Bioinformatics approaches to study antibiotics resistance emergence across levels of biological organization.

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M. Sc. Anna Górska
aus Warschau

Tübingen 2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 31.01.2019

Dekan: Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter: Prof. Dr. Daniel Huson
2. Berichterstatter: Prof. Dr. Kay Nieselt
3. Berichterstatter: Prof. Dr. Knut Reinert

Abstract

The *Review on Antimicrobial Resistance* predicts that in thirty years infections with antibiotic-resistant microorganisms will become one of the leading causes of death. The discovery of new antibiotics has so far been too slow to ensure continuous use of antibiotics in the face of growing resistance. Therefore, efforts to curb resistance emergence gain in importance. These efforts comprise two complementary strategies. The first focuses on the mechanisms of resistance emergence, in the hope that it would enable development of pharmacological agents constraining resistance emergence. The second aims at improving antibiotic use practices, based on studies of the impact of antibiotics on resistance emergence within patient populations. Antibiotic resistance emerges in bacterial cells, negatively influences the human gut microbiome, and transfers between people. Hence, antibiotic resistance has impacts across several levels of biological organization.

This thesis describes four projects, which concerned various aspects of antibiotics resistance. The first two projects deal with basic resistance emergence mechanisms, on the level of bacterial strains and bacterial consortia, whereas the other two deal with finding better practices for antibiotic use on a population level.

During the first project, I analyzed changes in genomes of MRSA strains isolated from several patients throughout antibiotic therapies and developing MRSA infections. I observed changes in number and types of virulence factors responsible for interacting with the human body, which are attributed to mobile genetic elements. In the second project, I showed that, prompted by antibiotic therapy, within the human gut microbiome resistance transfers from bacterial genomes onto plasmids, prophages, and free phages. Hence, resistance emergence depends not only on the antibiotic therapy but also on the state of the gut microbiome, which again results from the patients' overall health and previous antibiotic therapies.

The third project, SATURN, employed machine learning methods for a large set of data regarding patients' demographics, comorbidities, antibiotic therapies, surgeries, and colonization with multi-drug resistant bacteria. The final classifiers were made available on the `AskSaturn` website where the doctors can compare antibiotic therapies based on the probability of colonization with multi-drug resistant bacteria. The fourth project, Tübiom, focused on the antibiotic-influenced gut microbiomes of the healthy population.

The first two projects rely on genome and metagenome sequencing data. For them, I designed specialized bioinformatics analysis pipelines. The latter two projects use mixed data, which were analyzed with machine learning algorithms. These projects also involved web development and data visualization. Although each of the projects requires different data and methods, each of them provides a crucial part in a pipeline aiming at utilizing gut microbiome information in medical practice to constrain resistance emergence.

Acknowledgements

Foremost, I would like to thank my *Doktorvater* Prof. Daniel Huson, for creating a great lab and for all his guidance throughout my time as his student. I hope one day I would have a chance to be as great a mentor as Prof. Huson was to me.

Completing work on any of the projects would not be possible without the expertise and patience of all of the collaborators: Dr. med. Silke Peter and Prof. Matthias Willmann, Prof. Evelina Tacconelli and Dr. Primrose Beryl, and Prof. Mihai Pop.

Further, I would like to thank members of my advisory committee, especially Prof. Dr. Richard Neher, for the insightful discussions and perspective.

I would also like to acknowledge to the International Max Planck Research School, for providing the funding, and the program coordinator Sarah Danes for her support during this time.

I thank my fellow lab mates and friends from Zentrum für Bioinformatik for all of their help in research and teaching, and, even more so, coffee-centered emotional support.

Finally, a special thanks to my husband, writing this thesis would not have been possible without his love and care.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The *Review on Antimicrobial Resistance* predicts that in 2050 ten million people worldwide will lose their life due to infections with antibiotic-resistant microorganisms (AMR) [1]. It is an extreme increase in comparison to $\sim 700,000$ deaths caused by infections with AMR in 2014. This increase is driven by the overuse of antibiotics in medicine and agriculture causing emergence of antibiotic resistance [2]. Already in 2014 the WHO classified antibiotics resistance as a severe threat to public health [3], as resistance causes billions of euros of healthcare costs and millions of excessive days of hospital stay for patients in the US and EU alone [4].

The fight against antibiotics resistance is carried out on two fronts: the discovery of novel antibiotics and constraining resistance emergence. New antibiotic discovery is so slow and laborious, that out of the eighteen biggest pharmaceutical companies, only three keep working in this field [5]. Even when new compounds are introduced, resistance soon curbs their therapeutic potential. Consequently, we are in a desperate need to advance in the latter front. Combating the emergence of resistance would ensure continuous safe use of known antibiotics and the future use of those newly developed.

On the one hand, resistance emergence can be restrained by introducing better antibiotic use practices and education of both doctors and patients. Those actions are termed antibiotic-stewardship. Since therapeutic and resistance-emergence effects depend on the patients' characteristics, data-driven solutions to suggest therapies minimizing the probability of antibiotic resistance emergence are needed. On the other hand, future pharmacological solutions could potentially decrease resistance. Still, details of mechanisms of resistance emergence remain unknown.

Concentrations of therapeutic antibiotics in the human body are orders of magnitude higher than those occurring naturally in the environment [6]. Therefore, they exert high ecological pressure on the gut microbiomes resulting in a series of undesirable side effects. First, they cause perturbations of the taxonomic composition of the human gut microbiome and its internal metabolic processes. Second, they decrease colonization resistance, freeing space for potential

pathogens [7, 8], and cause a rapid increase in both the overall diversity of antibiotic resistance genes, termed resistome and the frequency of horizontal gene transfer (HGT).

Strengthening of the resistome worsens the prognosis for future antibiotic therapies and facilitates the accumulation of antibiotic resistance genes (ARG) in bacterial cells, leading to the emergence of multi-drug resistant organisms (MDR), which could include dangerous pathogens. Since the human organism is not a closed system, MDR transfer to other people, causing resistance in the population. Consequently, the phenomenon of antibiotic resistance affects all levels of biological organization.

Projects presented in this thesis furthered both approaches of the fight against resistance emergence. The first two projects focused on resistance emergence mechanisms, the other two projects on developing better practices of antibiotic usage. The projects concerned multiple levels of biological organization, as presented in Fig. 1.1.

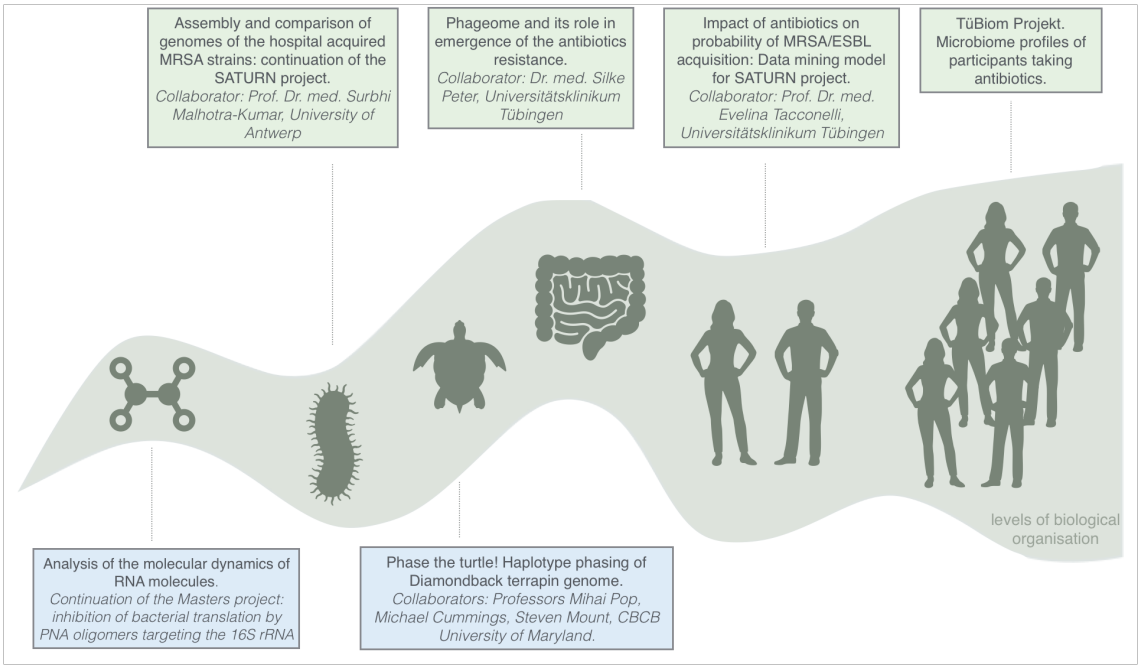


Figure 1.1: Projects discussed in the thesis ordered by levels of biological organization. The top row denotes the main projects concerning antibiotic resistance emergence. Bottom row enumerates other projects unrelated to antibiotic resistance described in Chapter 7.

On the lowest level, an antibiotic interacts with its molecular target in a bacterial cell. The target could be modified so that it disables that interaction. Other cell-level resistance mechanisms, such as chemical modification and efflux prevent an antibiotic from meeting its target. Resistant genes are encoded on bacterial genomes

or plasmids. Their quality and quantity determine the survival or death of cells. Changes in the bacterial genomes during antibiotic therapy are the focus of the project described in Chapter 3. I investigated the resistance and virulence genes in the genomes of the hospital-acquired methicillin-resistant *Staphylococcus aureus* (MRSA) strains isolated from patients undergoing antibiotic therapies. The strains were isolated from patient samples and sequenced. The project relied on the genome assembly and annotation.

The next organization level concerns bacterial consortia, such as the gut microbiome, which plays a central role in the emergence of resistance. Resistant genes transfer between bacterial cells through horizontal gene transfer and as a result are retained in the patient's gut. That is the focus of the second project described in Chapter 4. I analyzed how the resistance transfers between the bacterial genomes, and mobile genetic elements, during antibiotic therapy and 30 days of recovery. I analyzed the whole-genome-sequencing and phageome-only sequencing of the gut microbiome.

The two next projects concern the population level. The SATURN project described in Chapter 5 analyzed data of 10,000 hospitalized patients. I attempted to determine a relationship between patient's characteristics and the administered treatments to the probability of colonization with the multi-drug resistant bacteria. The second project, Tübiom aimed at investigating the healthy population. I wanted to check if there are lifestyle choices promoting microbiomes that were more resistant towards the antibiotic therapy (Chapter 6).

Although not all of the projects shared the same biological question, they were often connected by methodological aspects. Each project required the development of a novel pipeline. However, sometimes it was possible to cross-apply some of the methods, e.g., the k-mer processing required in the *Phase Turtle!* project was also applied to analyze the phageome and microbiome datasets in Chapter 4. Therefore, in the last Chapter 7 the two side projects, dealing with other biological questions than antibiotic-resistance are shortly described.

Chapter 2

Background

2.1 Mechanisms of antibiotics action and resistance

The term *antibiotic* refers to natural or synthetic compounds that in low concentrations restrict the growth of bacterial cells, or cause their death [9]. Since their discovery, antibiotics have become an indispensable tool in everyday medical practice, employed to fight pathogenic bacteria infecting the human body. The antibiotic era started with a series of lucky coincidences in the lab of Alexander Flemming in 1928. It took another twelve years before Penicillin was commercially available, and before antibiotics could transform medical practice. At that time, it seemed that bacterial infections were contained forever [10]. However, for the majority of antibiotics, resistance was observed shortly after their introduction into commercial usage (Fig. 2.1).

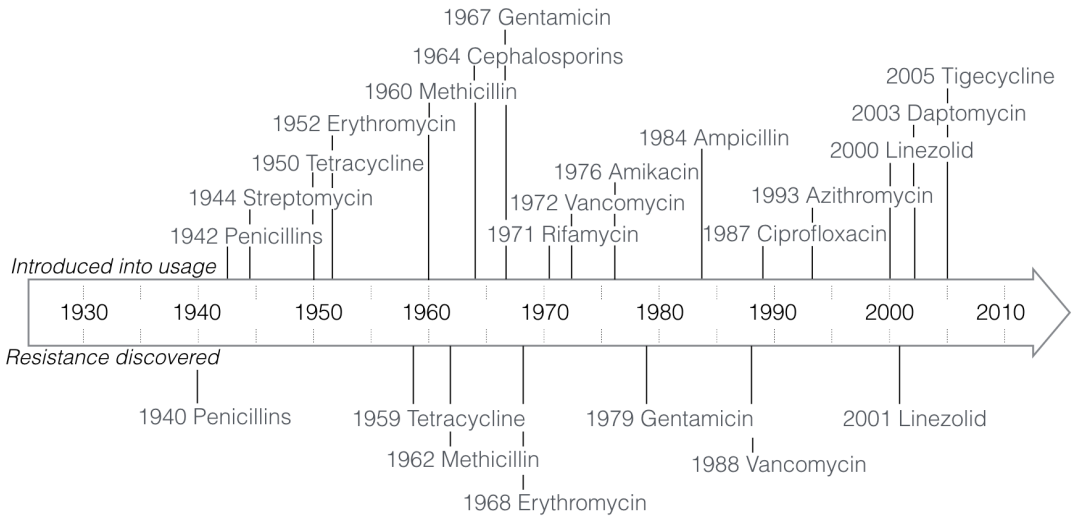


Figure 2.1: Timeline of antibiotic introduction to commercial usage and discovery of resistance [3, 11].

Nature has invented many diverse antimicrobial compounds, that scientists have extended and modified. They are classified according to their chemical makeup, mechanism of action, effect on the bacterial cell, spectrum of activity and medical application. Bacteria fall into two broad groups according to the structure of the cell envelope [12]. Gram-positive bacteria have a thick peptidoglycan cell wall (retaining the crystal violet stain), while Gram-negative bacteria have a thin wall with an outer membrane. The two main groups of bacteria also determine the first classification of antibiotics: broad-spectrum antibiotics which are active against both Gram-positive and -negative bacteria, and narrow spectrum which work only against a specific group of bacteria.

The second classification is determined by an effect the antibiotic has on the bacterial cell. Antibiotic is classified as bactericidal when it causes the cell death or bacteriostatic if it inhibits essential processes causing bacteria to stop growing and dividing [13]. When prescribing bacteriostatic antibiotic doctors rely on the patient's immune system to destroy the pathogenic cells. Researchers criticize this classification as too simplistic since none of the antibiotics can be fully either bacteriostatic or bactericidal. In the end, the antibiotic effect depends not only on the molecular mechanism of action but also on the bacterial species. Moreover, antibiotics are classified based on *in vitro* studies, which might not correspond with their *in vivo* activity. Therefore, rather than just belonging to one of the classes, each antibiotic has potential to be both bactericidal and bacteriostatic depending on the microbial and non-microbial conditions [14].

Lastly, antibiotics are grouped based on their molecular targets and action mechanisms. This classification is the most important for this thesis as an antibiotic resistance mechanism depends on the mechanism of action. Fig. 2.2 presents major chemical groups of antibiotics separated by the general mechanism of action discussed in the following subsections.

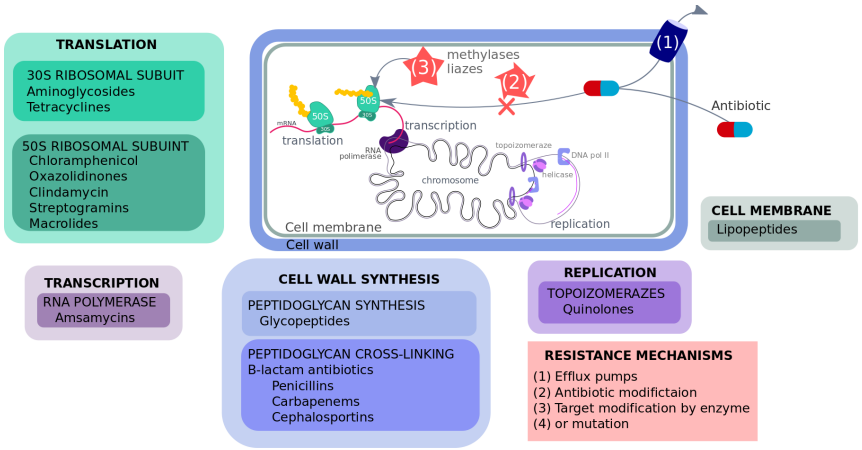


Figure 2.2: Classification of antibiotics according to the chemical groups, effects on the bacterial cell and action mechanisms.

Transcription inhibitors

Transcription and translation are the two steps of gene expression, both of which are vital and conserved processes in bacteria, and consequently, are excellent targets for antibiotics [15]. Transcription is a process of synthesizing an RNA molecule complementary to one of the DNA strands, performed by a complex of RNA polymerase (RNAP) consisting of $\alpha_2\beta\beta'\omega$ protein subunits [16, 17]. The RNA polymerase moves along a DNA molecule, unwinds the helix and catalyzes the formation of the bonds between nucleotides to form a new RNA molecule.

The transcripts can be mRNAs, substrates for translation, but also other functional RNAs. The antibiotics that target transcription bind to the β subunit and block either elongation of the RNA chain, like Rifamycins or interaction of RNAP with a promoter on the DNA molecule [18, 19].

Translation inhibitors

Translation involves a single mRNA molecule, multiple tRNAs carrying amino acids, and protein factors facilitating initiation (IF1, IF2, IF3), elongation (EF-G and EF-Tu), and release (RF1, RF2, RF3), and, most importantly, a ribosome. It is a complex of the large (50S) and small (30S) subunits, which consist of a rRNA-based scaffold and several accompanying proteins. The large subunit is built of two rRNA molecules (5S and 23S rRNAs) and 31 proteins, and the small subunit has one rRNA molecule (16S rRNA) and 21 proteins. RNA constitutes a majority of the ribosome's weight and is the primary target for a large group of chemically diverse antibiotics.

Translation consists of three steps: initiation, elongation, and termination [20]. In the initiation step, the ribosome assembles on the start codon on the mRNA molecule. First, the IF3 binds to the 30S ribosomal subunit. Next, the 3' end of the 16S rRNA molecule pairs with the Shine-Dalgarno sequence upstream from the start codon, and the 30S/mRNA complex is formed. Finally, the binding of the complex to the 50S ribosomal subunit facilitated by the IF2 [21].

A single elongation cycle, adds one amino acid to the new peptide. It takes three steps [22]. The first step is decoding, when aminoacyl-tRNA is placed on the ribosomal A aminoacyl site (A-site), with the help of the EF-Tu factor. Next, the peptidyl-transferase center (PTC) catalyzes the formation of the peptide bond between an amino acid on the tRNA in the A-site and the last amino acid of the new polypeptide chain attached to the tRNA in the P-site. The second step is transferring the new polypeptide from the tRNA bound in the peptidyl site (P-site) onto the tRNA in the A-site. Eventually, in the last step, the EF-G facilitates translocation of the tRNAs between A and P, and P and E sites, leaving the A-site empty so another cycle can begin. Elongation continues until a stop codon is encountered, it is terminated with the help of the two protein release factors 1 and 2 (RF1 and RF2). They recognize the stop codons and cleave the polypeptide from the tRNA in the P site.

Antibiotics targeting translation had been studied for decades, but their molecular action mechanisms could be finally described in detail using crystallography. Almost all translation inhibitors interact with rRNA and bind to, or near one of the significant sites: A-site on the 16S rRNA, the catalytic center of the ribosome, the PTC located on the 23S rRNA and peptide exit tunnel on the 23S rRNA [23, 24]. Binding sites of the several antibiotic groups overlap with each other [20], so they share action and resistance mechanisms.

Aminoglycosides and typical tetracyclines constitute large classes of broad-spectrum antibiotics that target the A-site [25]. Aminoglycosides are bacteriostatic against Gram-negative and bacteriocidal against Gram-positives. Typical tetracyclines are primarily bacteriostatic. They include several naturally occurring compounds, like chlortetracycline or oxytetracycline, and a large group or semisynthetic compounds. Typical tetracyclines are used in therapy and as a growth promoter in agriculture [26].

Streptogramins consist of two molecules: streptogramin A and B, each having a different binding site, but working together. First, streptogramin A binds in the proximity of the P-site on the large ribosomal subunit, what introduces a conformational change enabling streptogramin B to block the exit tunnel [27]. Streptogramins were used in agriculture all over the world, but since 1999 their use has been banned in Europe [28]. Streptogramins used in therapy, like pristinamycin, fight infections with methicillin-resistant *Staphylococcus Aureus* (MRSA). Other antibiotics targeting the ribosomal exit tunnel include lincosamides, and macrolides, such as erythromycin and azithromycin. Some lincosamides, like clindamycin, have two binding sites, one in the PTC and around the exit tunnel.

Linezolid is an important antibiotic effective against bacteria resistant to other antibiotics, e.g., vancomycin- or methicillin-resistant *Staphylococcus aureus* [29]. It targets the PTC site [30, 31], but is active only against Gram-positive bacteria since the Gram-negatives are resistant due to the efflux pump [32].

The most straightforward resistance mechanism against all of the antibiotics targeting rRNAs is mutations in the rRNA genes. However, bacterial genomes often carry multiple copies of rRNA operons, hence for the high level of resistance, mutation of the majority of those repeats is required. Since rRNA is a ribozyme, mutations in vital parts of the ribosome have high fitness costs. Therefore, enzymes modifying the rRNAs post-transcription like methyltransferases constitute popular resistance mechanisms [20]. Methyltransferases confer resistance to macrolides and linezolid [33, 34]. Since binding sites of some antibiotics overlap, the methyltransferase can cause cross-resistance, like in the case of *cfr* methyltransferase that confers resistance not only to linezolid but also streptogramins, chloramphenicol, and clindamycin.

The second relevant mechanism of resistance employs ribosomal protection proteins (RPPs). RPPs are homologous to the elongation factors, and like them, they are GTPases [35] so they can actively remove an antibiotic bound to the ribosome.

RPPs are the primary resistance mechanism for tetracyclines [36]. Other mechanisms of resistance include decreased membrane permeability [37] and efflux pumps, ABC-transporters [34], enzymatic inactivation of an antibiotic by acetyltransferases or phosphotransferases [38].

Targeting cell envelope

The bacterial cell envelope is the first line of protection against environmental conditions. The cell envelope is composed of a membrane(s), a peptidoglycan cell wall and other outer structures. Its composition differs in Gram-negative and -positive bacteria. In Gram-negatives, there are two membranes separated by a cell wall and a periplasm. Whereas in the Gram-positives there is no outer membrane, but the cell wall is $\sim$ twenty times thicker [39]. The cell envelope is crucial for bacterial survival, accessible, and their synthesis is conserved. Therefore, the cell envelope is a target for several antibiotics including the highly relevant group of β-lactams and glycopeptides.

The cell membrane of Gram-positive bacteria is targeted by daptomycin, the first introduced lipopeptide antibiotic [40]. Although the particularities of daptomycin's action mechanism remain unclear, it appears to be binding to the membrane. Daptomycin invades the membrane it with its lipophilic tail. Next, more antibiotic molecules oligomerize in the bacterial membrane so that they finally destabilize it and cause an efflux of potassium ions [41, 42]. The refined model proposes that daptomycin molecules assemble on the inner and outer leaflets of the membrane and create the pore that causes ion leakage [43]. The role of the phospholipid phosphatidylglycerol molecules is unknown. However, they are crucial for the daptomycin's binding to the membrane. The FDA approved daptomycin in 2003 for medical therapy of skin infections caused by Gram-positive pathogens. Nowadays, it is used intravenously mostly to treat MRSA infections.

After the integrity of the cell membrane is lost, an extensive cell envelope stress response is activated [44]. The *lia* genes driving the response network are among the daptomycin resistance genes. Other resistance mechanisms concern genes responsible for producing envelope components, e.g., enzymes modifying phospholipids or producing phospholipid phosphatidylglycerol [45, 46].

Atypical tetracyclines are chemical derivatives of typical tetracyclines [47]. Like typical tetracyclines, they are also classified as broad-spectrum but have a different action mechanism and are bacteriocidic. Chelocardin depolarizes membrane and disturbs the integrity of the cell membrane, so far neither the details of action mechanism nor of the resistance are known [48].

Bacterial cell wall is a rigid external skeleton composed of muramic acid held together by pentapeptide side chains. Its role is not only protection but also withstanding the turgor pressure of the cell [39]. The process of cell wall synthesis has three stages. Firstly, *mur* enzymes synthesize peptidoglycan units in the cytoplasm.

In the second stage, peptidoglycans are transported to the membrane where they are linked into strands by a transglycosylase. Finally, in the third stage, the strands are cross-linked by one of the transpeptidases and form a two-dimensional sheet of murein. Almost all steps of cell wall synthesis and assembly are targets of various classes of antibiotics, especially the final crosslinking that is targeted by β-lactams [49, 50].

β-lactam antibiotics target transpeptidases responsible for cross-linking peptidoglycan peptide side-chains. They mimic D-alanyl-D-alanine that is a substrate for the transpeptidases. They modify the serine residue of the active site preventing further catalytic activity of those enzymes [51]. Targets of the penicillin constitute a broad family of enzymes termed penicillin-binding proteins (PBPs), they are located in the bacterial periplasm and fulfill wall-related functions [52, 53].

