):629–655, 2022.

[30] E. A. Bancroft, "Antimicrobial resistance: it's not just for hospitals", *JAMA*, 298(15):1803–1804, 2007.

[31] A. Kauter et al., "The gut microbiome of horses: current research on equine enteral microbiota and future perspectives", *Animal Microbiome*, 1(1):14, 2019.

[32] S. Naushad et al., "Comprehensive phylogenetic analysis of bovine non-aureus Staphylococci species based on whole-genome sequencing", *Frontiers in Microbiology*, 7:1990, 2016.

[33] J. N. van Spijk, S. Schmitt, and A. Schoster, "Infections caused by multidrug-resistant bacteria in an equine hospital (2012-2015)", *Equine Veterinary Education*, 31(12):653–658, 2019.

[34] C. Hoang Quoc et al., "Carbapenemase genes and multidrug resistance of Acinetobacter baumannii: a cross sectional study of patients with pneumonia in southern Vietnam", *Antibiotics*, 8(3):148, 2019.

[35] A. F. Shorr, "Epidemiology and economic impact of meticillin-resistant Staphylococcus aureus: review and analysis of the literature", *PharmacoEconomics*, 25(9):751–768, 2007.

[36] J. Handelsman et al., "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products", *Chemistry & Biology*, 5(10):R245–R249, 1998.

[37] M. Touchon et al., "Phylogenetic background and habitat drive the genetic diversification of Escherichia coli", *PLoS Genetics*, 16(6):e1008866, 2020.

[38] M. W. Nachman and S. L. Crowell, "Estimate of the mutation rate per nucleotide in humans", *Genetics*, 156(1):297–304, 2000.

[39] I. Ezkurdia et al., "Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes", *Human Molecular Genetics*, 23(22):5866–5878, 2014.

[40] M. Ghalayini et al., "Evolution of a dominant natural isolate of Escherichia coli in the human gut over the course of a year suggests a neutral evolution with reduced effective population size", *Applied and Environmental Microbiology*, 84(6):e02377–17, 2018.

[41] A. Scally, "The mutation rate in human evolution and demographic inference", *Current Opinion in Genetics & Development*, 41:36–43, 2016.

[42] B. Gibson et al., "The distribution of bacterial doubling times in the wild", *Proceedings of the Royal Society B*, 285(1880):20180789, 2018.

[43] F. Berglund et al., "Identification and reconstruction of novel antibiotic resistance genes from metagenomes", *Microbiome*, 7(1):52, 2019.

[44] G. Arango-Argoty et al., "DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data", *Microbiome*, 6(1): 23, 2018.

[45] Y. Yang et al., "ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database", *Bioinformatics*, 32(15):2346–2351, 2016.

[46] E. Doster et al., "MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data", *Nucleic Acids Research*, 48(D1):D561–D569, 2020.

[47] C. Munck et al., "Limited dissemination of the wastewater treatment plant core resistome", *Nature Communications*, 6:8452, 2015.

[48] B. Liu and M. Pop, "ARDB-antibiotic resistance genes database", *Nucleic Acids Research*, 37(D1):D443–D447, 2009.

[49] L. Gerzova et al., "Characterization of microbiota composition and presence of selected antibiotic resistance genes in carriage water of ornamental fish", *PLoS One*, 9(8):e103865, 2014.

[50] J. Guo et al., "Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements", *Water Research*, 123:468–478, 2017.

[51] A. Karkman, K. Pärnänen, and D. Larsson, "Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments", *Nature Communications*, 10(1):80, 2019.

[52] V. Guernier-Cambert et al., "Diverse bacterial resistance genes detected in fecal samples from clinically healthy women and infants in Australia-a descriptive pilot study", *Frontiers in Microbiology*, 12:596984, 2021.

[53] A. S. R. Duarte et al., "Addressing learning needs on the use of metagenomics in antimicrobial resistance surveillance", *Frontiers in Public Health*, 8:38, 2020.

[54] S. Ben Maamar et al., "Mobilizable antibiotic resistance genes are present in dust microbial communities", *PLoS Pathogens*, 16(1):e1008211, 2020.

[55] G. Steinkraus, R. White, and L. Friedrich, "Vancomycin MIC creep in non-vancomycin-intermediate Staphylococcus aureus (VISA), vancomycin-susceptible clinical methicillin-resistant S. aureus (MRSA) blood isolates from 2001-05", *Journal of Antimicrobial Chemotherapy*, 60(4):788–794, 2007.

[56] C. L. Gupta et al., "Longitudinal study on the effects of growth-promoting and therapeutic antibiotics on the dynamics of chicken cloacal and litter microbiomes and resistomes", *Microbiome*, 9(1):178, 2021.

[57] Centers for Disease Control and Prevention, "Antibiotic resistance threats in the United States", *U.S. Department of Health and Human Services*, 2013.

[58] D. Wallinga et al., "A review of the effectiveness of current US policies on antimicrobial use in meat and poultry production", *Current Environmental Health Reports*, 9(2):339–354, 2022.

[59] P. Liu et al., "Different impacts of manure and chemical fertilizers on bacterial community structure and antibiotic resistance genes in arable soils", *Chemosphere*, 188:455–464, 2017.

[60] R. S. Hendriksen et al., "Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage", *Nature Communications*, 10(1):1124, 2019.

[61] S. M. Lakin et al., "MEGARes: an antimicrobial resistance database for high throughput sequencing", *Nucleic Acids Research*, 45(D1):D574–D580, 2017.

[62] WHO Regional Office for Europe/European Centre for Disease Prevention and Control, "Antimicrobial resistance surveillance in Europe 2022 - 2020 data", *Copenhagen: WHO Regional Office for Europe*, 2022.

[63] S. Scheithauer et al., "Etablierung der Genomischen Erreger-Surveillance zur Stärkung des Pandemie-und Infektionsschutzes in Deutschland", *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 66(4):443–449, 2023.

[64] F. S. Codjoe and E. S. Donkor, "Carbapenem resistance: a review", *Medical Sciences*, 6(1):1, 2017.

[65] H. S. Gweon et al., "The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples", *Environmental Microbiome*, 14(1):7, 2019.

[66] V. D. Andersen et al., "Robustness in quantifying the abundance of antimicrobial resistance genes in pooled faeces samples from batches of slaughter pigs using metagenomics analysis", *Journal of Global Antimicrobial Resistance*, 24:398–402, 2021.

[67] V. Bortolaia et al., "ResFinder 4.0 for predictions of phenotypes from genotypes", *Journal of Antimicrobial Chemotherapy*, 75(12):3491–3500, 2020.

[68] M. Scholz et al., "Strain-level microbial epidemiology and population genomics from shotgun metagenomics", *Nature Methods*, 13(5):435–438, 2016.

[69] T. Dickerson et al., "Metagenomic profiling and identification of antimicrobial resistance genes from airborne microbial communities", *Online Journal of Public Health Informatics*, 7(1):e18, 2015.

[70] J. A. Gilbert and C. L. Dupont, "Microbial metagenomics: beyond the genome", *Annual Review of Marine Science*, 3:347–371, 2011.

[71] M. H. Norouzi-Beirami et al., "CAMAMED: a pipeline for composition-aware mapping-based analysis of metagenomic data", *NAR Genomics and Bioinformatics*, 3(1):lqaa107, 2021.

[72] P. Hugenholtz and G. W. Tyson, "Metagenomics", *Nature*, 455:481–483, 2008.

[73] V. Kunin et al., "A bioinformatician's guide to metagenomics", *Microbiology and Molecular Biology Reviews*, 72(4):557–578, 2008.

[74] S. Y. Ho and S. Duchêne, "Dating the emergence of human pathogens", *Science*, 368(6497):1310–1311, 2020.

[75] Q. Xu et al., "Microbial generalists and specialists differently contribute to the community diversity in farmland soils", *Journal of Advanced Research*, 40:17–27, 2022.

[76] H.-X. Chang et al., "Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity", *Frontiers in Microbiology*, 8:519, 2017.

[77] M. F. Dias et al., "Exploring the resistome, virulome and microbiome of drinking water in environmental and clinical settings", *Water Research*, 174 :115630, 2020.

[78] A. S. R. Duarte et al., "Metagenomics-based approach to source-attribution of antimicrobial resistance determinants-identification of reservoir resistome signatures", *Frontiers in Microbiology*, 11:601407, 2021.

[79] K. R. Chng et al., "Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment", *Nature Medicine*, 26 (6):941–951, 2020.

[80] C. Quince et al., "Extensive modulation of the fecal metagenome in children with Crohn's disease during exclusive enteral nutrition", *The American Journal of Gastroenterology*, 110(12):1718–1729, 2015.

[81] R. De, A. K. Mukhopadhyay, and S. Dutta, "Metagenomic analysis of gut microbiome and resistome of diarrheal fecal samples from Kolkata, India, reveals the core and variable microbiota including signatures of microbial dark matter", *Gut Pathogens*, 12:32, 2020.

[82] A. L. Lind and K. S. Pollard, "Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing", *Microbiome*, 9 (1):58, 2021.

[83] Y.-Z. Zhang, M. Shi, and E. C. Holmes, "Using metagenomics to characterize an expanding virosphere", *Cell*, 172(6):1168–1172, 2018.

[84] M. Shi, Y.-Z. Zhang, and E. C. Holmes, "Meta-transcriptomics and the evolutionary biology of RNA viruses", *Virus Research*, 243:83–90, 2018.

[85] P. J. Turnbaugh et al., "The human microbiome project", *Nature*, 449 (7164):804–810, 2007.

[86] D. B. Rusch et al., "The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific", *PLoS Biology*, 5(3): e77, 2007.

[87] D. Danko et al., "A global metagenomic map of urban microbiomes and antimicrobial resistance", *Cell*, 184(13):3376–3393, 2021.

[88] K. Kang et al., "The environmental exposures and inner- and intercity traffic flows of the metro system may contribute to the skin microbiome and resistome", *Cell Reports*, 24(5):1190–1202, 2018.

[89] F. P. Breitwieser and S. L. Salzberg, "Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification", *Bioinformatics*, 36(4):1303–1304, 2020.

[90] J. Vollmers, S. Wiegand, and A.-K. Kaster, "Comparing and evaluating metagenome assembly tools from a microbiologist's perspective-not only size matters!", *PLoS One*, 12(1):e0169662, 2017.

[91] N. Segata et al., "Metagenomic microbial community profiling using unique clade-specific marker genes", *Nature Methods*, 9(8):811–814, 2012.

[92] I. Moreno-Indias et al., "Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions", *Frontiers in Microbiology*, 12:635781, 2021.

[93] M. Ayling, M. D. Clark, and R. M. Leggett, "New approaches for metagenome assembly with short reads", *Briefings in Bioinformatics*, 21(2):584–594, 2020.

[94] G. Gautreau et al., "PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph", *PLoS Computational Biology*, 16(3):e1007732, 2020.

[95] S. Krakau et al., "nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning", *NAR Genomics and Bioinformatics*, 4(1): lqac007, 2022.

[96] A. Milanese et al., "Microbial abundance, activity and population genomic profiling with mOTUs2", *Nature Communications*, 10(1):1014, 2019.

[97] E. Afshinnekoo et al., "Geospatial resolution of human and bacterial diversity with city-scale metagenomics", *Cell Systems*, 1(1):72–87, 2015.

[98] S.-C. Wu et al., "Metagenome analysis as a tool to study bacterial infection associated with acute surgical abdomen", *Journal of Clinical Medicine*, 7 (10):346, 2018.

[99] N. D. Grubaugh et al., "An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar", *Genome Biology*, 20(1):8, 2019.

[100] J. Qin et al., "A metagenome-wide association study of gut microbiota in type 2 diabetes", *Nature*, 490(7418):55–60, 2012.

[101] B. Rimal and A. D. Patterson, "A centenarian entourage of bile acids and gut bacteria", *Nature*, 599(7885):380–381, 2021.

[102] A. Sundquist et al., "Bacterial flora-typing with targeted, chip-based pyrosequencing", *BMC Microbiology*, 7:108, 2007.

[103] A. D. Rowan-Nash et al., "Antimicrobial resistance gene prevalence in a population of patients with advanced dementia is related to specific pathobionts", *iScience*, 23(3):100905, 2020.

[104] J. Lu et al., "Bracken: estimating species abundance in metagenomics data", *PeerJ Computer Science*, 3:e104, 2017.

[105] Q. Zhou, X. Su, and K. Ning, "Assessment of quality control approaches for metagenomic data analysis", *Scientific Reports*, 4(1):1–11, 2014.

[106] H. Y. Simon et al., "Benchmarking metagenomics tools for taxonomic classification", *Cell*, 178(4):779–794, 2019.

[107] S. Lindgreen, K. L. Adair, and P. P. Gardner, "An evaluation of the accuracy and speed of metagenome analysis tools", *Scientific Reports*, 6:19233, 2016.

[108] S. Kieser et al., "ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data", *BMC Bioinformatics*, 21(1):257, 2020.

[109] S. Nurk et al., "metaSPAdes: a new versatile metagenomic assembler", *Genome Research*, 27(5):824–834, 2017.

[110] T. H. Dadi et al., "SLIMM: species level identification of microorganisms from metagenomes", *PeerJ*, 5:e3138, 2017.

[111] F. P. Breitwieser, J. Lu, and S. L. Salzberg, "A review of methods and databases for metagenomic classification and assembly", *Briefings in Bioinformatics*, 20(4):1125–1136, 2019.

[112] R. Ranjan et al., "Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing", *Biochemical and Biophysical Research Communications*, 469(4):967–977, 2016.

[113] K. Banerjee et al., "An adaptive multivariate two-sample test with application to microbiome differential abundance analysis", *Frontiers in Genetics*, 10:350, 2019.

[114] R. Bharti and D. G. Grimm, "Current challenges and best-practice protocols for microbiome analysis", *Briefings in Bioinformatics*, 22(1):178–193, 2021.

[115] J. Mas-Lloret et al., "Gut microbiome diversity detected by high-coverage 16S and shotgun sequencing of paired stool and colon sample", *Scientific Data*, 7(1):92, 2020.

[116] J. Lu and S. L. Salzberg, "Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2", *Microbiome*, 8(1):124, 2020.

[117] M. S. Beaudry et al., "Improved microbial community characterization of 16S rRNA via metagenome hybridization capture enrichment", *Frontiers in Microbiology*, 12:644662, 2021.

[118] M. L. Calle, "Statistical analysis of metagenomics data", *Genomics & Informatics*, 17(1):e6, 2019.

[119] Y. S. Bukin et al., "The effect of 16S rRNA region choice on bacterial community metabarcoding results", *Scientific Data*, 6:190007, 2019.

[120] R. J. Case et al., "Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies", *Applied and Environmental Microbiology*, 73 (1):278–288, 2007.

[121] T. A. K. Freitas et al., "Accurate read-based metagenome characterization using a hierarchical suite of unique signatures", *Nucleic Acids Research*, 43 (10):e69, 2015.

[122] N. J. Ames et al., "The human microbiome and understanding the 16S rRNA gene in translational nursing science", *Nursing Research*, 66(2):184–197, 2017.

[123] T. Rochegüe et al., "Impact of antibiotic therapies on resistance genes dynamic and composition of the animal gut microbiota", *Animals*, 11(11): 3280, 2021.

[124] S. Chen et al., "fastp: an ultra-fast all-in-one FASTQ preprocessor", *Bioinformatics*, 34(17):i884–i890, 2018.

[125] T. Wang et al., "The Human Pangenome Project: a global resource to map genomic diversity", *Nature*, 604(7906):437–446, 2022.

[126] Q. Wang et al., "Single-cell transcriptome sequencing on the Nanopore platform with ScNapBar", *RNA*, 27(7):763–770, 2021.

[127] A. Amir et al., "Deblur rapidly resolves single-nucleotide community sequence patterns", *mSystems*, 2(2):e00191–16, 2017.

[128] H. Mbareche et al., "An overview of bioinformatics tools for DNA metabarcoding analysis of microbial communities of bioaerosols: digest for microbiologists", *Life*, 10(9):185, 2020.

[129] B. J. Kelly et al., "Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA", *Bioinformatics*, 31(15):2461–2468, 2015.

[130] F. Cottier et al., "Advantages of meta-total RNA sequencing (MeTRS) over shotgun metagenomics and amplicon-based sequencing in the profiling of complex microbial communities", *npj Biofilms and Microbiomes*, 4:2, 2018.

[131] M. R. McLaren, A. D. Willis, and B. J. Callahan, "Consistent and correctable bias in metagenomic sequencing experiments", *eLife*, 8:e46923, 2019.

[132] G. Arango-Argoty et al., "MetaStorm: a public resource for customizable metagenomics annotation", *PLoS One*, 11(9):e0162442, 2016.

[133] F. Plaza Oñate et al., "MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data", *Bioinformatics*, 35 (9):1544–1552, 2019.

[134] S. K. Rajan et al., "Phylogenetic microbiota profiling in fecal samples depends on combination of sequencing depth and choice of NGS analysis method", *PLoS One*, 14(9):e0222171, 2019.

[135] A. T. Reese and R. R. Dunn, "Drivers of microbiome biodiversity: a review of general rules, feces, and ignorance", *mBio*, 9(4):e01294–18, 2018.

[136] A. Brady and S. L. Salzberg, "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models", *Nature Methods*, 6(9):673–676, 2009.

[137] S. Gupta et al., "Identification of discriminatory antibiotic resistance genes among environmental resistomes using extremely randomized tree algorithm", *Microbiome*, 7(1):123, 2019.

[138] S. Das et al., "Natural mutations in a Staphylococcus aureus virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation", *PNAS*, 113(22):E3101–E3110, 2016.

[139] C. Stathopoulos et al., "Secretion of virulence determinants by the general secretory pathway in gram-negative pathogens: an evolving story", *Microbes and Infection*, 2(9):1061–1072, 2000.

[140] Q. Liu et al., "Type I fimbriae subunit fimA enhances Escherichia coli biofilm formation but affects L-threonine carbon distribution", *Frontiers in Bioengineering and Biotechnology*, 10:904636, 2022.

[141] I. B. Slizovskiy et al., "Mobilization of antibiotic resistance: are current approaches for colocalizing resistomes and mobilomes useful?", *Frontiers in Microbiology*, 11:1376, 2020.

[142] P. S. Krawczyk, L. Lipinski, and A. Dziembowski, "PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures", *Nucleic Acids Research*, 46(6):e35, 2018.

[143] Z. Fang et al., "PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning", *GigaScience*, 8(6):giz066, 2019.

[144] G. Liu et al., "Antibiotic-induced, increased conjugative transfer is common to diverse naturally occurring ESBL plasmids in Escherichia coli", *Frontiers in Microbiology*, 10:2119, 2019.

[145] A. Kothari et al., "Large circular plasmids from groundwater plasmidomes span multiple incompatibility groups and are enriched in multimetal resistance genes", *mBio*, 10(1):e02899–18, 2019.

[146] L. Xu et al., "The effect of antibiotics on the gut microbiome: a metagenomics analysis of microbial shift and gut antibiotic resistance in antibiotic treated mice", *BMC Genomics*, 21(1):263, 2020.

[147] N. LaPierre et al., "MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction", *Methods*, 166: 74–82, 2019.

[148] C. Zhong et al., "Integrating pan-genome with metagenome for microbial community profiling", *Computational and Structural Biotechnology Journal*, 19:1458–1466, 2021.

[149] M. Shahab and N. Shahab, "Coevolution of the human host and gut microbiome: metagenomics of microbiota", *Cureus*, 14(6):e26310, 2022.

[150] A.-K. Lederer et al., "Postoperative changes of the microbiome: are surgical complications related to the gut flora? A systematic review", *BMC Surgery*, 17(1):125, 2017.

[151] C.-H. Chen et al., "Diversity of nasal microbiota and its interaction with surface microbiota among residents in healthcare institutes", *Scientific Reports*, 9(1):6175, 2019.

[152] R. Sender, S. Fuchs, and R. Milo, "Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans", *Cell*, 164(3):337–340, 2016.

[153] K. Dougal et al., "Characterisation of the faecal bacterial community in adult and elderly horses fed a high fibre, high oil or high starch diet using 454 pyrosequencing", *PLoS One*, 9(2):e87424, 2014.

[154] N. Mach et al., "Priming for welfare: gut microbiota is associated with equitation conditions and behavior in horse athletes", *Scientific Reports*, 10(1):8311, 2020.

[155] M. Reyman et al., "Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life", *Nature Communications*, 10 (1):4997, 2019.

[156] R. K. Dudek-Wicher, A. Junka, and M. Bartoszewicz, "The influence of antibiotics and dietary components on gut microbiota", *Gastroenterology Review*, 13(2):85–92, 2018.

[157] E. Venable et al., "Role of the gut microbiota in equine health and disease", *Animal Frontiers*, 6(3):43–49, 2016.

[158] B. K. Kuntal et al., "'NetShift': a methodology for understanding 'driver microbes' from healthy and disease microbiome datasets", *The ISME Journal*, 13(2):442–454, 2019.

[159] O. Manor and E. Borenstein, "Revised computational metagenomic processing uncovers hidden and biologically meaningful functional variation in the human microbiome", *Microbiome*, 5(1):19, 2017.

[160] Z. Liu et al., "Acute sleep-wake cycle shift results in community alteration of human gut microbiome", *mSphere*, 5(1):e00914–19, 2020.

[161] A. Curry, "Ancient poop reveals extinction in gut bacteria", *Science*, 372 (6543):671, 2021.

[162] J. Kaminski et al., "High-specificity targeted functional profiling in microbial communities with ShortBRED", *PLoS Computational Biology*, 11(12): e1004557, 2015.