β-lactams include penicillins, cephalosporins, and carbapenems, all of them are equipped with a β-lactam ring and belong to broad-spectrum antibiotics. They are the most used antibiotics since they constitute more than a half of the antibiotic market in the US [54]. Their extensive usage caused an increase in the resistance, primarily mediated by enzymes disrupting the β-lactam ring, i.e., β-lactamases [55].

β-lactamases are structurally related to the penicillin-binding proteins (PBPs), and similarly, they constitute a large and diverse group of enzymes, with various ranges of activity regarding substrates and conditions. They evolve quickly in response to numerous derivatives of the β-lactam antibiotics [56]. Among thousands of known β-lactamases, especially dangerous are the extended-spectrum β-lactamases (ESBLs), that confer resistance to multiple β-lactams at once, e.g., the penicillins, all-generation cephalosporins, and aztreonam [57]. ESBLs are classified into nine families: TEM, SHV, CTX-M, PER, VEB, GES, TLA, BES, and OXA, depending on the profile of resistance and protein sequence [58]. Other mechanisms of β-lactam resistance include mutated inherently resistant PBPs, found, among others, in methicillin-resistant *Staphylococcus Aureus* (MRSA), reduced number of porins in the outer membrane of the Gram-negative bacteria and a variety of efflux pumps [59].

Glycopeptides are defined as narrow-spectrum antibiotics since they affect only Gram-positive bacteria [60]. They inhibit cell wall synthesis through binding to the peptide elements of peptidoglycan [61]. The most popular glycopeptide is vancomycin. Vancomycin resistance is conferred by the dehydrogenase *VanH* and ligase *VanA* that together provide an alternative path for peptidoglycan cross-linking [62]. Vancomycin is used in MRSA treatment. However, nowadays doctors observe a rise of vancomycin-resistant *Staphylococcus aureus* (VRSA) [63].

Replication inhibitors

First step of bacterial cell division is DNA replication. It is an extremely controlled process as its fidelity is vital for the survival of the bacterial species, and as such, it is another conserved process targeted by antibiotics. Replication happens in two Y-like replication forks, where the DNA is unwinded, and two new strands are synthesized [64]. As the replication forks move along the bacterial genome, a DNA super tension builds ahead of them. Since bacterial genomes are circular, the DNA has no free end to turn and remove the tension, but if it is not removed, the replication cannot proceed. Therefore, bacteria produce enzymes for removing tension called topoisomerases. The topoisomerase I cuts one of the DNA strands, moves it and ligases it again so that there is no tension. The type II topoisomerase cuts both strands and is, therefore, able to remove supercoiling [65].

In Gram-negative bacteria, the type II topoisomerase, also called gyrase, is the primary target for quinolones [66]. Quinolones stabilize binding between the gyraze and DNA, which leads to fragmentation of the DNA and cell death [67]. When the replication forks meet, the two new DNA molecules are separated by the type II topoisomerase. The Topoisomerase IV that also removes the positive supercoils and constitutes the primary target of the quinolones in Gram-positive bacteria [68].

Quinolones are widely used broad-spectrum antibiotics. Nowadays, they are equipped with an additional fluorine atom that increases their affinity to topoisomerases [69]. Hence they are referred to as fluoroquinolones. The main fluoroquinolone resistance mechanism relies on mutations in the genes encoding the targeted topoisomerases. Both topoisomerases consist of two copies of each of the two subunits encoded by *gyrA* and *gyrB*, *parC* and *parE* for gyraze and topoisomerase IV respectively [70]. Other mechanisms of quinolone resistance encoded on plasmids include topoisomerase protection proteins, quinolone modification enzymes, and efflux pumps [68].

Overview of antibiotic resistance

Definitions of antibiotic resistance differ depending on the context. The most general one defines resistance as an ability of bacteria to grow and divide in the presence of an antibiotic. In microbiology, resistance is a non-binary value relative to the antibiotic concentration. In medicine, resistance is defined with respect to the success of the antibiotic treatment, but in fact, each particular therapeutic outcome depends on microbiological and non-microbiological factors [71]. The resistance rate is a portion of resistant isolates among all of the isolates, so in environmental studies, those measurements are much less accurate. For this thesis, antibiotics resistance refers to the binary state of presence of molecular mechanisms potentially enabling bacteria to withstand antibiotic presence.

Resistance mechanisms are classified depending on their origin and molecular background. According to the first classification, resistance mechanisms can be

intrinsic, acquired and adaptive. Intrinsic resistance includes general features of bacteria that make them immune to an antibiotic, e.g., cell membrane permeability, porins, and efflux pumps. Acquired resistance denotes features transferable by HGT, i.e., enzymes modifying antibiotic or its target. Adaptive resistance by definition depends on the presence of antibiotics and works through epigenetic modifications [72]. According to the second classification, resistance genes fall into four classes: enzymes neutralizing antibiotics, enzymes modifying antibiotic's targets, mutated target genes and efflux pumps [73]. Table 2.1 outlines dominant resistance genes for the most popular antibiotic classes.

As presented in Table 2.1, there are multiple resistance mechanisms for each antibiotic. However, that is not the only complication. Firstly, our knowledge is incomplete as resistance mechanisms are still being discovered, and not necessarily in the context of new or modified antibiotics. Secondly, resistance mechanisms can comprise multiple steps, like in the case of the MarA protein, which upregulates expression of quinolone efflux pumps and at the same time suppresses the expression of porin proteins reducing quinolone intake [86]. On the one hand, resistance mechanisms can be synergistic. On the other hand, a single resistance mechanism can provide resistance to multiple antibiotics, as in the case of rRNA methylation which can cause cross-resistance to some antibiotics targeting ribosome [81]. Multiple resistance mechanisms contribute in various degrees to a resistance phenotype - depending on antibiotic and bacterial species. Thus, it is difficult to infer a phenotypical resistance solely from the genetic background.

Antibiotics resistance emergence (AR emergence)

Antibiotic resistance could be contained if it did not spread amongst bacteria, especially pathogens. A bacterium can acquire resistance in two ways: through mutations in the antibiotic target or by horizontal gene transfer (HGT). The mutation rate of the bacterial genomes is estimated to be one change per 10^9 nucleotides per cell generation [64]. This is both high and low. High, because it just takes around 200 new bacterial cells (for a bacterium with a genome size of $5 * 10^6$) for a mutation to appear. However, low, since it would have to be an extremely lucky mutation to cause resistance. Thus horizontal gene transfer (HGT) is the primary mechanism of resistance emergence [89].

Table 2.1: The primary resistance mechanisms according to the antibiotic class.

Antibiotics	Antibiotic modification	Target of modification or protection	Efflux pumps
Aminoglycosides [74, 75]	N-acetyltransferases: *AAC*, O-nucleotidyltransferases: *ANT*, O-phosphotransferases: *APH*	16S rRNA mutations, RNA-methylotransferases: *armA, rmtA-H*	*AcrAD, MexXY-OprM, EmrE, LmrA, MdfA*
Tetracyclines [36, 26]	monooxygenase: *Tet(X)*	RPPs: *Tet(O,M,S,W,Q,T), otr(A)*	*Tet(A-E,G-J,V,Z)*
Chloramphnicol [76]	acetyltransferases: *cat*, phosphotransferase: *cmlv*	methyltransferase: *clbC*	*cml, cmr, fex, flo, cmx, MdfA*
Oxazolidinones [77, 78, 79]		mutations of 23S rRNA, L3 and L4 ribosomal proteins, methyltransferase: *cfr*	ABC transporters: *optrA*,
Streptogramis [80]	acetyltransferasses: *vat*	RNA methylase: *erm*	putative ABC transporters: *vga, msr*
Macrolides [81, 82]	esterase	RNA methylase: *erm*	*mef, msr, lmrP, srmB, tlrC*, ABC transporters: *carA, msrE*
Lincosamides [81]	nucleotidyltransferases: *lnu*	RNA methylase: *erm*	
Rifamycin [83, 84]	monooxygenase: *rox*, phosphotransferase: *rph*, glycosyltransferase	mutations of *rpoB*, rpoB protection protein *RbpA*	
Glycopeptides [62]		dehydrogenase: *VanH*	
β-lactams [85, 59]	β-lactamases, ESBLs		
Quinolones [86, 87]		mutations in *gyrA, parC, parE*, protection protein *Qnr*	*norA, MexAB-OprM, oqxAB*
Lipopeptides [88]		*rpoB, rpoC, MprF, liaR, CdsA*	

Horizontal gene transfer (HGT)

Although events of HGT are both random and rare, bacterial evolution is governed by inter- and intracellular movements of genetic material. Intercellular movement is facilitated by three processes: transformation, conjugation, and transduction. Transformation happens when a bacterium incorporates naked DNA directly from the environment. For a natural transformation to occur, two conditions must be met. First, there has to be naked DNA in the environment originating either from lysed bacteria or exported in a mesosome. Second, the recipient bacterial cell needs to be competent, able to intake extracellular DNA. This process starts with a double-stranded DNA binding to the surface of a bacterium. Next, it is fragmented and transported - those processes differ between Gram-positive and -negative bacteria, but none of them is known in full detail [90]. However, we know the state of competency is regulated by a network involving more than twenty genes. If the imported DNA is not a plasmid, it has to integrate into the bacterial genome in the process of recombination. There are multiple bacterial species among known human pathogens that are both donors and acceptors of the natural transformation [91].

Conjugation is a mechanism for transferring plasmids between bacteria. Unlike natural transformation, during conjugation bacteria are in physical contact, connected with a pilus. In theory, via means of conjugation, an entire chromosome could be transferred. However, in practice it is a slow process that would require bacteria to remain in contact for a long time, therefore transferring an entire chromosome is extremely rare [89, 92]. Conjugation is controlled by $\sim$ 40 genes encoded in the transfer region of the conjugative elements, such as plasmids or transposons. Conjugation happens between two bacteria, out of which one possesses the conjugative elements, and another is deprived of it [93].

Transduction is intermediated by bacterial viruses, or phages for short (Fig. 2.3). After it attaches to a bacterial cell, the phage introduces its genome into the bacterium. Subsequently, the phage genome integrates into the bacterial chromosome. A phage can remain integrated, and be passed to daughter cells as during division (lysogenic cycle). Otherwise, a phage can enter a lytic lifecycle. After its proteins are expressed, phage particles are composed, resulting in lysis of the bacterial cell and release of all new phage particles. Some of the new phage particles carry not only the phage genome but also fragments of host genomes. Those particles continue to infect further cells and transferring genetic material between bacteria.

Researchers estimate that in the majority of environments phage particles greatly outnumber bacterial cells [94]. Their complex lifecycles relying on the arms-race with bacteria cause phage genomes to be remarkably diverse and variable [95, 96], and poorly characterized. In the NCBI database are only $\sim$2,000 complete annotated phage genomes in comparison to the $\sim$35,000 bacterial genomes (as of January 2018).

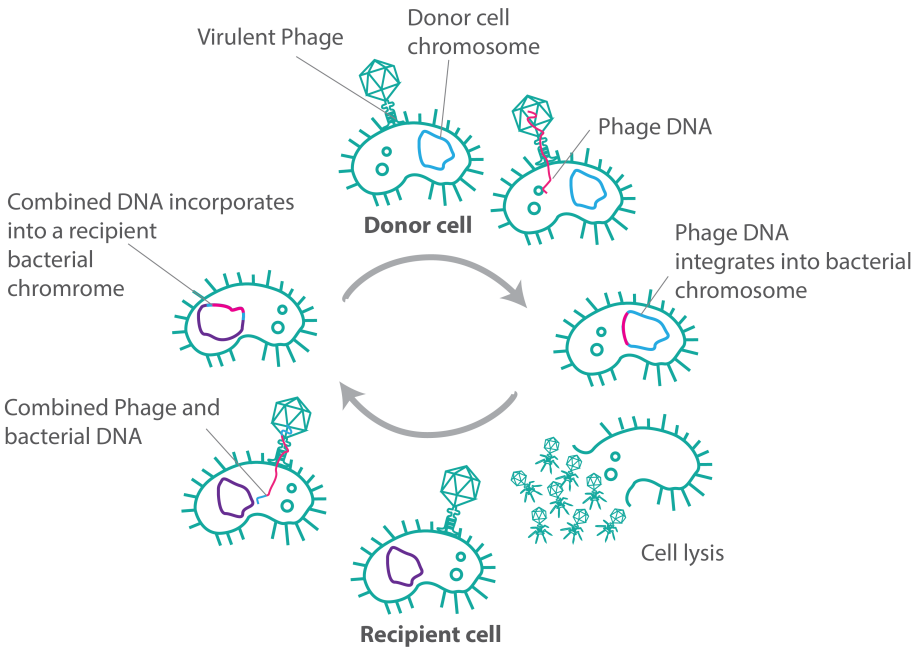


Figure 2.3: Mechanism of transduction.

Phage genomes are hugely versatile. Phages can carry any genetic material: dsDNA, ssDNA, dsRNA, and ssRNA, circular or linear and code from four up to hundreds of genes. The smallest regarding genome size are phages with a single-stranded RNA (ssRNA). Genomes of such phages, MS2 or Qβ, are ~4Kb long and carry four proteins. The genome of the exemplary dsRNA phage ϕ6 is ~13Kb bases long but comprises of 3 segments. Phage ϕX174 is a representative of the ssDNA phages. Its genome is ~5.5Kb long and contains 11 genes. dsDNA phages have the largest genomes. They include the famous phage λ with the genome that is 50Kb long and carries 92 genes, and massive T4 with its 170Kb genome, 288 genes, among which only about half has known function [97], including eight tRNA genes [98].

Phages, like other viruses, are a part of the taxonomic tree. The taxonomy of phages is based on capsid shape and host range. It has been criticized as it does not reflect the actual relationship between phages. As an alternative, researchers proposed various reticulate network classifications, based on the genetic similarities. For the reticulate networks, phage genomes are mapped with many features such as genes, k-mers, or modules, so that distances between them are computed, based on which a reticulate network is constructed [99].

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)

Bacteria have developed the CRISPR system, so-called *immune system*, to protect themselves against unknown DNA. CRISPR are cassettes of short, 26-72 bp long,

signature fragments of foreign DNA (spacers) interleaved with 21-48 bp long identical direct repeats. Typically not more than 50 fragments are deposited within a single cassette, but bacteria can have multiple cassettes. CRISPR loci are adaptable, old sequences are removed, and new can be incorporated [100]. The majority of CRISPR spacers originate from phages, but they can also protect against plasmids [101].

CRISPR loci are transcribed together with an upstream leader sequence and cut into small RNA molecules (crRNAs) containing a single spacer and a part of a direct repeat [102]. Depending on the type, different mechanisms and actors are involved, but the common feature is that crRNA guides the degradation machinery towards the DNA molecules complementary to their sequence.

Mobile genetic elements (MGEs)

Mobile genetic elements (MGEs) include plasmids, phages, and a large class of smaller MGEs, i.e., DNA fragments that change locations within bacterial genomes and between bacterial cells [103]. They can be located on plasmids and bacterial chromosomes. They are classified based on their size and genetic structure.

Transposons are MGEs composed of genetic cassettes flanked with insertion sequences. An insertion sequence (IS) is an ORF coding for a transposase with direct inverted repeats on each of its ends [104]. ISes are ∼0.3 kb long, and transposons are between 2.5 and 60 kbp long. They can carry any gene and move within bacterial genomes. Transposons can also be conjugative and transfer themselves or mobilize plasmids to transfer between cells [105].

Integrons are ancient genetic elements found in many bacterial genomes. They are hotspots for genetic variance, capturing potentially useful genes also those without a promotor (gene cassettes), and securing their expression [106]. A minimal integron consists of three elements: an integrase, a recombination site *att*, and a promoter *Pc*. As they do not encode any mobility systems, they often couple with other MGEs and get transferred. Integrons contribute to AR emergence since they are known to accumulate ARGs.

Multi-drug resistance (MDR)

Overall, horizontal gene transfer enables accumulation of the MGEs with antibiotic-resistant genes which leads to the emergence of multi-drug resistant bacteria [107]. Definition of MDR used to include all organisms resistant to more than one antimicrobial agent. Nowadays, multi-drug resistance is defined separately for each bacterial species and in the context of antibiotic groups [108]. Especially important are human pathogens, e.g., Methicillin-resistant *Staphylococcus Aureus* (MRSA) and ESBL-producing bacteria. Such *super-bugs* are responsible for increased mortality among patients and prolongation of hospitalization stays [109, 110].

Methicillin-resistant *Staphylococcus Aureus* (MRSA)

Gram-positive MRSA is one of the most clinically significant MDRs due to its high pathogenicity and resistance level [107]. Depending on the location, MRSA can cause skin, joint and pulmonary infections [111]. The MRSA was first described in the 1960's. Since then, its resistance potential has been expanding. MRSAs are resistant to β-lactams, glycopeptides, quinolones, aminoglycosides, oxazolidinones, tetracyclines, chloramphenicol, and others [112]. Until 1980's vancomycin remained the main viable therapy for MRSA, what quickly lead to the emergence of VRSA, a vancomycin-resistant *S. Aureus* strain in the early 1990's [113].

MRSA colonization is prevalent in hospitals, nursing homes, and other long-term care facilities (HA-MRSA). Currently, MRSAs are also found in the community (CA-MRSA). The CA-MRSA is less resistant and toxic than HA-MRSA. Colonization of the nose is the primary factor increasing the possibility of infection [114]. Also, patients suffering from diabetes, using invasive devices, with a weakened immune system or of older age are more susceptible to the MRSA-infection. It is estimated, that as much as half of population carry *S. Aureus* in their nose. Currently, the vast majority of *S. Aureus* isolates are resistant to penicillins.

Since the year 2000 European Centre for Disease Prevention and Control (ECDC) [115] has been systematically monitoring levels of antibiotic resistance within European hospitals. Level of antibiotic resistance is defined as a proportion of the resistant strains among all of the isolates. The EU-wide average of MRSA resistance has been slowly decreasing (Fig. 2.4), suggesting that the stewardship efforts have been successful. However, there is significant variance among the European countries, and in many of them, MRSA still poses a significant threat.

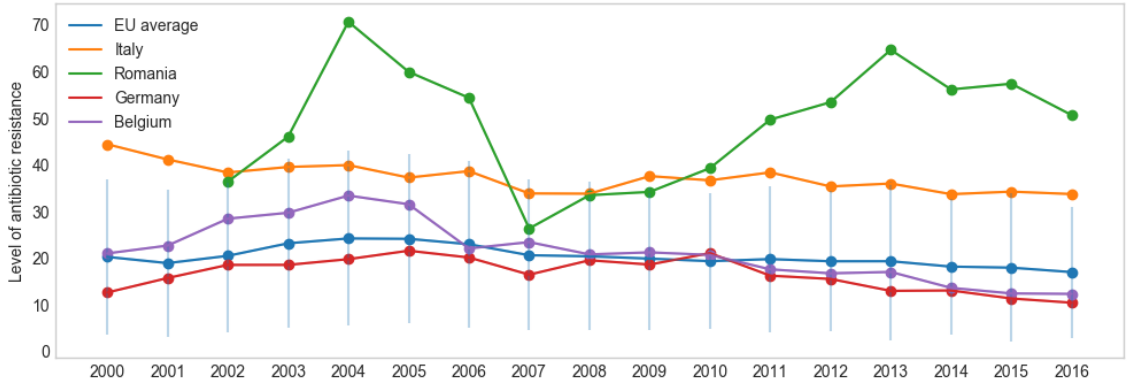


Figure 2.4: Levels of MRSAs among the *Staphyloccocus* isolates in chosen european countries.

Methicillin resistance is conferred by the *mecA* gene located on the MRSA's chromosome in the SCC*mec* cassette. *mecA* encodes a transpeptidase which performs

peptidoglycan cross-linking in the presence of the β-lactam antibiotics [116]. It is not clear how exactly did *S. Aureus* obtain the *mecA* gene. Close homologs of the *mecA* were identified in a number of other staphylococci [117]. Other resistant genes are localized on chromosomes and plasmids, some of them within transposons. Nowadays MRSAs are equipped with an arsenal of the resistance mechanisms including β-lactamases, modified gyraze, and topoisomerase subunits, and rRNA modification enzymes. Moreover, MRSAs are equipped with a broad arsenal of bacterial toxins.

ESBL-producing bacteria (ESBL)

Extended-Spectrum Beta-Lactamases (ESBL) were discovered in the 1980's. Since then, scientists and doctors all over the world observed the rise of ESBL producing bacteria (ESBLs). ESBL are mostly present in common human gut bacteria *Enterobacteriaceae*. However, more importantly, they were found in dangerous pathogens such as *E. coli*, *K. pneumoniae* and *S. enterica*. Average resistance levels of among *E. coli* isolates across the EU had been steadily increasing (Fig. 2.5).

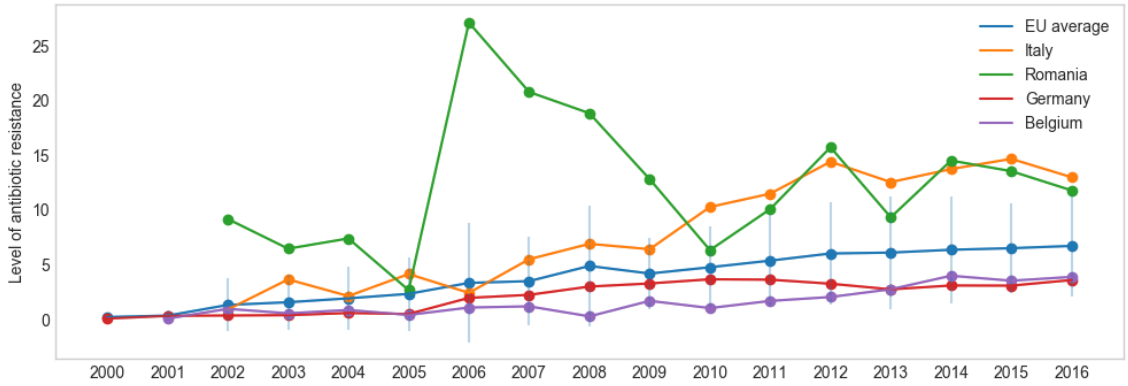


Figure 2.5: Levels of resistance among the *E. coli* isolates in selected european countries.

Extended-Spectrum Beta-Lactamases are subjected to horizontal gene transfer (HGT) as they are located on plasmids and bacterial chromosomes in the proximity of transposons [118]. Therefore, they are often found in the company of other antibiotic-resistant genes, even on a single plasmid [119]. Consequently, they evolve dynamically, spread between bacteria and patients extremely fast. Foremost, ESBL-producing bacteria have been found in hospitals. The class profile of the Beta-Lactamases differs depending on the antibiotic usage. However, in countries with unrestraint access to antibiotics, as much as half of the tested volunteers carry ESBL-producing bacteria [120].

2.2 Human gut microbiome

Bacteria inhabit all sites on the human body [121], but exceptionally rich, diverse and vital is the gut microbiome, i.e., bacteria, viruses, fungi, and protozoa inhabiting gastrointestinal tract [122]. The gut microbiota remain in a mutualistic relationship with the host [123]. An organism provides habitat and nutrition for the microbiota which contribute their diverse metabolic pathways, they ferment carbohydrates, produce vitamins and metabolize various xenobiotics [124]. Researchers describe the human gut microbiome with great numbers. Trillions of bacteria inhabit the human gut. They are divided in into $\sim 1,800$ genera, $\sim 20,000$ species which collectively comprise an enormous genetic diversity estimated to contain ~ 10 million unique genes. Therefore there are ~ 450 times more bacterial genes than human genes within the human body [125, 126, 127]. However, the human gut bacteria fall mostly into the two phyla: the Gram-negative *Bacteroides* and Gram-positive *Firmictues* [128].

Decreasing sequencing costs enable reliable taxonomic and functional profiling of gut microbiomes for thousands of samples. This enabled linking the divergence of the human gut microbiome profiles to diseases ranging from gut-related cancer [129] and inflammatory bowel disease [130], through diabetes [131] and heart-related diseases [132], to Alzheimer's [133] and psychiatric disorders, such as depression [134]. The main characteristic speaking to the overall *health* of the gut microbiome is a ratio of the *Firmictues* to *Bacteroides* abundance. Scientists have observed it increases in obese mice [135, 136] and obese patients [137].

Overall, the taxonomic structure of the gut microbiome is highly variable from person to person. However, the gut metagenomic taxonomic profiles cluster into three enterotypes driven mostly by a domination of one of the genera *Bacteroides*, *Prevotella*, or *Ruminococcus* [138]. However, the concept of enterotypes is criticized for being not comprehensive, as the original dataset included mostly European participants [139].