[163] H. Nalluri et al., "Peri-operative antibiotics acutely and significantly impact intestinal microbiota following bariatric surgery", *Scientific Reports*, 10(1): 20340, 2020.

[164] F. Lindenberg et al., "Development of the equine gut microbiota", *Scientific Reports*, 9(1):14427, 2019.

[165] D. E. Wilson and D. M. Reeder, "Mammal species of the world: a taxonomic and geographic reference", *Johns Hopkins University Press*, 1:630–631, 2005.

[166] K. J. Kelly, L. A. McDuffee, and K. Mears, "The effect of human-horse interactions on equine behaviour, physiology, and welfare: a scoping review", *Animals*, 11(10):2782, 2021.

[167] R. Gilroy et al., "Metagenomic investigation of the equine faecal microbiome reveals extensive taxonomic diversity", *PeerJ*, 10:e13084, 2022.

[168] M. Costa et al., "Characterization and comparison of the bacterial microbiota in different gastrointestinal tract compartments in horses", *The Veterinary Journal*, 205(1):74–80, 2015.

[169] J. S. Weese et al., "Changes in the faecal microbiota of mares precede the development of post partum colic", *Equine Veterinary Journal*, 47(6):641–649, 2015.

[170] M. Lindefeldt et al., "The ketogenic diet influences taxonomic and functional composition of the gut microbiota in children with severe epilepsy", *npj Biofilms and Microbiomes*, 5(1):5, 2019.

[171] L. Ang et al., "Gut microbiome characteristics in feral and domesticated horses from different geographic locations", *Communications Biology*, 5(1): 172, 2022.

[172] A. Bandla et al., "910 metagenome-assembled genomes from the phytobiomes of three urban-farmed leafy Asian greens", *Scientific Data*, 7(1): 278, 2020.

[173] N. J. Gotelli and R. K. Colwell, "Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness", *Ecology Letters*, 4:379–391, 2001.

[174] S. D. Stöckle et al., "A pilot randomised clinical trial comparing a short-term perioperative prophylaxis regimen to a long-term standard protocol in equine colic surgery", *Antibiotics*, 10(5):587, 2021.

[175] A. Kauter et al., "Frequency, local dynamics, and genomic characteristics of ESBL-producing Escherichia coli isolated from specimens of hospitalized horses", *Frontiers in Microbiology*, 12:671676, 2021.

[176] A. Ravi et al., "Loss of microbial diversity and pathogen domination of the gut microbiota in critically ill patients", *Microbial Genomics*, 5(9):e000293, 2019.

[177] C. Ierano et al., "Appropriateness of surgical antimicrobial prophylaxis practices in Australia", *JAMA Network Open*, 2(11):e1915003, 2019.

[178] S. S. Magill et al., "Prevalence of antimicrobial use in US acute care hospitals, May-September 2011", *JAMA*, 312(14):1438–1446, 2014.

[179] S. K. Reed et al., "Effects of phenylbutazone alone or in combination with flunixin meglumine on blood protein concentrations in horses", *American Journal of Veterinary Research*, 67(3):398–402, 2006.

[180] C. Arnold et al., "Alterations in the fecal microbiome and metabolome of horses with antimicrobial-associated diarrhea compared to antibiotic-treated and non-treated healthy case controls", *Animals*, 11(6):1807, 2021.

[181] M. C. Costa et al., "Changes in the equine fecal microbiota associated with the use of systemic antimicrobial drugs", *BMC Veterinary Research*, 11:19, 2015.

[182] H. Yakkala et al., "Comparative genome analysis reveals niche-specific genome expansion in Acinetobacter baumannii strains", *PLoS One*, 14(6): e0218204, 2019.

[183] H. Lin and S. D. Peddada, "Analysis of microbial compositions: a review of normalization and differential abundance analysis", *npj Biofilms and Microbiomes*, 6(1):60, 2020.

[184] G. Fillion-Bertrand et al., "Lung microbiome is influenced by the environment and asthmatic status in an equine model of asthma", *American Journal of Respiratory Cell and Molecular Biology*, 60(2):189–197, 2019.

[185] R. A. Quinn et al., "Biogeochemical forces shape the composition and physiology of polymicrobial communities in the cystic fibrosis lung", *mBio*, 5 (2):e00956–13, 2014.

[186] P. Vonaesch et al., "Stunted childhood growth is associated with decompartmentalization of the gastrointestinal tract and overgrowth of oropharyngeal taxa", *PNAS*, 115(36):E8489–E8498, 2018.

[187] S. A. Kim et al., "Global gene expression patterns and induction of innate immune response in human laryngeal epithelial cells in response to Acinetobacter baumannii outer membrane protein A", *FEMS Immunology & Medical Microbiology*, 54(1):45–52, 2008.

[188] G. Wilharm et al., "Relatedness of wildlife and livestock avian isolates of the nosocomial pathogen Acinetobacter baumannii to lineages spread in hospitals worldwide", *Environmental Microbiology*, 19(10):4349–4364, 2017.

[189] P. Klotz et al., "Seasonal occurrence and carbapenem susceptibility of bovine Acinetobacter baumannii in Germany", *Frontiers in Microbiology*, 10:272, 2019.

[190] A. Kauter et al., "Antibiotic prophylaxis and hospitalization of horses subjected to median laparotomy: gut microbiota trajectories and abundance increase of Escherichia", *bioRxiv*, 2023.05.24.542119, 2023.

[191] R. D. Stewart et al., "Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen", *Nature Communications*, 9(1):870, 2018.

[192] R. Gilroy et al., "Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture", *PeerJ*, 9:e10941, 2021.

[193] B. J. Callahan et al., "DADA2: high-resolution sample inference from Illumina amplicon data", *Nature Methods*, 13(7):581–583, 2016.

[194] M. Achtman and Z. Zhou, "Metagenomics of the modern and historical human oral microbiome with phylogenetic studies on Streptococcus mutans and Streptococcus sobrinus", *Philosophical Transactions of the Royal Society B*, 375(1812):20190573, 2020.

[195] T. Z. DeSantis et al., "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB", *Applied and Environmental Microbiology*, 72(7):5069–5072, 2006.

[196] J. R. Cole et al., "Ribosomal Database Project: data and tools for high throughput rRNA analysis", *Nucleic Acids Research*, 42(D1):D633–D642, 2014.

[197] G. Henderson et al., "Improved taxonomic assignment of rumen bacterial 16S rRNA sequences using a revised SILVA taxonomic framework", *PeerJ*, 7:e6496, 2019.

[198] A. Dhariwal et al., "ResistoXplorer: a web-based tool for visual, statistical and exploratory data analysis of resistome data", *NAR Genomics and Bioinformatics*, 3(1):lqab018, 2021.

[199] D. Jiang et al., "Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities", *Frontiers in Genetics*, 10:995, 2019.

[200] M. B. Pereira et al., "Comparison of normalization methods for the analysis of metagenomic gene abundance data", *BMC Genomics*, 19(1):274, 2018.

[201] M. Vellend, "Do commonly used indices of β-diversity measure species turnover?", *Journal of Vegetation Science*, 12(4):545–552, 2001.

[202] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", *Bioinformatics*, 26(1):139–140, 2010.

[203] L. Jost, "The relation between evenness and diversity", *Diversity*, 2(2):207–232, 2010.

[204] R. H. Whittaker, "Vegetation of the Siskiyou mountains, Oregon and California", *Ecological Monographs*, 30(3):279–338, 1960.

[205] I. F. Spellerberg and P. J. Fedor, "A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener'Index", *Global Ecology and Biogeography*, 12(3): 177–179, 2003.

[206] E. H. Simpson, "Measurement of diversity", *Nature*, 163(4148):688, 1949.

[207] C. E. Shannon, "A mathematical theory of communication", *The Bell System Technical Journal*, 27(3):379–423, 1948.

[208] B. D. Wagner et al., "On the use of diversity measures in longitudinal sequencing studies of microbial communities", *Frontiers in Microbiology*, 9 :1037, 2018.

[209] J. R. Bray and J. T. Curtis, "An ordination of the upland forest communities of southern Wisconsin", *Ecological Monographs*, 27(4):326–349, 1957.

[210] S. Smith, "phylosmith: an R-package for reproducible and efficient microbiome analysis with phyloseq-objects", *Journal of Open Source Software*, 4 (38):1442, 2019.

[211] E. A. Dinsdale et al., "Multivariate analysis of functional metagenomes", *Frontiers in Genetics*, 4:41, 2013.

[212] F. Meyer et al., "MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis", *Briefings in Bioinformatics*, 20(4):1151–1159, 2019.

[213] R. Schmieder and R. Edwards, "Quality control and preprocessing of metagenomic datasets", *Bioinformatics*, 27(6):863–864, 2011.

[214] D. Morais et al., "MEDUSA: a pipeline for sensitive taxonomic classification and flexible functional annotation of metagenomic shotgun sequences", *Frontiers in Genetics*, 13:814437, 2022.

[215] K. D. Curry et al., "Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data", *Nature Methods*, 19(7):845–853, 2022.

[216] D. H. Parks et al., "Evaluation of the microba community profiler for taxonomic profiling of metagenomic datasets from the human gut microbiome", *Frontiers in Microbiology*, 12:643682, 2021.

[217] J. Tamames, M. Cobo-Simón, and F. Puente-Sánchez, "Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes", *BMC Genomics*, 20(1):960, 2019.

[218] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with Kraken 2", *Genome Biology*, 20(1):257, 2019.

[219] P. F. Harrison et al., "Topconfects: a package for confident effect sizes in differential expression analysis provides a more biologically useful ranked gene list", *Genome Biology*, 20(1):67, 2019.

[220] F. Beghini et al., "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3", *eLife*, 10:e65088, 2021.

[221] R. Van Damme et al., "Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN)", *PLoS Computational Biology*, 17(2):e1008716, 2021.

[222] R. Sajulga et al., "Survey of metaproteomics software tools for functional microbiome analysis", *PLoS One*, 15(11):e0241503, 2020.

[223] P. E. Compeau, P. A. Pevzner, and G. Tesler, "How to apply de Bruijn graphs to genome assembly", *Nature Biotechnology*, 29(11):987–991, 2011.

[224] M. R. Olm et al., "dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through dereplication", *The ISME Journal*, 11(12):2864–2868, 2017.

[225] D. Li et al., "MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices", *Methods*, 102 :3–11, 2016.

[226] R. Méheust et al., "Groundwater Elusimicrobia are metabolically diverse compared to gut microbiome Elusimicrobia and some have a novel nitrogenase paralog", *The ISME Journal*, 14(12):2907–2922, 2020.

[227] M. Montoliu-Nerin et al., "Building de novo reference genome assemblies of complex eukaryotic microorganisms from single nuclei", *Scientific Reports*, 10(1):1303, 2020.

[228] J. Tamames and F. Puente-Sánchez, "SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline", *Frontiers in Microbiology*, 9: 3349, 2019.

[229] R. M. Bowers et al., "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea", *Nature Biotechnology*, 35(8):725–731, 2017.

[230] M. Kolmogorov et al., "metaFlye: scalable long-read metagenome assembly using repeat graphs", *Nature Methods*, 17(11):1103–1110, 2020.

[231] A. Mikheenko, V. Saveliev, and A. Gurevich, "MetaQUAST: evaluation of metagenome assemblies", *Bioinformatics*, 32(7):1088–1090, 2016.

[232] A. Sczyrba et al., "Critical assessment of metagenome interpretation-a benchmark of metagenomics software", *Nature Methods*, 14(11):1063–1071, 2017.

[233] E. L. Moss, D. G. Maghini, and A. S. Bhatt, "Complete, closed bacterial genomes from microbiomes using nanopore sequencing", *Nature Biotechnology*, 38(6):701–707, 2020.

[234] D. D. Kang et al., "MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies", *PeerJ*, 7:e7359, 2019.

[235] A. Mu et al., "Reconstruction of the genomes of drug-resistant pathogens for outbreak investigation through metagenomic sequencing", *mSphere*, 4 (1):e00529–18, 2019.

[236] P.-A. Chaumeil et al., "GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database", *Bioinformatics*, 36(6):1925–1927, 2019.

[237] S. Nayfach et al., "A genomic catalog of earth's microbiomes", *Nature Biotechnology*, 39(4):499–509, 2021.

[238] B. Jia et al., "CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database", *Nucleic Acids Research*, 45 (D1):D566–D573, 2017.

[239] D. Pellow, I. Mizrahi, and R. Shamir, "PlasClass improves plasmid sequence classification", *PLoS Computational Biology*, 16(4):e1007781, 2020.

[240] S. Wu et al., "WebMGA: a customizable web server for fast metagenomic sequence analysis", *BMC Genomics*, 12:444, 2011.

[241] D. P. Depledge, I. Mohr, and A. C. Wilson, "Going the distance: optimizing RNA-Seq strategies for transcriptomic analysis of complex viral genomes", *Journal of Virology*, 93(1):e01342–18, 2018.

[242] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research", *Journal of Chiropractic Medicine*, 15(2):155–163, 2016.

[243] F. Mölder et al., "Sustainable data analysis with Snakemake", *F1000Research*, 10:33, 2021.

[244] E. Bolyen et al., "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2", *Nature Biotechnology*, 37(8):852–857, 2019.

[245] W. Shen et al., "SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation", *PLoS One*, 11(10):e0163962, 2016.

[246] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST", *Bioinformatics*, 26(19):2460–2461, 2010.

[247] A. Oren and G. M. Garrity, "Valid publication of the names of forty-two phyla of prokaryotes", *International Journal of Systematic and Evolutionary Microbiology*, 71(10):005056, 2021.

[248] P. J. McMurdie and S. Holmes, "phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data", *PLoS One*, 8(4):e61217, 2013.

[249] J. Reeder et al., "MicrobiomeExplorer: an R package for the analysis and visualization of microbial communities", *Bioinformatics*, 37(9):1317–1318, 2021.

[250] P. Ewels et al., "MultiQC: summarize analysis results for multiple tools and samples in a single report", *Bioinformatics*, 32(19):3047–3048, 2016.

[251] A. Telatin, P. Fariselli, and G. Birolo, "SeqFu: a suite of utilities for the robust and reproducible manipulation of sequence files", *Bioengineering*, 8 (5):59, 2021.

[252] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a web browser", *BMC Bioinformatics*, 12:385, 2011.

[253] M. Kokot, M. Długosz, and S. Deorowicz, "KMC 3: counting and manipulating k-mer statistics", *Bioinformatics*, 33(17):2759–2761, 2017.

[254] W. Gish and D. J. States, "Identification of protein coding regions by database similarity search", *Nature Genetics*, 3(3):266–272, 1993.

[255] D. Hyatt et al., "Prodigal: prokaryotic gene recognition and translation initiation site identification", *BMC Bioinformatics*, 11:119, 2010.

[256] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2", *Nature Methods*, 9(4):357–359, 2012.

[257] P. Danecek et al., "Twelve years of SAMtools and BCFtools", *GigaScience*, 10(2):giab008, 2021.

[258] A. M. Kozlov et al., "RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference", *Bioinformatics*, 35(21):4453–4455, 2019.

[259] I. Letunic and P. Bork, "Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation", *Nucleic Acids Research*, 49 (W1):W293–W296, 2021.

[260] L. Chen et al., "VFDB: a reference database for bacterial virulence factors", *Nucleic Acids Research*, 33(D1):D325–D328, 2005.

[261] J. Morales et al., "A joint NCBI and EMBL-EBI transcript set for clinical genomics and research", *Nature*, 604(7905):310–315, 2022.

[262] M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets", *Nature Biotechnology*, 35 (11):1026–1028, 2017.

[263] J. Ramirez et al., "Antibiotics as major disruptors of gut microbiota", *Frontiers in Cellular and Infection Microbiology*, 10:572912, 2020.

[264] C. E. Arnold et al., "The effects of signalment, diet, geographic location, season, and colitis associated with antimicrobial use or Salmonella infection on the fecal microbiome of horses", *Journal of Veterinary Internal Medicine*, 35(5):2437–2448, 2021.

[265] J. Barroso-Batista et al., "The first steps of adaptation of Escherichia coli to the gut are dominated by soft sweeps", *PLoS Genetics*, 10(3):e1004182, 2014.

[266] H. L. Stewart et al., "Changes in the faecal bacterial microbiota during hospitalisation of horses with colic and the effect of different causes of colic", *Equine Veterinary Journal*, 53(6):1119–1131, 2021.

[267] S. Sriswasdi, C.-c. Yang, and W. Iwasaki, "Generalist species drive microbial dispersion and evolution", *Nature Communications*, 8(1):1162, 2017.

[268] V. T. N. Bich et al., "Moderate and transient impact of antibiotic use on the gut microbiota in a rural Vietnamese cohort", *Scientific Reports*, 12: 20189, 2022.

[269] C. Ruppen et al., "Is penicillin plus gentamicin synergistic against clinical group B Streptococcus isolates?: an in vitro study", *Frontiers in Microbiology*, 7:1680, 2016.

[270] T. Bajpai et al., "Prevalence of TEM, SHV, and CTX-M beta-lactamase genes in the urinary isolates of a tertiary care hospital", *Avicenna Journal of Medicine*, 7(1):12–16, 2017.

# 7 List of Publications

A. Kauter, J. Brombach, A. Lübke-Becker, D. Kannapin, C. Bang, S. Franzenburg, S. D. Stoeckle, A. Mellmann, N. Effelsberg, R. Köck, S. Günther, L. H. Wieler, H. Gehlen, T. Semmler, S. A. Wolf, and B. Walther, "Antibiotic prophylaxis and hospitalization of horses subjected to median laparotomy: gut microbiota trajectories and abundance increase of Escherichia", *bioRxiv*, 2023.05.24.542119, 2023

A. Kauter, L. Epping, F. Ghazisaeedi, A. Lübke-Becker, S. A. Wolf, D. Kannapin, S. D. Stoeckle, T. Semmler, S. Günther, H. Gehlen, and B. Walther, "Frequency, local dynamics, and genomic characteristics of ESBL-producing Escherichia coli isolated from specimens of hospitalized horses", *Frontiers in Microbiology*, 12: 671676, 2021

# A Appendix

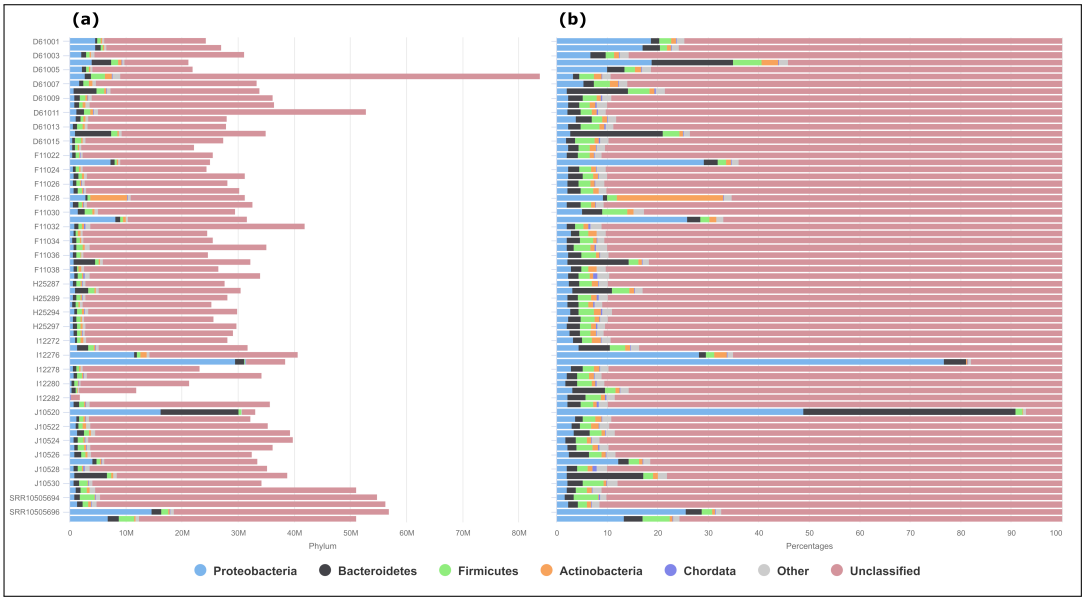


**Figure A.1: Taxonomic Quality Control (Gut Microbiome).** Visualized are stacked bar charts illustrating read counts for each sample of gut metagenome data set ($n = 68$) and coloured according to the abundance of the 5 major phyla. Minor ranks are grouped into 'other', while reads unable to be typed are represented by 'unclassified'. (a) Illustrated are the raw count abundances for the major taxonomic phyla across samples produced by Kraken2. Samples visibly differ in sequencing depth and taxonomic content, requiring additional normalization through rarefaction before further analyses can be conducted. (b) Relative abundances in percentages of the composition per sample. These were computed sample-wise by normalizing absolute counts through the corresponding sequencing depth. A majority of all reads cannot be directly assigned to a taxonomic rank (unclassified).

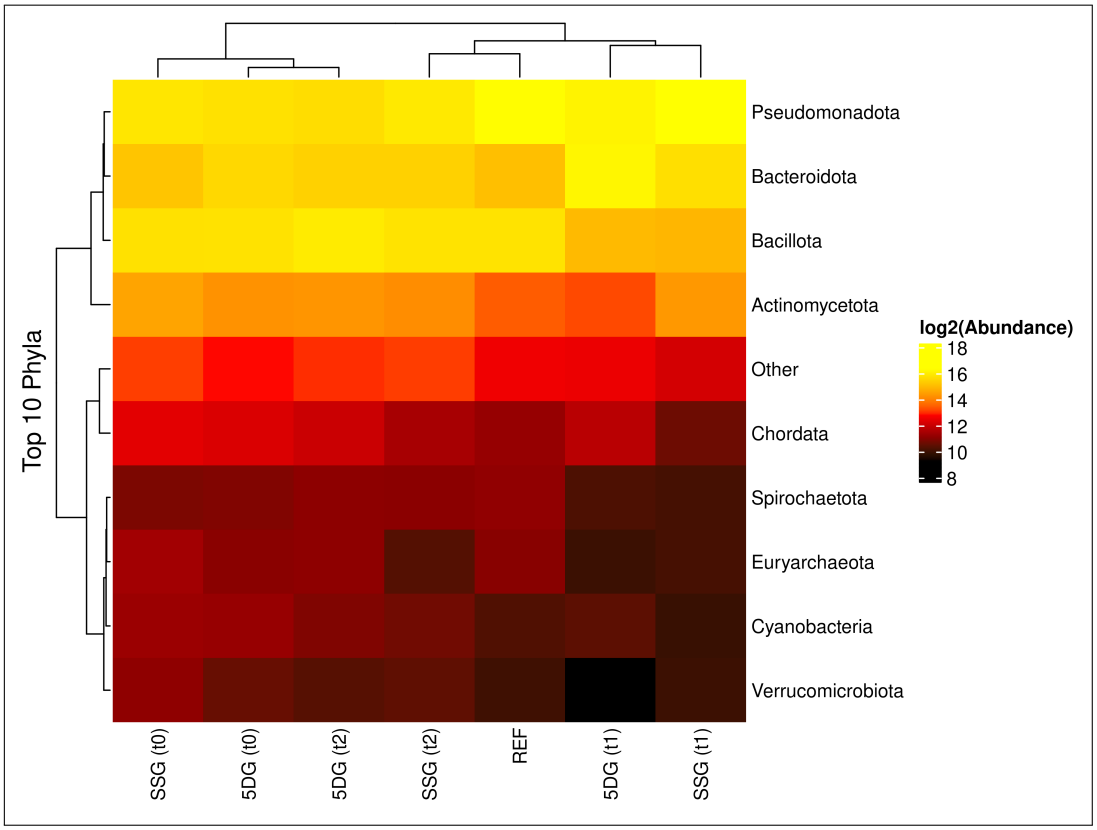


**Figure A.2: Clusters of Phylum-Level Abundances (Gut Microbiome).** Taxonomic profiles were summarized at phylum-level and averaged across the sample groups. The heatmap visualizes the abundance ($\log_2$ of normalized counts) from low (dark red) to high (yellow) across the data set. Columns represent the different study groups at specific time points. Rows illustrate the 9 most abundant phyla, with minor ranks being grouped into 'other'. Phyla are clustered by their similarity to each other in the represented abundance profiles (dendrogram on the left). Columns are additionally grouped by similarity across these (dendrogram on top). REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

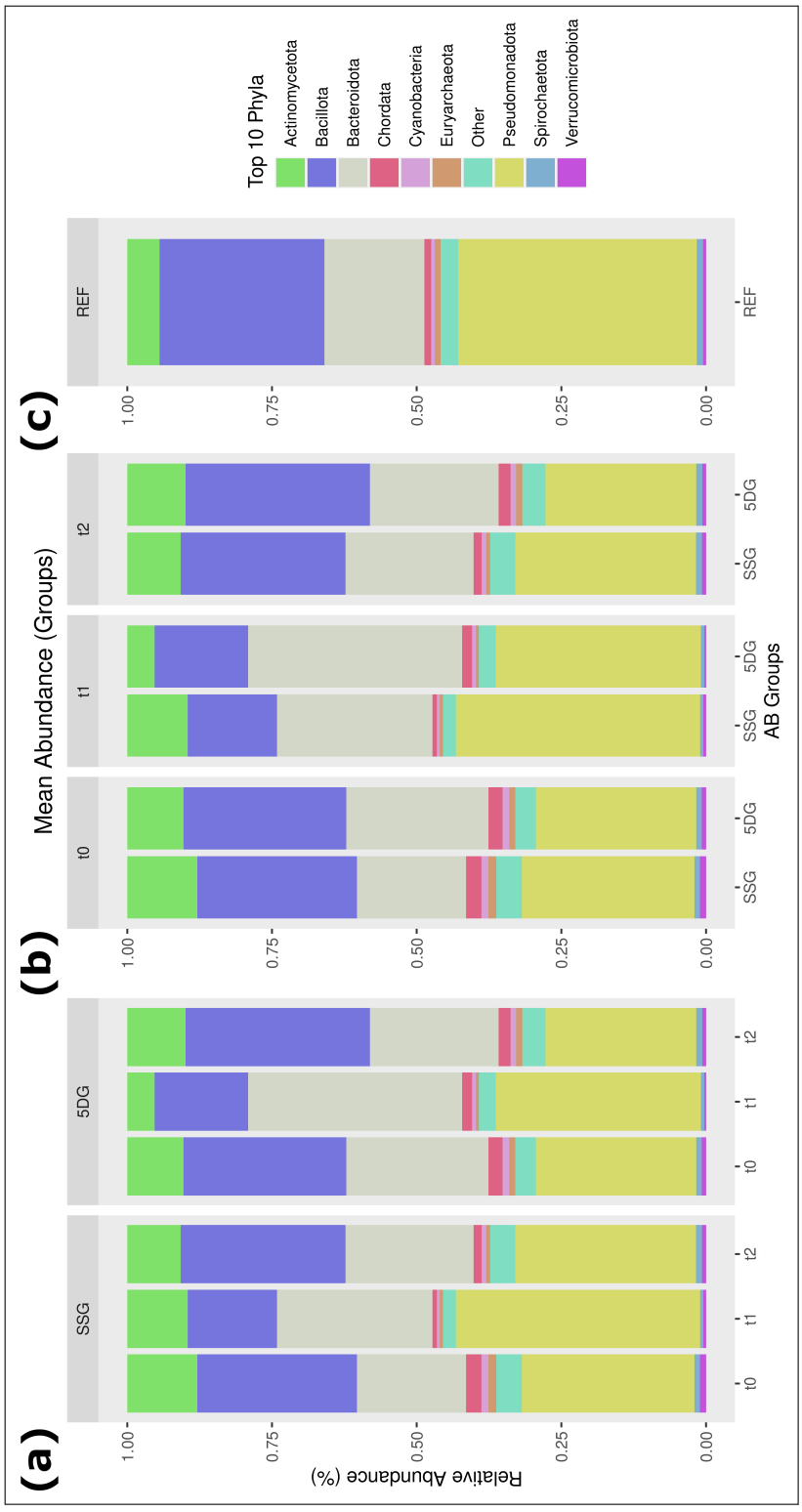


**Figure A.3: Summarized Bar Charts at Phylum-Level (Gut Microbiome).** Stacked bar charts representing the averaged composition of each sample group on the phylum-level. The 9 most abundant phyla are visualized as percentages in composition for (a) the prophylaxis groups, (b) across the time points and (c) for the NCBI reference set. Minor taxonomic ranks are grouped into 'other'. REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

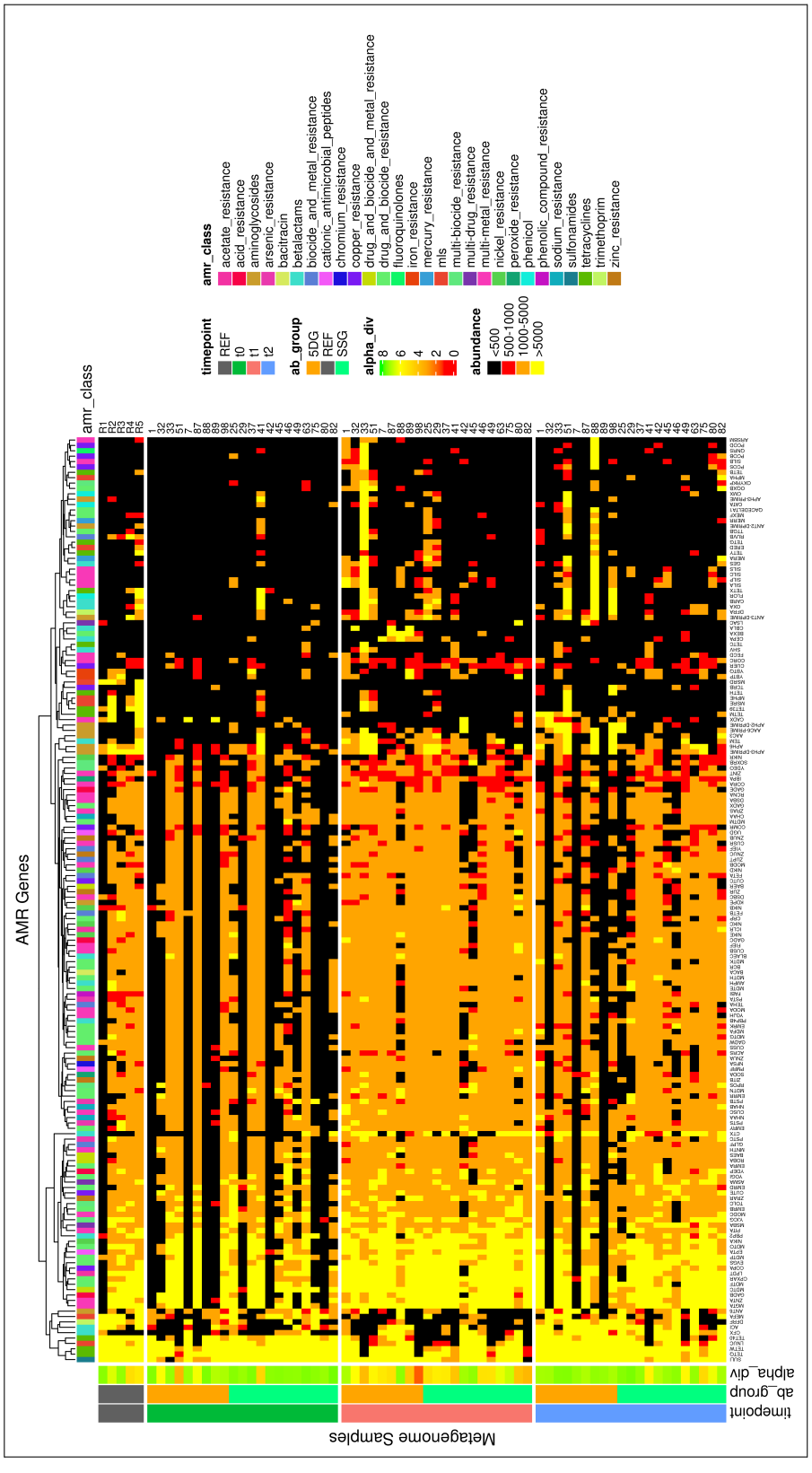


**Figure A.4: Low-Abundance Resistance Components (Gut Microbiome).** Heatmap illustrating the occurrence of resistance gene fragments which could not be reconstructed through metagenomic assembly. These include ARGs, as well as genes conferring resistance against biocides and heavy metals. Samples (right) are divided horizontally by groups and time points (left). Columns represent individual resistance elements (bottom) across different classes of components (top) and clustered by similarity in abundance profiles (dendrogram). The scale from black (less than 500 fragments) to yellow (over 5000 fragments) illustrates the overall abundance of resistance elements within the gut samples. Additional metadata, including alpha diversity and IDs of the equine patients are plotted alongside the heatmap (left and right). AMR: Antimicrobial Resistance, REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

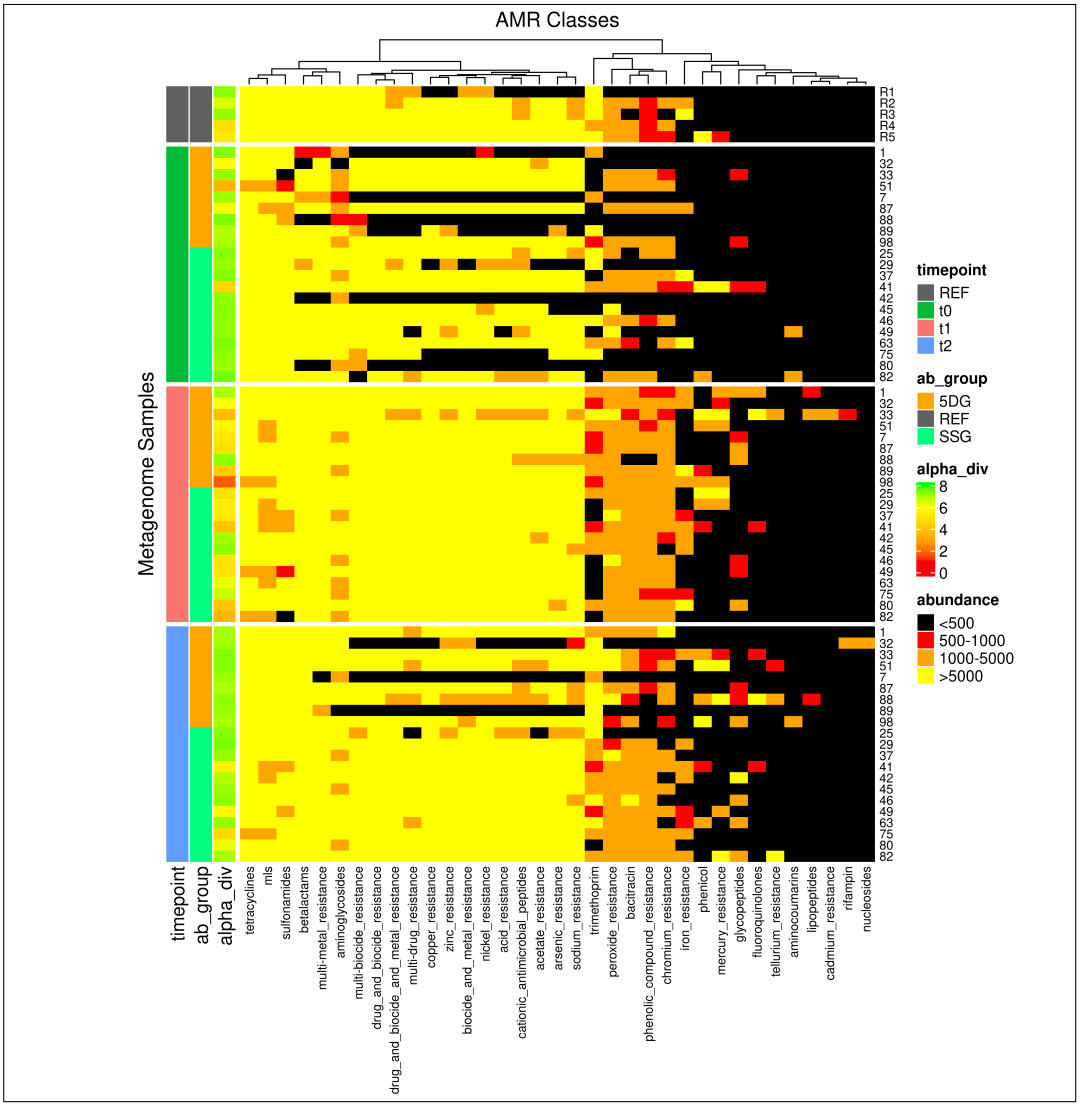


**Figure A.5: Resistance Class Abundances (Gut Microbiome).** Heatmap of resistance gene abundances, summarized to the level of drug classes. These include ARGs, as well as genes conferring resistance against biocides and heavy metals. Samples (right) are divided horizontally by groups and time points (left). Columns represent individual resistance classes (bottom), clustered by similarity in abundance profiles (dendrogram). The scale from black (less than 500 fragments) to yellow (over 5000 fragments) illustrates the overall abundance of resistance elements within the gut samples. Additional metadata, including alpha diversity and IDs of the equine patients are plotted alongside the heatmap (left and right). AMR: Antimicrobial Resistance, REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

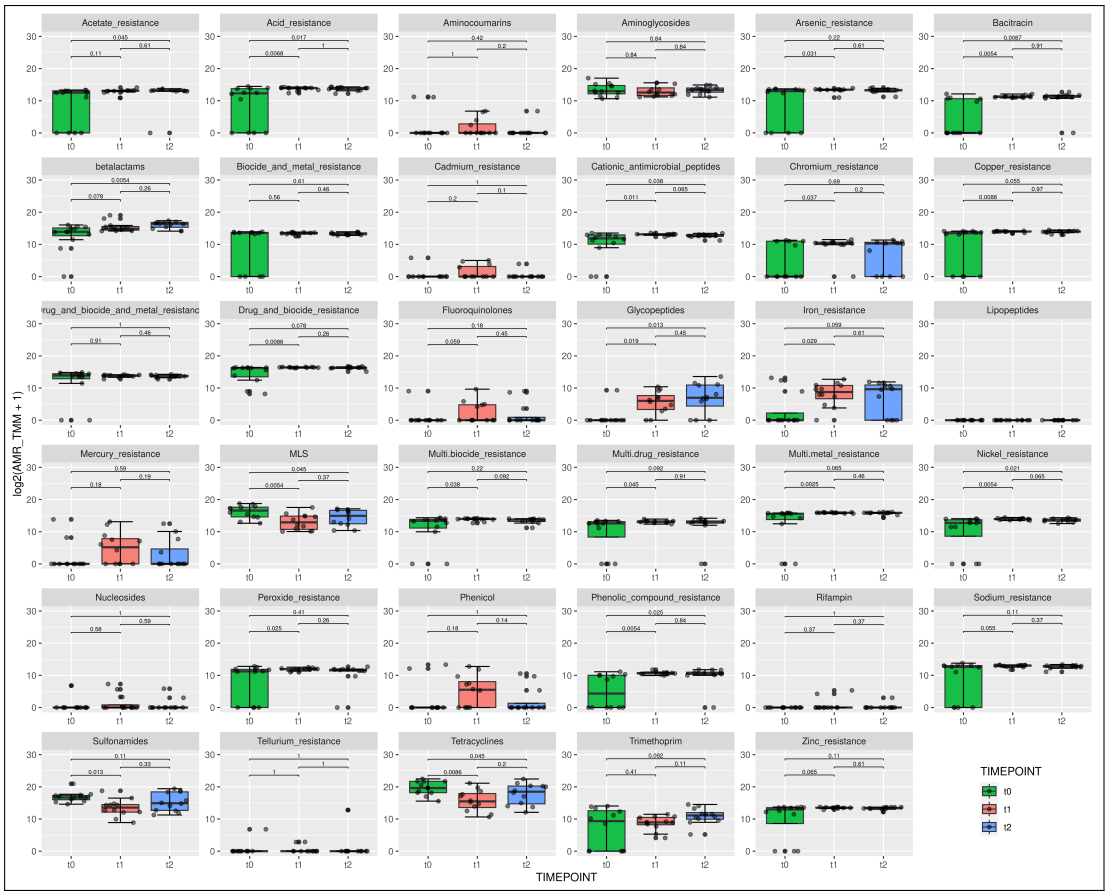


**Figure A.6: Resistance Classes SSG (Gut Microbiome).** Differential analysis of resistance gene classes across time for equine patients of the SSG ($n = 15$). Resistance gene abundances were summarized to class level and tested for significant changes over time (colours) per class (each box plot comparison). Wilcoxon rank-sum tests were performed to infer significant differences. A p-value $\leq 0.05$ was defined as statistically significant. AMR: Antimicrobial Resistance, TMM: Trimmed Mean Of M Values.

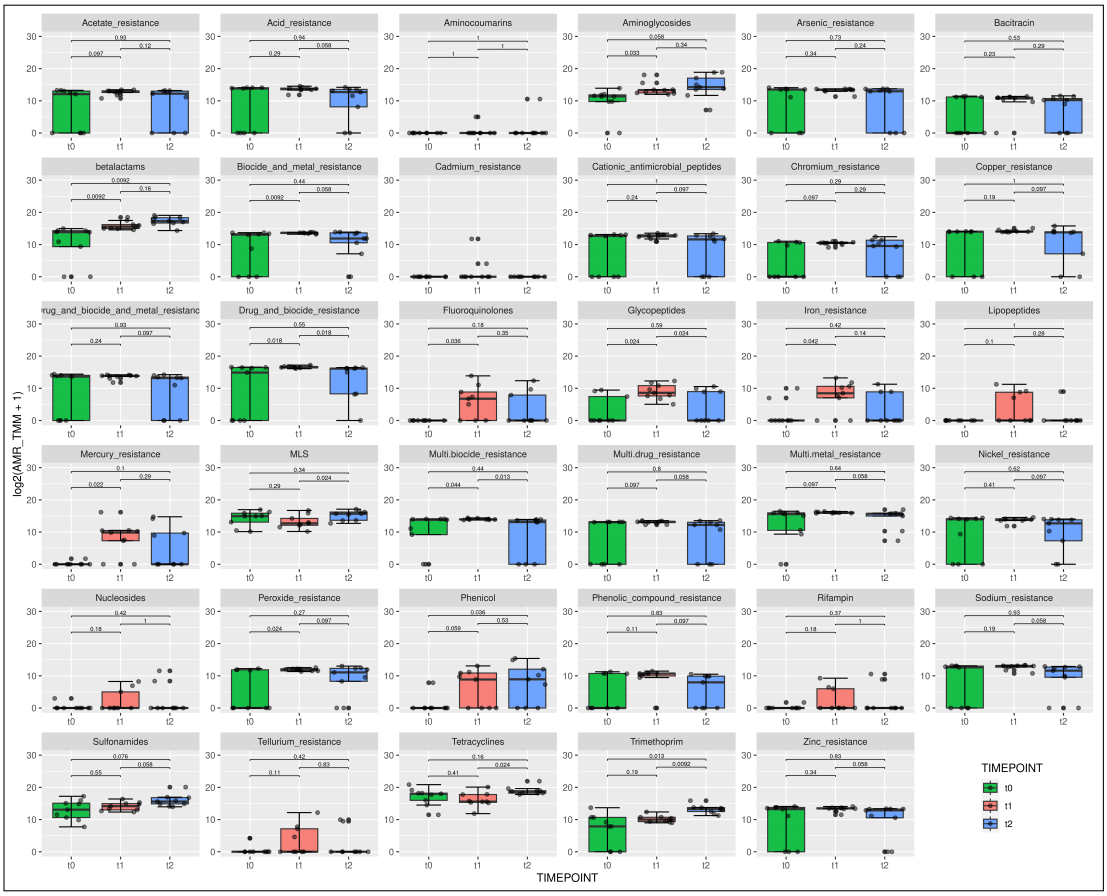


**Figure A.7: Resistance Classes 5DG (Gut Microbiome).** Differential analysis of resistance gene classes across time for equine patients of the 5DG ($n = 11$). Resistance gene abundances were summarized to class level and tested for significant changes over time (colours) per class (each box plot comparison). Wilcoxon rank-sum tests were performed to infer significant differences. A p-value $\leq 0.05$ was defined as statistically significant. AMR: Antimicrobial Resistance, TMM: Trimmed Mean Of M Values.