Although scientists have reliably observed gut microbiome structure depends on a person's health, age [140], lifestyle, diet [141], and travel, nature of interactions between the gut microbiome and the human organism are not fully understood [142]. It is not clear which of the sides in the microbiome-organism interaction dominates, or if the disease drives the change of the microbiome or the other way round. Nevertheless, it is becoming evident the gut bacteria remain in the dynamic homeostasis amongst themselves and with the host [143].

The gut-brain axis exemplifies the bidirectional nature of the microbiome-organism interactions. Researchers identified several biochemical pathways within the gut microbiome, producing endocrine, neurological and immunological metabolites [144]. As a result, microbial dysbiosis contributes to disease, and oppositely, healthy microbiome could protect from it. Therefore, monitoring, preservation and maybe eventually manipulation of the gut microbiome, established essential new aspects of medical therapies.

2.2.1 Studying the gut microbiome

The gut microbiomes have been studied since the 1970's. The first methods were culture-based. The bacteria were cultured in anaerobic conditions using different media so that they could be characterized [145]. The introduction of next-generation sequencing (NGS) technologies in the 2000s opened up opportunities for large-scale, fast, reliable and culture-free studies of microbiomes.

Nowadays, there are numerous multistep protocols for studying different aspects of the microbiome. E.g., the functional metagenomics used for quantification of antibiotic resistance. There, fragments of DNA extracted from the microbiome are cloned into the *E. coli* cells that are next plated on selective media. Later only those of interest, e.g., showing resistance, are sequenced [146].

Another class of methods relies on sequencing of a single gene. Especially important is the conserved 16S rDNA gene, as its sequence is used in the taxonomy of bacteria. Although, the overall structure of rRNA is quite conserved, locally conservation levels vary throughout the length of 16S rDNA. Consequently, structure of the 16S rDNA enables both efficient design of primers, which are complementary to the conserved regions, and taxonomic profiling, which is based in the variable regions. Bioinformatics analysis relies on the alignment of sequencing reads against the database of the used gene. The portion of the reads aligned to different hits in the database constitutes a profile, in the case of 16S rRNA, the hits point to the taxa, and the result is a taxonomic profile [147].

The most straightforward approach is sequencing of the entire DNA extracted from a microbiome sample with little or no pre-sequencing processing. This approach is termed shotgun whole-genome sequencing (WGS). The standard bioinformatics analysis pipeline starts from the alignment of reads to a database, in most cases, of protein sequences such as the NCBI's NR. The alignment, together with the hierarchical classification of the records enables both a taxonomic and functional profiling [148].

2.2.2 Impact of antibiotics on the human gut microbiome

An antibiotic-associated perturbation of the gut microbiome occurs already on the second or third day of an antibiotic therapy. Consequently, antibiotics constitute severe disturbance factors. Susceptivity to all antibiotic classes is well characterized to some of the gut bacterial species. Despite that, the impact of the specific antibiotic on the entire gut microbiome is unpredictable. Although the majority of gut bacteria can be cultured, still a considerable portion of them are unknown [149]. Furthermore, the microbiome is not merely a group of bacteria. It is a dynamic network where eliminating some of the taxa might cause indirect and unpredictable shifts.

Antibiotics differ by their spectrum of activity, consequently so does their impact on the gut microbiome and the adverse effects they cause [150]. Ciprofloxacin causes

a reduction of the abundance of the Firmicutes *Faecalibacterium*, *Ruminococcus* and *Alistipes* from phylum Bacteriodetes, an increase in the Bacteroides *Bacteroides* and Firmicutes *Lachnospiraceae* [151], and a general increase of the Firmicutes to Bacteroides ratio. At times the antibiotic effect is more specific, reducing some of the genera and promoting the others [152, 153].

Numerous studies conducted in human and animals showed that regardless of antibiotic class, a therapy reduces microbial diversity within the gut [154]. Also, by killing commensal bacteria, antibiotics decrease colonization resistance, making space for potential pathogens [7, 8]. Although researchers discovered general laws governing the taxonomic shifts under antibiotic pressure, several studies reported, that changes in the microbiome also depend on the state of the microbiome in the first place, and hence they are highly individual [155].

Restoration of the taxonomic structure takes from 1 to 6 months. The speed of the restoration is individual. However, it is facilitated by good overall health and negative history of antibiotic therapy [156]. Therefore, depending on various factors the effects of the particular antibiotic therapy in particular patient could be long-term. The speed of restoration can be treated as a functional definition of the *healthiness* of the gut microbiome [142, 157]. In some cases, the changes persisted for years, like in the case of clindamycin therapy, after which the taxonomic structure within *Bacteroides* took two years to restore [158].

Antibiotics promote bacterial strains harboring ARGs but also strengthen overall AR-reservoir [159]. An increase in resistance is rarely specific towards the antibiotic. However, both ARG abundance and their diversity increase with each therapy. The ARG abundance normalizes after therapy, but the diversity persists, so when a pathogen comes along it has access to a broad resistance reservoir. This worsens prognosis for future antibiotic therapies and promotes the MDR emergence. MDRs have also been detected up to two years after the therapy [160].

Antibiotic-resistant genes are found in microbiomes across all environments [161], and in the gut microbiomes of healthy people all around the world. The resistome diversity within the population positively correlates with the uncontrolled access to the antibiotics [162, 163]. Healthy human gut microbiomes on average contain β-lactamases, tetracycline, and aminoglycoside ARGs [160].

Although antibiotics are active against bacteria, they are in fact therapeutics of humans. Antibiotics have severe adverse effects for the gut microbiome including perturbations of taxonomic structure and metabolic processes, the increase in resistome diversity, and the increase in HGT frequency. Metabolism of an antibiotic within the gut also disturbs its pharmacokinetics affecting the final therapeutic effect [152]. Research into those interdependencies between perturbations of the microbiome and the host's health is crucial, so that safe antibiotics usage is possible.

2.2.3 Large-scale human gut microbiome studies

To fully understand inter- and intrapersonal diversity, several large-scale projects investigating microbiomes of sizeable groups of participants have been undertaken. The first that paved the path was the Human Microbiome Project (HMP) launched in 2007. The project received $150 million in funding for five years. For each of the 300 non-hospitalized volunteers who took part in the project, samples from five body parts and over several time points were collected [164].

The Human Microbiome Project (HMP) provided 16S rRNA and WGS sequencing samples for the multiple body sites, 3000 reference bacterial genomes and laid the groundwork for future microbiome projects. It described the structure of the most common human gut microbiome [165]. Currently, the HMP is in its second phase: integrative HMP, that using multi-omics approaches investigates microbiomes during pregnancy and for patients suffering from IBD and diabetes.

The American Gut Project started in November 2012. In 2014, a collaborative sister-project British Gut Project was launched. In 2017 they collected gut microbiome samples from 11,336 participants, in the majority from the USA, but also Europe and Australia. The participants filled in a questionnaire, providing details about their health, diet, and lifestyle. It is a crowd-funded project, as the participants pay for the analysis of their samples, but are later have access to the results. The samples were analyzed regarding taxonomic diversity using 16S rRNA sequencing, functional diversity with metabolomics studies [166].

In those studies, a significant number of samples was influenced by recent antibiotic therapy. The American Gut Project reported that metabolome diversity was higher in the group of participants who took antibiotics in the previous month (139 people) in comparison to those participants who declared not taking antibiotic in the previous year (117 people). However, the taxonomic diversity was lower in the antibiotic-influenced microbiomes. The authors pinpoint the diversity of the antibiotic therapies in their set has prevented them from performing an antibiotic-specific analysis.

The Guangzhou Cohort Study is the newest exciting large-cohort microbiome project. It is in its initial stage. So far they have recruited 17,214 pregnant women. They are planning to track their microbiomes during pregnancy and compare them to those of their children. The results will be correlated with the extensive metadata [167].

2.3 Human gut mobileome

The gut microbiome is a dynamic network of bacteria, connected by HGT. One of the most known examples of HGT-mediated acquisition of relevant genes is a study of gut bacteria of several Japanese volunteers. Taxonomic profiles of their microbiomes were similar to those of other populations. However, the genomes of

their gut bacteria were enriched with genes enabling degradation of substances found solely in algae [168].

Mobile genetic elements, including (pro)phages, plasmids, transposons, integrons and insertion sequences for all bacterial cells in the microbiome constitute a mobilome [169]. Small-scale methods to study MGEs such as plaque-based and microscopy analysis of phages, PCR-based sequencing, plasmid or transposon capture, provide a limited picture, and are inapplicable in metagenomics studies [170, 171]. Phages are mostly studied via sequencing, using either a specialized protocol where the complete viruses are first isolated from the sample [172, 173] or through screening of whole-genome shotgun (WGS) metagenomic data [174]. The latter method enables studying of all MGEs and bacteria within a microbiome simultaneously.

Mobile genetic elements often have a mosaic genetic structure and no distinct genetic characteristics. They are versatile, underrepresented and misannotated in the databases. In the majority, their genes are also found outside of MGEs. Therefore, alignment-free MGE identification methods were developed, to be employed in the analysis of WGS datasets. VirFinder [175] and PlasFlow [176] employ k-mer based machine learning to identify phages and plasmids respectively. However, as the tools were first trained on the k-mers extracted from known sequences, comprehensiveness of the database limits their sensitivity. Therefore, MGE *in silico* identification within metagenomic sequencing or assembly is computationally expensive and error-prone [177].

The microbiome network changes under stress. The frequency of all types of HGT within a microbiome increases under environmental pressure caused by disturbances such as diet change, inflammation or antibiotic therapy [178]. Additionally, since phages infect and lyse bacteria, they prolong the environmental pressure prompted by the first factor. Phage impact depends on their host range, which is a result of the constant double-sided evolution. On the one hand, phages benefit from a broad-range host spectrum, and on the other hand, it comes at the cost of reduced efficiency [179].

Overall phage population in the gut microbiome, termed phageome, consists mostly of temperate phages of the most abundant bacterial phyla: Bacteroides and Firmicutes. Much like microbiome, phageome taxonomic profile remains stable in an undisturbed gut, but it responds to diet changes [180]. However, phageomes are more diverse and variant between people than the bacterial portion of the microbiome [181]. Phageomes are diverse also in a genetic sense, as they harbor a sizeable functional diversity, including ARGs [182, 183].

2.3.1 Impact of antibiotics on the human gut mobileome

The fate of MGEs in the face of antibiotic therapy depends on the fate of their bacterial hosts. Although phages are not directly susceptible to antibiotics, the antibiotic pressure on the bacteria causes shifts in the genetic landscape of

phages [184], regarding both phage taxonomy and functional profiles. Consequently, an impact of the particular antibiotic on phageome depends also on the antibiotic class [185, 173].

Besides the environmental-selection of resistant strains, HGT is the secondary factor driving AR emergence [186], especially in such a dense and diverse environment as the gut microbiome [187]. Studies of plasmids and transposons have been focused on known elements. Within the gut microbiome, researchers observed conjugative transposons and plasmids transferring erythromycin resistance genes between *Bacteroides* [188, 189]. However, methods for MGE identification within microbiome WGS sequencing, enabling large-scale analysis have only been developed recently (2017). Also, the structure and behavior of MGEs are quite complex, e.g. plasmids can comprise transposons, that can also move between a plasmid and a chromosome within a bacterial cell.

For those reasons, not much is known about MGEs and changes of their genetic structure within the gut microbiome under antibiotic pressure. However, it is clear that antibiotics affect entire microbiomes in all their aspects. They prompt massive taxonomic and genetic changes. Therefore, accurate description of MGE dynamics within microbiomes constitute crucial groundwork for developing methods for controlling AR emergence.

2.4 Combating resistance emergence

Governments and global regulators such as the European Union, WHO, CDC, and FDA have recognized the danger posed by rising antibiotic resistance. They have developed several strategies to counteract resistance emergence treating all aspects of the problem: promoting novel antibiotic discovery, antibiotic stewardship, education and regulation of antibiotic usage in medicine and farming.

Some experts estimate we need as much as twenty novel antibiotic classes to remain ahead of resistance for the next fifty years [190]. Bacteria are incredibly diverse and fast evolving. Therefore, to be effective an antibiotic needs to target a vital process of bacterial metabolism and a conserved molecular target so that its mutation rate is restricted. An antibiotic also needs to reach an adequate cellular compartment, i.e., be able to get through a bacterial wall and a membrane(s), to a proper compartment and finally bind to its target. On top of that, an antibiotic has to avoid possible resistance mechanisms such as efflux pumps or modification enzymes and finally, it has to be non-toxic for human cells. All of those restrictions make discovery of novel antibiotics an arduous process.

In 2010, the *10 by '20* initiative was put in motion [191]. It was set to discover ten novel antibiotic classes by 2020, by labs and companies in the USA and Europe. The rate of discovery has already been slower than expected, with just a few novel antibiotics introduced during this period, including several

derivatives of cephalosporin and vancomycin, and fidaxomicin that is only active against *Clostridium difficile* [192]. Recently a promising novel class of antibiotics, cyclic peptide malacidins, was discovered in the metagenomic sequencing of the environmental samples. Mencidins are active against MDRs such as MRSA [193]. However, they are far from being ready for introduction into medical practice.

The next strategy for combating antibiotic resistance is antibiotic stewardship, i.e., rationalization of antibiotic usage in the treatment of patients within hospitals, especially ICUs where gross of antibiotic usage takes place. Stewardship has multiple forms ranging from a manual expert review of prescriptions to automatized computer systems. The former method works *post factum*, and the latter method, automatic systems assist physicians during the decision-making process. Independently on the form, antibiotic stewardship programs work, as it was proven they improve therapeutic effects for patients [194, 195].

Antibiotic stewardship refers to the institutional efforts to improve compliance with guidelines optimized towards efficiency against the particular infection. The guidelines so far are not directly optimized for the resistance emergence and colonization with the MDR bacteria. However, the choice of therapy and its success relies foremost on the adequate diagnosis and pathogen identification. Sequencing and bioinformatics tools could also assist that.

Another massive problem is the use of antibiotics in agriculture. Farming is a significant source of antibiotic pollution in the environment. However, it is beyond the scope of this theses. Still, I should mention, the global regulators such as the EU are gradually restricting antibiotic usage in farming, as it became clear without those restrictions, emergence is not ever going to be defeated.

2.5 SATURN project

The European Union Council advised by the European Centre for Disease Prevention and Control (ECDC) perceives AR emergence as a severe threat. Consequently, EU funded 38 research projects to investigate this phenomenon, through the 7^{th} framework program (FP7) [196]. So far the successor programme Horizon2020 funded 19 projects to investigate AR.

One of the FP7-funded projects was the SATURN project. The name stands for *impact of Specific Antibiotic Therapies on the prevalence of hUman host ResistaNt bacteria*. The project lasted 60 months, costed €7.8M and involved 13 institutions located in 11 countries: Switzerland, Italy, Israel, the Netherlands, Belgium, Poland, France, Spain, Germany, Serbia, and Romania. SATURN gathered specialists in microbiological molecular, epidemiological and clinical fields, who worked towards a comprehensive model of multi-drug resistance for medicine [197], and providing scientific evidence for improving antibiotic use practices.

SATURN investigated antibiotic resistance on three levels of biological organization. The project consisted of six work packages (WP), levels of bacterial cells, of patients and patient groups. WP3, WP4, and WP5 were the main observational clinical studies, WP2 was an intervention study, and WP1 and WP6 relied on the samples and data collected by the WP2-WP5 and supported the main packages. WP1 dealt with bacterial genetics and the WP6 with antibiotic pharmacodynamics (Fig. 2.6).

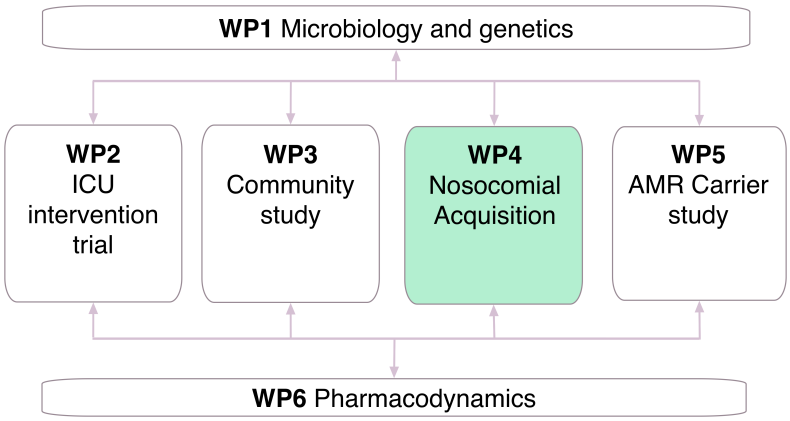


Figure 2.6: Scheme of SATURN work packages. Color highlights WP4 that we worked on. adapted from the Saturn project website [197].

WP2 attempted to shed light on antibiotic prescription practices. It focused on the question if it is better to mix antibiotics or cycle them. In cycling, all patients in a ward take at the time the same antibiotic. Antibiotics are rotated periodically [198]. In theory, this strategy should decrease the overall AR in hospital, in comparison to an alternative approach that is antibiotic mixing, i.e., prescribing random antibiotics. $\sim 10,000$ patients were involved in the WP2 study, among which $\sim 4,000$ underwent antibiotic cycling and another $\sim 4,000$, the antibiotic mixing. There was no difference in the resistance rates between the two groups [199]. Therefore both cycling and mixing strategies select resistant bacteria equally.

WP3 investigated resistance levels in the community, i.e., not hospitalized participants treated for urinary tract infections, along with their families. Researchers recruited two groups of participants from three EU countries: treated and non-treated with antibiotics. The researchers observed high levels of resistance in the European community. Fluoroquinolones turned out to be particularly harmful. They influenced gut microflora in the early days of therapies so that the overall duration of therapy had no impact on AR emergence. This study confirmed that resistance transfers among members of households [200].

WP4 focused on how the rates of AR emergence among hospitalized patients relate to patients' features and the antibiotic therapies [201]. Data comprised records of $\sim 10,000$ people, collected over the three years in Italy, Serbia, and Romania. I

analyzed WP4 data as a part of my Ph.D. work. The project is described in detail in Chapter 5.

WP5 investigated progression of the infection with Carbapenem-resistant *Enterobacteriaceae* (CRE). For the infected patients, the researchers collected rectal swabs. Researchers established that usage of the antibiotics, especially fluoroquinolones after the positive rectal swab drives the development of the infection [202].

The SATURN project lasted for 54 months, and besides scientific project planning, it required great effort in coordination. During the first 18 months, the partners spent on unification of the protocols for microbiological tests and data collection. During the following 18 months the data and samples were collected, and in the last phase, finally they were analyzed, and the collaborators were consulted.

2.6 Machine learning

Clinical studies primarily fall into two classes: interventional and observational [203]. In the former, researcher designs an experiment within clinical context that means deciding on, at least a part of, patient therapies. That is rarely possible. The majority of clinical studies are observational. They boil down to detailed observations of patient cohorts undergoing treatments ordered by their doctors, which results in large, complex datasets, where the data are often internally correlated, much like SATURN WP4 dataset. Such studies constitute potential applications of machine learning (ML) methods as they enable analysis of large datasets considering all features simultaneously.

The term *machine learning* refers to algorithms that perform tasks that are not explicitly programmed. Simple algorithm filtering data points based on a cutoff does not fall into that category, however, if the cutoff is determined automatically, that would be sufficient to classify the algorithm as ML. A typical example is a program playing checkers. The program does not contain direct instructions on what to do in each of the cases, nevertheless, it can play. First, the program analyzed a large number of games of checkers, based on which it *learned* how to play [204], so that it can also play in the situations not included in the training dataset. Analogously the algorithm choosing the cutoff had to be instructed on how to do so, based on some datasets of the known cases. However, it should be able to select a cutoff for the entirely new data.

The most appealing advantage of ML is their ability to provide an answer for new data. This feature is termed *generalization*. In many applications, the real-world questions refer to the future. We provide currently known historical information and ask the classifier about the future. Hence, the widely used term *predict*. Classifiers, *learn* on training sets, so they can *generlize* and *predict* outcome for new datapoints.

Depending on the field and application, ML methods are referred to by a different term including big data analytics, data mining, pattern recognition, statistical learning or artificial intelligence. All of those terms relate to the same class of methods, however, used within different contexts and various goals. *Data mining* relates to explorative applications of ML methods, aiming at the identification of hidden patterns. Often, the researchers start with data mining to understand the dataset at hand. Later those initial classifiers can be used to stratify data before the final predictors are developed.

2.6.1 Machine learning algorithms and their classification

The group of machine learning algorithms is expanding as various algorithms, that primarily belonged to statistics or data analysis, are being used as predictors. Fig. 2.7 presents three main ML tasks: clustering, classification, and regression. All three speak to the structure of the datasets, and all three can be used as predictors. The task depends on the data.

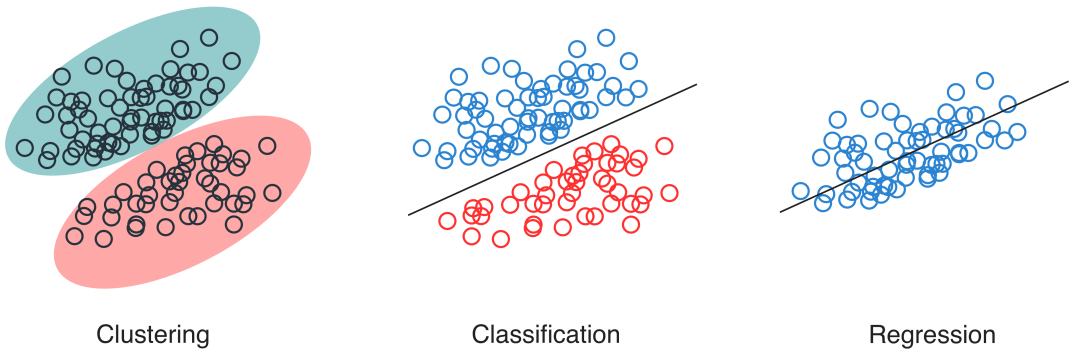


Figure 2.7: Three tasks of the Machine Learning. The clustering answers to a question *what grouping is in the data?*, the classification: *are the classes of the datapoints separable?* and regression *what function is underneath the data?*.

Machine learning algorithms fall into *supervised* and *unsupervised* classes. The supervised ML algorithms require an input of a complete dataset containing both data points and output values. *Supervised* methods return classifiers able to provide output values for new data [205]. The *unsupervised* algorithms do not use output values. Instead, they analyze the structure of the input data through clustering or density estimation. For new data points, the classifier returns clusters for the new data. If a dataset is only partially labeled, the algorithm is classified as *semi-supervised* [206]. There, the combination of ML technics is used. Firstly, to discover the missing labels based on the labeled part, and secondly, to construct the final model.

The ML methods further divide by the main mathematical approach they employ. Some of the aspects of the several ML algorithm classes are discussed below along with the data pre-processing and classifier performance measurement methods that were used in the thesis.

Generalized linear models

The algorithms based on linear models assume an outcome (y) depends on the linear combination of the input variables ($x_1..x_n$), like in the equation:

$$f(x) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n, \tag{2.1}$$

where n is the number of features in the dataset. The objective of these algorithms is discovering the $\beta_0..\beta_n$ regression coefficients, as they enable computation of the outcome function $f(x)$. The majority of the algorithms using linear models are primarily suited to solve the regression task, and with the modifications, they can also be used to solve the binary classification task.

Learning the regression coefficients is posed as a minimization problem in respect to the distance of the solution $f(x)$ to the actual data. One of the most straightforward algorithms is Least squares (LS), often used to solve regression problems. LS minimizes global distance (S) of the model provided by the linear function to the data points:

$$S = \sum_{i=1}^{n} r_i^2, \tag{2.2}$$

where n is the number of data points, and r_i is the distance of the single point to the solution of the linear function ($f(x)$) like in Equation 2.1, with the current candidates for the regression coefficients β:

$$r_i = y_i - f(x_i, \beta). \tag{2.3}$$

Finally, the regression coefficients for which the gradient of S equals zero are computed. However, finding the perfect β in one step is hard and computationally expensive, especially for complex datasets. Therefore, the iterative algorithms were introduced, where each step of the iteration further approximates the solution. The program finishes once the solution converges, the β vector stops changing, or after the maximal number of iterations is reached. One of such algorithms is the Gradient Descent algorithm. It uses the loss function that conveys how much the prediction with the currently considered regression coefficients differs from the output. In one iteration all of the dimensions of the β vector are updated in the direction of the steepest descent of the cost function, across all points in the training data. The cost function for a regression coefficient vector $J(\beta)$ for the linear model $f(x)$ (2.1), and

m data points in the training set is defined as a sum of squared distances between the $f_\beta(x^{(i)})$ and the solution $y^{(i)}$:

$$J(\beta) = \frac{1}{2m}\sum_{i}^{m}(f_\beta(x^{(i)} - y^{(i)})^2. \tag{2.4}$$

Therefore, the iteration step that updates the β vector is expressed by the equation:

$$\beta_i^{(i+1)} := \beta_i^{(i)} - \alpha\frac{1}{m}\sum_{j=1}^{m}(f_\beta(x^j) - y^j)x_i^j \qquad \text{for} \quad j = 0, .., n, \tag{2.5}$$

where α is the learning rate. In the case of the large training set m that could take too much time. Therefore, a modification of this algorithm Stochastic Gradient Descent (SGD) was introduced. Dataset size m is reduced by randomly selecting the subset of training points in each iteration step, instead of using the full set. Consequently, the SGD algorithm requires a modified cost function defined for a single data point $(x^{(i)}, y^{(i)})$:

$$J_{SGD}(\beta, (x^{(i)}, y^{(i)})) = \frac{1}{2}(f_\beta(x^{(i)} - y^{(i)}))^2. \tag{2.6}$$

In SGD, the β vector is updated with every training datapoint:

$$\beta_i^{(i+1)} := \beta_i^{(i)} - \alpha(f_\beta(x^j) - y^j)x_i^j \qquad \text{for} \quad j = 0, .., n, \tag{2.7}$$

Although SDG solves the regression problem, it can also solve the classification task, when $y(x)$ consists of binary class labels. The left side of the regression equation 2.1 is replaced with a chance function of belonging to one of the classes. Since the chance function is logistic, the algorithm is called Logistic Regression (LR). The function $f(x)$ with the best regression coefficients β computed for the new data point returns a probability value of belonging to the chosen class [207, 208].

Support Vector Machines (SVM)

The Support Vector Machines (SVMs) algorithm is one of the most popular supervised ML methods [209]. SVMs were used for discovery of gene-gene interactions, from Gene Wide Association Studies datasets [210], for cancer-driver gene discovery [211] and prediction of diabetes [212]. The idea of the support vectors is inherent to other algorithms like Support Vectors Regression (SVR), which solves the regression problem [213] and Support Vector Clustering (SVC) solving the clustering task.

In the basic configuration, SVMs distinguishes between two classes. The algorithm solves a regression equation of a hyperplane separating the data points belonging to the two classes while maximizing margin between them. The margin is

defined as a sum of the *support vectors* lengths. Each support vector starts in one of the data points and is perpendicular to the hyperplane separating the data points. Consequently, a result of SVM is a maximum margin classifier [214, 215].

If the data points are linearly separable the *hard margin* SVMs can be used. However, they rarely are. *Soft margin* SVMs allow some data points to violate the hyperplane, or in other words, they relax the definition of the margin to allow for errors. The error is parameterized so that the tradeoff between the accuracy and overfitting can be controlled [205].

However, independently on the margin, linear separation is often impossible to find. Therefore, the SVM algorithm uses a *kernel* function to transform data into another space, where hopefully the linear separation would be possible. Kernels transform the *attribute space* of the original dataset into the *feature space*. Thanks to kernel functions SVMs constitute a class of extremely versatile ML tools. Popular ML toolkits include large sets of kernels such as linear, polynomial or radial basis.

Ensemble methods

The ensemble machine learning algorithms classified do not share a mathematical core, rather they share a general approach. A single ensemble method classifier consists of multiple base classifiers. Their rationale is that a large number of weak classifiers improves generalization while reducing overfitting. Ensemble methods are further divided into averaging and boosting. Averaging methods attempt to construct a large group of independent base classifiers whereas the boosting methods build a collection of classifiers sequentially, with each classifier attempting to improve the previous one.

Random forest (RF) is an ensemble, bagging algorithm. Single RF classifier consists of multiple classification trees [216]. In a classification tree, each node represents a single decision on how to separate the input data. Splits consist of a feature and its value, based on which separation of the data points results in the two maximally clean groups in respect to the two classes. The features and values can be used multiple times within one tree, therefore, trees can grow quite large. Size of a tree is limited by parameters specifying the maximal depth and the minimal number of data points in a leaf. A tree is built by recursive data partitioning so that every split maximally decreases impurity of data subsets represented by subsequent nodes (CART algorithm for building trees [217]). While building a single tree, the algorithm uses a random subset of samples and selects the best feature from a randomly selected subset of features. The two-tier randomization ensures trees in an RF are independent. In the Extra-RF algorithm, the value of a split is also randomly selected. RF is parameterized by a number of trees in a forest and a number of features available for selection at each split. RF classifies a new observation via *asking* each tree for prediction and returning the class with the most votes.

While randomly choosing a subset of observations for each tree in the RF, there is on average 36.8% of observations that are never used. This subset of observations is called *Out of the bag* (OOB) and is used to compute accuracy. The RF algorithm enables independent variable importance measurements. For each feature, the algorithm will analyze splits that use it, and measure how efficient they are in the partitioning of the data using the Gini index.

The Adaptive Boosting algorithm (ADA boost) is another ensemble method. Similarly to RF, ADA boost makes a prediction based on the majority vote of all base classifiers. However, every classifier is *boosted*. Boosting lies in the subsequent modifications of training data. ADA boost assigns weights to each of the samples and modifies them with each iteration. The weight of the data point is increased if the classifiers failed to predict a proper class for it. Therefore, each classifier is consequently more fitted towards those problematic cases [218]. The base classifiers are often small decision trees.

Gradient Boosting algorithm approaches boosting as the optimization task in respect to the loss function. Algorithm modifies each sequential base classifier to minimize the value of the loss function [219].

Artificial neural networks

The Artifical Neural Networks algorithm (NN) constitutes a separate class of ML. A single Neural Network consists of layers of *neurons* or *perceptions*: a single input layer, several hidden layers, and one output layer. There might be multiple hidden layers, and each can comprise hundreds of perceptrons, depending on the complexity of the learning problem. Single perceptron accepts multiple inputs of real numbers and returns a numerical output. Perceptions of one layer pass values to those in the following layer, in the one to multiple manners. A perceptron is parametrized by weights, modifying inputs, bias, and activation [220].

The backpropagation algorithm enables finding the weights, biases, and activation values for the given neural network [221]. The backpropagation algorithm uses similar concepts as SGD. However, there are more dimensions to it, since the weights and activation values depend on the values from the previous layer. Hence, the propagation term, as it starts from the output layers, and moves towards input layers. It also uses the averaged quadratic cost function, that is, like in SGD, computed separately for each neuron.

2.6.2 Classifiers performance measurements

The input dataset defines the ML-task, and therefore, the main class of algorithms to be used: supervised or unsupervised. However, there remains an ample space of algorithms and parameters to search. Therefore, the methods enabling comparison of the performance of the classifiers are used so that selecting of the best-performing

combination of algorithm and parameters is possible. Below, several methods for classifier performance measurement are discussed.

Confusion matrices Confusion matrix refers to four values often computed to describe a performance of any tests. Confusion matrix contains a number of true positives (TP), the number of genuinely positive data points that was also predicted as such by the classifier. True negatives (TN), genuinely negative data points also predicted as belonging to the negative class. False positives (FP), negative data points predicted as positive, and false negatives (FN), positive data points that were falsely marked by positive. Confusion matrix enables computations of the other values such as accuracy and sensitivity. Accuracy is a portion of the correctly called data points of both classes ($\frac{TP+TN}{TP+TN+FP+FN}$). Sensitivity is a portion of the correctly called positive class data points out of the positive class data points predicted by the classifier ($\frac{TP}{TP+FN}$).

Cross-validation To test the performance of the classifier with a confusion matrix, first, the data has to be divided into train and test subsets. Often the dataset is divided into a numb (N) of subsets, the classifier is trained using the data of all but one sets ($N-1$) and tested on the one not used for training. These steps is repeated N times, and the average statistics for the runs is computed. This procedure is termed *cross-validation*. The final classifier is typically built using all of the available data [222].

Receiving Operator Characteristics (ROC) Receiving Operator Characteristics (ROC) is one of the classical methods of quantifying the performance of the classifier [223]. The ROC curve is defined as the relationship of the False Positive rate to the True Positive rate, depending on the thresholds. For RF, the threshold is the portion of trees needed to classify a data point as one of the classes. The area under the ROC curve (AUROC) denotes the overall performance of the classifier. The larger the AUROC, the better.

Overfitting Classifier training is typically optimized regarding accuracy. That strategy often leads to *overfitting*. This happens when classifiers learned so much on the training data they eventually fail to *generalize*. To control it, one needs to measure prediction accuracy for both train and test subsets. The difference between them favoring of the training set denotes overfitting, the larger the difference, the worse the classifier.

Permutation significance test Permutation significance measurement denotes if the classifier learns from the data at all. First, it trains the classifier on the training set and measures the accuracy based on the test set. Next, the class labels

are permuted, and the learning and testing steps are repeated. The proportion of runs where the classifiers with the mixed-labeled dataset achieved better accuracy than the native dataset is reported. The significant portion of the runs when the mixed dataset performed better denotes the classifier failed to learn.

2.6.3 Data pre-processing

The data preprocessing comes typically before or in-between algorithm selection steps, as feature selection and other data manipulation modify the dataset so that the algorithm selection needed to be repeated. This section outlines several aspects of the datasets and their pre-processing.

Curse of dimensionality ML often deals with datasets of high dimensionality. However, the more dimensions (features) dataset has, the more data points are needed to generalize. This relationship is exponential. This is known as the *curse of dimensionality* [224]. The *unsupervised* ML algorithms can deal with higher dimensionalities in respect to the number of data points, in comparison to the *supervised* methods [222]. This is one of the aspects that govern the choice of algorithms.

Variable importance and dimensionality reduction The supervised classifiers enable feature importance measurement called *permutation accuracy importance*. For a single variable at a time, its values are permuted among observations. The classifier is trained again for such dataset. The difference in accuracy between the classifier trained on the regular and permutated datasets speaks to how informative or *important* is the feature. *Permutation accuracy importance* was reported to be less biased then Gini-based method [225], in case of the RF algorithms.

The feature importance measurements provide a unique insight into the data. They can also drive the *feature selection* process that aims at removing uninformative features to reduce dimensionality and limit the complexity of the data. There are numerous strategies for dimensionality reduction, the choice of which depends on the data type [226].

Informative value of the features can be estimated with a univariate approach by evaluating the distribution of each feature and their relationship to the output vector. The most natural approach is removing features with low variance. Other approaches rely on testing if the distributions of the feature values differ between classes, with methods such as Analysis of variance (ANOVA) or chi-square tests. Features for which distributions of their values differ significantly between the classes are highly scored.

Data encoding Although the size of the dataset is an essential factor, researchers used to say *the data on its own is not enough, no matter how much of it you have.* Data for machine learning are typically in the form of a numerical matrix. However, real-world high dimensional datasets are almost never entirely numerical. There are no general rules on how to encode the specific data for the specific ML algorithms, other than that the encoding should faithfully represent the real world data values.

Unbalanced data Often the class of interest is the minority class, like the infected patients among all people admitted to the hospital, or MGE reads among WGS datasets. Unbalanced data is a common problem in medical-informatics [227]. A substantial difference in representation of the classes in a training set might cause a classifier to achieve high accuracy with zero sensitivity, and prevent it from using minority class. There are four strategies for dealing with the problem of unbalanced data. First is data stratification. This relies on finding a subset of the data with lower complexity. Other strategies include over-sampling of the minority class, under-sampling the majority class, and weighting of the data points.

Missing data No dataset is perfect. The majority of datasets is burdened with missing data, i.e., records lacking information for some of the features. Strategy for dealing with missing data depends on their pattern. There are three main categories of missing data: data missing completely at random (MCAR), at random (MAR) or not at random (NMAR). Data MCAR emerge from non-systematic errors. In data MAR, there is an underlying distribution that governs the distribution of the missing data. Oftenly data collected via questionnaires are burdened with MAR, where some people are more prone to leave specific questions unanswered. The last type is NMAR, where the pattern follows a clear structure, or there is a strong correlation between one of the features in the dataset and the missing values.

There are several strategies for dealing with MCAR and MAR, among which the most straightforward is *listwise deletion*, namely removing all incomplete data points. It does not introduce bias for MAR and MCAR. However, of course, those might be potentially valuable data points. In the case of NMAR, the reason for missing data determines the strategy to deal with it. Even removing all of the records with missing data in the case of NMAR might introduce bias.

Other methods rely on some forms of educated guessing. So that it does not introduce bias but allows to use those data points. Standard methods include average or median computation, maximum likelihood, or randomly drawing from a picked distribution. Naturally, the more records there are in the data set, with more confidence the value can be chosen as the distribution is more evident [228, 229].

Chapter 3

MRSA from colonization to infection

3.1 Introduction

In this project, the changes in the genomes of the Methicillin-resistant *Staphylococcus Aureus* (MRSA) isolates were investigated. The strains were isolated from the SATURN patients undergoing antibiotic therapy, who were negative at hospital admission, but had at least two or more positive MRSA samples, and developed an MRSA infection within 30 days of the follow-up.

Five patients fulfilled that criteria: SE1582, SE1884, SE1890, SE1895, and SE2054. All of them were treated in the same hospital. Each of the patients was characterized by individual features and underwent unique antibiotic treatment (Table 3.1). Here I compared the genetic makeup of the MRSA strains isolated from different sites: the nose, lungs, and at different points during the therapy (Fig. 3.1).

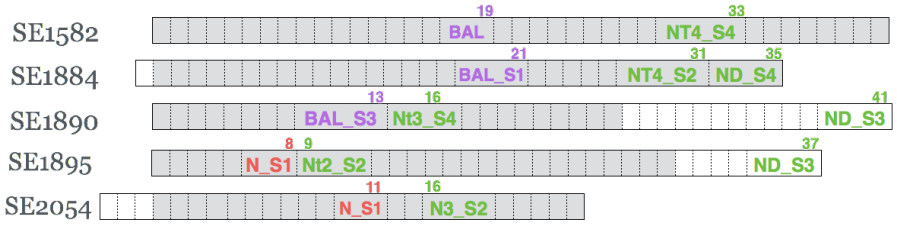


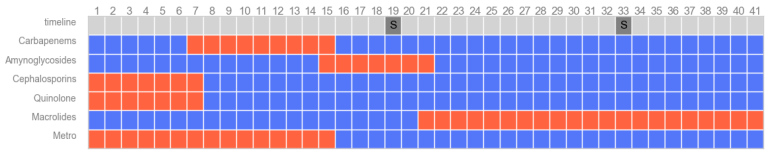
Figure 3.1: Sampling pattern in relation to antibiotic therapy (gray fields) and overall hospital stay (white fields). The nasal strains (N) were the first colonization strains, and those found in sputum (BAL) were directly responsible for the infection. Numbers denote days from the beginning of therapy.

The group of Prof. Surbhi Malhotra-Kumar from the University of Antwerp isolated MRSA strains from the clinical samples of the selected patients and performed sequencing of the DNA isolated from the colonies identified as MRSA. My role was to analyze the differences in the genomes of the MRSAs while the developing infection, and throughout the treatment.

Table 3.1: Patients' features and treatment schedules. Subplots are labeled by patient identifier (e.g. SE1582). Days on which samples were taken are denoted with an `S` in the timeline track.

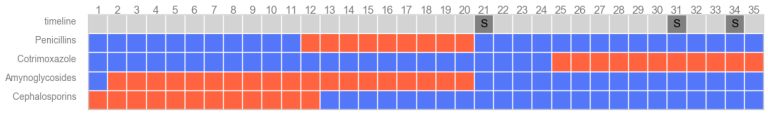
SE1582

- 66-year-old man
- Underwent a surgery
- MRSA-positive roommate
- Admitted from LTCF



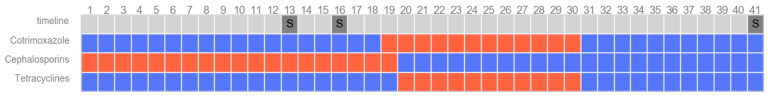
SE1884

- 79-year-old woman
- Admitted from acute care
- Underwent a surgery
- Cardiovascular disease
- Diabetes



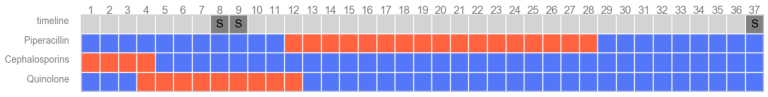
SE1890

- 54-year-old man
- Admitted from acute care
- Overweight
- Cardiovascular disease



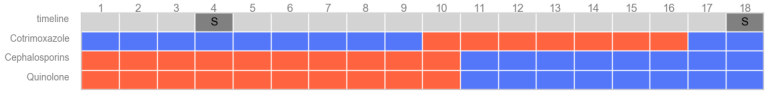
SE1895

- 46-year-old woman
- Admitted from acute care
- Underwent a surgery
- Skin lesions



SE2054

- 62-year-old man
- Admitted from home
- Cardiovascular disease
- Myocardial infarction
- Underwent a surgery



All patients were characterized by the multiple features known to increase the probability of the MRSA infection. All of the patients underwent an extensive antibiotic therapy, which was prolonged by the MRSA respiratory infection (Table 3.1), and none of them took antibiotics before the hospital admission. The majority of patients came to the hospital from another care facility. Four patients underwent surgery. Three patients had cardiovascular disease and one suffered from diabetes.

3.2 Bioinformatics pipeline

The dataset consisted of MiSeq sequencing of the MRSA genomes isolated from thirteen samples of the five patients. The size of the dataset and the high variability of features impeded comparison between the patients. Therefore, an analysis focused on the intra-patient differences. Two strategies were employed: reference-guided assembly and variant calling (Fig. 3.2).

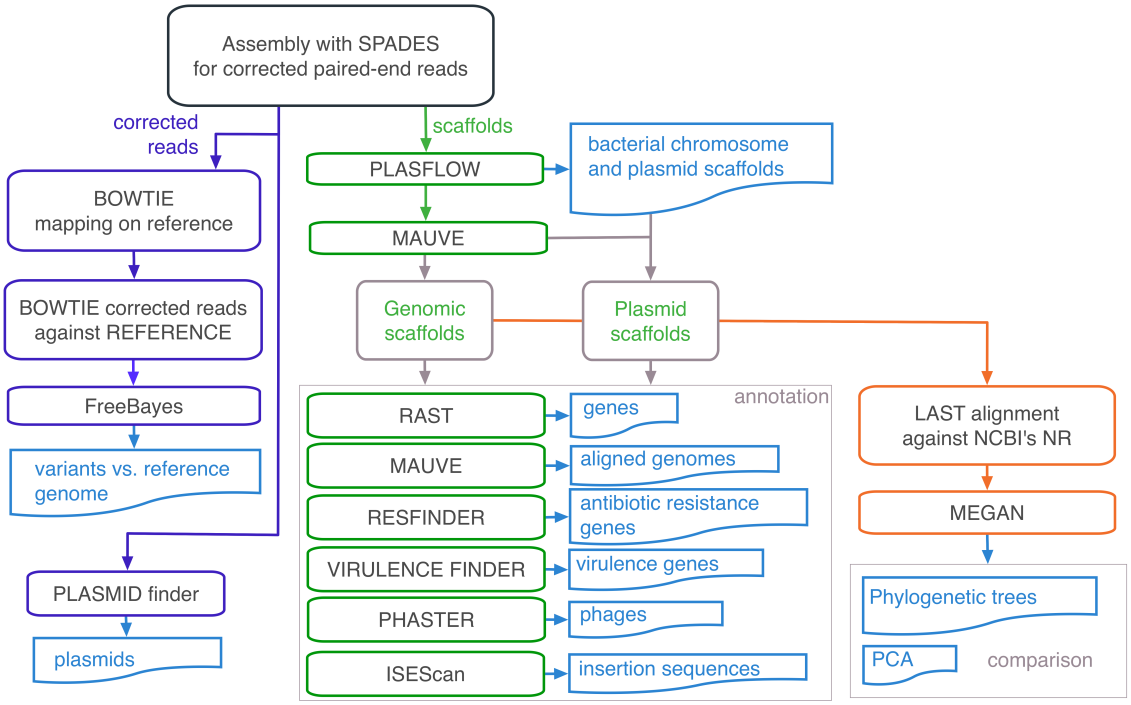


Figure 3.2: Overview of the MRSA's genome assembly and analysis pipeline. Green color denotes the main assembly and annotation and violet the variant calling parts of the pipeline. Blue boxes describe what information was obtained at each of the steps.

Operons of rRNA genes constitute the largest repetitive region in the MRSA genome. A single rRNA operon consists typically of multiple copies of genes encoding 16S, 23S, 5S rRNAs, and tRNAs. As assembly of repetitive regions from short sequencing reads poses a computational problem, rRNA reads were filtered out for a separate assembly. First, in the reference genome regions annotated as RNA were identified. They were merged if they were located less then 1000 bp apart. Next, the corrected reads were mapped onto them. The mapped reads, along with their mates regardless whether they were mapped or not, were separated into two sets. They were assembled separately with Spades, rRNAs with parameters: `-careful-k 19,21,33,35 -cov-cutoff 1` and the rest of the reads with parameters: `-careful-k 19,21,33,35 -cov-cutoff 10`.

The contigs were aligned against the NCBI NT database with BLAST so that the potential contaminants could be removed. The filtered scaffolds were then ordered using a reference genome and the scaffold-builder web-server. Next, RAST was used for annotation. Several web servers were used to answer specific questions: ResFinder for ARGs identification, PHAST for phage identification and VirulenceFinder to find virulence genes. Below the methods used in the pipeline (Fig. 3.2) are briefly described. The pipeline starts from the scaffolds of the two separated assembly runs merged into one file.

- Read mapping with Bowtie [230]
  Bowtie was used to map the reads against the MRSA reference genome *Staphylococcus aureus* subsp. NCTC 8325 chromosome (NC_007795.1). Next, samtools [231] were used for manipulating and analysis of the output files.

- Assembly with Spades [232]
  There are so many assembly programs that several publications were written to compare them [233, 234, 235]. Most of the assemblers implement de Bruijn graphs but vary in their post-processing. Spades was selected because it included the reads correction procedure. It also implements a *multisized de Bruijn graph* (using different k-mer sizes) and good algorithms for dealing with bulge/tip and chimeric sequences. Additionally, Spades was described as exceptionally efficient in utilizing pair-end reads.

- Plasflow [176]
  Plasflow is one of the recently released tools using k-mer and machine learning methods to classify sequences as bacterial chromosomes or plasmids. It was used to remove plasmids before scaffold reordering.

- Contig reordering and merging with Scaffold Builder [236]
  Scaffold Builder is a web-server for ordering scaffolds or contigs using the reference genome. Unlike Mauve the Scaffold Builder can connect sequential scaffolds if their terminal sequences are highly similar.

- Assembly quality measurements
  For the single-strain assembly task addressed in this project, the quality measurements relying on the size of the scaffolds and comparisons to the reference could be used. The most popular characteristics of assembly quality are N50 and N90 [237]. N50 is the highest threshold, such that all contigs longer than the threshold cover $\geqslant 50\%$ of the reference genome, for N90 the contigs cover $\geqslant 90\%$ of the reference. Therefore the best is the assembly with the largest N50 or N90 characteristics.

- Genome alignment with Mauve [238]
  Mauve stans for Multiple Alignment of Conserved Genomic Sequence With Rearrangements. The Mauve program is often used for whole-genome comparison. It computes Locally Collinear Blocks (LCB) that indicate corresponding fragments of assemblies. Mauve was used for scaffold reordering and alignment of the genomes of strains isolated from the same patient.

- Genome annotation with RAST [239, 240, 241]
  RAST stands for Rapid Annotations using Subsystems Technology. It is an annotation server for genome assemblies of known bacteria. An input to RAST consists of contigs and a taxonomic identifier of the closest relative of the assembled bacteria. Briefly, the RAST pipeline relies on gene prediction and annotation of gene function with BLASTP. However, it is quite precise since it performs multiple iterations of filtering and comparisons with the reference. The genes are assigned to pathways that in RAST are termed subsystems.

- Genome annotation with myRAST [239, 240, 241]
  myRAST is a desktop edition of RAST, which also enables annotation of genomes. However, instead of using the provided reference, it computes the taxonomic correspondences locally. Therefore it can annotate fragments of genomes, like metagenomic assemblies. In this project, myRAST was used to annotate subsystems on all sequences, and to annotate plasmid sequences.

- Annotation of resistance genes, plasmids and virulence genes, with ResFinder [242], PlasmidFinder [243], VirulenceFinder [244]
  These tools employ BLAST programs to compare the input genomes or assemblies with the curated protein databases. In the case of PlasmidFinder and ViruelnceFinder, the databases are species-specific. Only the hits above 90% of similarity and covering more than 60% of length are taken into account.

- ISEScan [245]
  ISEScan detects insertion sequences (ISs). An insertion sequence consists of an ORF encoding a transposase, and two flanking inverted repeats (IR). ISes contribute to HGT. ISEScan is an HMM-based tool. It aligns protein sequences of the predicted genes to the transposase HMM profile. In the second step, ISEScan finds the IRs.