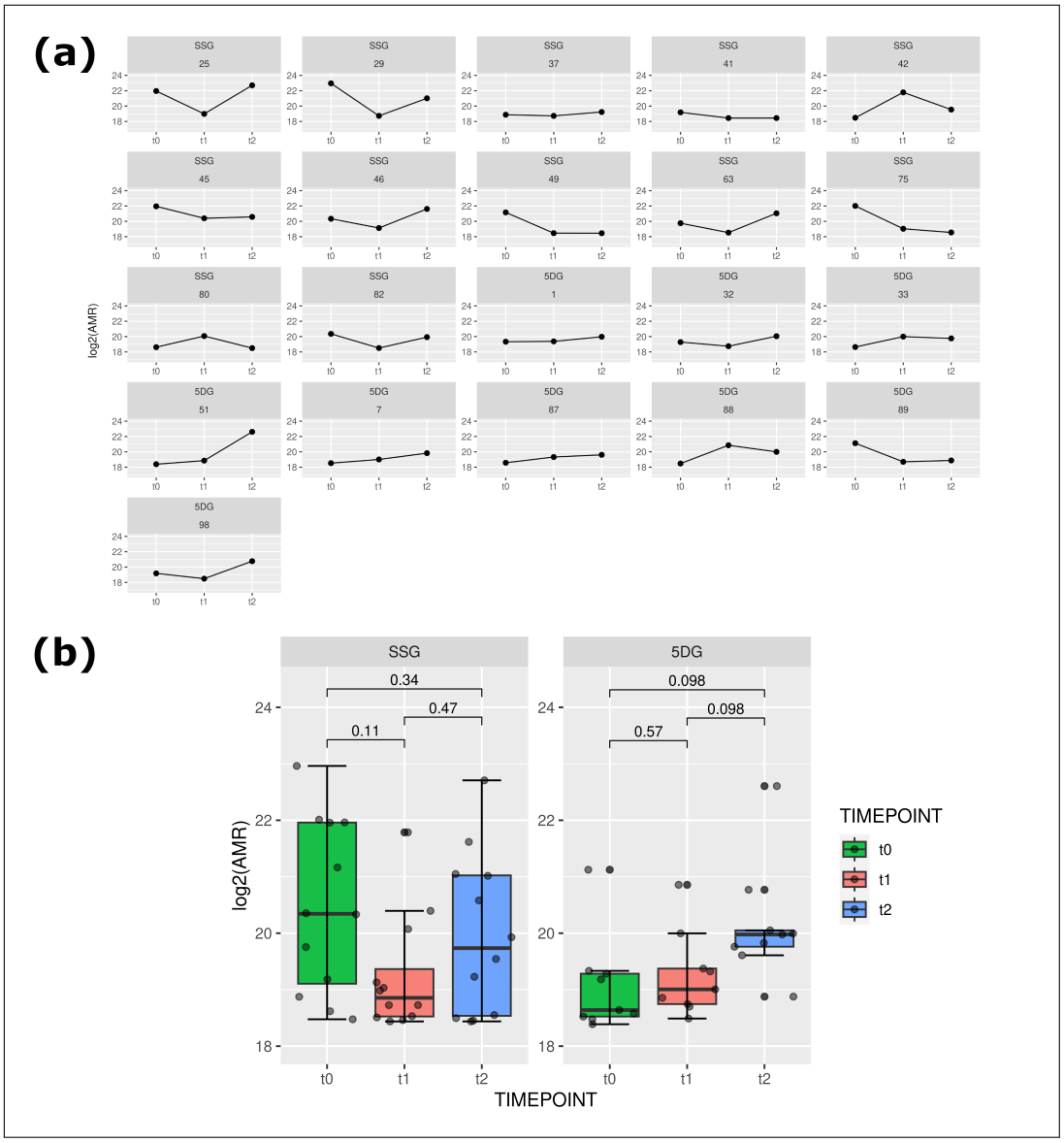


**Figure A.8: Total Resistance Abundance (Gut Microbiome).** Total resistance abundance within the gut metagenome samples. Mapped gene fragments were summarized to total resistance hits. (a) These were first visualized patient-wise over time and (b) statistically compared between time points (colours) for each prophylaxis group (SSG: $n = 12$, 5DG: $n = 9$). Wilcoxon rank-sum tests were performed to infer significant differences. A p-value $\leq 0.05$ was defined as statistically significant. AMR: Antimicrobial Resistance, SSG: Single-Shot Group, 5DG: 5-Day Group.

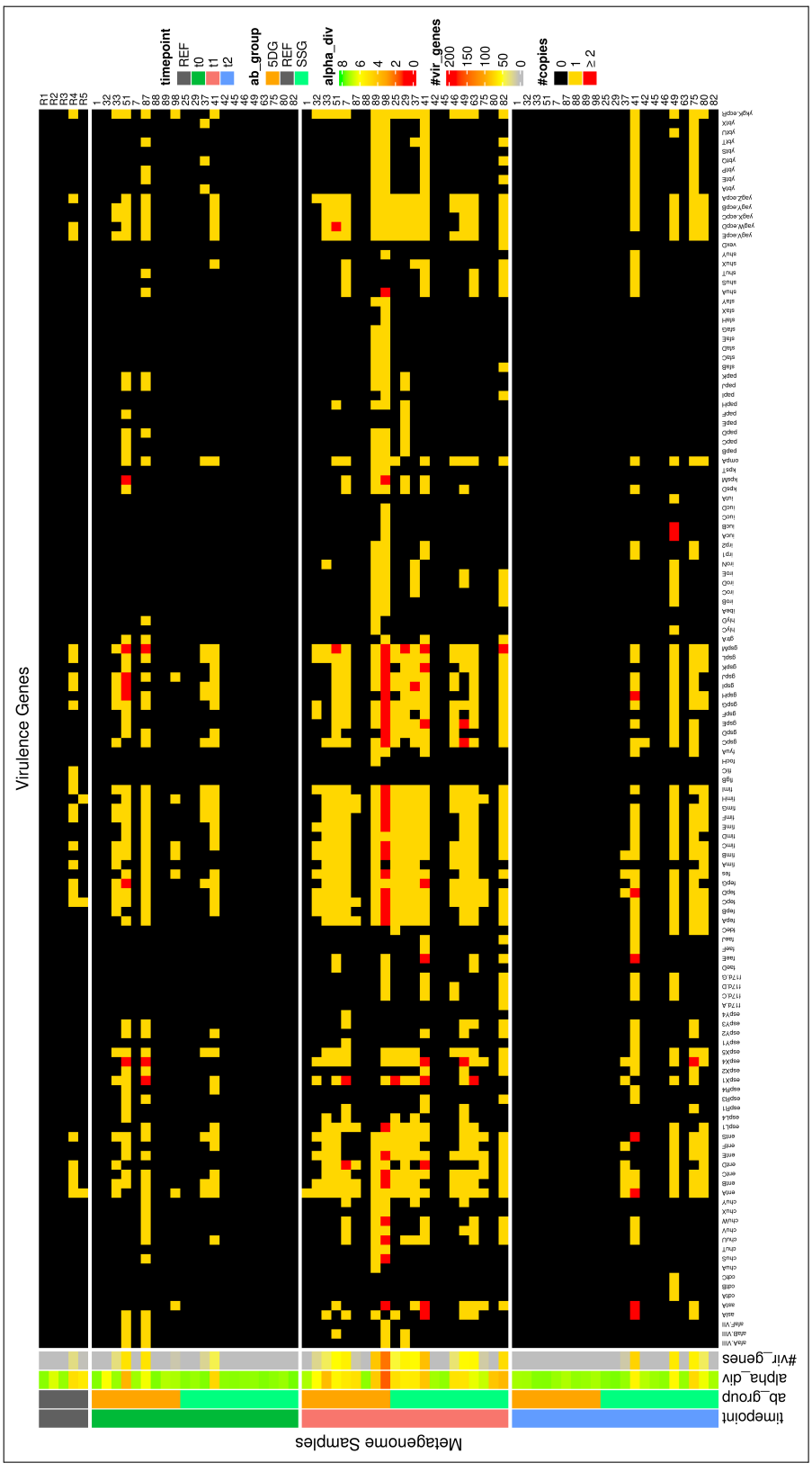


**Figure A.9: Virulence Components (Gut Microbiome).** Visualized are the abundances of genetic virulence components across the sample set. Samples (right) are divided horizontally by groups and time points (left). Columns represent individual virulence components. The heatmap colours indicate absence (black) or abundance (low to high, yellow to red) of specific elements. Additional metadata, including the total abundance of virulence genes, as well as alpha diversity and IDs of the equine patients are plotted alongside the heatmap (left and right). REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

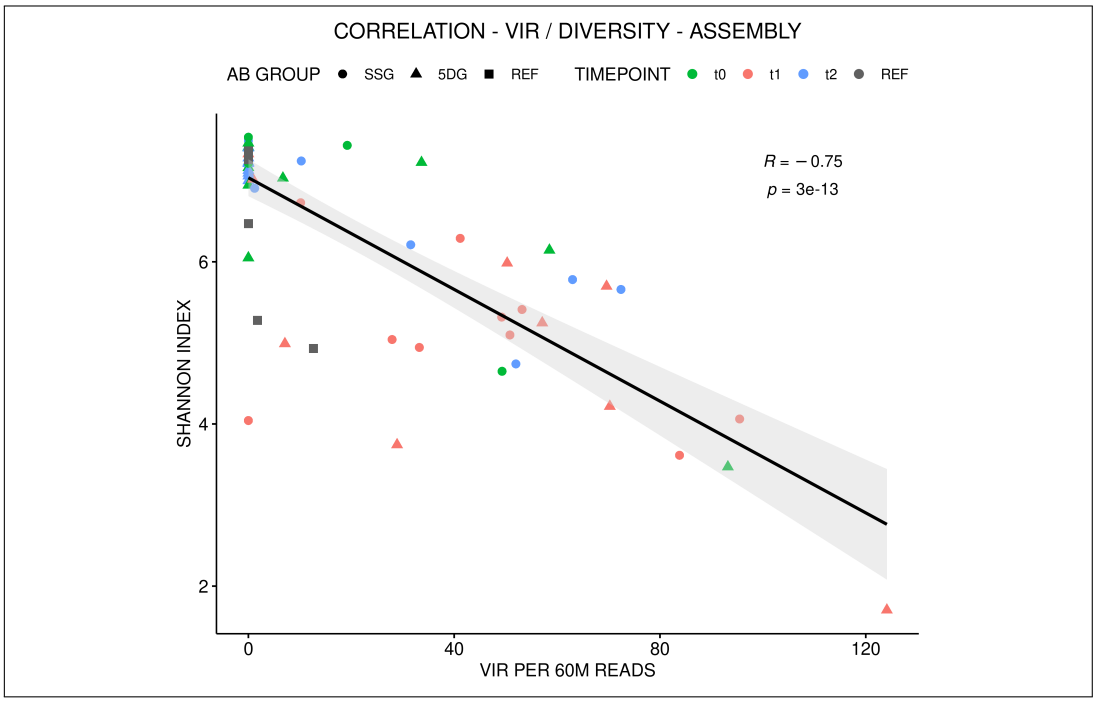


**Figure A.10: Correlation of Virulence & Diversity (Gut Microbiome).** Spearman correlation of normalized virulence components and microbiome alpha diversity across the sample set. Individual samples were coloured according to their time points, with shapes illustrating the distinct study groups. Samples with low alpha diversity were more likely to carry high amounts of virulence elements. Most individuals with high virulence gene counts were associated with the 5DG. REF: NCBI Group, SSG: Single-Shot Group, 5DG: 5-Day Group.

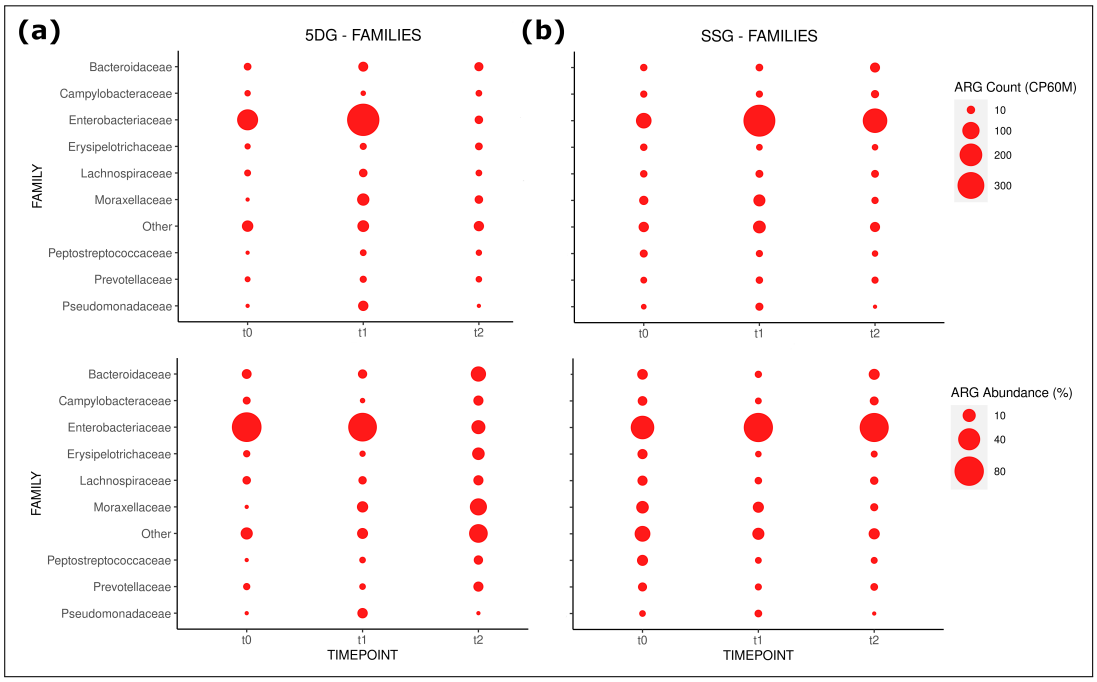


**Figure A.11: Resistance-Associated Families (Gut Microbiome).** Bubble plots visualizing the abundance of resistance genes across bacterial families. Families containing the highest counts of resistance genes were visualized for the 5DG (a) and comparatively for the SSG (b). The size of the bubbles correlates to the normalized resistance genes counts (top) and as percentages (bottom). ARG: Antibiotic Resistance Gene, SSG: Single-Shot Group, 5DG: 5-Day Group.

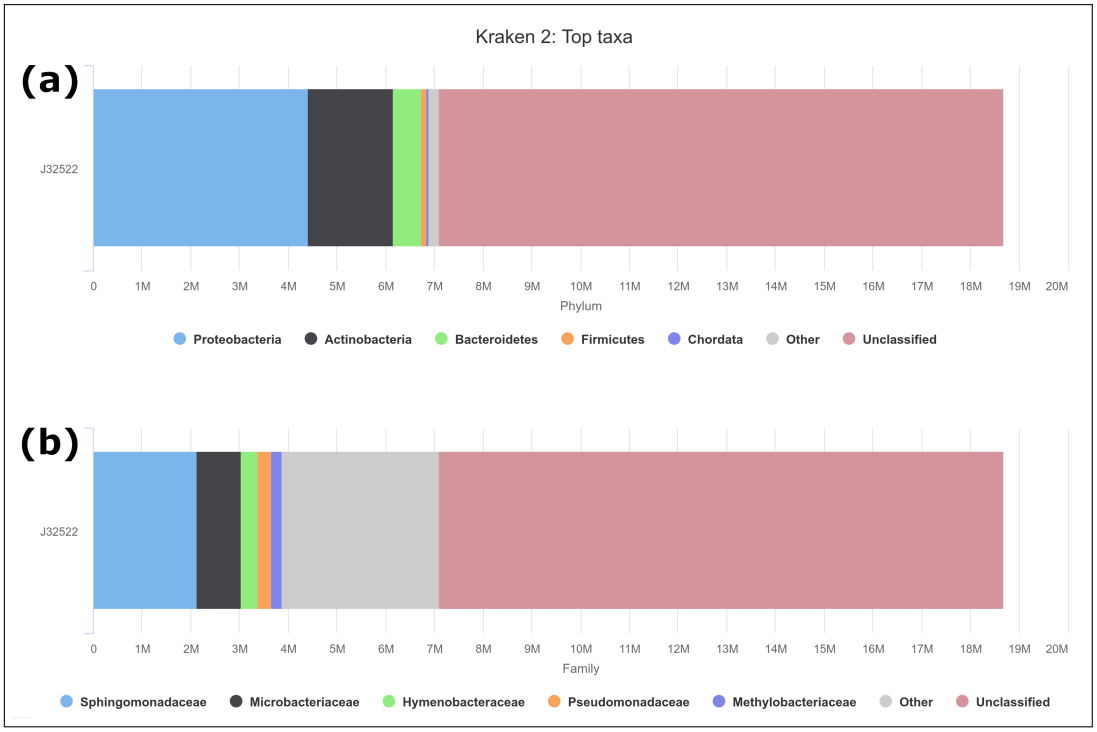


**Figure A.12: Taxonomic Quality Control (Nasal Microbiome).** Visualized are stacked bar charts illustrating read counts for the nasal shotgun metagenome sample (J32522) and coloured according to the abundance of the 5 major taxonomic groups. Minor ranks are grouped into 'other', while reads unable to be typed are represented by 'unclassified'. (a) Bar chart visualization for the phylum-level, (b) for the family-level based on raw counts produced by Kraken2. A majority of all reads cannot be directly assigned to a taxonomic rank (unclassified).

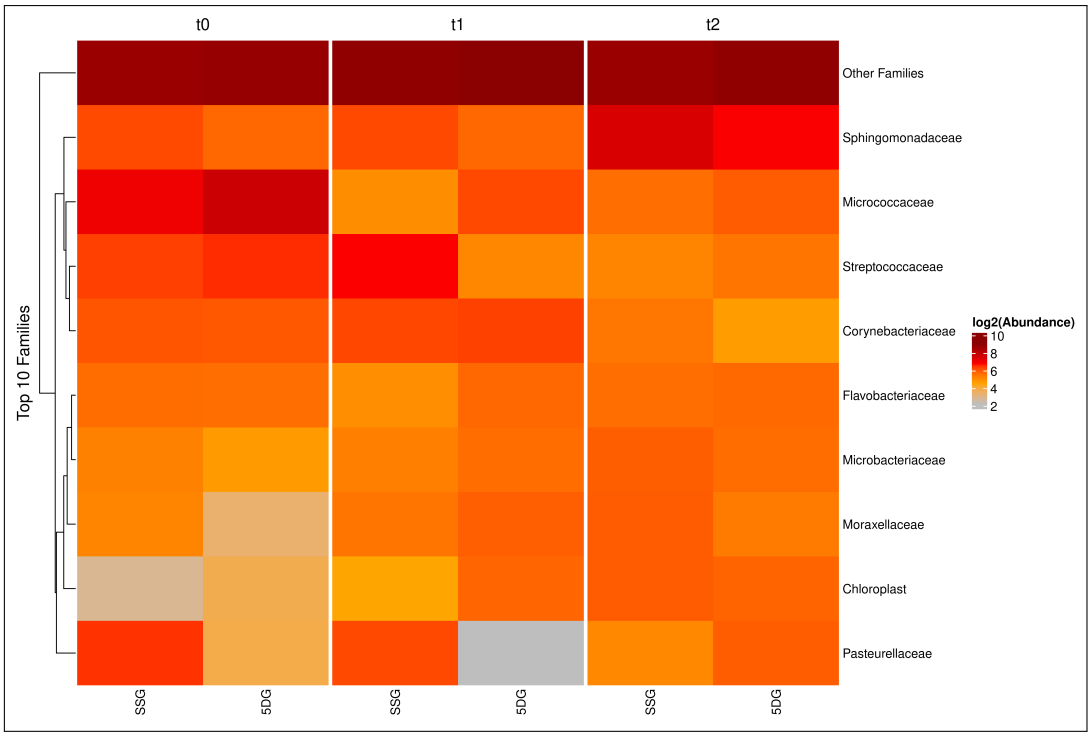


**Figure A.13: Heatmap of Family-Level Abundances (Nasal Microbiome).** Taxonomic profiles were summarized at family-level and averaged across the sample groups. The heatmap visualizes the abundance ($\log_2$ of normalized counts) from low (yellow) to high (dark red) across the data set. Columns represent the different study groups at specific time points. Rows illustrate the 9 most abundant families, with minor ranks being grouped into 'other'. Families are clustered by their similarity to each other in the represented abundance profiles (dendrogram on the left). SSG: Single-Shot Group, 5DG: 5-Day Group.