- PHASTER [246]
  The Phage Search Tool PHASTER is a web-server for finding pro-phage sequences in bacterial genomes. An input can be a genome sequence or a scaffold file. It uses GLIMMER for gene prediction, and BLASTP to compare translated ORFs with the prophage protein database. A cluster of phage-like genes is denoted a prophage if it has at least six genes annotated as one of the protease, integrase or tail, and contains an integrase and a potential phage attachment site (att). For each site, a completeness score is assigned. Finally, a list of phages and scores is returned, along with the classification of the hit's incompleteness: *questionable*, *incomplete* or *intact*.

- Variant calling with FreeBayes [247]
  FreeBayes generates a variant call format file (VCF) from an alignment of reads onto the reference genome. It uses a probabilistic model to decide whether the identified difference is a SNP or an INDEL.

- Function-based comparisons with LAST+MEGAN-LR [248, 148]
  The assembled scaffolds can be treated as long reads and the MRSA genome sequencing datasets as easy metagenomic datasets. The assembled scaffolds are aligned with LAST against the NCBI NR database. The MRSA scaffolds were analyzed with MEGAN-LR pipeline and compared with PCAs and phylogenetic trees using functional annotations.

The computations were performed using in-house servers and Baden-Württemberg's HPC services of the BwUniCluster. For the pipeline, plotting, and analysis the Python3.6 programming language was used [249, 250].

3.3 Results of assembly and annotation

3.3.1 Raw data and contamination

The dataset consists of thirteen MiSeq sequencing samples. Reads were pair-ended and 120 nucleotides long. Regarding read numbers the samples were quite variable, as the smallest had $\sim 65,000$ reads and the largest $\sim 300,000$ reads (Table D.1).

The corrected reads were mapped against the reference genome, and the reads that did not map were aligned against the NCBI NT database with BLAST. Only one hit with a very low $e - value \leqslant 10^{-4}$ was allowed for each read. The most represented species were listed. On the one hand, I expected the unmapped reads to align well to genomes of other *S. aureus* strains. On the other hand, if the majority of the reads mapped to a single species, then that would be a strong indication the sample was contaminated. Table 3.2 outlines the most common BLAST hits among the unmapped reads. Among the most hits were transposons, plasmids, and phage sequences. Therefore, the samples were not contaminated.

Table 3.2: Twenty most represented hits among the BLAST alignment of the reads without mapping to the reference genome.

Hit name	Read count	Samples
S. aureus subsp. aureus strain FORC_001	905,883	Present in all samples
S. aureus strain CA15	750,043	Present in all samples
S. capitis CR01 complete genome	573,553	Present in all samples
S. aureus strain RKI4	540,755	Present in all samples
S. aureus strain M121	428,873	Present in all samples
S. aureus strain FCFHV36	391,144	Present in all samples
S. aureus genome assembly NCTC13435	303,944	Present in all samples
S. aureus subsp. aureus strain Gv69	298,285	Present in all samples
S. aureus DNA	538,486	Present in all samples
E. faecium strain E240 transposon Tn5801	333,206	Not present in SE2054
Staphylococcus phage B166	246,640	Not present in SE2054
S. aureus subsp. aureus Z172 plasmid pZ172_1	280,506	Not present in SE1890
S. aureus plasmid SAP104A	531,380	Not present in SE1890, SE2054
S. aureus plasmid rep	360,092	Only in SE1582BAL, SE1890*

3.3.2 Assembly quality

Table D.2 presents the basic assembly statistics. For each sample, the rRNA and non-rRNA scaffolds were put together and subsequently reordered against the reference genome (Table D.3). In the majority, the samples were characterized by similar values. They had between 500 and 700 scaffolds, with an overall size of $\sim$ 2.8Mb, which was close to the size of the reference genome. The largest scaffold was $\sim$ 1.0Mb long and an average scaffold length was $\sim 5,600$bp. Samples SE1582BAL_S1, SE1884NT4_S2, and SE1890ND_S3 deviated from those average values. More sequence and more scaffolds characterized the first two samples, and the last sample assembled quite well into $\sim$ 300 scaffolds.

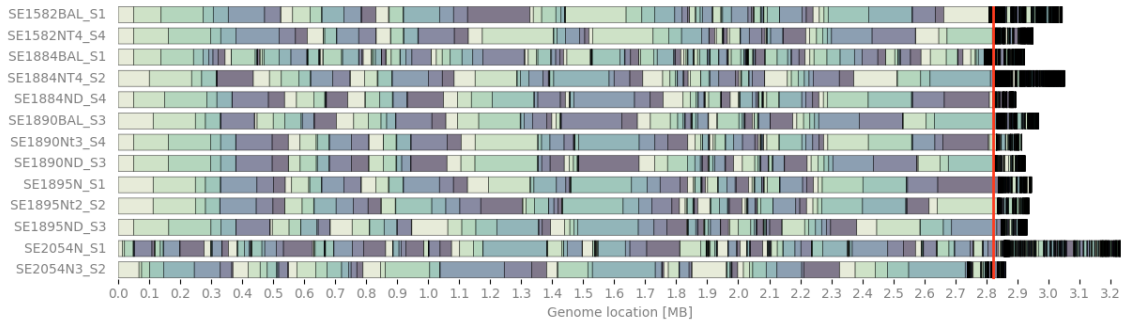


Figure 3.3: Assembly fragmentation. Each box denotes a scaffold. The red line denotes size of the reference genome.

The assembly fragmentation was similar for all samples, except the assemblies of the samples SE1582BAL_S1 and SE2054N_S1, which were visibly more fragmented (Fig. 3.3). All of the assemblies had a tail of small-sized scaffolds, but the SE2054N_S1 sample had an exceptionally long one and was visibly more

fragmented. The more significant amount of sequence and fragmentation of the assembly of some samples might suggest the isolates carried plasmids.

3.3.3 Plasmid identification

A single MRSA cell can contain up to eight plasmids [251] with a wide range of sizes (Fig. D.1). Plasmid sequences are versatile, and their fragments are often found in bacterial chromosomes. Therefore reads cannot be universally filtered based on their mapping to known plasmid sequences. Accordingly, a multistage pipeline was employed to exclude potential plasmid scaffolds. First, the PlasFinder for reads was executed.

Plasmid Finder was only able to assemble fragments of the known plasmids (Fig. D.2). Those fragmentary plasmid hits enabled a comparison between the samples. For the majority of the patients, the plasmid pattern did not change during the sampling. Therefore, there was no instance of plasmid acquisition, which would have made a strain infectious. The only exception was the SE1582NT4_S4 sample, which had fragments of pKH3 and pKH12 plasmids, where the sample of the previous timestep (SE1582BAL_S1) contained fragments of pDLK1, pNE131, and pKH12 plasmids.

3.3.4 Reference-based scaffold ordering

Next, the scaffolds were reordered again, with two different programs Scaffold builder and Mauve. Several selections of scaffolds such as excluding low coverage, short or plasmid scaffolds, were used. However, these reordering attempts failed, as either the programs finished with an error or the alignment of the assembly to the reference genome showed too many rearrangements. Therefore, the plasmid identification was insufficient. Consequently, I used the k-mer-based tool Plasflow [176], to identify and remove plasmid scaffolds from the assembly.

The Plasflow program split scaffolds into three classes: *plasmids*, *bacterial chromosomes* and *not classified*. The chromosome and unclassified classes were merged as they both could include prophages. Next, the bacterial chromosomes and unclassified scaffolds were reordered with Mauve. The resulting LCB-based alignments to the reference genomes are presented in Fig. 3.4.

The alignments show evidence for rare and small rearrangements with respect to the reference genome. Neither of the assemblies contained the $\sim$ 0.5 Mb fragment of the reference located in the middle of the genome between $\sim$ $1.4 and 1.5$ Mb. Conversely, all assemblies, excluding those of the patient SE2054 isolates, contained a large block in the beginning, which did not have a counterpart in the reference. At this point, after Plasflow-based scaffold separation, Mauve reordering and the alignment, the assembled scaffolds fall into two categories: plasmid scaffolds and genome scaffolds (Fig. 3.5).

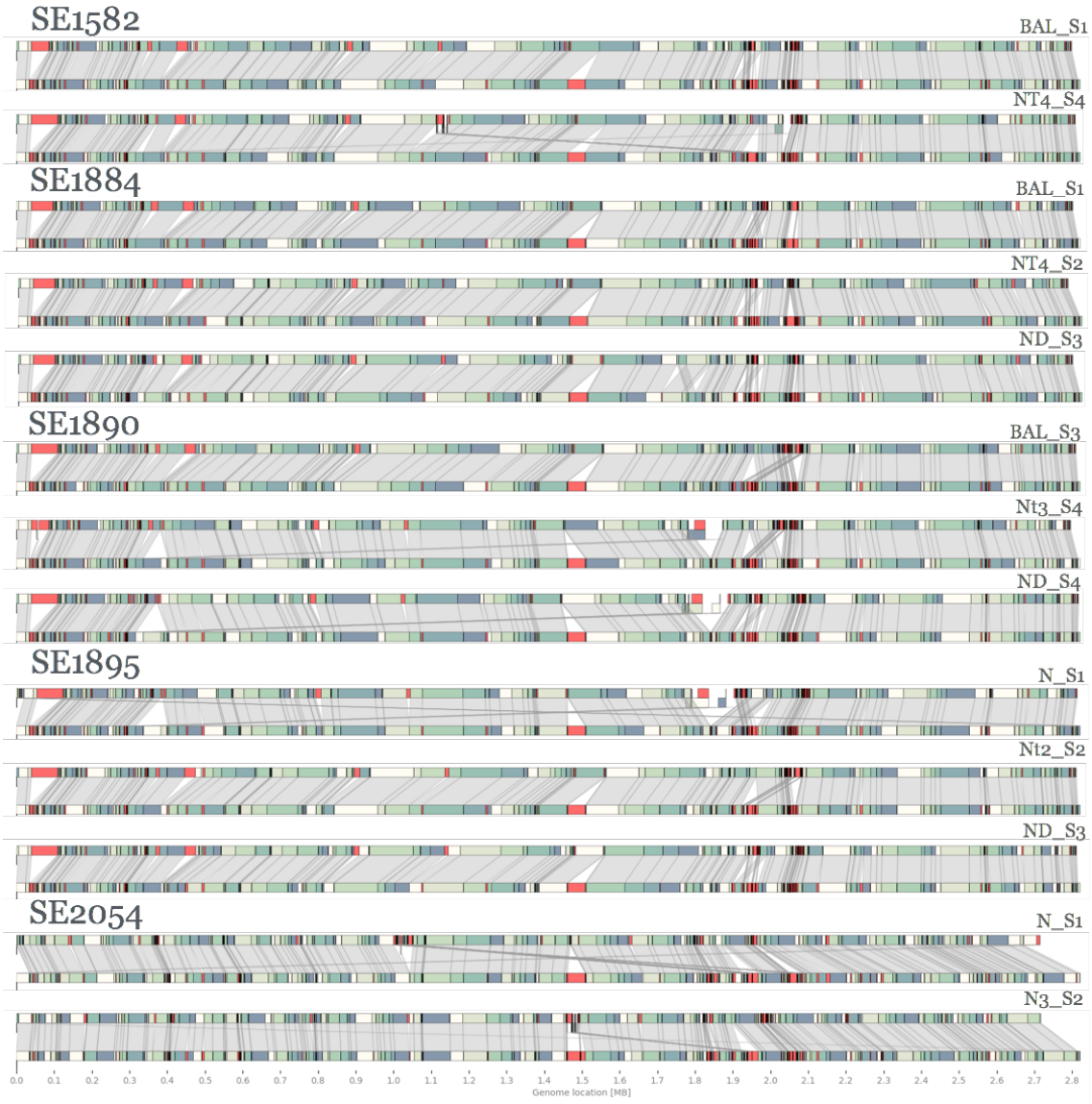


Figure 3.4: The LCB-alignments of the assembled genomes and the reference located on the latter bar. Red color denotes LCBs without a corresponding block.

The plasmid scaffold group consisted of the scaffolds annotated as plasmids by Plasflow and chromosome scaffolds with conflicting rearrangements. The genome scaffold group contained scaffolds used for rearrangement regardless whether they had a correspondence in the reference. The final group comprised fragments of the reference sequence that were not covered by the assembly. For the assembled genomes, those were regarded as deletions. Analogously, the sequences in the assembly without a corresponding block in the reference was termed insertions.

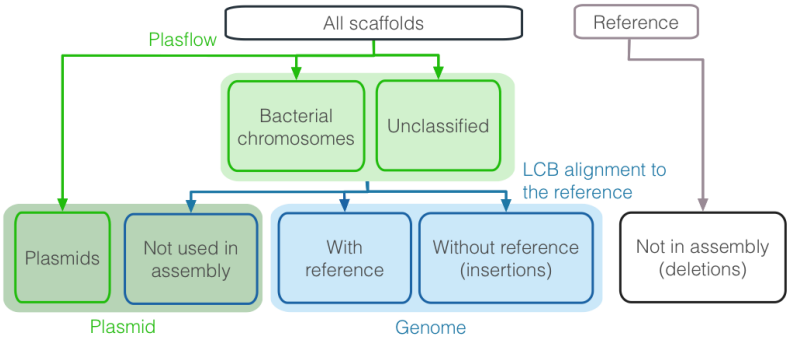


Figure 3.5: Scaffold groups identified through LCB alignment to the reference.

Characteristics of deletions and insertions

The deletions carried genes belonging to multiple subsystems (Fig. 3.6). In all of the samples, the most significant number of genes referred to the Phage subsystem, suggesting that the strains differed from the reference genome in a large portion by prophages.

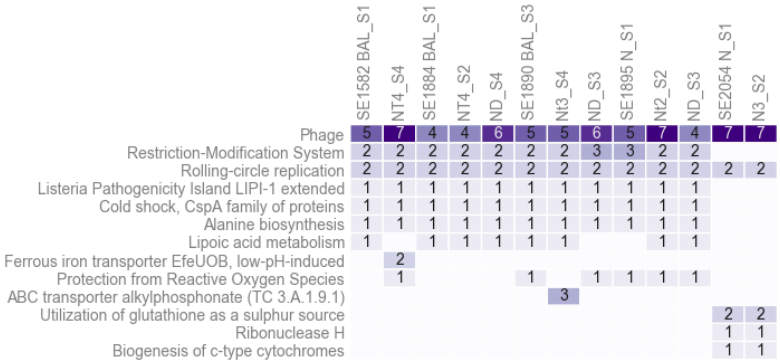


Figure 3.6: The number of genes found in the deletions based on the myRAST subsystems.

The patterns fall into two groups. Samples of the four patients SE1582, SE1884, SE1890 and SE1895 had a similar pattern, whereas samples of the SE2054 patient differed visibly. This suggests the patient SE2054 was infected with different MRSA strain than the other patients, for which the reference genome was not the closest known relative.

The gene content of the insertions was much less consistent. The insertions were characterized by a unique set of subsystems represented by a single gene (Fig. 3.7). There was no correspondence between the samples of the same patients. This suggests that the insertions were random, and therefore, were signs of the errors in sequencing and assembly process, rather than of a biological phenomenon. Therefore,

from now on, the analysis focused on the three groups of scaffolds: genome, plasmid, and deletions.

Figure 3.7: Numbers of genes found in the insertions. The labels denote myRAST subsystems.

3.3.5 Phage detection and characteristics

The previous step revealed that phage genes constituted the most significant portion of the genetic content of the deletions. Hence, the next steps included phage taxonomic annotation. First, VirFinder was used to determine the relative abundance of prophages in different scaffold groups. The distribution of the resulting p-values (Fig. 3.8), showed that scaffolds of the deletions and plasmids were enriched with phages in comparison to the genomic scaffolds.

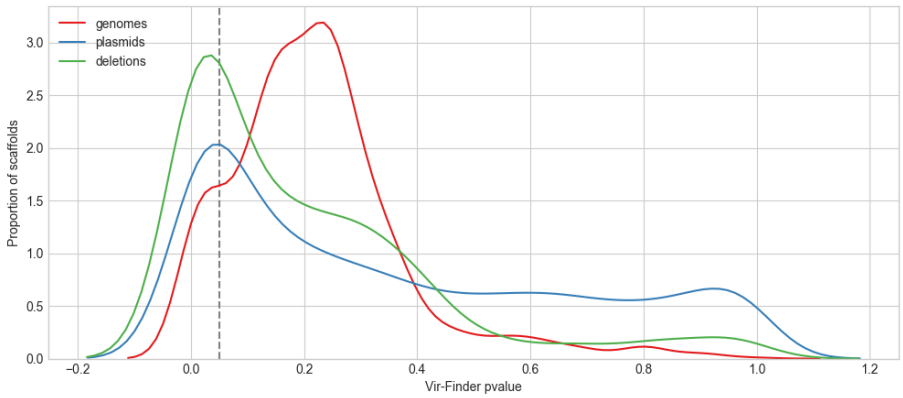


Figure 3.8: VirFinder p-values for scaffolds for all samples, separated by the scaffolds groups. The dashed line represents the p-value cutoff (0.05), below which the scaffolds are assumed to contain phages.

These results supported the separation of the scaffolds, but also the choice of the reference. Phages constitute a vital part of the MRSA's genome and contribute to the virulence. Several phage strains encode toxins, such as toxin A and leukocidin [252]. In the next step, PHASTER web server was used to confirm the results of VirFinder and to identify the phages (Fig. 3.9).

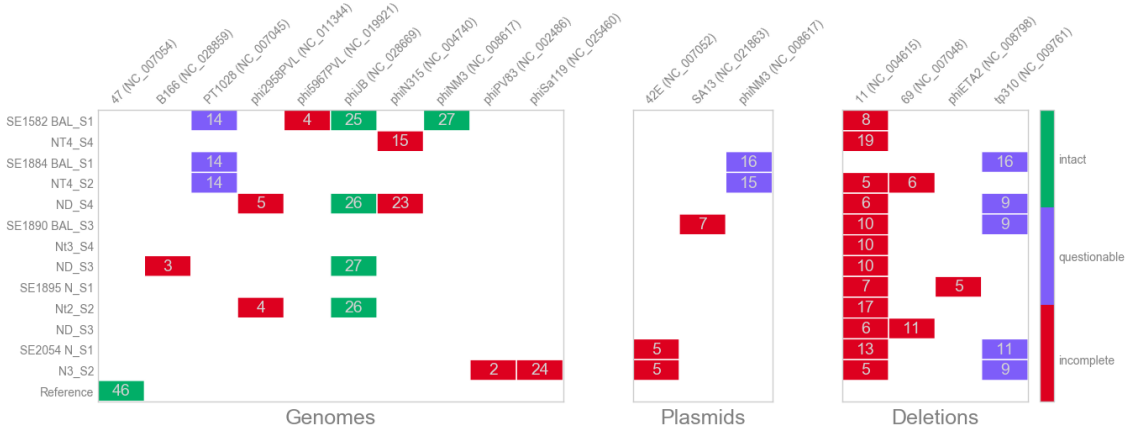


Figure 3.9: Phages identified by PHASTER in the three groups of scaffolds: the assembled genomes, plasmids, and deletions. Numbers on the heatmap represent the number of proteins. There were multiple phage taxa assigned to every region. In the figure, only the most frequent phage name is shown.

PHASTER identified phage regions in the deletions and plasmids, which confirmed the results of VirFinder. However, none of them included an intact phage. The majority of samples contained an incomplete Phage 11 in the deletions. In the plasmid sequences, PHASTER found several phage proteins, which is not uncommon, as the phage and plasmid proteins are often mislabeled in the database.

The samples of the same patient did not share the phage pattern. This was unexpected, as the rest of the features like resistant genes or insertion sequences, showed a strong clustering within the single patient. Only two phages: phiJB and phiNM3 were found intact, and they were located only in the genomic sequences. Both of them are transducing, and able to transfer resistance genes, which has been shown before [253]. They could refer to one phage since they share 93% of the sequence.

The reference genome contained only one intact phage 47, which was not found in any of the sequencing samples. This means that the incomplete and questionable phages were artifacts of the sequencing, assembly, and reordering. Nevertheless, almost every sample had between 10 and 40 phage proteins.

3.3.6 Insertion sequences

Fig. 3.10 presents the number of detected insertion sequences by family and scaffold group. Insertion sequences are the smallest possible MGE. They flank transposons. All of the deletions, except those of the SE2054 patient, had two IS30. With those, the phage proteins found by PHASTER, and the high proportion of phage sequence detected by VirFinder, it is safe to say the gaps in the assembly in comparison to the reference sequence, were mostly due to the MGEs.

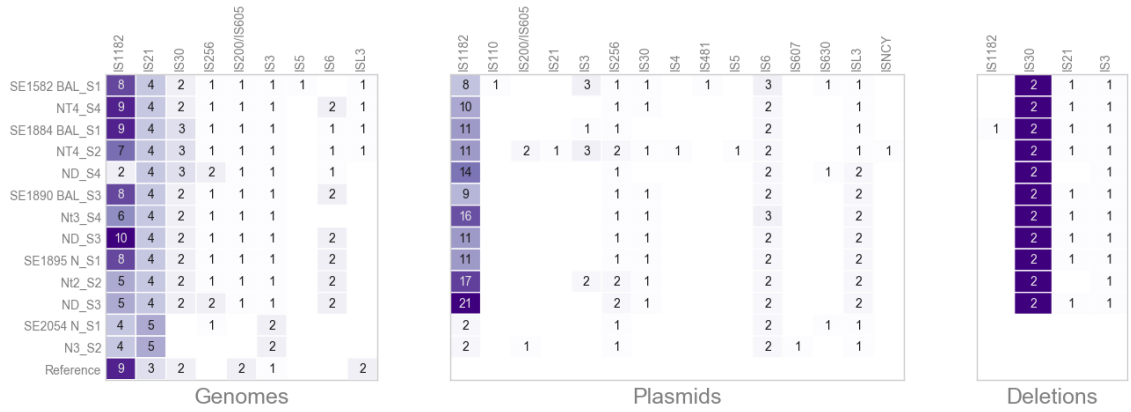


Figure 3.10: The composition of the IS families.

Unlike in the case of phages, the pattern of ISes was much more consistent within the patient samples. In each of the scaffold group, two IS families dominated the distribution of the IS families. IS1182 and IS21 dominated the genome sequences, and IS1182 dominated plasmid sequences. In the genomes, the IS family distribution resembled the IS distribution of the reference. This confirms the correctness of the assembly.

3.3.7 Genetic rearrangements between isolates

The next step was an alignment of the assembled and ordered MRSA genomes to each other within a single patient. The alignment revealed small rearrangements, and almost no insertions or deletions (Fig. 3.11).

The rearrangements were mostly localized around the 2.0 Mb position, and right where the average VirFinder p-values decreased (Fig. 3.12). Suggesting this region was enriched with phages. This could explain the rearrangements but also the misassembly. Therefore, in the following step focused on the differences between the annotations.

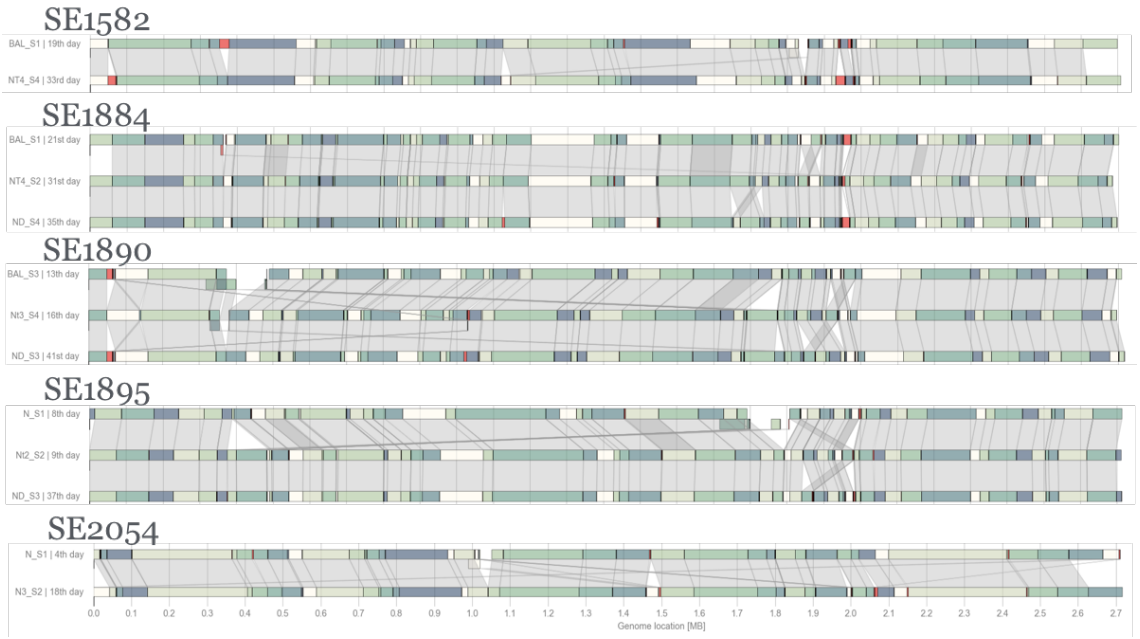


Figure 3.11: Mauve alignment of samples within one patient. Red color denotes the elements that lacked in the reference in at least one of the other sequences.