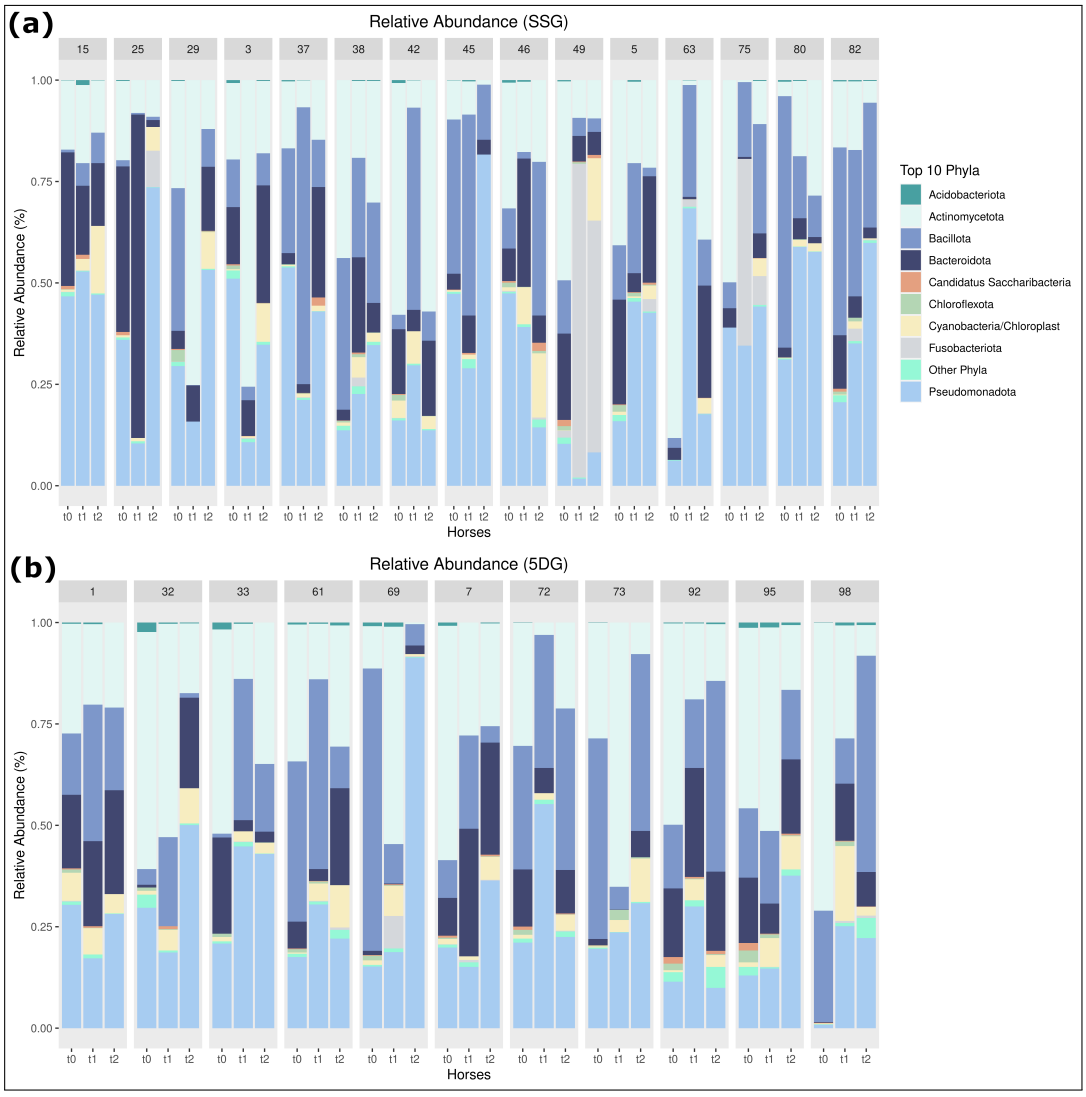


**Figure A.14: Bar Charts of Phylum-Level Compositions (Nasal Microbiome).** Stacked bar charts representing the composition of individual samples on the phylum-level for the 9 most abundant taxonomic groups, with minor ranks being grouped into 'other' (colours). Individual horses ($n = 26$) are clustered into the two prophylaxis groups (a) SSG and (b) 5DG. Samples were collected for each equine patient (top row IDs) across three time points ($t_0$, $t_1$ and $t_2$). Visualized are the percentages of phyla per group over time. SSG: Single-Shot Group, 5DG: 5-Day Group.

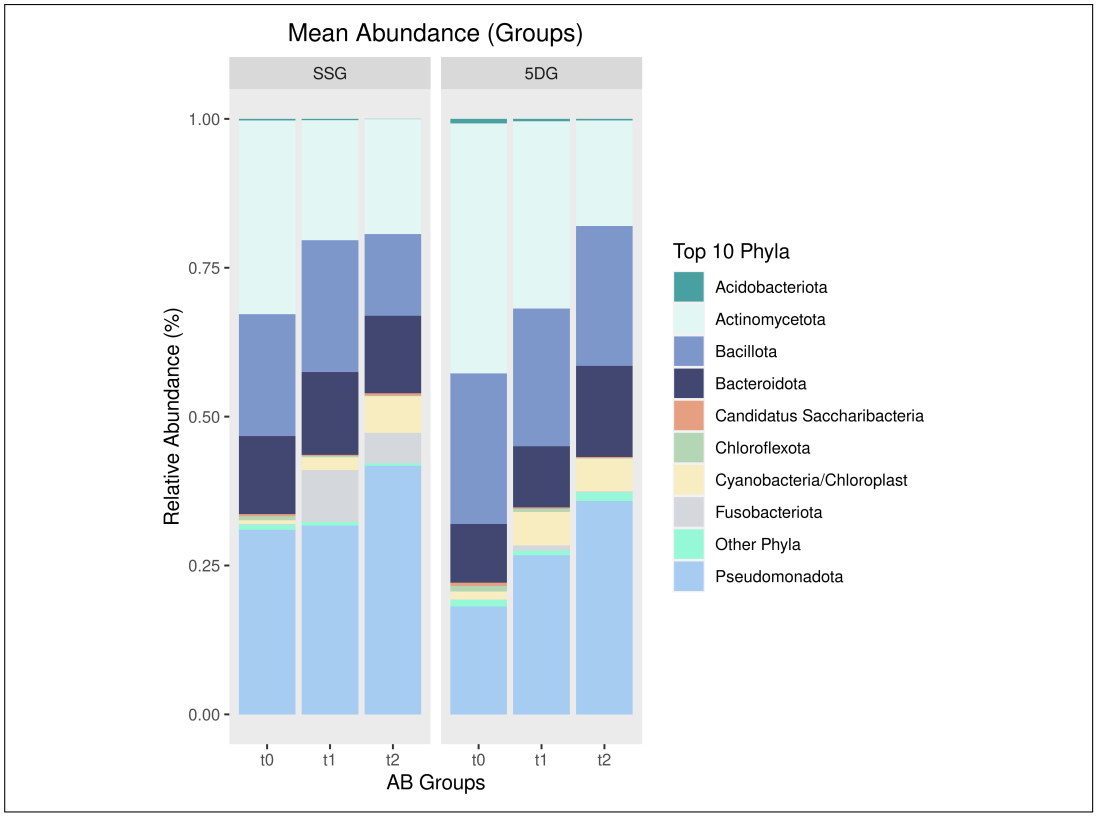


**Figure A.15: Summarized Bar Charts at Phylum-Level (Nasal Microbiome).** Stacked bar charts representing the averaged composition of the samples over different groups on the phylum-level. The 9 most abundant phyla are visualized as percentages in composition for (a) the prophylaxis groups and (b) across the time points. Minor taxonomic ranks are grouped into 'other'. SSG: Single-Shot Group, 5DG: 5-Day Group.

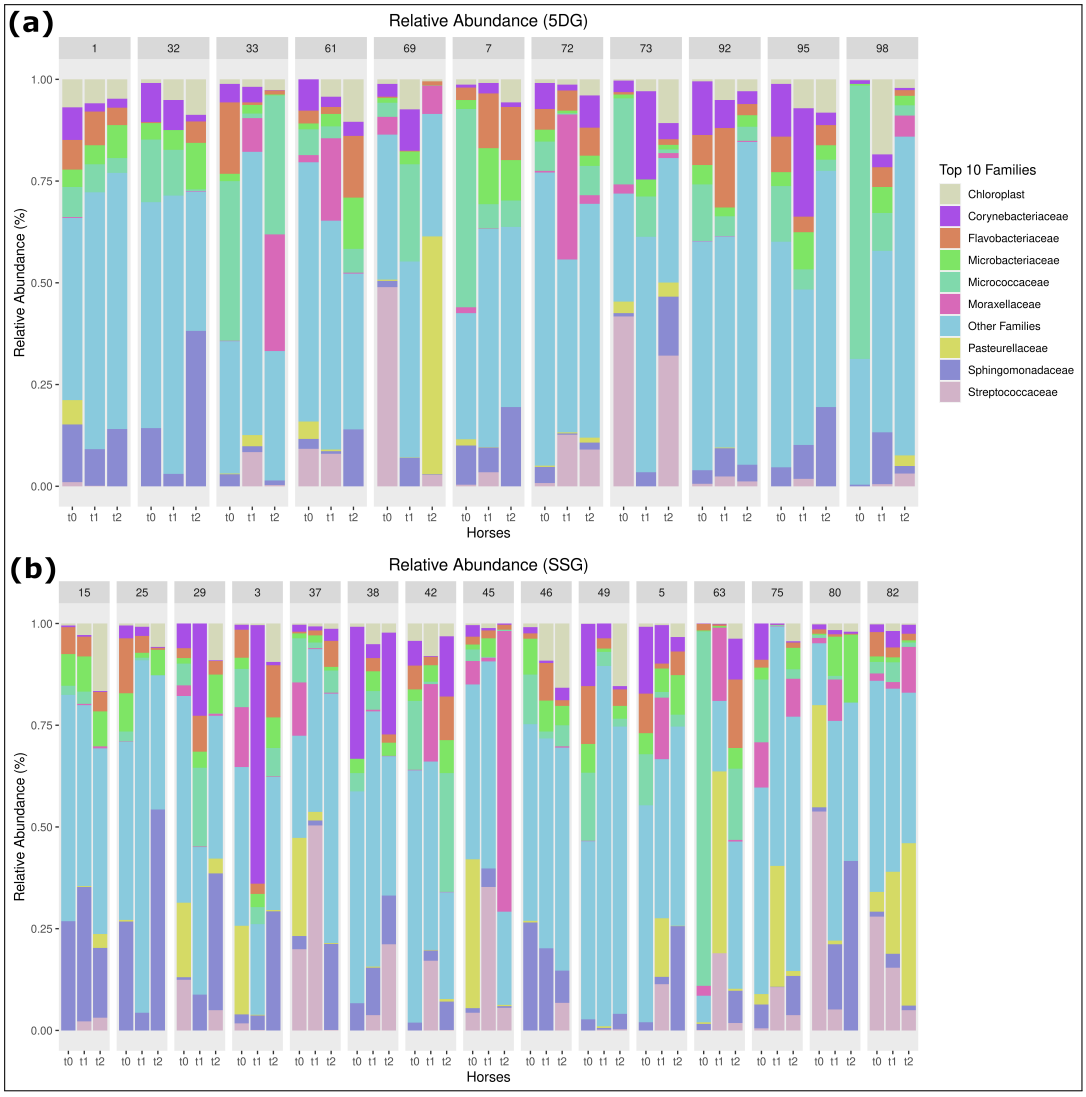


**Figure A.16: Bar Charts of Family-Level Compositions (Nasal Microbiome).** Stacked bar charts representing the composition of individual samples on the phylum-level for the 9 most abundant taxonomic groups, with minor ranks being grouped into 'other' (colours). Individual horses ($n = 26$) are clustered into the two prophylaxis groups (a) SSG and (b) 5DG. Samples were collected for each equine patient (top row IDs) across three time points ($t_0$, $t_1$ and $t_2$). Visualized are the percentages of families per group over time. SSG: Single-Shot Group, 5DG: 5-Day Group.

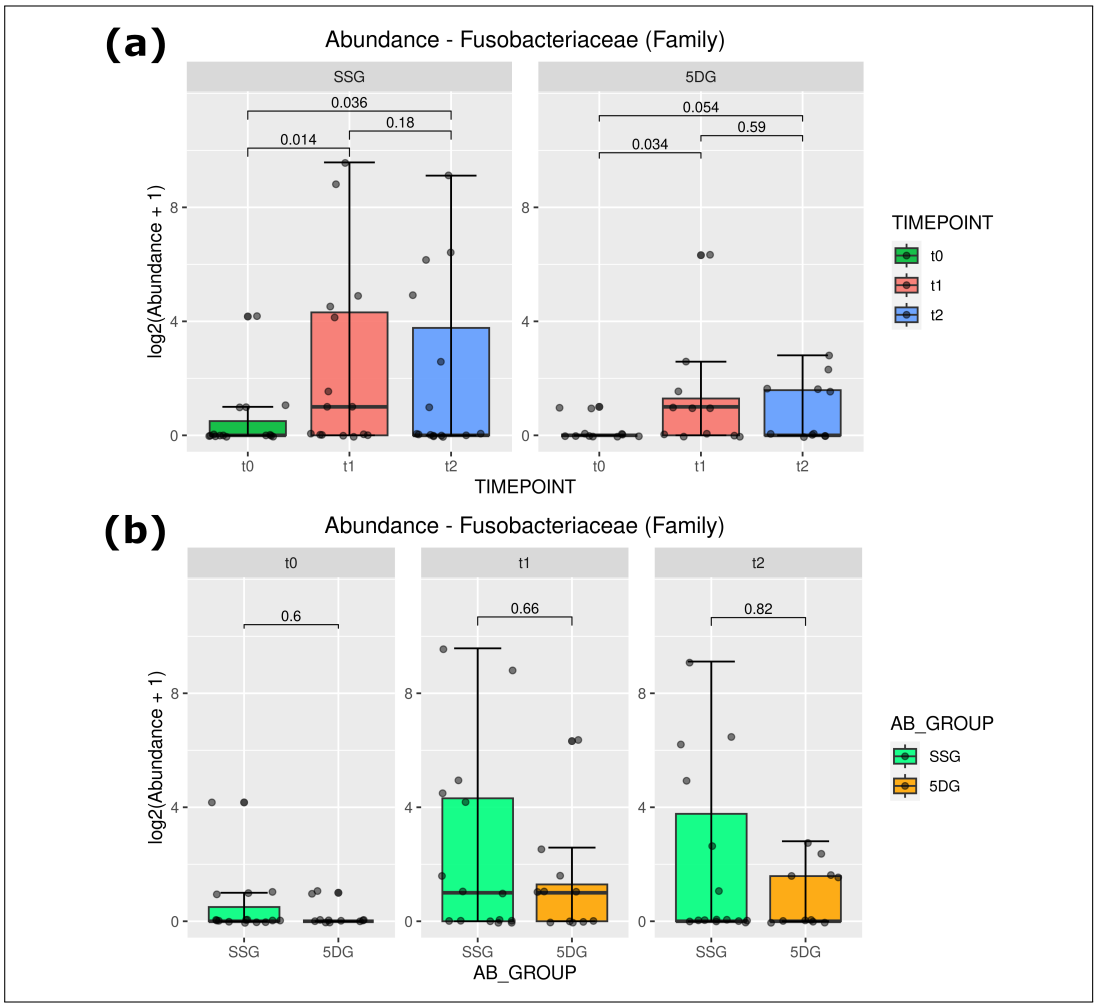


**Figure A.17: Abundance of *Fusobacteriaceae* (Nasal Microbiome).** The abundance of the family of *Fusobacteriaceae* was assessed and comparatively investigated across the sample set. Individual abundances were summarized as box plots for each prophylaxis group (a) and subsequently compared over time (b) using Wilcoxon rank-sum tests (SSG: $n = 15$, 5DG: $n = 11$). A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

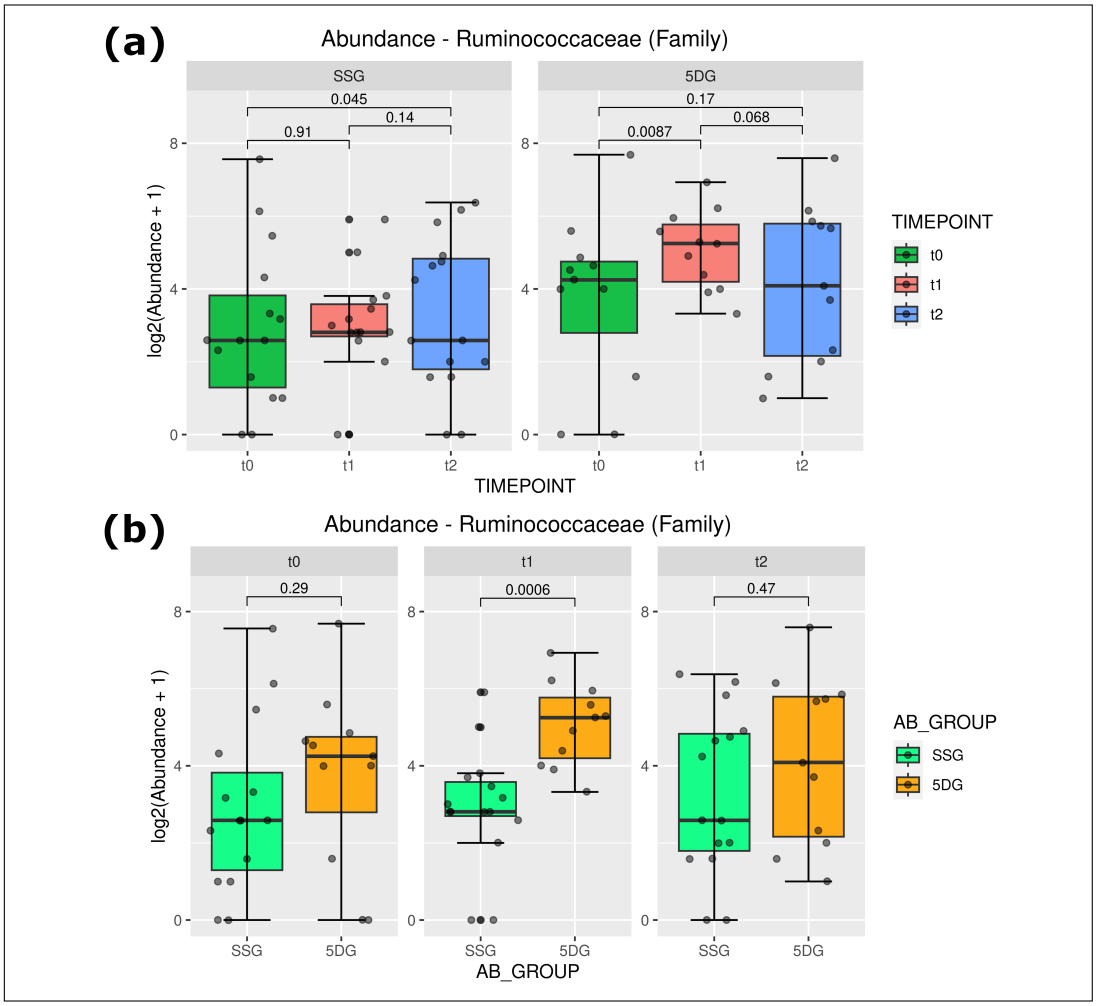


**Figure A.18: Abundance of *Ruminococcaceae* (Nasal Microbiome).** The abundance of the family of *Ruminococcaceae* was assessed and comparatively investigated across the sample set. Individual abundances were summarized as box plots for each prophylaxis group (a) and subsequently compared over time (b) using Wilcoxon rank-sum tests (SSG: $n = 15$, 5DG: $n = 11$). A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

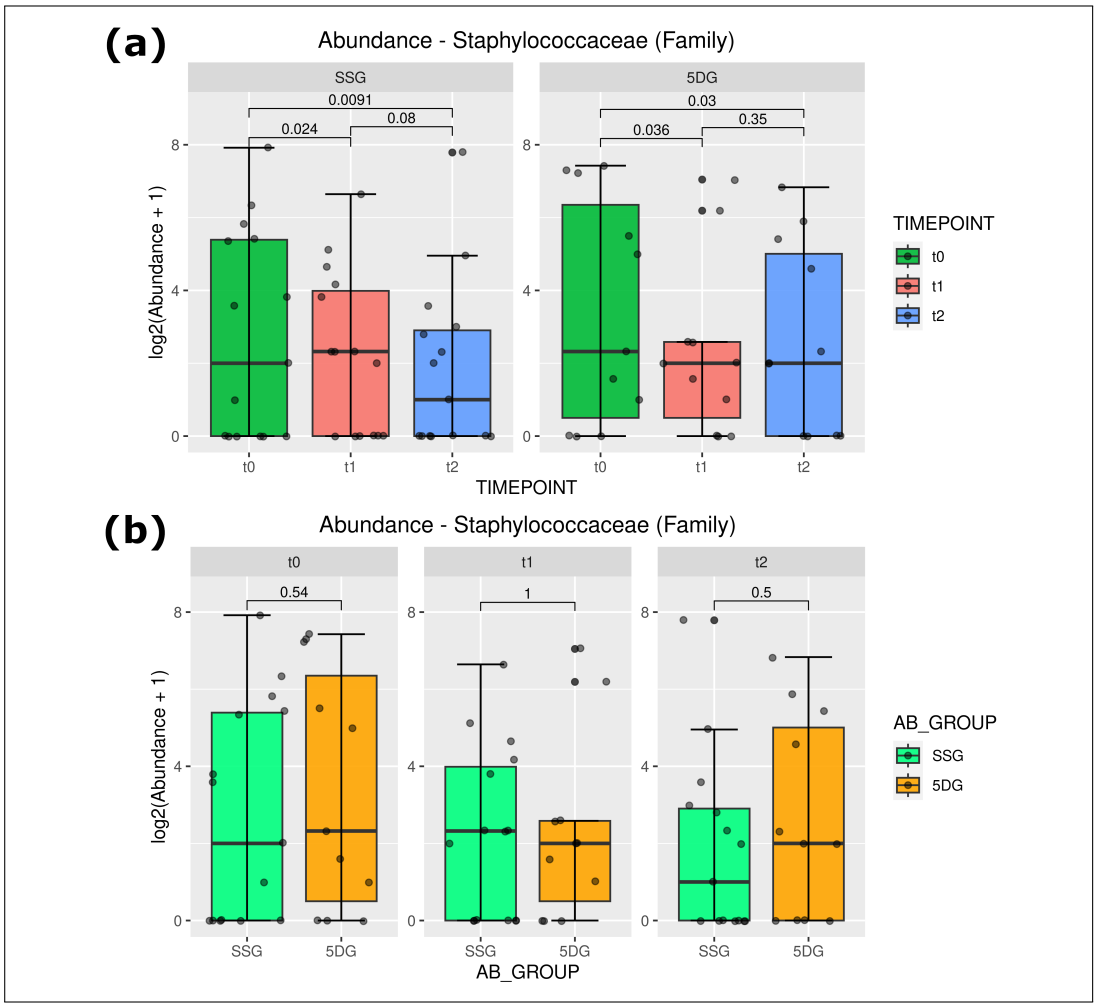


**Figure A.19: Abundance of *Staphylococcaceae* (Nasal Microbiome).** The abundance of the family of *Staphylococcaceae* was assessed and comparatively investigated across the sample set. Individual abundances were summarized as box plots for each prophylaxis group (a) and subsequently compared over time (b) using Wilcoxon rank-sum tests (SSG: $n = 15$, 5DG: $n = 11$). A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

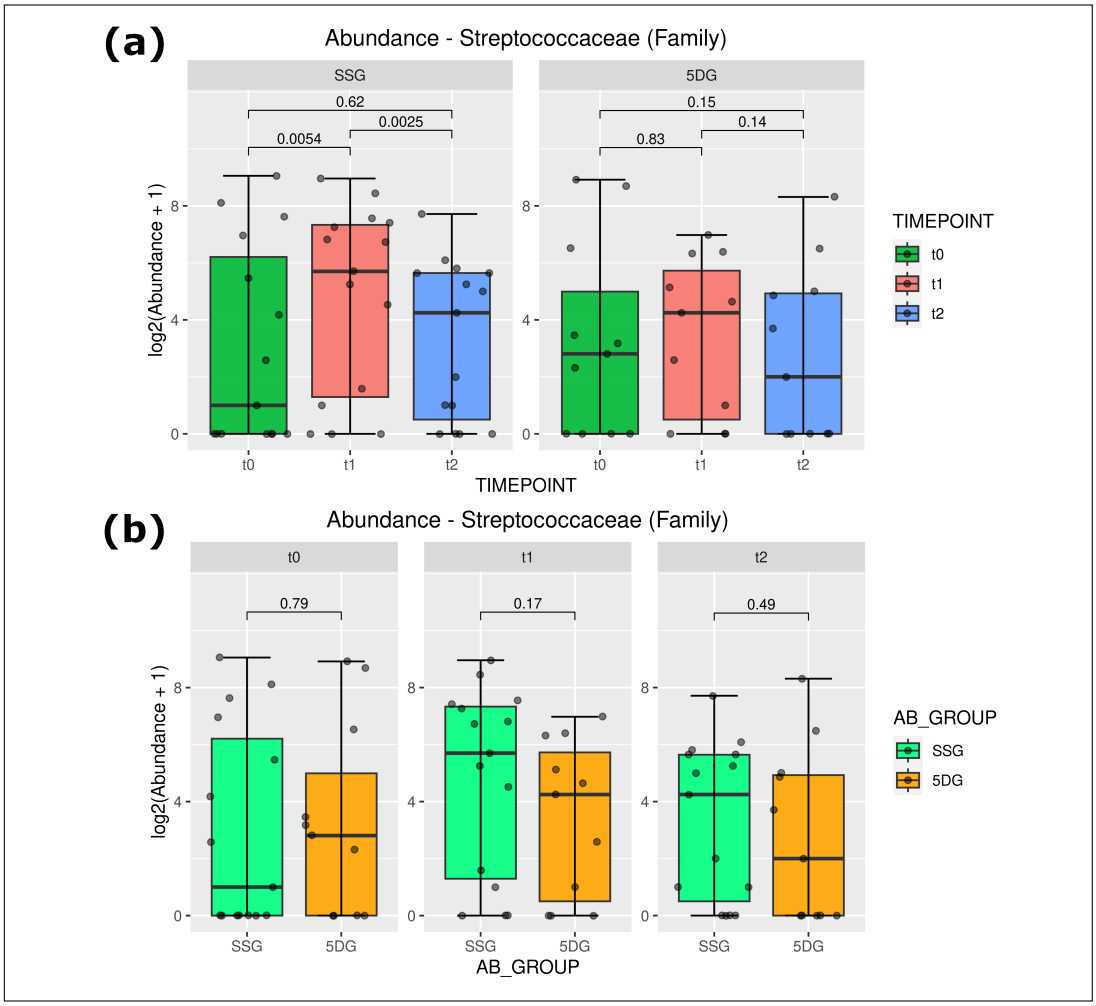


**Figure A.20: Abundance of *Streptococcaceae* (Nasal Microbiome).** The abundance of the family of *Streptococcaceae* was assessed and comparatively investigated across the sample set. Individual abundances were summarized as box plots for each prophylaxis group (a) and subsequently compared over time (b) using Wilcoxon rank-sum tests (SSG: $n = 15$, 5DG: $n = 11$). A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

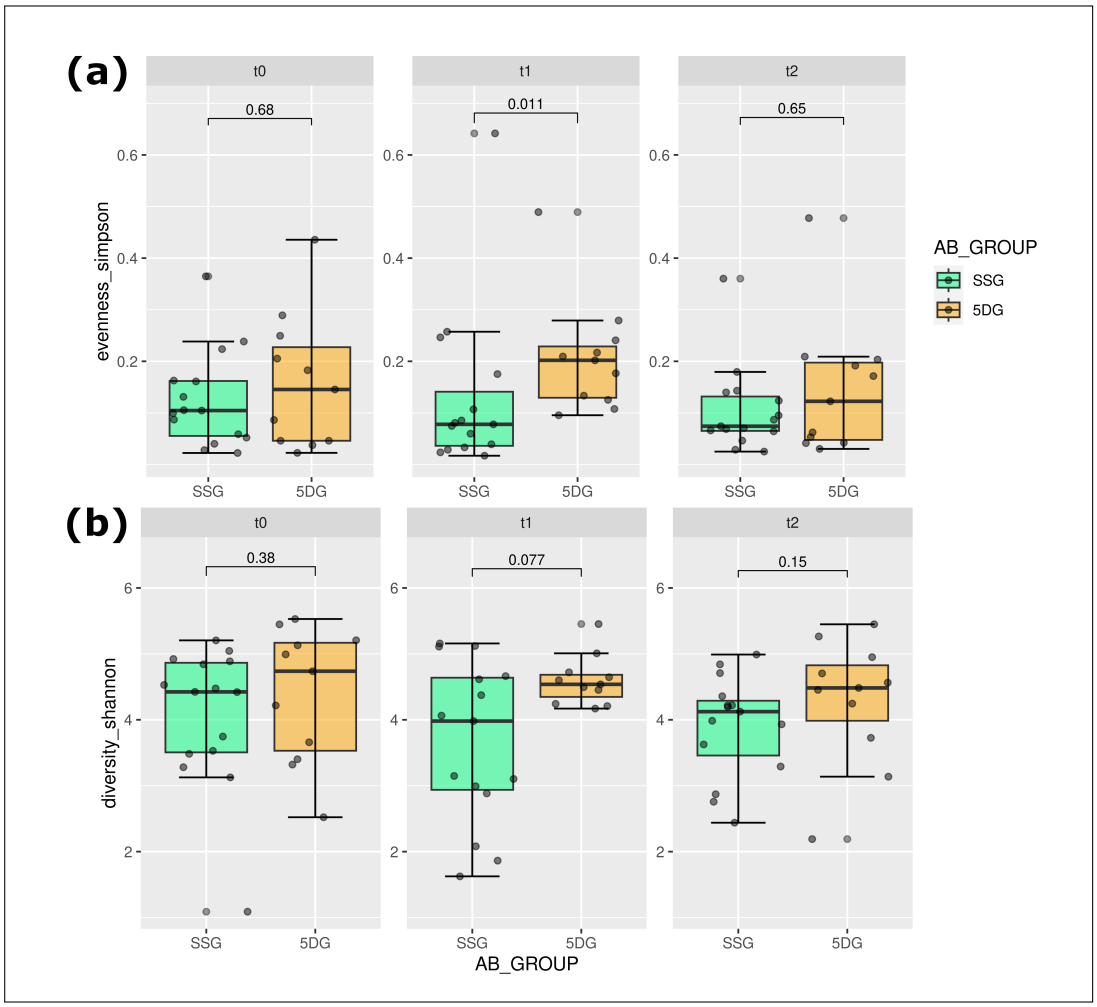


**Figure A.21: Temporal Alpha Diversity (Nasal Microbiome).** Comparison of diversity between sample groups (colours) across time points using box plots (SSG: $n = 15$, 5DG: $n = 11$). (a) Illustrates the evenness, while (b) represents the Shannon diversity over time. Both values decrease from $t_0$ to $t_1$ and recover towards $t_2$. Wilcoxon rank-sum tests were performed to infer significant differences. A p-value $\leq 0.05$ was defined as statistically significant. SSG: Single-Shot Group, 5DG: 5-Day Group.

| Microbiome | Group | Time point | Diversity (Shannon) | Evenness (Simpson) | Richness (Chao1) |
|---|---|---|---|---|---|
| Gut | 5DG | $t_0$ | 6.54 | 0.019 | 5,721 |
| | | $t_1$ | 5.10 | 0.009 | 5,347 |
| | | $t_2$ | 7.20 | 0.033 | 6,222 |
| | SSG | $t_0$ | 7.13 | 0.029 | 6,216 |
| | | $t_1$ | 5.41 | 0.009 | 6,016 |
| | | $t_2$ | 6.69 | 0.022 | 6,098 |
| | NCBI | - | 6.26 | 0.018 | 6,213 |
| Nasal | 5DG | $t_0$ | 4.38 | 0.159 | 559 |
| | | $t_1$ | 4.59 | 0.207 | 386 |
| | | $t_2$ | 4.29 | 0.146 | 453 |
| | SSG | $t_0$ | 4.07 | 0.125 | 520 |
| | | $t_1$ | 3.65 | 0.130 | 363 |
| | | $t_2$ | 3.90 | 0.105 | 377 |

**Table A.1: Diversity Indices (Gut & Nasal Microbiome).** All metagenome shotgun and 16S rRNA sequences were analysed for ecological diversity using MetaGEN and Meta16s, respectively. Diversity measures were averaged for the specific groups and time points and summarized within this table. SSG: Single-Shot Group, 5DG: 5-Day Group.

| HorseID | Group | Time point | Day | Gut (WGS) | Nasal (16S) | Nasal (WGS) |
|---|---|---|---|---|---|---|
| 1 | 5DG | $t_0$ | 0 | H25290 | DL218 | - |
| 1 | 5DG | $t_1$ | 3 | H25286 | DL214 | - |
| 1 | 5DG | $t_2$ | 10 | H25296 | N00678 | - |
| 3 | SSG | $t_0$ | 0 | - | DL215 | - |
| 3 | SSG | $t_1$ | 3 | - | N00641 | - |
| 3 | SSG | $t_2$ | 10 | - | N00671 | - |
| 5 | SSG | $t_0$ | 0 | - | N00656 | - |
| 5 | SSG | $t_1$ | 3 | - | N00627 | - |
| 5 | SSG | $t_2$ | 9 | - | N00648 | - |
| 7 | 5DG | $t_0$ | 0 | H25289 | N00682 | - |
| 7 | 5DG | $t_1$ | 3 | H25288 | N00683 | - |
| 7 | 5DG | $t_2$ | 10 | H25297 | N00684 | - |
| 15 | SSG | $t_0$ | 0 | - | N00657 | - |
| 15 | SSG | $t_1$ | 3 | - | N00647 | - |
| 15 | SSG | $t_2$ | 10 | - | N00675 | - |
| 25 | SSG | $t_0$ | 0 | H25294 | N00685 | - |
| 25 | SSG | $t_1$ | 3 | I12276 | N00686 | - |
| 25 | SSG | $t_2$ | 10 | H25287 | N00687 | - |
| 29 | SSG | $t_0$ | 0 | F11029 | N00628 | - |
| 29 | SSG | $t_1$ | 3 | F11031 | N00649 | - |
| 29 | SSG | $t_2$ | 10 | F11033 | N00637 | - |
| 32 | 5DG | $t_0$ | 0 | I12282 | N00688 | - |
| 32 | 5DG | $t_1$ | 3 | I12281 | N00689 | - |
| 32 | 5DG | $t_2$ | 10 | I12280 | N00690 | J32522 |
| 33 | 5DG | $t_0$ | 0 | I12278 | N00691 | - |
| 33 | 5DG | $t_1$ | 3 | I12277 | N00692 | - |
| 33 | 5DG | $t_2$ | 10 | I12279 | N00693 | - |
| 37 | SSG | $t_0$ | 0 | I12272 | N00694 | - |
| 37 | SSG | $t_1$ | 3 | I12273 | N00695 | - |
| 37 | SSG | $t_2$ | 10 | I12271 | N00696 | - |
| 38 | SSG | $t_0$ | 0 | - | N00673 | - |
| 38 | SSG | $t_1$ | 3 | - | N00661 | - |
| 38 | SSG | $t_2$ | 10 | - | N00645 | - |
| 41 | SSG | $t_0$ | 0 | D61001 | - | - |
| 41 | SSG | $t_1$ | 3 | D61002 | - | - |
| 41 | SSG | $t_2$ | 10 | D61003 | - | - |
| 42 | SSG | $t_0$ | 0 | F11032 | DL213 | - |
| 42 | SSG | $t_1$ | 3 | F11034 | N00667 | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| 42 | SSG | $t_2$ | 10 | F11036 | N00655 | - |
| 45 | SSG | $t_0$ | 0 | F11021 | N00622 | - |
| 45 | SSG | $t_1$ | 3 | F11024 | N00666 | - |
| 45 | SSG | $t_2$ | 10 | F11027 | DL129 | - |
| 46 | SSG | $t_0$ | 0 | F11035 | N00606 | - |
| 46 | SSG | $t_1$ | 3 | F11037 | N00669 | - |
| 46 | SSG | $t_2$ | 10 | F11022 | N00617 | - |
| 49 | SSG | $t_0$ | 0 | F11025 | N00676 | - |
| 49 | SSG | $t_1$ | 3 | F11028 | N00677 | - |
| 49 | SSG | $t_2$ | 10 | F11030 | N00672 | - |
| 51 | 5DG | $t_0$ | 0 | D61004 | - | - |
| 51 | 5DG | $t_1$ | 3 | D61005 | - | - |
| 51 | 5DG | $t_2$ | 10 | D61006 | - | - |
| 61 | 5DG | $t_0$ | 0 | - | N00616 | - |
| 61 | 5DG | $t_1$ | 3 | - | DL207 | - |
| 61 | 5DG | $t_2$ | 10 | - | N00679 | - |
| 63 | SSG | $t_0$ | 0 | J10522 | N00668 | - |
| 63 | SSG | $t_1$ | 3 | J10523 | N00644 | - |
| 63 | SSG | $t_2$ | 10 | J10524 | N00640 | - |
| 69 | 5DG | $t_0$ | 0 | - | N00621 | - |
| 69 | 5DG | $t_1$ | 3 | - | N00681 | - |
| 69 | 5DG | $t_2$ | 10 | - | N00614 | - |
| 72 | 5DG | $t_0$ | 0 | - | N00639 | - |
| 72 | 5DG | $t_1$ | 3 | - | DL202 | - |
| 72 | 5DG | $t_2$ | 10 | - | N00660 | - |
| 73 | 5DG | $t_0$ | 0 | - | N00659 | - |
| 73 | 5DG | $t_1$ | 3 | - | N00610 | - |
| 73 | 5DG | $t_2$ | 10 | - | N00653 | - |
| 75 | SSG | $t_0$ | 0 | J10525 | N00658 | - |
| 75 | SSG | $t_1$ | 3 | J10526 | DL203 | - |
| 75 | SSG | $t_2$ | 10 | J10527 | N00643 | - |
| 80 | SSG | $t_0$ | 0 | J10528 | DL144 | - |
| 80 | SSG | $t_1$ | 3 | J10529 | DL209 | - |
| 80 | SSG | $t_2$ | 10 | J10530 | N00601 | - |
| 82 | SSG | $t_0$ | 0 | F11038 | DL210 | - |
| 82 | SSG | $t_1$ | 3 | F11023 | DL217 | - |
| 82 | SSG | $t_2$ | 10 | F11026 | DL128 | - |
| 87 | 5DG | $t_0$ | 0 | D61007 | - | - |
| 87 | 5DG | $t_1$ | 3 | D61008 | - | - |
| 87 | 5DG | $t_2$ | 10 | D61009 | - | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| 88 | 5DG | $t_0$ | 0 | D61010 | - | - |
| 88 | 5DG | $t_1$ | 3 | D61011 | - | - |
| 88 | 5DG | $t_2$ | 10 | D61012 | - | - |
| 89 | 5DG | $t_0$ | 0 | D61013 | - | - |
| 89 | 5DG | $t_1$ | 3 | D61014 | - | - |
| 89 | 5DG | $t_2$ | 10 | D61015 | - | - |
| 92 | 5DG | $t_0$ | 0 | - | DL146 | - |
| 92 | 5DG | $t_1$ | 3 | - | DL211 | - |
| 92 | 5DG | $t_2$ | 10 | - | DL204 | - |
| 95 | 5DG | $t_0$ | 0 | - | DL206 | - |
| 95 | 5DG | $t_1$ | 3 | - | N00626 | - |
| 95 | 5DG | $t_2$ | 10 | - | DL133 | - |
| 98 | 5DG | $t_0$ | 0 | J10519 | N00602 | - |
| 98 | 5DG | $t_1$ | 3 | J10520 | DL219 | - |
| 98 | 5DG | $t_2$ | 10 | J10521 | N00652 | - |
| R1 | NCBI | - | - | SRR10505693 | - | - |
| R2 | NCBI | - | - | SRR10505694 | - | - |
| R3 | NCBI | - | - | SRR10505695 | - | - |
| R4 | NCBI | - | - | SRR10505696 | - | - |
| R5 | NCBI | - | - | SRR10505697 | - | - |

**Table A.2: Overview of the Sample Collection (Gut & Nasal Microbiome).** Details of the equine patients, as well as an overview of all samples utilized for the characterization of the gut and nasal microbiomes. While faecal and nasal samples were collected from every equine patient across three time points ($t_0$, $t_1$, $t_2$), only those selected for sequencing and analysis are listed within this table. A large amount of nasal swabs were collected for 16s rRNA amplicon sequencing, with refined subsets chosen for metagenome shotgun sequencing. Raw data has been deposited in the SRA under the BioProject ID PRJNA998844. Sequences from NCBI were received from BioProject PRJNA590977. SSG: Single-Shot Group, WGS: Whole Genome Sequencing, 5DG: 5-Day Group.