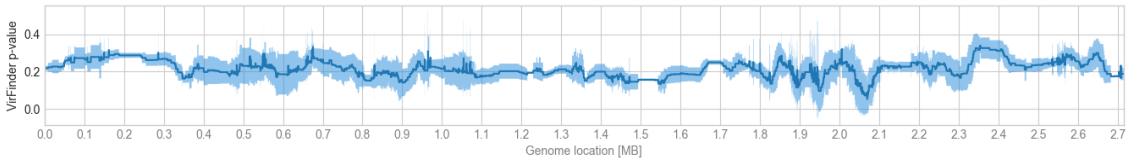


Figure 3.12: The line represents VirFinder p-value averaged across all samples, and the light blue area indicates the standard deviation.

3.3.8 Basic annotation statistics

The assembled genomes were annotated with the myRAST web-server using default settings. Table 3.3 presents the numbers of genes per sample in the genomic sequences and plasmid sequences.

In the majority of the samples, the rRNA operon read filtering seemed not to work correctly, as there were several rRNA operons found in the non-RNA read assembly (not shown). Although the separation was imperfect, it improved the assembly. In the majority of the samples, the numbers of rRNAs, tRNAs, and CDSs were similar to those of the reference genome. The exceptions were samples SE1582NT4_S4 with a lower number of tRNAs and SE1890Nt3_S4 with a low number of rRNAs. Sample SE1582NT4_S4 also had abnormally low N90, which suggests that it did not assemble well.

Table 3.3: Numbers of annotated tRNAs, rRNAs and genes (CDS) in all of the samples for the two scaffold groups: Plasmids and Genomes.

	Plasmids			Genomes		
Sample	CDS	tRNA	rRNA	CDS	tRNA	rRNA
SE1582						
BAL_S1	150	18	4	2,767	33	20
NT4_S4	77	30	2	3,127	64	25
SE1884						
BAL_S1	102	19	1	2,760	32	20
NT4_S2	180	29	2	2,754	30	20
ND_S4	87	35	3	2,759	33	18
SE1890						
BAL_S3	101	22	2	3,204	61	23
Nt3_S4	91	23	3	3,054	66	23
ND_S3	79	0	1	2,803	41	19
SE1895						
N_S1	86	0	0	2,787	33	19
Nt2_S2	83	22	1	2,764	33	21
ND_S3	105	40	3	2,781	23	19
SE2054						
N_S1	128	29	0	2,646	35	22
N3_S2	121	25	0	2,637	37	22
Reference				2,796	69	34

3.3.9 Plasmids

Several subsystems found in the plasmid sequences supported the results that those sequences contained MGEs (Fig 3.13). The transposon Tn522 subsystem was well represented in the samples of the patients SE1582, SE1895, and two samples of the SE1890 patient. Two first samples of the SE1884 patient had some phage genes. In the majority of the samples β-lactamase was found. Plasmid sequences in all of the samples had a gene conferring resistance to streptothricin. Finally, samples of the SE2054 contained a well-represented subsystem responsible for cadmium resistance.

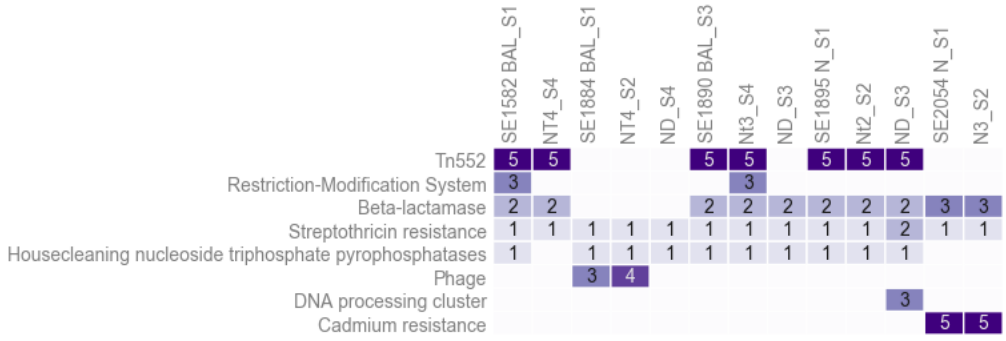


Figure 3.13: Subsystem content in the plasmid sequences.

3.3.10 Gene-level differences across sampling

Subsystems provide a high-level classification of the annotated functions. However, only between 30-40% of the annotated genes were assigned to a subsystem. This was also the case for the reference genome. Therefore, by focusing on the subsystems, much information could have been lost. Consequently, the gene-level annotations were compared, although for visualization some of the genes were grouped.

Figure 3.14: Heatmaps presenting the number of genes that differ between the samples within a single patient. Some of the genes, like synthetases, phages or prophage-associated genes were grouped.

Fig. 3.14 presents the genes, numbers of which differ between samples. The majority of those genes were enzymes of the basal metabolism, such as synthases, transferases, and non-enzymes such as tRNAs. Also, large portions of the differentiating genes in almost all of the samples were phage-related, which agrees with the results discussed previously.

The number of differentiating genes corresponded to neither the time between the samples nor the level of the LCB-based rearrangements. The samples of the SE1895 patient had the most time between them (29 days) also turned out to have the smallest number of differentiating genes. However, the SE1582 patient took the most significant number (six) of varying antibiotic classes. One of the patients took four antibiotics, and the rest of the patients' antibiotic therapies comprised three antibiotic classes.

For all but the SE1895 patient, differences in the multiple resistance genes such as gentamycin, penicillin, erythromycin, tetracyclines, and methicillin resistance genes were observed. In many samples, the number of virulence genes such as exotoxins and serine proteases fluctuated. However, overall no pattern in the changes of gene numbers was observed.

3.3.11 Antibiotics resistance and virulence factors

Two gene classes determine the overall pathogenicity of MRSA strains: virulence factors such as toxins, adhesins, and antibiotic resistance genes. Two specialized tools to annotate those features VirFinder and ResFinder were used. Since the previous steps showed that the genome/plasmid separation could be imperfect, the tools were applied to both of those sets of scaffolds (Fig. 3.15).

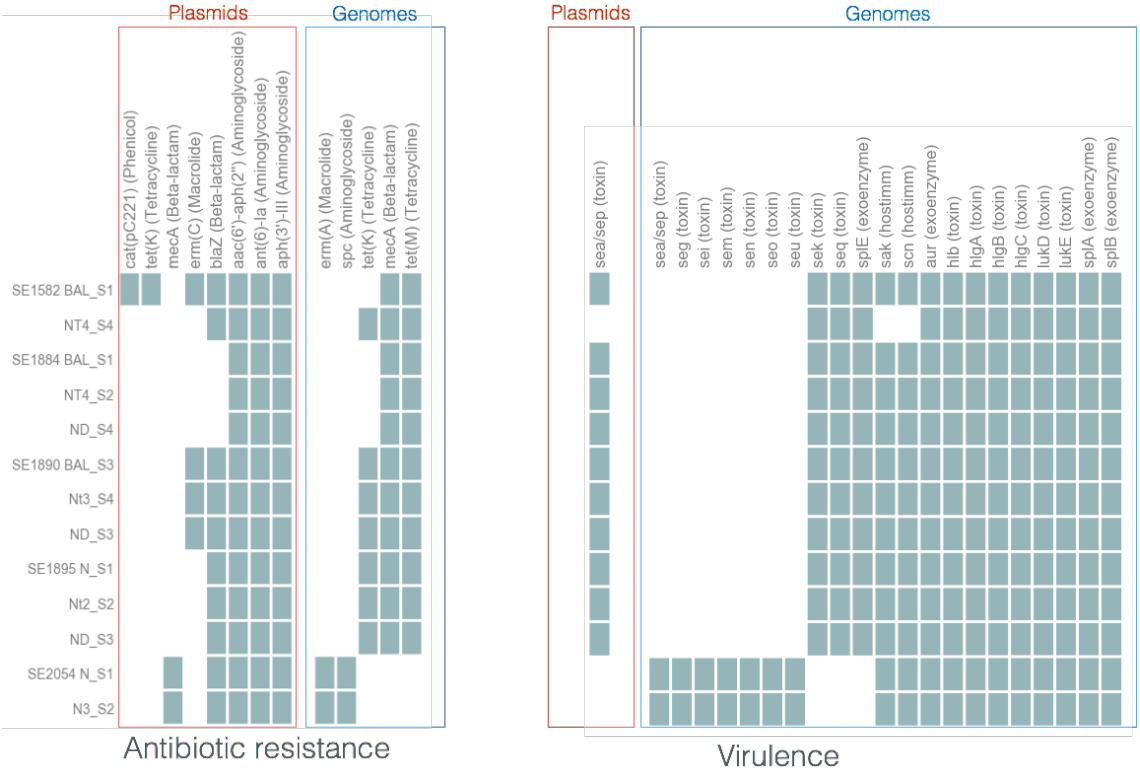


Figure 3.15: Presesnce/absence matrix of the resistance genes and virulence factors.

All of the isolates carried a typical for MRSA arsenal of toxins including hemolysins (hlg), leukotoxins (luk), staphylokinase (sak) and aureolysin (aur). With them, MRSA cells can attack the red blood cells, white blood cells, plasminogen, and fight inflammation, respectively. Enterotoxins such as sei, sem, sen, seo that attack the intestine, were observed in samples of the SE2054 patient.

The MRSAs proved to be true MDRs as they contained tetracycline, macrolide-resistant genes, and beta-lactamases. In most of the cases, patterns of the virulence and resistance factors cluster by the patient. All strains contained mecA conferring resistance to methicillin. The mecA is a marker for MRSA. Therefore, it is worrisome the mecA of the SE2054 patient was localized on the plasmid scaffolds. This supports the hypotheses that the strains isolated from the SE2054 patient were more distant from the reference than the other isolates.

3.4 MEGAN analysis

To determine relationships between the isolates the LAST+MEGAN-LR pipeline was performed separately for the genomic and plasmid sequences. MEGAN assigned all of the genomic scaffolds to the *Staphylococcus aureus* node or to species nodes below it. The plasmid scaffolds went mostly to the bacterial node, which was to be expected, however, unexpectedly there were also viruses and Eukaryota found (Fig. D.4).

On the one hand, the genomes of the isolates from the SE2054 patient were similar to each other but distant from all of the other isolates (Fig. 3.16). The majority of the isolates did not group by the patient but remained close to each other. On the other hand, in the plasmid tree, all of the samples were equally distant from each other. The trees agreed with the results of other methods.

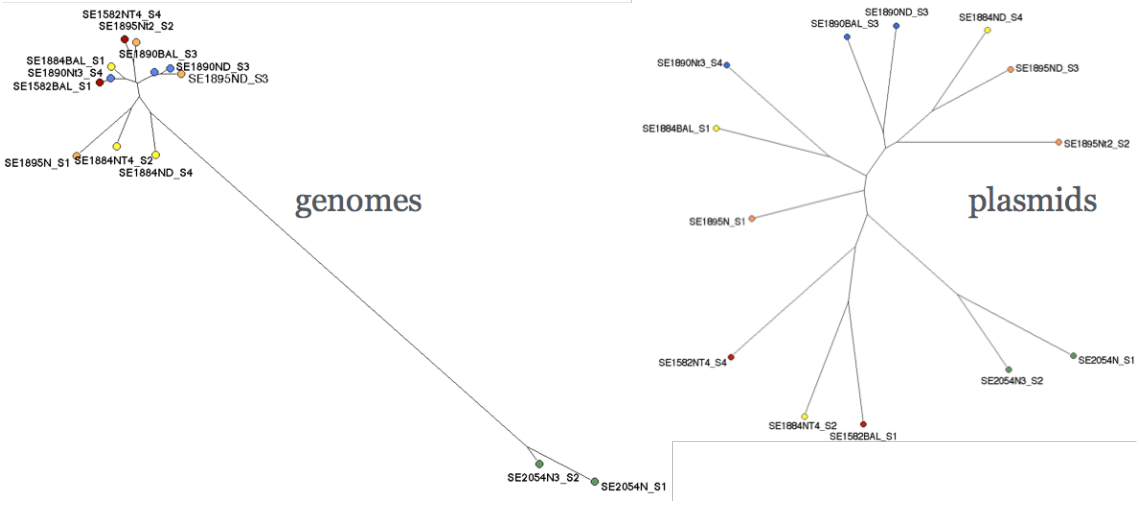


Figure 3.16: Neighbor-Joining tree based on the functional assignments (Interpro2Go).

3.5 Variant calling

Variant calling was performed to reveal nucleotide-level differences in the strains. A typical variant calling protocol consists of mapping reads onto a reference genome and its analysis with a statistical model. The FreeBayes program identifies two types of variants: single nucleotide polymorphisms (SNPs) and short insertions/deletions (INDELS).

For the cleaned read mapping against the reference genome, the FreeBayes program was used. The SNPs and INDELs with the lowest quality were excluded (Fig. D.3). Fig. 3.17 shows the number of SNPs and INDELs per gene and sample. Once more, the samples of the SE2054 patient largely differed than the rest of the isolates.

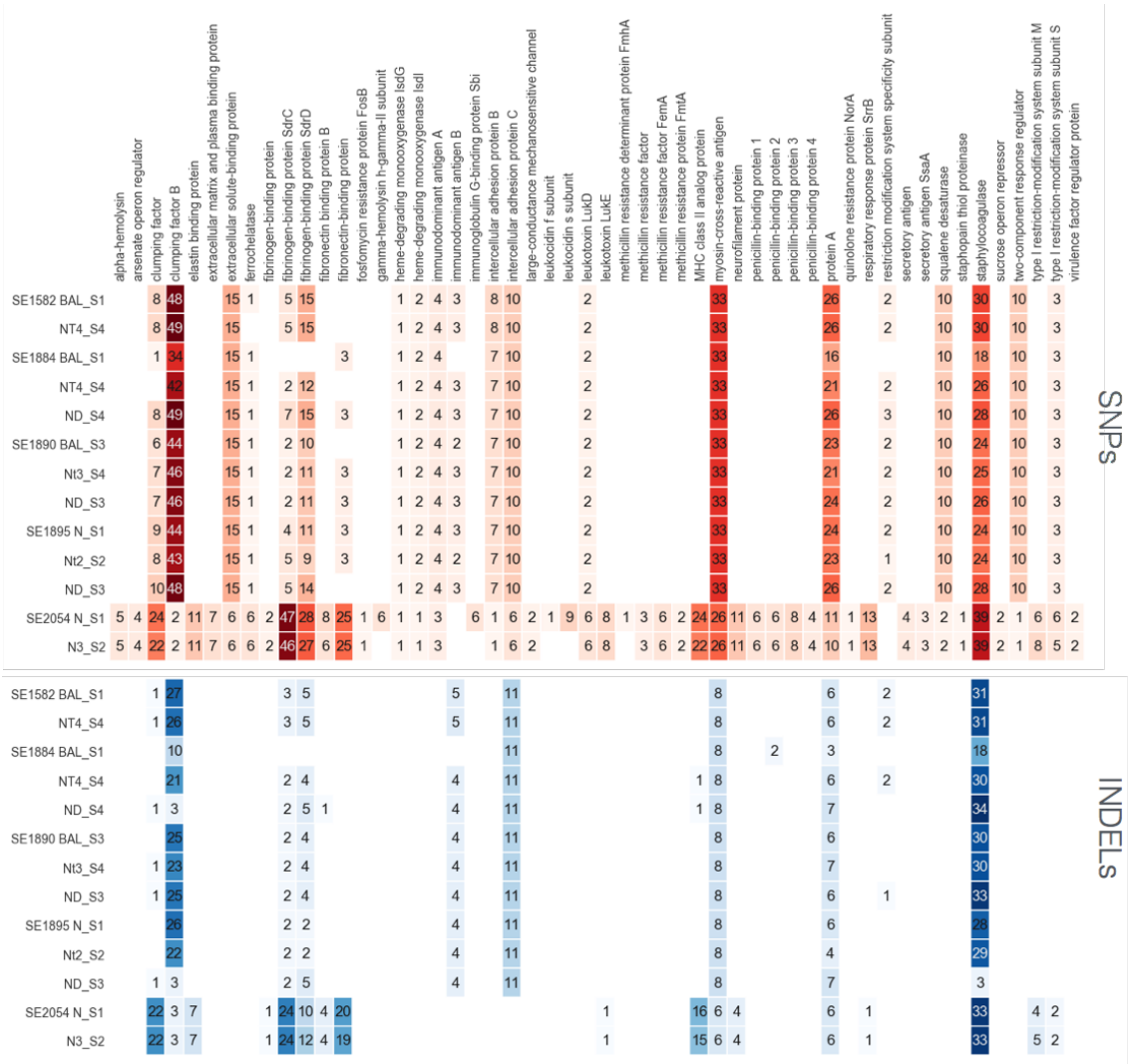


Figure 3.17: Numbers of SNPs (red) and INDELs (blue).

The SNP/INDEL patterns for the samples of all but the SE2054 patient were quite similar. This means the genetic differences found by the assembly and annotation could be located on MGEs, primarily since there were multiple lines of evidence presented before for the presence of MGEs in the MRSA genomes and on the presumed plasmids.

The majority of the proteins with SNPs and INDELs were virulence factors, namely the proteins responsible for direct contact with the cells of the human body. The most significant number of SNPs and INDELs was found in the clumping factor B, myosin-cross-reactive antigen, intercellular adhesin proteins, staphylocoagulasses, leukotocixns and fibrinogen-binding proteins.

3.6 Summary and conclusions

All samples were successfully assembled and annotated with multiple tools. The results confirmed and complemented each other. The most variability was observed in the sequences of proteins located on the surface of the cell, participating in contact with the host. This makes sense since the time-distance between the samples is not that large. All of the variability, from genome rearrangements to genetic alterations, plasmids, and bacteriophages are a function of time, the pressure of the host immune system and the antibiotics therapy. Unfortunately, the small number of patients and the complexity of the therapies made it difficult to correlate the variability to any particular feature of metadata.

The hypotheses that antibiotics therapy drives the emergence of resistance on the cellular level could not be tested in this project. There were too few patients, with too complex and unique antibiotic therapies. Nevertheless, in this study, the MRSA isolates of the patient with the most elaborate therapy had the highest variability.

The MRSA isolates were rich in virulence factors and antibiotic-resistant genes. Moreover, they were riddled with MGE-related regions and proteins, consequently, they had a high potential for driving AR emergence. The results can surely be improved with deeper sequencing or using a technology providing longer reads, or optical maps to enable better reordering rather than using an arbitrary reference genome. Both parts of this project, sequencing and bioinformatics analysis, were performed before the long-read sequencing was accessible.

Chapter 4

Gut mobileome under antibiotics

4.1 Introduction

The gut microbiomes of two healthy individuals were studied throughout six-days long ciprofloxacin therapy and subsequent 28 days of recovery. Authors analyzed the abundance of antibiotics resistance genes (ARGs) and confirmed that antibiotic pressure causes AR emergence in gut bacterial communities [254].

In their discussion, the authors point at horizontal gene transfer (HGT) as an essential factor in AR emergence [255]. This chapter describes a follow-up study, aiming at characterizing HGT within the gut microbiome under pressure of ciprofloxacin, focusing on the plasmids, transposons, and especially bacteriophages.

Usually, microbiome sequencing samples contain bacterial, phage and plasmid DNA. The constant genetic exchange between bacterial chromosomes and mobile genetic elements (MGEs) makes them difficult to quantify in the microbiome sequencing data. To analyze the phage fraction of the gut microbiome, the phage-only sequencing was carried out alongside the standard whole-genome microbiome sequencing [172, 173].

Dr. Silke Peter and Prof. Matthias Willmann planned and carried out the collection of samples and sequencing. My role was to develop an analysis scheme to describe mobile genetic elements and their dynamics within the gut microbiome, focusing on their role in AR emergence. The long-term goal is to develop methods that do not require specialized phage-only sequencing. This setup provided a unique opportunity to compare both datasets across multiple timesteps.

This chapter is divided into three parts: Methods, Results, and Conclusion. The Methods section outlines the technicalities of the pipeline: the programs, parameters, databases and statistical methods. In the following section, the results are presented and discussed. Some results relied on multiple methods, and therefore the order they are presented differs from the one in the Methods section.

4.2 Methods

4.2.1 Sequencing

Two healthy volunteers were administered 500 mg Ciprofloxacin twice daily orally. The stool samples were collected at six different timepoints: day 0 (before treatment), days 1, 3 and 6 (during antibiotic treatment) and days +2 and +28 (after treatment). Samples were processed, stored, the DNA extraction for the stool metagenome was performed as described before [254], and the sequencing was performed at GATC Biotech AGusing a paired-end sequencing with a read length of 2x300 bp on an Illumina MiSeq with an insert size of 550 bp. The enrichment and extraction of virus-like particles (VLPs) was performed as described previously [172, 173]. The phage sequencing was performed on a NextSeq 500 system (mid-output kit, 2x150).

4.2.2 Bioinformatics pipeline

The sequences of mobile genetic elements (MGEs) are changeable, versatile, and poorly represented in the current databases. Therefore, the classical read-based analysis was not well applicable. However, the alternative approach based on reads assembly and annotation is burdened with assembly and database biases. Therefore, both types of analysis were employed (Fig. 4.1).

The read-based approach provides a global picture of the shifts in the data, and the assembly-based a detailed analysis of the bacterial chromosomes, MGEs, and their genes. To minimize potential bias, the parameters were chosen conservatively. So, the assembly-based analysis was burdened by false negative, rather than a false positive error.

First, the reads of the Phageome set were preprocessed with CutAdapt [256]. The reads of the Microbiome set were merged, trimmed and filtered. Next, the analysis pipeline was applied to all samples of both sets: the sequencing of the isolated VLPs (Phageome set) and metagenomic sequencing of the WGS set (Microbiome set).

As both sets were human gut samples, the sequencing most probably contained human sequences. However, the Phageome set was also potentially contaminated with bacterial sequences. The contamination identification pipeline employed sections of the both read- and assembly- based analysis. It is described in detail in Section 4.2.5. The reads with an alignment to the identified contaminants were filtered out. Finally, three read sets were created: all reads *Raw*, filtered reads *Cleaned* and reads mapped to the assembled and filtered scaffolds *Assembly*.

4.2.3 Read-based analysis

The read-based analysis consisted of two parts: the classical metagenomics approach based on the read alignment to a database, and a database-free approach based on

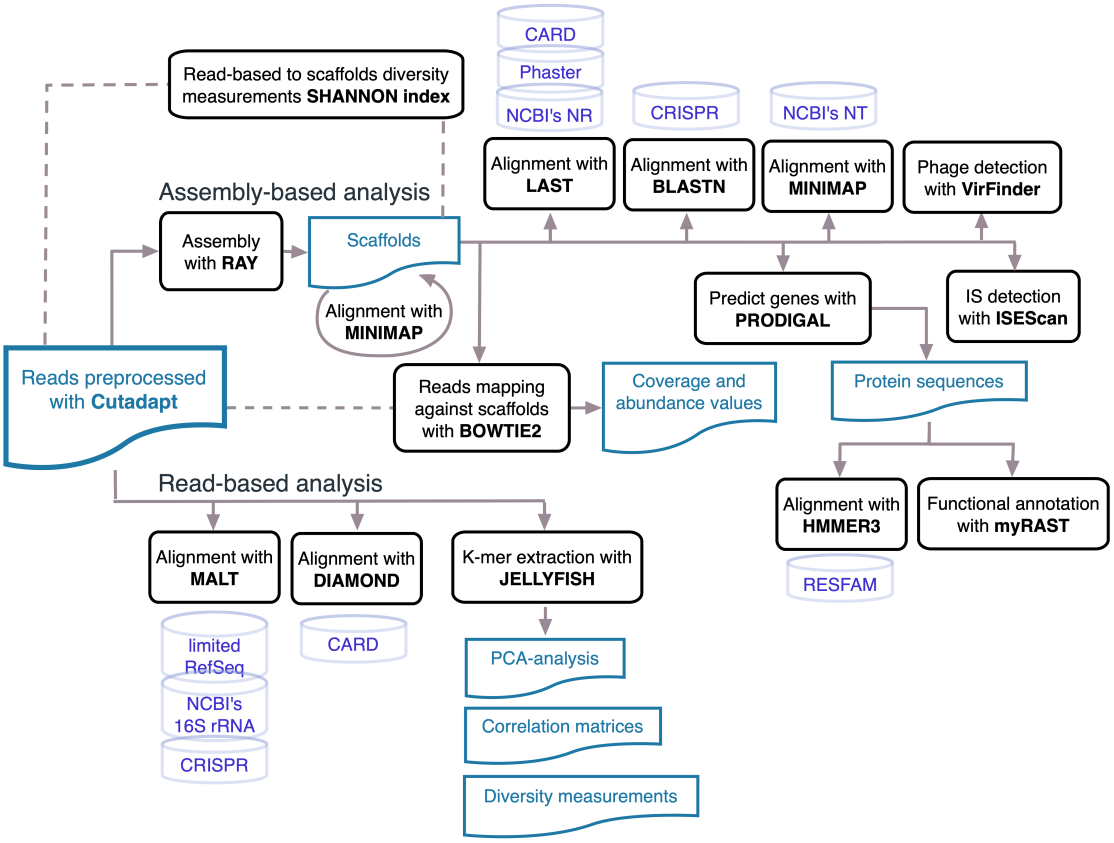


Figure 4.1: Data analysis pipeline. Reads preprocessing was done with Cutadapt [256]. Each sample was analyzed in two ways: assembly-based and read-based.

a k-mer analysis. The first path of the pipeline provided an overview of the data and a direct comparison to the first study.