| MBO | Group | Time | Family | logFC | FDR |
|---|---|---|---|---|---|
| Gut | 5DG | $t_0$ vs $t_1$ | *Aeromonadaceae* | 2.17 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Alloherpesviridae* | 4.46 | 0.002 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Autographiviridae* | 5.94 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Bacteriovoracaceae* | 0.70 | 0.035 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Bacteroidaceae* | 6.53 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Bruguierivoracaceae* | 2.88 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Budviciaceae* | 2.53 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Candidatus Methanomethylophilaceae* | 1.64 | 0.012 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Caulobacteraceae* | 1.63 | 0.046 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Chromatiaceae* | 1.49 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Clostridiaceae* | 1.29 | 0.039 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Coriobacteriaceae* | -2.62 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Demerecviridae* | 4.16 | 0.034 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Eggerthellaceae* | -0.95 | 0.039 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Enterobacteriaceae* | 5.50 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Enterococcaceae* | 1.60 | 0.004 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Erwiniaceae* | 6.55 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Flavobacteriaceae* | 1.94 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Fusobacteriaceae* | 1.37 | 0.035 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Hafniaceae* | 2.83 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Halieaceae* | 0.60 | 0.047 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Herelleviridae* | 4.55 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Inoviridae* | 3.37 | 0.015 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Jonesiaceae* | 1.58 | 0.003 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Lachnospiraceae* | 2.28 | 0.001 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Microviridae* | 5.95 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Moraxellaceae* | 9.12 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Morganellaceae* | 1.51 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Myoviridae* | 7.80 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Neisseriaceae* | 1.24 | 0.003 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Oceanospirillaceae* | 0.63 | 0.039 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Odoribacteraceae* | 2.97 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Oleiphilaceae* | 0.95 | 0.014 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Oxalobacteraceae* | 4.98 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Pasteurellaceae* | 2.21 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Pectobacteriaceae* | 3.53 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Peribunyaviridae* | 2.86 | 0.012 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Podoviridae* | 5.63 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Pseudomonadaceae* | 7.10 | 0.000 |

| | | | | | |
|---|---|---|---|---|---|
| Gut | 5DG | $t_0$ vs $t_1$ | *Siphoviridae* | 2.83 | 0.001 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Sphingobacteriaceae* | 1.74 | 0.001 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Staphylococcaceae* | -2.41 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Tannerellaceae* | 4.55 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Veillonellaceae* | 2.35 | 0.031 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Weeksellaceae* | 4.61 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Xanthomonadaceae* | 1.97 | 0.037 |
| Gut | 5DG | $t_0$ vs $t_1$ | *Yersiniaceae* | 4.58 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_2$ | *Atopobiaceae* | -1.73 | 0.013 |
| Gut | 5DG | $t_0$ vs $t_2$ | *Coriobacteriaceae* | -2.57 | 0.000 |
| Gut | 5DG | $t_0$ vs $t_2$ | *Fortieaceae* | -9.24 | 0.001 |
| Gut | 5DG | $t_0$ vs $t_2$ | *Muribaculaceae* | -2.35 | 0.007 |
| Gut | 5DG | $t_0$ vs $t_2$ | *Staphylococcaceae* | -2.17 | 0.002 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Aeromonadaceae* | -2.30 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Alcanivoracaceae* | -0.58 | 0.049 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Alloherpesviridae* | -4.17 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Autographiviridae* | -5.24 | 0.001 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Bacteroidaceae* | -8.24 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Bruguierivoracaceae* | -2.79 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Budviciaceae* | -2.36 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Candidatus Methanomethylophilaceae* | -1.69 | 0.006 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Caulobacteraceae* | -1.78 | 0.021 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Chromatiaceae* | -1.48 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Chromobacteriaceae* | -0.77 | 0.023 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Chuviridae* | -2.85 | 0.001 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Closteroviridae* | -2.04 | 0.039 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Clostridiaceae* | -1.35 | 0.023 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Dissulfurispiraceae* | 0.83 | 0.025 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Drexlerviridae* | -7.26 | 0.010 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Endozoicomonadaceae* | -0.76 | 0.049 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Enterobacteriaceae* | -8.80 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Enterococcaceae* | -1.70 | 0.002 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Erwiniaceae* | -6.44 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Flavobacteriaceae* | -1.56 | 0.007 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Fortieaceae* | -9.49 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Hafniaceae* | -2.94 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Herelleviridae* | -4.53 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Hytrosaviridae* | -2.11 | 0.023 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Inoviridae* | -3.79 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Jonesiaceae* | -1.52 | 0.005 |

| | | | | | |
|---|---|---|---|---|---|
| Gut | 5DG | $t_1$ vs $t_2$ | *Lachnospiraceae* | -2.05 | 0.004 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Lipothrixviridae* | -2.70 | 0.004 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Microviridae* | -4.45 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Moraxellaceae* | -5.97 | 0.002 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Morganellaceae* | -1.35 | 0.002 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Muribaculaceae* | -1.95 | 0.008 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Myoviridae* | -6.22 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Narnaviridae* | -2.70 | 0.001 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Neisseriaceae* | -1.42 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Odoribacteraceae* | -3.45 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Oleiphilaceae* | -0.93 | 0.021 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Oxalobacteraceae* | -3.41 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Pasteurellaceae* | -1.87 | 0.002 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Pectobacteriaceae* | -3.54 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Peribunyaviridae* | -3.25 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Phenuiviridae* | -1.92 | 0.027 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Podoviridae* | -5.87 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Polydnaviridae* | -3.31 | 0.001 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Pseudomonadaceae* | -6.73 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Rudiviridae* | -2.47 | 0.032 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Secoviridae* | -1.85 | 0.040 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Siphoviridae* | -2.18 | 0.021 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Sphingobacteriaceae* | -1.63 | 0.003 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Tannerellaceae* | -5.12 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Tobaniviridae* | -2.33 | 0.006 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Weeksellaceae* | -4.59 | 0.000 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Xanthomonadaceae* | -2.08 | 0.021 |
| Gut | 5DG | $t_1$ vs $t_2$ | *Yersiniaceae* | -4.60 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Ackermannviridae* | 5.35 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Alcaligenaceae* | 0.96 | 0.049 |
| Gut | SSG | $t_0$ vs $t_1$ | *Anaeromyxobacteraceae* | -0.64 | 0.049 |
| Gut | SSG | $t_0$ vs $t_1$ | *Bacteroidaceae* | 3.19 | 0.001 |
| Gut | SSG | $t_0$ vs $t_1$ | *Bicaudaviridae* | 3.46 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Bogoriellaceae* | 1.06 | 0.037 |
| Gut | SSG | $t_0$ vs $t_1$ | *Caulobacteraceae* | 2.55 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Cellulomonadaceae* | 2.60 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Comamonadaceae* | 2.97 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Cyclobacteriaceae* | 0.66 | 0.023 |
| Gut | SSG | $t_0$ vs $t_1$ | *Enterobacteriaceae* | 3.96 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Fortieaceae* | -10.69 | 0.000 |

| | | | | | |
|---|---|---|---|---|---|
| Gut | SSG | $t_0$ vs $t_1$ | *Gordoniaceae* | 1.02 | 0.033 |
| Gut | SSG | $t_0$ vs $t_1$ | *Kiritimatiellaceae* | -0.90 | 0.049 |
| Gut | SSG | $t_0$ vs $t_1$ | *Marinifilaceae* | 0.72 | 0.031 |
| Gut | SSG | $t_0$ vs $t_1$ | *Methanocorpusculaceae* | -2.18 | 0.049 |
| Gut | SSG | $t_0$ vs $t_1$ | *Methanomicrobiaceae* | -1.38 | 0.013 |
| Gut | SSG | $t_0$ vs $t_1$ | *Methanoregulaceae* | -1.89 | 0.001 |
| Gut | SSG | $t_0$ vs $t_1$ | *Methanospirillaceae* | -1.69 | 0.011 |
| Gut | SSG | $t_0$ vs $t_1$ | *Microbacteriaceae* | 1.74 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Micrococcaceae* | 2.28 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Microscillaceae* | 2.93 | 0.003 |
| Gut | SSG | $t_0$ vs $t_1$ | *Mycobacteriaceae* | 1.08 | 0.026 |
| Gut | SSG | $t_0$ vs $t_1$ | *Myoviridae* | 3.96 | 0.001 |
| Gut | SSG | $t_0$ vs $t_1$ | *Nocardiaceae* | 4.94 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Nocardioidaceae* | 1.69 | 0.001 |
| Gut | SSG | $t_0$ vs $t_1$ | *Nostocaceae* | -1.27 | 0.039 |
| Gut | SSG | $t_0$ vs $t_1$ | *Picornaviridae* | -3.21 | 0.001 |
| Gut | SSG | $t_0$ vs $t_1$ | *Pithoviridae* | -2.89 | 0.006 |
| Gut | SSG | $t_0$ vs $t_1$ | *Planococcaceae* | 1.38 | 0.049 |
| Gut | SSG | $t_0$ vs $t_1$ | *Podoviridae* | 3.66 | 0.001 |
| Gut | SSG | $t_0$ vs $t_1$ | *Porphyromonadaceae* | 0.99 | 0.033 |
| Gut | SSG | $t_0$ vs $t_1$ | *Prolixibacteraceae* | 0.68 | 0.031 |
| Gut | SSG | $t_0$ vs $t_1$ | *Promicromonosporaceae* | 1.98 | 0.000 |
| Gut | SSG | $t_0$ vs $t_1$ | *Sanguibacteraceae* | 1.67 | 0.001 |
| Gut | SSG | $t_0$ vs $t_1$ | *Saprospiraceae* | 0.84 | 0.024 |
| Gut | SSG | $t_0$ vs $t_1$ | *Siphoviridae* | 2.19 | 0.005 |
| Gut | SSG | $t_0$ vs $t_1$ | *Sphingobacteriaceae* | 1.39 | 0.005 |
| Gut | SSG | $t_0$ vs $t_1$ | *Succinivibrionaceae* | 1.70 | 0.001 |
| Gut | SSG | $t_0$ vs $t_1$ | *Tannerellaceae* | 2.37 | 0.003 |
| Gut | SSG | $t_0$ vs $t_1$ | *Tolecusatellitidae* | 1.97 | 0.020 |
| Gut | SSG | $t_0$ vs $t_1$ | *Tsukamurellaceae* | 0.88 | 0.044 |
| Gut | SSG | $t_0$ vs $t_1$ | *Veillonellaceae* | 2.60 | 0.003 |
| Gut | SSG | $t_0$ vs $t_1$ | *Virgaviridae* | 2.51 | 0.005 |
| Gut | SSG | $t_0$ vs $t_1$ | *Xanthomonadaceae* | 3.62 | 0.000 |
| Gut | SSG | $t_0$ vs $t_2$ | *Drexlerviridae* | -6.05 | 0.032 |
| Gut | SSG | $t_0$ vs $t_2$ | *Fortieaceae* | -10.77 | 0.000 |
| Gut | SSG | $t_0$ vs $t_2$ | *Methanoregulaceae* | -1.65 | 0.023 |
| Gut | SSG | $t_0$ vs $t_2$ | *Methanospirillaceae* | -1.78 | 0.023 |
| Gut | SSG | $t_0$ vs $t_2$ | *Picornaviridae* | -3.29 | 0.004 |
| Gut | SSG | $t_0$ vs $t_2$ | *Planococcaceae* | 1.76 | 0.023 |
| Gut | SSG | $t_0$ vs $t_2$ | *Rhabdoviridae* | 3.13 | 0.023 |

| | | | | | |
|---|---|---|---|---|---|
| Gut | SSG | $t_0$ vs $t_2$ | *Salinivirgaceae* | 1.05 | 0.023 |
| Gut | SSG | $t_0$ vs $t_2$ | *Veillonellaceae* | 3.25 | 0.001 |
| Gut | SSG | $t_1$ vs $t_2$ | *Ackermannviridae* | -3.46 | 0.028 |
| Gut | SSG | $t_1$ vs $t_2$ | *Alcaligenaceae* | -1.15 | 0.018 |
| Gut | SSG | $t_1$ vs $t_2$ | *Bogoriellaceae* | -1.31 | 0.009 |
| Gut | SSG | $t_1$ vs $t_2$ | *Caulobacteraceae* | -2.68 | 0.000 |
| Gut | SSG | $t_1$ vs $t_2$ | *Cellulomonadaceae* | -2.90 | 0.000 |
| Gut | SSG | $t_1$ vs $t_2$ | *Comamonadaceae* | -3.24 | 0.000 |
| Gut | SSG | $t_1$ vs $t_2$ | *Dietziaceae* | -1.00 | 0.018 |
| Gut | SSG | $t_1$ vs $t_2$ | *Drexlerviridae* | -6.61 | 0.009 |
| Gut | SSG | $t_1$ vs $t_2$ | *Gordoniaceae* | -1.17 | 0.017 |
| Gut | SSG | $t_1$ vs $t_2$ | *Microbacteriaceae* | -1.95 | 0.000 |
| Gut | SSG | $t_1$ vs $t_2$ | *Micrococcaceae* | -2.43 | 0.000 |
| Gut | SSG | $t_1$ vs $t_2$ | *Mycobacteriaceae* | -1.31 | 0.006 |
| Gut | SSG | $t_1$ vs $t_2$ | *Nocardiaceae* | -5.10 | 0.000 |
| Gut | SSG | $t_1$ vs $t_2$ | *Nocardioidaceae* | -2.04 | 0.000 |
| Gut | SSG | $t_1$ vs $t_2$ | *Promicromonosporaceae* | -2.31 | 0.000 |
| Gut | SSG | $t_1$ vs $t_2$ | *Pseudomonadaceae* | -2.26 | 0.031 |
| Gut | SSG | $t_1$ vs $t_2$ | *Sanguibacteraceae* | -1.92 | 0.000 |
| Gut | SSG | $t_1$ vs $t_2$ | *Tsukamurellaceae* | -1.18 | 0.005 |
| Gut | SSG | $t_1$ vs $t_2$ | *Virgaviridae* | -2.28 | 0.023 |
| Gut | SSG | $t_1$ vs $t_2$ | *Xanthomonadaceae* | -3.93 | 0.000 |
| Gut | SSG vs 5DG | $t_0$ | *Atopobiaceae* | 1.50 | 0.020 |
| Gut | SSG vs 5DG | $t_0$ | *Coriobacteriaceae* | 2.42 | 0.000 |
| Gut | SSG vs 5DG | $t_0$ | *Enterobacteriaceae* | 2.91 | 0.046 |
| Gut | SSG vs 5DG | $t_0$ | *Moraxellaceae* | -5.54 | 0.020 |
| Gut | SSG vs 5DG | $t_0$ | *Muribaculaceae* | 2.54 | 0.000 |
| Gut | SSG vs 5DG | $t_0$ | *Staphylococcaceae* | 2.23 | 0.000 |
| Gut | SSG vs 5DG | $t_0$ | *Streptococcaceae* | 2.25 | 0.001 |
| Gut | SSG vs 5DG | $t_1$ | *Aeromonadaceae* | 1.84 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Alloherpesviridae* | 2.36 | 0.011 |
| Gut | SSG vs 5DG | $t_1$ | *Autographiviridae* | 4.59 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Bacteroidaceae* | 5.30 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Bogoriellaceae* | -1.52 | 0.002 |
| Gut | SSG vs 5DG | $t_1$ | *Brachyspiraceae* | 0.71 | 0.011 |
| Gut | SSG vs 5DG | $t_1$ | *Bruguierivoracaceae* | 2.37 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Budviciaceae* | 1.93 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Cellulomonadaceae* | -2.91 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Chromatiaceae* | 1.46 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Chuviridae* | 2.53 | 0.000 |

| | | | | | |
|---|---|---|---|---|---|
| Gut | SSG vs 5DG | $t_1$ | *Clostridiaceae* | 2.01 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Comamonadaceae* | -1.80 | 0.022 |
| Gut | SSG vs 5DG | $t_1$ | *Demerecviridae* | 3.30 | 0.017 |
| Gut | SSG vs 5DG | $t_1$ | *Dermabacteraceae* | -0.71 | 0.047 |
| Gut | SSG vs 5DG | $t_1$ | *Dietziaceae* | -1.26 | 0.001 |
| Gut | SSG vs 5DG | $t_1$ | *Enterobacteriaceae* | 4.45 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Enterococcaceae* | 1.25 | 0.016 |
| Gut | SSG vs 5DG | $t_1$ | *Erwiniaceae* | 6.03 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Flavobacteriaceae* | 1.48 | 0.004 |
| Gut | SSG vs 5DG | $t_1$ | *Fortieaceae* | 9.22 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Gordoniaceae* | -1.43 | 0.002 |
| Gut | SSG vs 5DG | $t_1$ | *Hafniaceae* | 2.62 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Herelleviridae* | 4.06 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Hominidae* | 1.16 | 0.005 |
| Gut | SSG vs 5DG | $t_1$ | *Hydrogenophilaceae* | 0.78 | 0.050 |
| Gut | SSG vs 5DG | $t_1$ | *Hytrosaviridae* | 1.78 | 0.020 |
| Gut | SSG vs 5DG | $t_1$ | *Inoviridae* | 2.05 | 0.014 |
| Gut | SSG vs 5DG | $t_1$ | *Jatrophihabitantaceae* | -0.89 | 0.019 |
| Gut | SSG vs 5DG | $t_1$ | *Ktedonosporobacteraceae* | -0.72 | 0.039 |
| Gut | SSG vs 5DG | $t_1$ | *Kytococcaceae* | -1.03 | 0.005 |
| Gut | SSG vs 5DG | $t_1$ | *Lachnospiraceae* | 2.78 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Lipothrixviridae* | 2.39 | 0.002 |
| Gut | SSG vs 5DG | $t_1$ | *Microbacteriaceae* | -2.03 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Micrococcaceae* | -2.44 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Microviridae* | 7.01 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Morganellaceae* | 1.27 | 0.001 |
| Gut | SSG vs 5DG | $t_1$ | *Mycobacteriaceae* | -1.15 | 0.021 |
| Gut | SSG vs 5DG | $t_1$ | *Narnaviridae* | 2.38 | 0.001 |
| Gut | SSG vs 5DG | $t_1$ | *Neisseriaceae* | 0.93 | 0.018 |
| Gut | SSG vs 5DG | $t_1$ | *Nimaviridae* | 1.92 | 0.009 |
| Gut | SSG vs 5DG | $t_1$ | *Nocardiaceae* | -5.00 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Nocardioidaceae* | -2.01 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Nostocaceae* | 1.29 | 0.035 |
| Gut | SSG vs 5DG | $t_1$ | *Nudiviridae* | 2.49 | 0.014 |
| Gut | SSG vs 5DG | $t_1$ | *Odoribacteraceae* | 2.99 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Oxalobacteraceae* | 4.45 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Pasteurellaceae* | 1.44 | 0.009 |
| Gut | SSG vs 5DG | $t_1$ | *Pectobacteriaceae* | 3.20 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Peribunyaviridae* | 2.92 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Phenuiviridae* | 1.58 | 0.029 |

| | | | | | |
|---|---|---|---|---|---|
| Gut | SSG vs 5DG | $t_1$ | *Phycodnaviridae* | 2.60 | 0.001 |
| Gut | SSG vs 5DG | $t_1$ | *Pithoviridae* | 3.19 | 0.002 |
| Gut | SSG vs 5DG | $t_1$ | *Podoviridae* | 2.97 | 0.008 |
| Gut | SSG vs 5DG | $t_1$ | *Polydnaviridae* | 3.01 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Promicromonosporaceae* | -2.50 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Pseudomonadaceae* | 4.65 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Sanguibacteraceae* | -1.74 | 0.001 |
| Gut | SSG vs 5DG | $t_1$ | *Secoviridae* | 1.52 | 0.044 |
| Gut | SSG vs 5DG | $t_1$ | *Succinivibrionaceae* | -1.53 | 0.011 |
| Gut | SSG vs 5DG | $t_1$ | *Tannerellaceae* | 2.90 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Tobaniviridae* | 1.99 | 0.005 |
| Gut | SSG vs 5DG | $t_1$ | *Tsukamurellaceae* | -1.20 | 0.004 |
| Gut | SSG vs 5DG | $t_1$ | *Weeksellaceae* | 3.59 | 0.000 |
| Gut | SSG vs 5DG | $t_1$ | *Xanthomonadaceae* | -1.80 | 0.040 |
| Gut | SSG vs 5DG | $t_1$ | *Yersiniaceae* | 3.98 | 0.000 |
| Gut | SSG vs 5DG | $t_2$ | - | - | - |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Bacillaceae 2* | 3.43 | 0.043 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Bacillales incertae sedis XI* | 3.90 | 0.012 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Bifidobacteriaceae* | -3.92 | 0.006 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Carnobacteriaceae* | 4.77 | 0.004 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Fusobacteriaceae* | 4.81 | 0.025 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Geobacteraceae* | -1.72 | 0.035 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Labilitrichaceae* | -1.89 | 0.023 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Marinilabiliaceae* | -2.64 | 0.006 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Pasteurellaceae* | 6.55 | 0.004 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Promicromonosporaceae* | -2.21 | 0.048 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Rhizobiales incertae sedis* | -2.07 | 0.006 |
| Nasal | 5DG | $t_0$ vs $t_1$ | *Rhodobiaceae* | -2.17 | 0.005 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Aeromonadaceae* | -2.31 | 0.018 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Bifidobacteriaceae* | -4.53 | 0.002 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Brucellaceae* | -2.27 | 0.042 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Campylobacteraceae* | -3.17 | 0.002 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Coxiellaceae* | 2.36 | 0.003 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Gaiellaceae* | 2.47 | 0.008 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Halomonadaceae* | 4.58 | 0.001 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Marinilabiliaceae* | -2.53 | 0.008 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Methylocystaceae* | -2.09 | 0.015 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Neisseriaceae* | 3.24 | 0.015 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Nitrosomonadaceae* | 4.46 | 0.000 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Porphyromonadaceae* | -2.87 | 0.008 |

| | | | | | |
|---|---|---|---|---|---|
| Nasal | 5DG | $t_0$ vs $t_2$ | *Rhizobiales incertae sedis* | -1.93 | 0.012 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Rhodobiaceae* | -1.71 | 0.022 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Solirubrobacteraceae* | 2.71 | 0.002 |
| Nasal | 5DG | $t_0$ vs $t_2$ | *Thermoactinomycetaceae 1* | 2.26 | 0.042 |
| Nasal | 5DG | $t_1$ vs $t_2$ | *Neisseriaceae* | 3.43 | 0.035 |
| Nasal | 5DG | $t_1$ vs $t_2$ | *Nitrosomonadaceae* | 4.46 | 0.000 |
| Nasal | 5DG | $t_1$ vs $t_2$ | *Solirubrobacteraceae* | 2.41 | 0.017 |
| Nasal | SSG | $t_0$ vs $t_1$ | *Bacillales incertae sedis X* | -4.70 | 0.000 |
| Nasal | SSG | $t_0$ vs $t_1$ | *Leptotrichiaceae* | -2.43 | 0.013 |
| Nasal | SSG | $t_0$ vs $t_1$ | *Micrococcaceae* | 2.26 | 0.047 |
| Nasal | SSG | $t_0$ vs $t_1$ | *Nocardiaceae* | 2.12 | 0.034 |
| Nasal | SSG | $t_0$ vs $t_1$ | *Sporichthyaceae* | -2.30 | 0.001 |
| Nasal | SSG | $t_0$ vs $t_1$ | *Succinivibrionaceae* | -4.02 | 0.000 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Acidaminococcaceae* | 1.69 | 0.035 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Bacillaceae 1* | -3.02 | 0.017 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Bacillales incertae sedis X* | -4.58 | 0.000 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Bacillales incertae sedis XII* | -2.07 | 0.011 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Bacteroidaceae* | 2.41 | 0.047 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Bogoriellaceae* | 2.12 | 0.035 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Burkholderiaceae* | -1.67 | 0.014 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Chitinophagaceae* | -1.92 | 0.035 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Chromatiaceae* | 2.73 | 0.003 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Clostridiales incertae sedis III* | -1.79 | 0.014 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Cryomorphaceae* | 1.96 | 0.035 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Dermatophilaceae* | 4.48 | 0.000 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Dietziaceae* | 2.32 | 0.035 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Enterococcaceae* | -2.45 | 0.019 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Fusobacteriaceae* | -3.27 | 0.036 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Gaiellaceae* | -2.07 | 0.011 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Gemmatimonadaceae* | -1.54 | 0.042 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Geodermatophilaceae* | 2.71 | 0.004 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Halomonadaceae* | 2.92 | 0.006 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Lactobacillaceae* | -2.83 | 0.025 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Leptotrichiaceae* | -2.48 | 0.004 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Micrococcaceae* | 2.13 | 0.025 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Oligosphaeraceae* | 1.50 | 0.037 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Rikenellaceae* | 2.78 | 0.001 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Sandaracinaceae* | 2.77 | 0.000 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Sanguibacteraceae* | 2.16 | 0.018 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Sphingobacteriaceae* | -1.93 | 0.035 |

| | | | | | |
|---|---|---|---|---|---|
| Nasal | SSG | $t_0$ vs $t_2$ | *Sporichthyaceae* | -2.26 | 0.001 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Streptomycetaceae* | -2.39 | 0.003 |
| Nasal | SSG | $t_0$ vs $t_2$ | *Succinivibrionaceae* | -3.08 | 0.002 |
| Nasal | SSG | $t_1$ vs $t_2$ | *Christensenellaceae* | -2.38 | 0.006 |
| Nasal | SSG | $t_1$ vs $t_2$ | *Chromatiaceae* | 3.77 | 0.000 |
| Nasal | SSG | $t_1$ vs $t_2$ | *Dermatophilaceae* | 3.44 | 0.000 |
| Nasal | SSG | $t_1$ vs $t_2$ | *Paenibacillaceae 1* | -2.31 | 0.015 |
| Nasal | SSG | $t_1$ vs $t_2$ | *Sandaracinaceae* | 2.66 | 0.000 |
| Nasal | SSG | $t_1$ vs $t_2$ | *Sphingobacteriaceae* | -2.25 | 0.032 |
| Nasal | SSG vs 5DG | $t_0$ | *Acetobacteraceae* | 1.77 | 0.031 |
| Nasal | SSG vs 5DG | $t_0$ | *Acholeplasmataceae* | 1.85 | 0.036 |
| Nasal | SSG vs 5DG | $t_0$ | *Bacillaceae 1* | -3.11 | 0.031 |
| Nasal | SSG vs 5DG | $t_0$ | *Bacillales incertae sedis X* | -4.42 | 0.000 |
| Nasal | SSG vs 5DG | $t_0$ | *Bacillales incertae sedis XI* | -3.22 | 0.021 |
| Nasal | SSG vs 5DG | $t_0$ | *Bacillales incertae sedis XII* | -2.27 | 0.015 |
| Nasal | SSG vs 5DG | $t_0$ | *Bacteroidaceae* | -2.84 | 0.039 |
| Nasal | SSG vs 5DG | $t_0$ | *Burkholderiaceae* | -2.36 | 0.003 |
| Nasal | SSG vs 5DG | $t_0$ | *Campylobacteraceae* | 2.57 | 0.003 |
| Nasal | SSG vs 5DG | $t_0$ | *Carnobacteriaceae* | -4.34 | 0.003 |
| Nasal | SSG vs 5DG | $t_0$ | *Clostridiales incertae sedis III* | -1.84 | 0.023 |
| Nasal | SSG vs 5DG | $t_0$ | *Clostridiales incertae sedis XII* | 1.77 | 0.021 |
| Nasal | SSG vs 5DG | $t_0$ | *Cryomorphaceae* | 2.28 | 0.018 |
| Nasal | SSG vs 5DG | $t_0$ | *Enterococcaceae* | -2.67 | 0.025 |
| Nasal | SSG vs 5DG | $t_0$ | *Family IV* | 1.55 | 0.037 |
| Nasal | SSG vs 5DG | $t_0$ | *Fibrobacteraceae* | 1.97 | 0.023 |
| Nasal | SSG vs 5DG | $t_0$ | *Fusobacteriaceae* | -7.87 | 0.000 |
| Nasal | SSG vs 5DG | $t_0$ | *Geobacteraceae* | 1.81 | 0.015 |
| Nasal | SSG vs 5DG | $t_0$ | *Labilitrichaceae* | 2.18 | 0.003 |
| Nasal | SSG vs 5DG | $t_0$ | *Marinilabiliaceae* | 3.08 | 0.000 |
| Nasal | SSG vs 5DG | $t_0$ | *Methylocystaceae* | 1.79 | 0.018 |
| Nasal | SSG vs 5DG | $t_0$ | *Micromonosporaceae* | 1.55 | 0.037 |
| Nasal | SSG vs 5DG | $t_0$ | *Pasteurellaceae* | -7.85 | 0.000 |
| Nasal | SSG vs 5DG | $t_0$ | *Phyllobacteriaceae* | 1.79 | 0.044 |
| Nasal | SSG vs 5DG | $t_0$ | *Planctomycetaceae* | 1.87 | 0.037 |
| Nasal | SSG vs 5DG | $t_0$ | *Polyangiaceae* | 1.51 | 0.030 |
| Nasal | SSG vs 5DG | $t_0$ | *Porphyromonadaceae* | 2.31 | 0.014 |
| Nasal | SSG vs 5DG | $t_0$ | *Promicromonosporaceae* | 1.74 | 0.039 |
| Nasal | SSG vs 5DG | $t_0$ | *Pseudonocardiaceae* | 1.59 | 0.045 |
| Nasal | SSG vs 5DG | $t_0$ | *Rhizobiales incertae sedis* | 1.61 | 0.018 |
| Nasal | SSG vs 5DG | $t_0$ | *Rhodobiaceae* | 1.67 | 0.011 |

| | | | | | |
|---|---|---|---|---|---|
| Nasal | SSG vs 5DG | $t_0$ | *Rikenellaceae* | 2.02 | 0.031 |
| Nasal | SSG vs 5DG | $t_0$ | *Sanguibacteraceae* | 1.94 | 0.037 |
| Nasal | SSG vs 5DG | $t_0$ | *Sporichthyaceae* | -2.00 | 0.011 |
| Nasal | SSG vs 5DG | $t_0$ | *Succinivibrionaceae* | -3.65 | 0.002 |
| Nasal | SSG vs 5DG | $t_0$ | *Veillonellaceae* | -2.80 | 0.015 |
| Nasal | SSG vs 5DG | $t_0$ | *Xanthobacteraceae* | 1.56 | 0.031 |
| Nasal | SSG vs 5DG | $t_1$ | - | - | - |
| Nasal | SSG vs 5DG | $t_2$ | *Bacteroidaceae* | -4.21 | 0.003 |
| Nasal | SSG vs 5DG | $t_2$ | *Bogoriellaceae* | -3.69 | 0.002 |
| Nasal | SSG vs 5DG | $t_2$ | *Burkholderiales incertae sedis* | 1.90 | 0.023 |
| Nasal | SSG vs 5DG | $t_2$ | *Coriobacteriaceae* | -2.12 | 0.038 |
| Nasal | SSG vs 5DG | $t_2$ | *Coxiellaceae* | 2.32 | 0.001 |
| Nasal | SSG vs 5DG | $t_2$ | *Cytophagaceae* | 3.20 | 0.002 |
| Nasal | SSG vs 5DG | $t_2$ | *Dermatophilaceae* | -4.38 | 0.000 |
| Nasal | SSG vs 5DG | $t_2$ | *Dietziaceae* | -4.06 | 0.002 |
| Nasal | SSG vs 5DG | $t_2$ | *Gaiellaceae* | 2.85 | 0.000 |
| Nasal | SSG vs 5DG | $t_2$ | *Lactobacillaceae* | 3.08 | 0.023 |
| Nasal | SSG vs 5DG | $t_2$ | *Neisseriaceae* | 2.72 | 0.022 |
| Nasal | SSG vs 5DG | $t_2$ | *Nitrosomonadaceae* | 4.42 | 0.000 |
| Nasal | SSG vs 5DG | $t_2$ | *Sandaracinaceae* | -2.61 | 0.003 |
| Nasal | SSG vs 5DG | $t_2$ | *Solirubrobacteraceae* | 2.69 | 0.000 |
| Nasal | SSG vs 5DG | $t_2$ | *Sphingobacteriaceae* | 2.11 | 0.038 |

**Table A.3: Differential Families (Gut & Nasal Microbiome).** List of differential taxa calculated between the SSG and 5DG across all time points using edgeR. The overview contains all significant hits with adjusted p-value (FDR) $\leq 0.05$. FC: Fold Change, FDR: False Discovery Rate, MBO: Microbiome, SSG: Single-Shot Group, 5DG: 5-Day Group.

| Software | ARG | ARG Class | ID | COV/ PROB |
|---|---|---|---|---|
| ABRicate | aac3 | aminoglycoside | 99.87 | 100.00 |
| ABRicate | aac3 | aminoglycoside | 100.00 | 100.00 |
| ABRicate | aph2-dprime | aminoglycoside | 100.00 | 100.00 |
| ABRicate | aph3-dprime | aminoglycoside | 99.63 | 100.00 |
| ABRicate | aph4 | aminoglycoside | 99.90 | 100.00 |
| ABRicate | aph6 | aminoglycoside | 100.00 | 100.00 |
| ABRicate | carb | beta-lactam | 99.67 | 100.00 |
| ABRicate | ermc | MLS | 100.00 | 100.00 |
| ABRicate | suli | sulfonamide | 100.00 | 100.00 |
| ABRicate | sulii | sulfonamide | 100.00 | 100.00 |
| ABRicate | dfrg | trimethoprim | 100.00 | 94.00 |
| DeepARG | mdtb | aminocoumarin | 58.80 | 0.36 |
| DeepARG | abes | aminoglycoside | 57.40 | 0.57 |
| DeepARG | acra | aminoglycoside | 51.40 | 0.19 |
| DeepARG | acrb | aminoglycoside | 58.50 | 0.20 |
| DeepARG | acrb | aminoglycoside | 55.70 | 0.26 |
| DeepARG | adea | aminoglycoside | 59.20 | 0.26 |
| DeepARG | arlr | aminoglycoside | 50.40 | 0.29 |
| DeepARG | dfra3 | aminoglycoside | 50.30 | 0.19 |
| DeepARG | dfra3 | aminoglycoside | 50.90 | 0.19 |
| DeepARG | emrb | aminoglycoside | 50.00 | 0.25 |
| DeepARG | mexe | aminoglycoside | 53.00 | 0.22 |
| DeepARG | mtra | aminoglycoside | 66.80 | 0.59 |
| DeepARG | ompr | aminoglycoside | 50.20 | 0.07 |
| DeepARG | ompr | aminoglycoside | 50.60 | 0.07 |
| DeepARG | oprm | aminoglycoside | 51.50 | 0.13 |
| DeepARG | qach | aminoglycoside | 59.60 | 0.58 |
| DeepARG | smed | aminoglycoside | 53.70 | 0.11 |
| DeepARG | vans | aminoglycoside | 55.30 | 0.35 |
| DeepARG | mexe | aminoglycoside aminocoumarin | 66.40 | 0.33 |
| DeepARG | dfrg | diaminopyrimidine | 100.00 | 0.78 |
| DeepARG | baca | fluoroquinolone | 65.50 | 0.42 |
| DeepARG | baca | fluoroquinolone | 51.00 | 0.26 |
| DeepARG | emrb | fluoroquinolone | 54.00 | 0.07 |
| DeepARG | emrb | fluoroquinolone | 56.50 | 0.07 |
| DeepARG | emre | fluoroquinolone | 52.30 | 0.63 |
| DeepARG | lrfa | fluoroquinolone | 52.00 | 0.35 |

| | | | | |
|---|---|---|---|---|
| DeepARG | qach | fluoroquinolone | 71.90 | 0.35 |
| DeepARG | qach | fluoroquinolone | 58.90 | 0.67 |
| DeepARG | fosa5 | fosfomycin | 57.10 | 0.78 |
| DeepARG | carb | glycopeptide | 100.00 | 0.28 |
| DeepARG | mexd | MLS | 58.00 | 0.10 |
| DeepARG | acra | multidrug | 51.40 | 0.37 |
| DeepARG | acrb | multidrug | 58.50 | 0.72 |
| DeepARG | acrb | multidrug | 55.70 | 0.37 |
| DeepARG | adea | multidrug | 59.20 | 0.35 |
| DeepARG | aph(3")-i | multidrug | 100.00 | 0.04 |
| DeepARG | baca | multidrug | 50.40 | 0.25 |
| DeepARG | baca | multidrug | 51.00 | 0.35 |
| DeepARG | dfra22 | multidrug | 51.10 | 0.14 |
| DeepARG | emrb | multidrug | 50.00 | 0.34 |
| DeepARG | emre | multidrug | 52.30 | 0.24 |
| DeepARG | mdtb | multidrug | 58.80 | 0.64 |
| DeepARG | mtra | multidrug | 66.80 | 0.26 |
| DeepARG | oprm | multidrug | 51.50 | 0.72 |
| DeepARG | rbpa | multidrug | 52.60 | 0.23 |
| DeepARG | smed | multidrug | 53.70 | 0.64 |
| DeepARG | vans | multidrug | 55.30 | 0.32 |
| DeepARG | aac(3)-iiia | peptide | 59.70 | 0.04 |
| DeepARG | abes | peptide | 57.40 | 0.17 |
| DeepARG | baca | peptide | 65.50 | 0.15 |
| DeepARG | baca | peptide | 50.40 | 0.22 |
| DeepARG | carb | peptide | 100.00 | 0.28 |
| DeepARG | cpxr | peptide | 51.80 | 0.29 |
| DeepARG | cpxr | peptide | 52.60 | 0.07 |
| DeepARG | dfra22 | peptide | 51.10 | 0.34 |
| DeepARG | dfra3 | peptide | 50.30 | 0.30 |
| DeepARG | dfra3 | peptide | 50.90 | 0.30 |
| DeepARG | dfrg | peptide | 100.00 | 0.12 |
| DeepARG | emrb | peptide | 54.00 | 0.50 |
| DeepARG | emrb | peptide | 56.50 | 0.51 |
| DeepARG | fosa5 | peptide | 57.10 | 0.19 |
| DeepARG | mexe | peptide | 53.00 | 0.23 |
| DeepARG | qach | peptide | 71.90 | 0.25 |
| DeepARG | qach | peptide | 58.90 | 0.11 |
| DeepARG | qach | peptide | 59.60 | 0.13 |
| DeepARG | rbpa | peptide | 52.60 | 0.26 |

| | | | | |
|---|---|---|---|---|
| DeepARG | vatf | peptide | 68.30 | 0.04 |
| DeepARG | lrfa | tetracycline | 52.00 | 0.15 |
| DeepARG | mexe | tetracycline | 66.40 | 0.31 |
| DeepARG | arlr | unclassified | 50.40 | 0.34 |
| DeepARG | cpxr | unclassified | 51.80 | 0.28 |
| DeepARG | cpxr | unclassified | 52.60 | 0.74 |
| DeepARG.potential | aac(3)-ii | aminoglycoside | 68.50 | 1.00 |
| DeepARG.potential | aac(3)-ii | aminoglycoside | 100.00 | 1.00 |
| DeepARG.potential | aac(3)-iiia | aminoglycoside | 59.70 | 0.88 |
| DeepARG.potential | aac(3)-iv | aminoglycoside | 100.00 | 1.00 |
| DeepARG.potential | aadk | aminoglycoside | 61.30 | 1.00 |
| DeepARG.potential | aph(2")-iii | aminoglycoside | 100.00 | 1.00 |
| DeepARG.potential | aph(3")-i | aminoglycoside | 100.00 | 0.87 |
| DeepARG.potential | aph(3')-i | aminoglycoside | 99.60 | 1.00 |
| DeepARG.potential | aph(3')-i | aminoglycoside | 67.90 | 1.00 |
| DeepARG.potential | aph(4)-i | aminoglycoside | 99.70 | 1.00 |
| DeepARG.potential | aph(6)-i | aminoglycoside | 100.00 | 1.00 |
| DeepARG.potential | emre | aminoglycoside | 61.30 | 0.98 |
| DeepARG.potential | kdpe | aminoglycoside | 51.60 | 0.99 |
| DeepARG.potential | ksga | aminoglycoside | 80.00 | 1.00 |
| DeepARG.potential | ksga | aminoglycoside | 53.10 | 1.00 |
| DeepARG.potential | baca | bacitracin | 69.30 | 0.99 |
| DeepARG.potential | ctx-m | beta-lactam | 50.80 | 0.97 |
| DeepARG.potential | oxa | beta-lactam | 57.30 | 1.00 |
| DeepARG.potential | tem | beta-lactam | 100.00 | 0.97 |
| DeepARG.potential | rosa | fosmidomycin | 60.10 | 1.00 |
| DeepARG.potential | brp(mbl) | glycopeptide | 58.30 | 0.99 |
| DeepARG.potential | brp(mbl) | glycopeptide | 57.90 | 0.99 |
| DeepARG.potential | vanr | glycopeptide | 61.30 | 1.00 |
| DeepARG.potential | vanr | glycopeptide | 68.90 | 1.00 |
| DeepARG.potential | vanr | glycopeptide | 64.70 | 1.00 |
| DeepARG.potential | ermc | MLS | 100.00 | 1.00 |
| DeepARG.potential | vatf | MLS | 68.30 | 0.89 |
| DeepARG.potential | acra | multidrug | 50.80 | 1.00 |
| DeepARG.potential | acrb | multidrug | 53.20 | 0.92 |
| DeepARG.potential | gols | multidrug | 56.30 | 1.00 |
| DeepARG.potential | marr | multidrug | 84.50 | 1.00 |
| DeepARG.potential | mexd | multidrug | 58.00 | 0.85 |
| DeepARG.potential | mexe | multidrug | 51.20 | 1.00 |
| DeepARG.potential | mexe | multidrug | 50.30 | 0.99 |

| | | | | |
|---|---|---|---|---|
| DeepARG.potential | mexe | multidrug | 51.00 | 0.92 |
| DeepARG.potential | mexf | multidrug | 52.80 | 0.96 |
| DeepARG.potential | ompr | multidrug | 65.80 | 1.00 |
| DeepARG.potential | ompr | multidrug | 50.20 | 0.86 |
| DeepARG.potential | ompr | multidrug | 50.60 | 0.86 |
| DeepARG.potential | ompr | multidrug | 50.40 | 0.99 |
| DeepARG.potential | ompr | multidrug | 52.70 | 0.99 |
| DeepARG.potential | ompr | multidrug | 56.10 | 0.99 |
| DeepARG.potential | ompr | multidrug | 52.70 | 0.99 |
| DeepARG.potential | rpob2 | multidrug | 69.50 | 1.00 |
| DeepARG.potential | chloramphenicol_exporter | phenicol | 100.00 | 1.00 |
| DeepARG.potential | sul1 | sulfonamide | 100.00 | 0.99 |
| DeepARG.potential | sul2 | sulfonamide | 100.00 | 0.97 |
| DeepARG.potential | teta(48) | tetracycline | 50.60 | 0.92 |

**Table A.4: Resistance Components (Nasal Microbiome).** Overview of resistance elements identified within the nasal shotgun metagenome sample (J32522). Both ABRicate and DeepARG were utilized for AMR profiling. This table includes the software used to infer genotypic resistances, the identified AMR genes, their corresponding resistance classes, as well as the identity and likelihood of each hit (coverage/probability). ABRicate determines the identity and coverage as percentages of each hit compared to the MEGARes database, while DeepARG does not infer coverage and instead provides probabilities. DeepARG also provides novel resistance elements (labelled as potential within the table). ARG: Antibiotic Resistance Gene, ID: Identity, COV: Coverage, PROB: Probability.