Firstly, the reads were aligned with MALT [257] against the following databases: 16S rRNA NCBI, CRISPR-spacer (blastn and semi-global mode), CARD database [258] and selected bacterial genomes from RefSeq database (malt parameters: mq=1, blastx, semi-global mode). The alignments against 16S rRNA NCBI database and the selected bacterial genomes from RefSeq were used for estimation of contamination levels.

The alignment against the CRISPR database was used to estimate the phage fraction within the Microbiome dataset. The spacer abundance values were normalized by the number of the reads in the sample. Next, for each of the samples, a CRISPR profile was computed and used for PCA computations. The details of the CRISPR-spacer database construction are described in Section 4.2.4.

The alignment of the reads to the CARD database provided rough estimates of the ARGs abundances. The hits were filtered with two cutoffs: 90% identity, and

the coverage spanning at least 80% of the protein's length. Next, the number of the aligned reads was averaged over the protein's length and the number of reads in the sample.

Secondly, the k-mers for both DNA-strands of the reads were extracted with Jellyfish [259]. Each sample was represented by a vector of the relative k-mers counts. K-mers of the following sizes were used for analysis: 15, 25, and 55. The appropriate size of the k-mer needed to be used so that the most abundant k-mers encapsulate all of the diversity between the samples. Low-complexity and low-count k-mers were ignored. The numerical k-mer profiles were used to compute PCs. Next, the binarized profiles were used to compute pairwise Pearson product-moment correlation coefficients (with `numpy.corrcoeff` function).

4.2.4 Assembly-based analysis

On the one hand, phage genomes are relatively small, so they should assemble well, on the other hand, the phage community in the gut is expected to be quite diverse, so the depth of sequencing might turn out to be insufficient. Nevertheless, two assembly strategies were employed. First, the k-mer sizes were tested on the sets of all reads for each of the variants, i.e., the set and participant (*pooled assembly*). Assembly was performed with RayMeta (v. 2.3.1) [260] for the k-mer size with all of the odd numbers from 19 to 39, and 55.

Secondly, the best parameters of the *pooled assembly* were used to assemble each of the samples separately (*separated assembly*). Assembly runs of the Phageome set were performed with pair-ended information, but for the Microbiome set, the assembly was performed without it.

Evaluating assembly quality, especially in the case of metagenomics, poses a problem. Most common strategies rely on maximizing N50, so they favor assemblies with the most significant number of the longest scaffolds. However, since phage genome size spans from thousands to small hundreds of thousands of bases, the scaffolds size criterion is not applicable. In the case of this assembly, the quality analysis relied on the number of predicted genes and their distribution on the scaffolds. Basic assembly quality statistics were controlled with Quast [261], genes were predicted with Prodigal [262].

Every scaffold was annotated with a range of features informing on various aspects of the microbiome and phageome (Fig. 4.2). Below the methods are described.

Alignment to CARD and PHASTER databases The LAST program [248] was used to align scaffolds against the CARD database and phage protein database PHASTER [246]. Then the hits were filtered by the e-value ($\leqslant 0.001$) and separated into forward and reverse strands. For each of the strands, the hits were grouped into pileups: if the alignment coordinates were overlapping, I assumed the hits aligned to

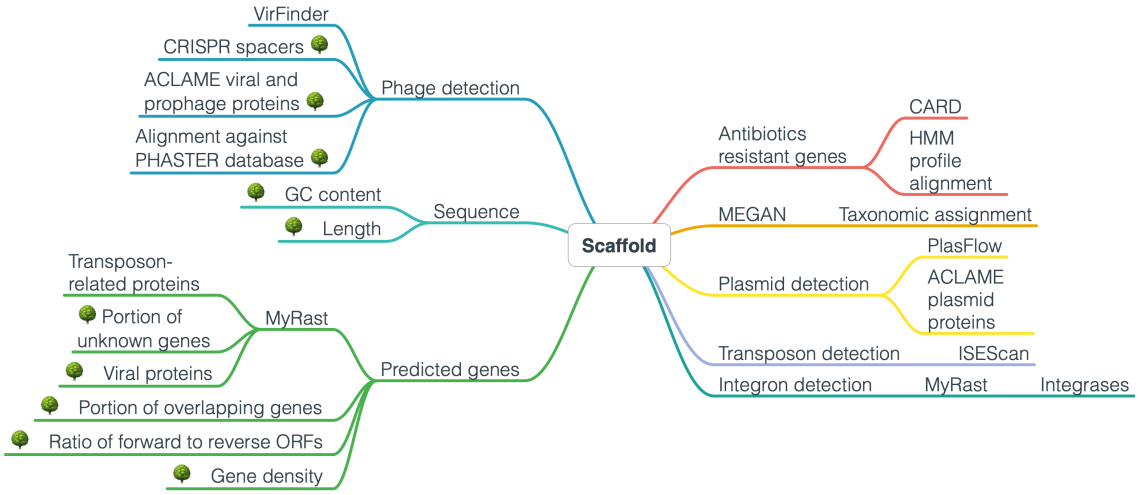


Figure 4.2: Features collected for scaffolds organized in a mindmap. The tree icon denotes the features used later in Random Forest runs.

a single ORF. For each pileup, the best alignment was chosen. Those were filtered by the coverage, span, bit-score, percentage identity and finally the lowest e-value. In the rare case that in the end there were multiple hits, their descriptions were merged.

Taxonomic profiling The scaffolds were aligned against the NCBI NR with LAST. The resulting files were meganized and analyzed using MEGAN [148]. MEGAN-LR [263] placed scaffolds on to the NCBI taxonomic tree. The resulting profile was not abundance-aware, as it used the assembled scaffolds. Nevertheless, it could be used as a general sanity check and for comparison with the previous study.

Gene prediction and annotation Genes were predicted with Prodigal [262]. They were first used to compute gene density, a portion of overlapping ORFs and portion of the ORFs with reverse orientation. Next, the protein sequences were extracted and input into myRAST [239] for the functional annotation and aligned against the ACLAME proteins. The alignments were filtered based on the e-value ($\leqslant 0.01$) and percentage identity ($\geqslant 60\%$). The scaffolds were annotated with a vector of the detected gene families. Each scaffold could have multiple proteins identified as one of the 32,000 families.

Gene density values were computed for all of the available phage genomes in the NCBI database and an equal number of the randomly selected bacterial genomes. Their distributions were compared to the density computed for the predicted genes for the scaffolds in the Microbiome and Phageome sets.

The ACLAME database contains 122,154 proteins of three MGE types: plasmids, viruses, and prophages, classifies into ~32,000 gene families. Gene families can be

attributed to single or multiple classes of mobile genetic elements. 29,816 proteins had a single class (plasmid: 16,632, vir: 9,124, proph: 4,060), 2,604 had two classes (proph and vir: 1,745, proph and plasmid: 592, plasmid and vir: 267) and 499 proteins were assigned to all three classes.

Cirpofloxacin resistance genes The ARG annotation was followed by the sequence-wise analysis of the ciprofloxacin-resistant genes. The sequence of the proteins conferring resistance to ciprofloxacin was analyzed in detail. The resistant gyrases were found on the assembled scaffolds through the multistep pipeline: gene prediction, extraction of the protein sequence, and their subsequent alignment to the CARD database. Next, the relevant mutations were confirmed at the nucleotide level through the analysis of the pileups of reads. Finally, the protein alignment of all of the detected proteins to the appropriate reference, along with the pileup codon counts were plotted.

CRISPR spacer database and annotation The CRISPR spacers are widely used as markers for phage sequences [264, 265]. This approach requires an extensive database of CRISPR spacer sequences. Two public CRISPR spacer databases were used: CRISPI [266] and CRISPR [267], the spacers were obtained with CRASS [268] from the Microbiome reads, and a number of public sets: samples downloaded from the SRA NCBI database (SRR [269], TS29 [270]) and fecal samples from the Human Microbiome Project (HMP [271]).

The CRISPR spacers were aligned against the assembled scaffolds with BLASTn [272]. Hits were only considered if the percent identity was larger than 90%, and the alignment covered over 90% of the spacer. On the other hand, scaffolds with a CRISPR cassette are probably bacterial. Therefore, the CRASS runs were performed on the scaffolds to exclude those that contain CRISPR cassettes. Filtering for phages based on the CRISPR spacer alignment has been used in various studies [265].

Phage scaffolds identification The phageome contained only phage sequences, whereas the Microbiome comprised both bacterial and phage sequences. Therefore a method identifying a scaffold as phage within a metagenomic dataset was needed. The tools to identify phage sequences have been developed mostly in the context of bacterial genome annotation, i.e., detecting prophages. However, there are two tools primary suited for detecting phage scaffolds within metagenomics assembly: VirSorter [273] and VirFinder [175]. VirSorter uses a wide range of metrics to rule whether a given sequence is viral or not. VirFinder uses k-mers and machine learning to identify phage scaffolds. In case of VirSorter the results are binary, but with VirFinder an arbitrary cutoff has to be defined (p-value $\leqslant 0.05$). Both methods were applied, but VirSorter returned no results. Hence only VirFinder was further used. VirFinder is trained on known phages deposited in the database. This means its machine-learning classifier does not encapsulate the entire space of phage genetic

diversity. Consequently, VirFinder's prediction for such a rich dataset like the human gut sequencing is burdened with a high rate of false negatives. To enrich the phage scaffold selection a Random Forest (RF) [216] was used.

The RF classifier was trained based on the positive class consisting of the scaffolds the VirFinder gave a low p-value ($\leqslant 0.05$). Each scaffold was encoded as a vector of numerical values representing the various phage, sequence and gene features presented in Fig. 4.2. A single cycle consisted of training and prediction steps. The classifier was trained on the undersampled dataset containing 90% of the positive data points and an equal number of randomly chosen negative data points. Hence, each run leaves a large number of unused scaffolds. For those, the new classifier was used to make a prediction. The cycle was repeated 500 times, so that in the end for all of the negative scaffolds the prediction will be done multiple times.

If at least 80% of the cycles resulted in a positive prediction by the RF classifier, the scaffold was denoted as a phage. The Out-Of-Bag accuracy (OOB) was used to evaluate the performance of the classifier, and mean decrease in accuracy to investigate feature importance. Further parameters of the RF classifiers were selected using the provided mechanism of parameter selection (grid_search).

The RF parameters had to be first selected so that the overfitting was minimized. The first test for overfitting measured the differences in accuracy between the train and test subsets across 500 runs under the cross-validation regime (40 to 60 %). Second test measures proportions of scaffolds denoted as phage by the entire RF set across a range of the cutoff values. The rationale is that better RF classifiers produce a flatter trajectory less dependent on the cutoff value.

Phage integration analysis Microbiome phage scaffolds were aligned against all scaffolds from the subsequent timestep within the Microbiome dataset. If the phage scaffold aligned entirely to a much longer non-phage scaffold, we assumed that was a possible integration occurrence. A new integration occurrence happened when there was no integration at the same time-point.

Phage taxonomic assignment For the scaffolds that have a valid alignment to one or more proteins from the PHASTER database, taxonomic identification was possible. The proteins in the PHASTER database had an NCBI phage taxonomic name. Hence, a single scaffold could have multiple PHASTER alignments with multiple taxa. The bit score sum was computed for each taxon from the proteins assigned to a scaffold. Next, a scaffold was assigned a phage taxon with the highest bit score sum. If there were more than one taxa with the highest bit-score sum, a viral name was not assigned.

MGE detection: Plasmids, Transposons and Integrons Other MGEs, such as plasmids, transposons, and integrons need to be identified within the metagenomic assembly. Plasmids were identified in two steps. Firstly, the PlasFlow program was

used with a 0.95 cutoff [176]. Secondly, scaffolds were filtered with at least one protein from a plasmid family in the ACLAME classification. Next, the insertion sequences were annotated on the scaffolds with ISEscan [245]. The scaffolds that were not classified as phages and had at least two IS found were classified as containing transposons. Analogously, scaffolds with integrons were defined as those that are not phage and have an integrase.

ACLAME protein family assignments enabled a parallel MGE classification. This was defined as the largest common subset of protein assignments for proteins on a single scaffold. However, provided all proteins had all annotations to all three families, this scaffold would appear as both phage and plasmid. Therefore, this classification is not independent.

4.2.5 Contamination assessment

Removing human sequences was relatively straightforward as they are not similar to those of the microbiome. However, separating phage from bacterial sequences posed a more significant challenge, as the phage genomes are often incorporated into bacteria. The bacterial sequence needed to be covered in a large portion and continuously by the phage reads so that it was evident that the bacterium is a contaminant. Fig. 4.3 presents the pipeline for contamination estimation.

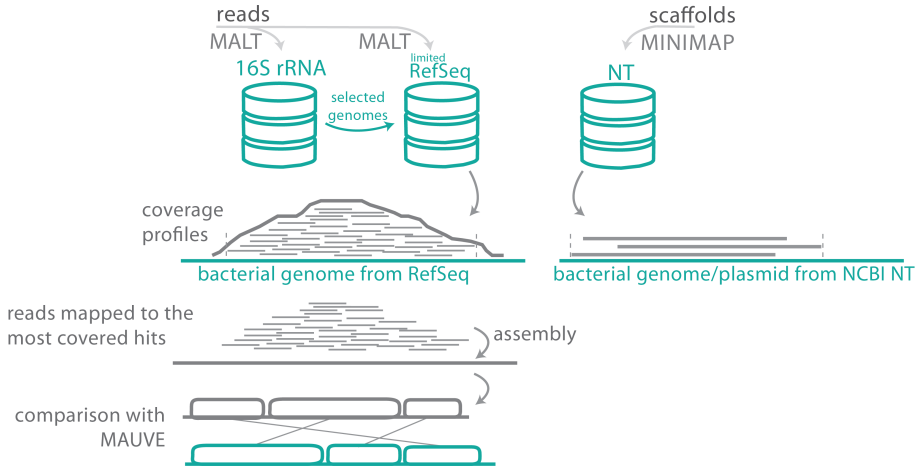


Figure 4.3: Contamination estimation pipeline.

Level of contamination was first assessed by the relative counts of the 16S rRNA reads in the Phageome and Microbiome sets. Potentially, the 16S rRNA gene or its fragments could be picked up by phages and consequently be present in the Phageome set. However, their proportion should be significantly lower than in the Microbiome. The proportions of the 16S rRNA reads were compared between the corresponding samples of the Phage and Microbiome sets.

The alignment of the reads to the 16S rRNA NCBI's database also provided the taxonomic profile for the Microbiome set. The full bacterial genomes of the taxa found in 16S rRNA alignment were selected from the RefSeq database to construct a limited RefSeq. The preprocessed Phageome reads were mapped against the sub-selected RefSeq. Next, the coverage profiles were computed, so that each record was represented by a vector of its length holding the number of reads covering each position. If the reads cover the entire length of the genome, it proves that the bacterial genomic DNA was present in the sample. Subsequently, the reads mapped to the most extensively covered genomes were extracted and assembled with Ray, separately for each sample. The resulting scaffolds were ordered with scaffold-builder [236] and compared to each other and the reference with MAUVE [238].

Assembled scaffolds were aligned against the NCBI nucleotide database (NCBI nt) with Minimap [274] (Fig. 4.3). The alignments were filtered by the coverage of the scaffold ($\geqslant 90\%$) and identity ($\geqslant 90\%$). If scaffolds collectively covered a significant portion of the record in the NCBI nt database given NCBI nt hit was denoted as contamination. The coverage threshold differed depending on the type of the record: for the bacterial genome ($\geqslant 20\%$), and for the plasmid sequence ($\geqslant 70\%$). Later, the scaffolds with a strong alignment to those were filtered out.

Finally, the preprocessed reads of both sets were mapped with Bowtie2 [230] to the database of potential contaminations: the human genome in case of Microbiome and human genome with the *Bacteroides* species *caccae* and *cellulosilyticus* identified in the previous step in the case of Phageome.

Because the contamination assessment used the scaffolds, I did not rerun the assembly for the Cleaned reads. Instead, the scaffolds were filtered. The filtering removed scaffolds with a high-quality alignment (coverage $\geqslant 80\%$ and identity $\geqslant 80\%$) to human contaminants in the case of the Microbiome set, and the identified bacterial genomes or plasmid sequences in the case of the Phageome set. The list of the contamination taxa was defined separately per sample. The subsequent steps of filtering included removing scaffolds with CRISPR cassettes and those shorter than 500 bp.

4.2.6 GC-content analysis

GC content was analyzed with the Kernel Density Estimation (KDE) function. KDE fits the density function to the given histogram, smooths the data, and presents only the relative values. The GC-content for the sequencing reads had two maxima in the KDE plots. The bacteria responsible for the GC-content shape were investigated. First, peaks were detected and the ratio of their heights was computed. Using the taxonomic assignments the sets of scaffolds were created so that they contain all but the selected taxa. If the peak ratios changed, the taxon in question was assumed to be responsible for it.

4.2.7 Abundance trajectory analysis

At the core of the analysis is the concept of a *abundance trajectory*, i.e., level of abundance for each of the timepoints. In the case of metagenomic sequencing the *abundance* cannot be measured directly - therefore, we used an approximation of the abundance. First, with Bowite2 [230] the cleaned reads were mapped to the scaffolds. Next, with samtools [275] the number of reads per position was computed. The average coverage (*cov*) per scaffold is computed according to the formula:

$$cov = \frac{\sum_{i=1}^{L} cov_i}{NL} \times 10^6 \tag{4.1}$$

here, L is the scaffold length, cov_i is coverage of the $i-th$ position and N is the total number of reads in the sample, the value was scaled by 10^6. The average coverage of the scaffolds was an approximation for abundance.

Feature abundance is the sum of the average coverage for all scaffolds with a defined feature and its value, e.g. GC $\geqslant$ 50%. Any annotated feature or its combination can be used (Fig. 4.2), e.g., antibiotic resistance for scaffolds longer than 1000 bp. *Feature abundance* for each of the consecutive timepoints constitutes the *feature abundance trajectory*.

The *feature abundance trajectories* were computed for all of the lowest-level features of the functional annotations, the ARGs from the CARD database, HMM profiles for the ARGs and functional gene assignments from myRAST. Separately, the trajectories were computed for the phage taxonomic assignments. The *feature abundance trajectory* for a single gene or a profile could also be computed for any on the scaffold selection. The scaffolds were selected so that they reflect the MGEs. In the Phageome, there was just one scaffold group: all scaffolds, encapsulating all free phages. In the Microbiome, there were several scaffold groups: phage, bacterial (non-phage), plasmid, transposon, and integrons. All possible *feature abundance trajectories* were computed for all features, scaffold selections, participants, and datasets.

Next, the trajectories were scaled (to range 0.0 to 1.0) and clustered with Agglomerative Clustering (four clusters, complete linkage, cosine affinity). For each cluster, the *average scaled trajectory* was computed. All observed *average scaled trajectories* were sorted, numbered and color-coded. Finally, for each participant, scaffold group and classification (e.g., all functional annotations), the profile of the trajectories were plotted in horizontal stack-charts.

The features were organized hierarchically. Therefore, the higher level trajectories comprised, the lower level trajectories. The first level constituted global analysis with no feature selection incorporating all scaffolds and resulting in a single *global abundance trajectory*. Next, the phylum-level analysis included a vast majority of the scaffolds described a division between Gram-positive and -negative bacteria. The third level: general MGE and antibiotics level, included selections of MGEs, carriage

of ARGs, the ACLAME classification and HMMs. The last gene-level analysis described the particular genes within the functional annotations, taxa within the taxonomic assignments, and profiles within the HMM profiles.

Abundance trajectories enabled analysis of the dynamics between the phages and the bacteria. However, the phage scaffolds needed to be paired with their hosts. WiSH [276] assigns a phage to its best host based on the shared k-mers. The sub-selected RefSeq served as a database of the potential host genomes. The scaled abundance trajectories of phages and their hosts along with their ratio were plotted. The ratio takes values from the range -1 to 1. The larger the positive value, the more the host abundance exceeds the phage abundance. Growing values denote the phage abundance decreasing in respect to its host's abundance.

Additionally the *diversity trajectories* were analyzed. The diversity was computed based on the *feature abundance*, with Shannon's alpha-diversity (H), according to the equation:

$$H = -\sum_{i=1}^{R} p_i \ln p_i, \tag{4.2}$$

where R is the number of features, and p_i the proportion of k-mers or of the feature abundance [277, 278, 279].

4.3 Results

4.3.1 Data and contamination

The Microbiome samples were comparable to the Phageome samples regarding the number of reads, but there was more sequence as the Microbiome reads were longer (Fig. 4.8). Overall, the samples had between 10 and 40 million reads each. There were substantial differences in the amount of sequence between them, even among those of the single participant and dataset. The differences were smaller in the Microbiome dataset than in the Phageome dataset. Those differences could affect the downstream assembly and the coverage analysis.

All steps from the contamination pipeline suggested the Phageome dataset was contaminated with bacteria. Firstly in participant A the 6^{th} and $+2^{nd}$ samples and participant B the 3^{rd} and $+28^{nd}$ samples had a higher proportion of the 16S rRNA reads that in the corresponding sample of the Metagenomic dataset (Table 4.1).

Secondly, the alignment of the Phageome scaffolds to the NCBI NT database revealed bacterial genomes and plasmids in the same samples as identified by the 16S rRNA analysis (Fig. 4.4). The reads mapping to the subselected RefSeq identified the *Bacteroides dorei* and *Bacteroides cellulosilyticus* in participants A and B, respectively.

Table 4.1: Numbers and proportions of reads aligned to the 16S rRNA database for both sets.

	Number of reads in A (‰)			Number of reads in B (‰)		
Day	Phageome		Microbiome	Phageome		Microbiome
0	8,547 (0.7)	$<$	25,494 (2.0)	6,606 (0.7)	$<$	21,802 (2.2)
1	7,811 (0.5)	$<$	12,074 (0.8)	7,804 (0.6)	$<$	12,479 (0.9)
3	**44,955 (1.1)**	$>$	25,765 (0.6)	**23,737 (1.7)**	$>$	13,346 (0.9)
6	7,122 (0.3)	$<$	24,224 (1.1)	15,647 (0.9)	$<$	17,156 (1.0)
+2	**32,258 (1.0)**	$>$	22,893 (0.7)	8,383 (0.3)	$=$	6,100 (0.3)
+28	5,965 (0.5)	$<$	15,456 (1.3)	**44,499 (1.5)**	$>$	20,596 (0.7)

Alignment of the reads to the bacterial genomes of the sub-selected RefSeq covered substantial portions of the bacterial genomes. However, the alignments were scattered and incomplete, in which case I am convinced the reads came from phages. Nevertheless, several bacterial genomes were covered throughout their entire length. In participant A the genome of the *Bacteroides caccae* was well covered in the 3^{rd}, 6^{th}, and $+2^{nd}$ samples, and in participant B *Bacteroides vulgatus*, *Bacteroides cellulosilyticus*, and *Bacteroides ovatus* were well covered by the aligned reads, in the 3^{rd}, 6^{th}, $+2^{nd}$ and $+28^{th}$ samples.

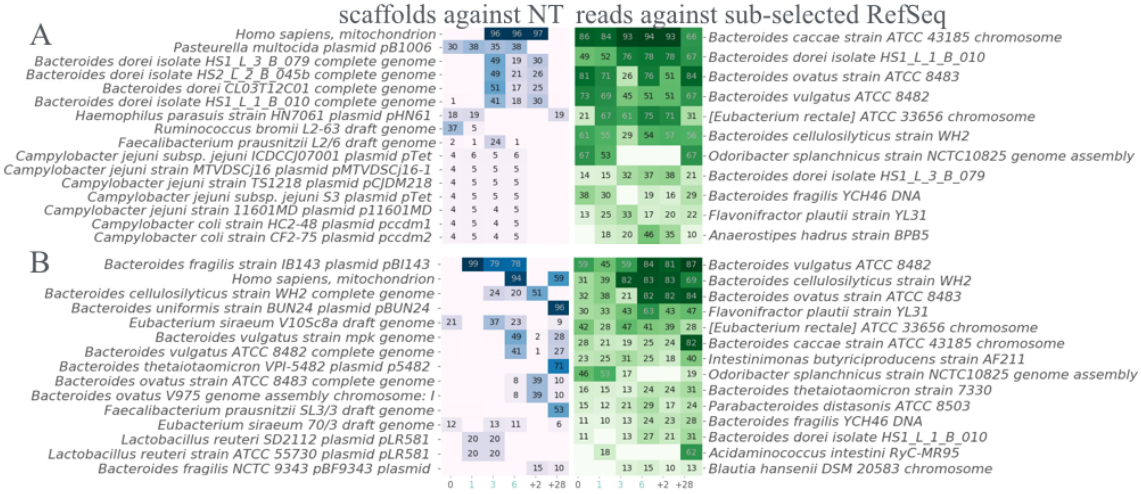


Figure 4.4: Contaminations in Phageome. Coverage of the NCBI-NT alignments with the scaffolds (a, c) and subselected RefSeq alignments with the reads (b, d).

Thirdly, *de novo* assembly of the bacterial genomes from the aligned reads worked only for the samples of two days per participant, i.e., 1^{st}, 6^{th} days in A and 3^{rd}, 6^{th} days in B. Mauve alignments of the assembled genomes to their corresponding reference (not shown) proved that those were high-quality assemblies although, in both participants, the assembly did not cover the entire reference. Especially as that was metagenomics sequencing, such high-quality assembly proved the bacterial contamination of the Phageome samples.

Other groups have also reported bacterial reads in the phageome sequencing projects [183]. This raised the question whether they are not a sign of interesting biological phenomena, such as sporulation. Spores are small enough to pass through the phage filter and could have undergone DNA-extraction alongside VLPs. However, sporulation is a complicated process controlled by an extensive network of genes. Therefore, I checked if any such genes were present in the annotation of the assembled scaffolds on the contaminations. The only gene corresponding to the sporulation found was the *Spore maturation protein A-like protein*. However, there were no alignments to the HMM of the Spo0A profile, which is a sporulation driver [280].

Finally, the reads were filtered based on the sample-specific list of contaminations. The set of reads without mapping to the identified contaminants make up the cleaned set. Scaffold filtering consisted of the three steps: length, *no CRISPR cassette*, and *no mapping to the contaminants*. Only a few scaffolds contained CRISPR cassettes in the pooled assembly of the Phageome set and even fewer in the separated assembly. No such scaffolds were found in the Microbiome set. The three-step scaffold filtering resulted in the greater convergence of the GC content distributions.

Analysis of the GC-content distributions is the first quality check for the sequencing data, as anomalies in the GC distributions can point to contamination. The microbiome GC-content distributions for the assembled scaffolds were characterized by a distinct secondary peak for both participants, but more distinct in participant B (Fig. 4.5). Interestingly the height of the peak changed throughout sampling. It is distinct on the first and last days of the therapy but disappeared on the 6^{th} and $+2^{nd}$ days. The secondary GC-peak decreased during therapy with ciprofloxacin, suggesting that it was composed of sequence coming from ciprofloxacin-susceptible bacteria, i.e., Gram-negatives.

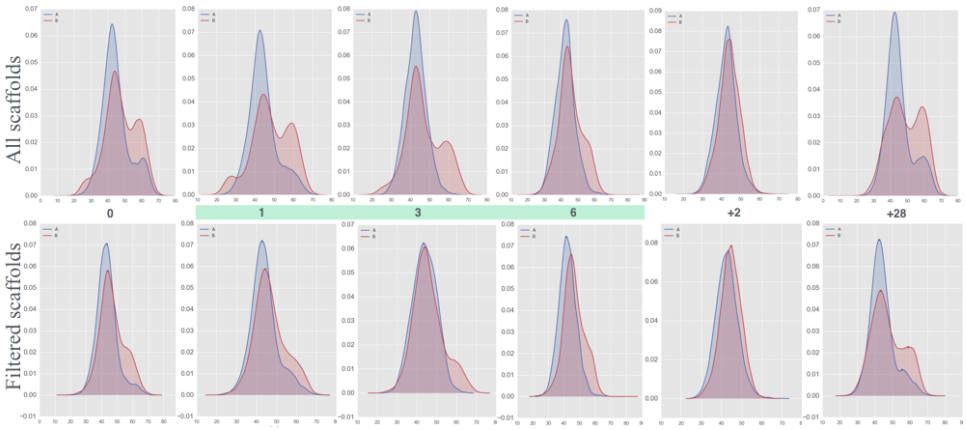


Figure 4.5: GC-content distribution for the Microbiome scaffolds. Filtered scaffolds denote scaffolds without taxonomic annotation to *Alistipes, Firmicutes, Subdoligranulumvariabile, Faecalibacterium, Clostridiales.*

Removing scaffolds with a taxonomic annotation to *Alistipes, Firmicutes, Subdoligranulumvariabile, Faecalibacterium, Clostridiales* bacteria resulted in decreasing of the secondary GC-peak. However, it was not completely removed, suggesting there were other taxa with high GC content. The *Alistipes* bacteria are Gram-negative, but *Firmicutes, Faecalibacterium, Clostridiales* are Gram-positive. Therefore it contradicted the hypothesis that Gram-negatives were responsible for the peak. Nevertheless, GC content is not a deterministic characteristic for bacterial taxa, and even small taxonomic groups of bacteria can include bacteria with a wide range of the GC content [281]. However, at this point, I assumed the secondary peak of the GC content was not a sign of the contamination.

4.3.2 Read-based analysis

Cleaned reads were aligned against the CARD database. Overall, ARGs were present in both datasets, and some their abundance changed in response to antibiotic therapy primarily in the Microbiome set. The results suggest that TetQ was the most abundant ARG (Fig. 4.6). The ARGs found in Phageome participant A were more diverse than those found in participant B, which, however, was not the case in the Microbiome dataset. The OXA-347 gene was found in participant A, in both datasets, which agreed with discoveries of the original study. There was a correspondence between the most abundant genes detected in both sets for the same participant. In participant A, among the first ten most abundant genes seven genes repeated in both datasets, whereas in B, six genes repeated.

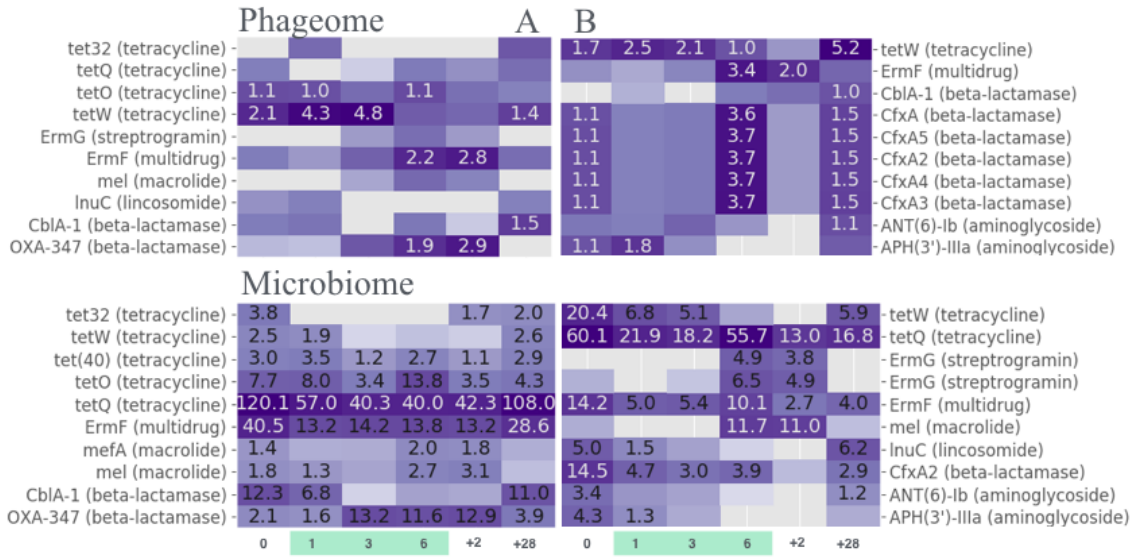


Figure 4.6: The read-based abundance of ARG for the most abundant genes.

The k-mer-based correlation showed the therapy-related change was not limited to the ARGs, as both sets changed under the pressure of antibiotics (Fig. 4.7). The correlation matrices were more comparable within a single patient than between the datasets. In participant A's Phageome and Microbiome, first two samples, and of the days from 3^{rd} and 6^{th}, were correlated with each other but anti-correlated with samples of the 0^{th} and 1^{st} days. In Phageome the next three days were anti- or weakly correlated to other samples. In Microbiome the samples of the 3^{rd} to $+2^{nd}$ days were correlated with each other. The last samples of Microbiome and Phageome correlated again with the first two days. Therefore, both Phageomes and Microbiomes in participant A restored their structure from the initial time-point after 28 days of recovery. In participant B, the first three days, and then the 6^{th} and $+2^{nd}$ days were weakly correlated with each other in both Phageome and Microbiome sets. The last sample was not correlated to any other samples. The microbiome of the participant B did not return to the initial time-point.

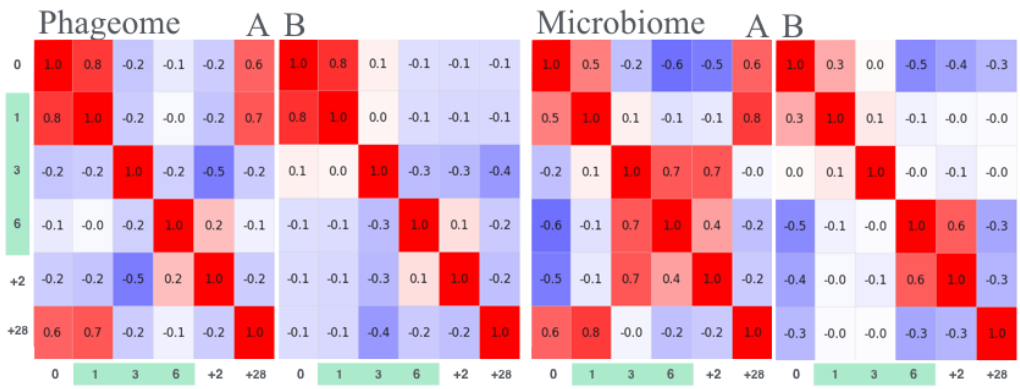


Figure 4.7: Correlation matrices for cleaned reads kmer size 25. The overall shape of the correlation matrices was independent of the underlying read selection and was robust towards the k-mer size.

In all variants, the communities reached semi-stable states under the antibiotic pressure. Changes in the Phageome went hand-in-hand with the changes in the Microbiome. After the therapy participant A's Microbiome and Phageome returned to the initial structure, which did not happen for participant B. The k-mer correlation pattern confirmed the findings of the original publication, which found similar patterns using PCA of taxonomic assignments.

4.3.3 Detection of mobile genetic elements

The *pooled assembly*, with k-mer size 25, resulted in the most scaffolds with both detected and predicted genes (Table E.4). Therefore, k-mer size 25 was used for the *separated assembly* without further testing. As expected, the separated assembly generated more sequence, in the form of shorter scaffolds, however, with

disproportionately more predicted genes. Therefore, the separated assembly was used for further analysis.

As expected, assembly caused information loss, as only between 40% to 82% of the reads were used to construct the assembly (Fig. 4.8). The proportion was lower in the Phageome than in the Microbiome, as was the number of scaffolds with meaningful annotation. This suggests that the Phageome had insufficient sequencing depth.

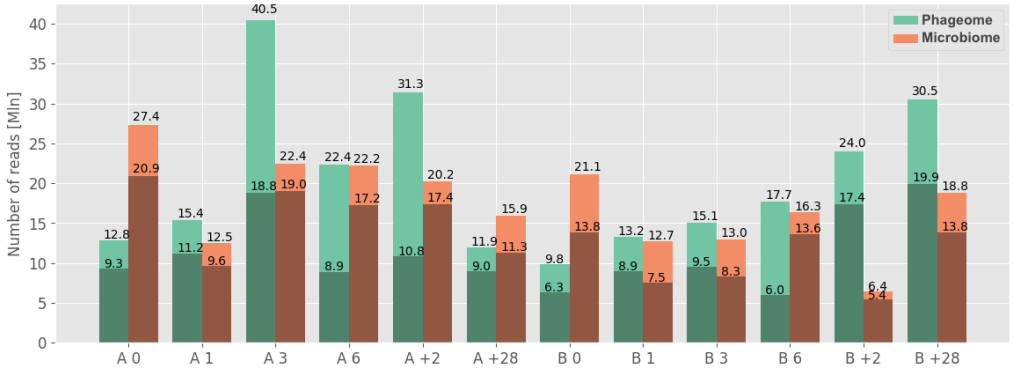


Figure 4.8: Number of reads aligned to the filtered scaffolds.

Plasmid identification

According to the ACLAME classification, obtained via protein alignment, in the Phageome dataset, viral and prophage scaffolds outnumbered those with the proteins assigned to plasmids (Fig. 4.9). In the Microbiome dataset, the proportions were reversed. There were much less viral or prophage and more plasmid scaffolds. An exception was the last sample of the Phageome participant B, which resembled more the microbiome than phageome profile, as it had a significant share of plasmid scaffolds. It is also one of the samples that were ruled as contaminated. In participant's B Phageome, the last time point appeared to be contaminated with plasmid sequences.

The results of Plasflow did not correlate with the ACLAME classification (Fig. 4.10). Across the three ACLAME MGE classes, the majority of scaffolds remained unclassified. In Phageome a relatively large portion of the scaffolds was classified as plasmids and bacterial chromosomes, especially in case of the ACLAME virus class. In Microbiome the PlasFlow annotations had similar proportions across all ACLAME classes.

The final plasmid classification was a two-step procedure: PlasFlow classification as plasmid and at least one plasmid ACLAME protein. The first step used k-mers and the second step a database-based protein classification. On its own, the PlasFlow method identified too many scaffolds, especially within the Phageome, and notably

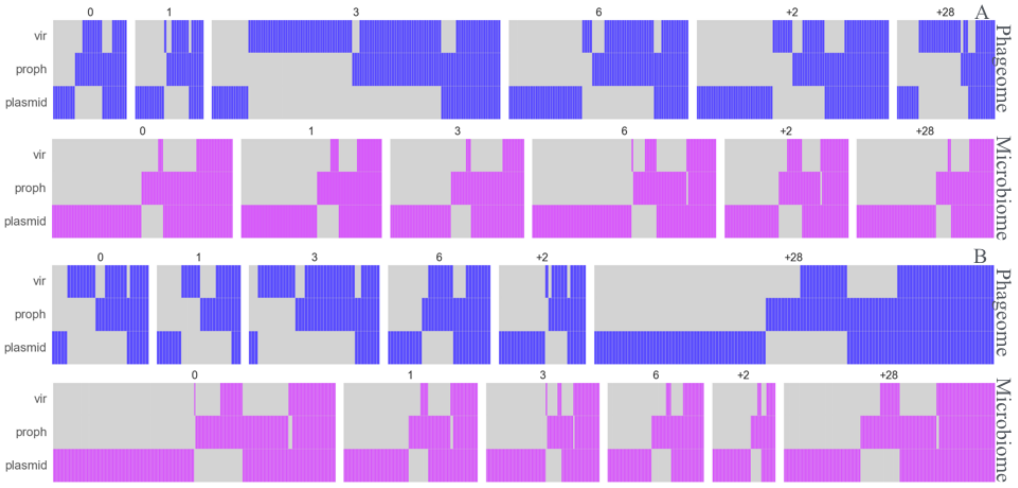


Figure 4.9: Distribution of the ACLAME protein classes for scaffolds with at least two ACLAME proteins. Columns represent scaffolds and rows correspond to ACLAME protein families: vir for phage, and prop for prophage.

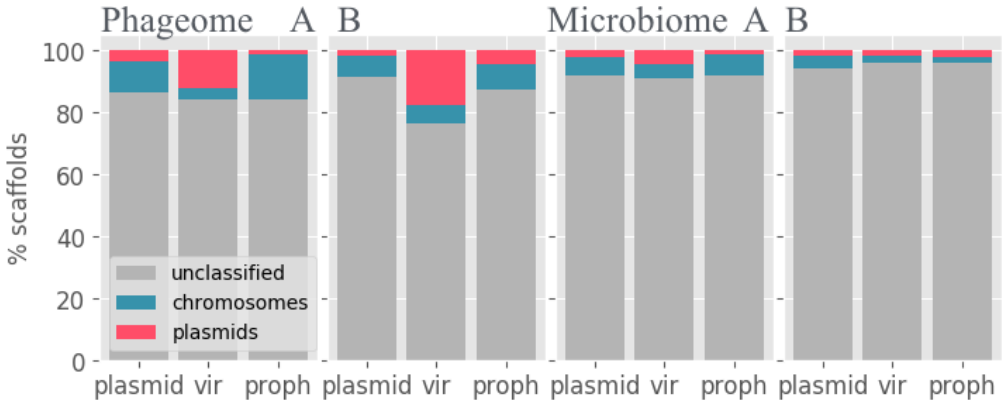


Figure 4.10: Results of PlasFlow in relation to the ACLAME classification. Each scaffold can be represented multiple times across all three selection.

the most among the scaffolds with ACLAME viral proteins. However, researchers have reported that phage genes are often miss-classified as plasmid [282]. In the end, $\sim 1.7\%$ of the Microbiome scaffolds were denoted as plasmids.

Transposon identification

The next MGE class is transposons. Since there were no dedicated tools for finding the transposons, the ISes were identified first. The most abundant IS class (Fig. 4.11), IS21 usually is associated with *Bacillus* and *Bacteroides* [283], which are normal human gut bacteria. However, the second most abundant class, IS66 is found in the soil bacteria *Agrobacterium* and *Rhizobium*, while the next most abundant IS3

is naturally present in both mentioned gut and soil bacteria. This suggests the IS assignment to families is burdened with an error, as the most abundant bacteria among the scaffolds with IS, were *Bacteroides* and *Firmicutes*. Nevertheless, there was a considerable number of scaffolds at least two identified and annotated ISes, and consequently transposons.

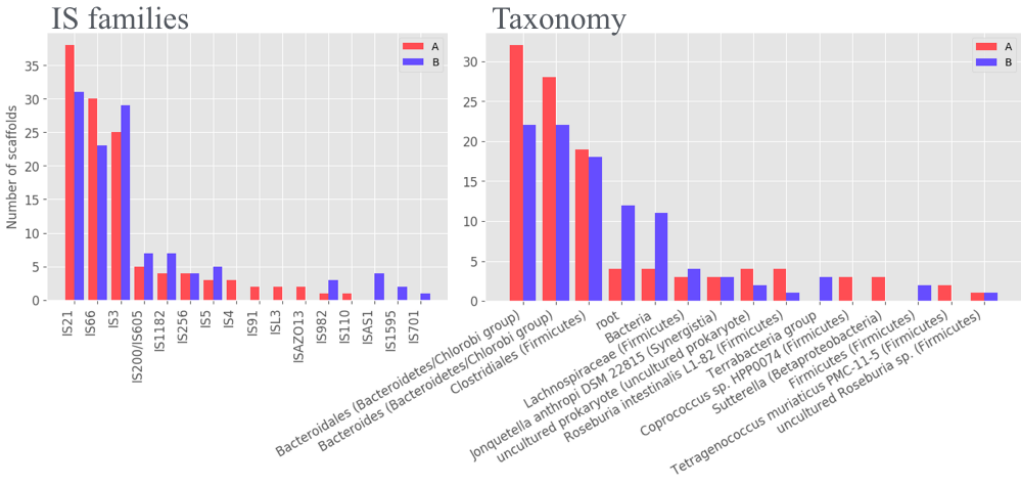


Figure 4.11: IS families and taxonomic profile for the transposon scaffolds.

Phage identification

Phages were detected with the five separate tools, which results were unified using the random forest. This section discusses first the results of the individual methods and the random forest in the end.

In the first place, the distribution of the gene densities were investigated. The gene densities of the phage genomes from the NCBI database were higher than for the bacterial genomes (Fig. 4.12). Independently of the assembly and dataset, the distribution of the density computed for predicted genes was nearly identical with a peak close to 100% - which was higher than the database derived gene density. Also, the gene density for the assembly was less concentrated suggesting there were errors in the gene prediction and miss-assemblies. Nevertheless, the overlap between the gene densities for the database records was too large for the gene density measurement to be used as the unequivocal identifying factor for the phage sequence.

The second phage identification method entailed using a CRISPR database. It firstly required collecting the database from various sources including existing public databases, and public sequencing datasets from which the spacers were extracted. The final CRISPR database contained 356,000 unique CRISPR spacers. Still, a relatively small portion of the scaffolds had an alignment to any of the known spacers (Table 4.2).

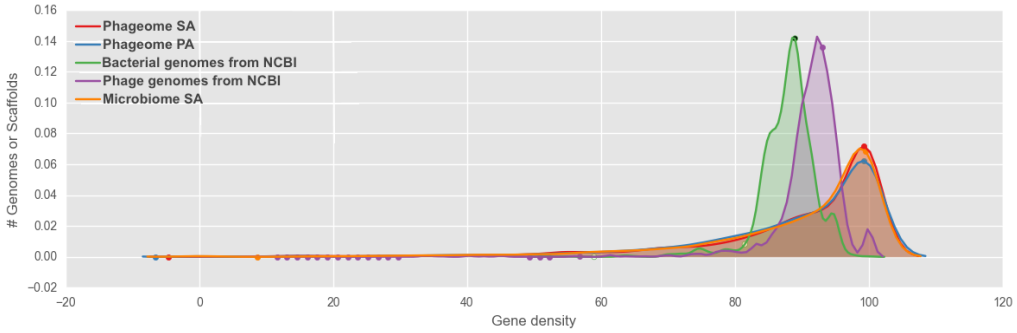


Figure 4.12: Distribution of the gene density for the bacterial and phage genomes from the database and the Phageome scaffolds in both strategies. SA stands for separated assembly and PA for pooled assembly.

Table 4.2: CRISPR spacers in all sets and number of scaffolds the spacers aligned to in pooled assembly (PA) and separated assembly (SA).

		Phageome (PA)		Phageome (SA)		Microbiome (SA)	
Spacer set	# spacers	A	B	A	B	A	B
HMP [271]	99,975	10,734	10,681	2,163	2,314	9,256	13,640
Microbiome [254]	6,883	1,008	1,188	222	229	558	1,078
SRR [269]	3,638	682	807	251	228	835	1,156
TS29 [270]	2,917	555	643	204	164	451	621
CRISPR [266]	125,495	8,469	9,440	413	347	1,293	2,484
CRISPI [267]	75,329	5,980	6,622	588	617	1,666	2,544
Total	355,999	15,356	16,520	2,869	3,108	12,486	19,373

The size of the CRISPR spacer sets did not correspond to the number of detected scaffolds. The most significant number of scaffolds aligned to the spacers extracted from the HMP datasets, although the group of spacers derived from HMP was not the largest. However, none of the rarefaction curves of the number of scaffolds *selected* by the increasing proportion of the CRISPR database, showed that the collected CRISPR spacer database was incomplete (Fig. E.5).

The next method VirFinder was based solely on the sequence of the scaffolds, so it did not require a database. Scaffolds with the VirFinder p-value $\leqslant 0.05$ were denoted as phage. Among the participant's A Microbiome 7-8% scaffolds were denoted as phage, and in participant B, it was 5-6% (Fig. 4.13). In the Phageome dataset the percentage varied across timepoints but overall was about twice as large as for the Microbiome.

The p-values provided by VirFinder were informative on their own. They differed between the datasets, participants and across the individual time-steps (Fig. 4.14). All distributions were skewed towards the low p-values, suggesting the majority of the scaffolds in both datasets were phage. The pattern of p-value distributions correlated to the antibiotic therapy. Overall, the maxima of the p-value distributions were roughly located in the same positions in both Microbiome and the Phageome.

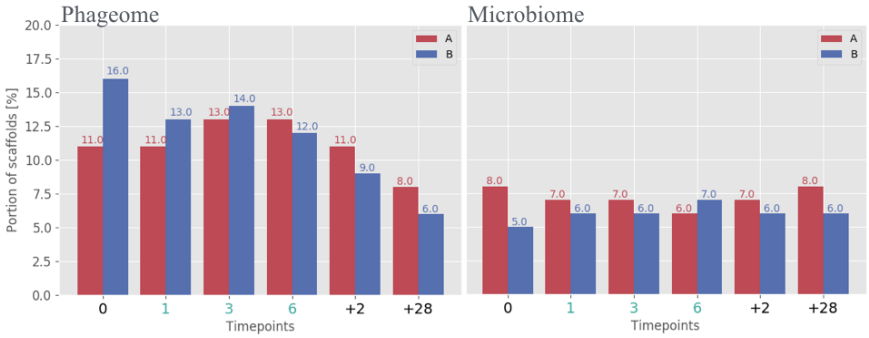


Figure 4.13: Portion of scaffolds denoted by VirFinder as p-value $\leqslant 0.05$ for both datasets.

The distributions of the first and last samples were quite similar to each other but differed between the participants. In the Phageome dataset the p-value distributions were uniformed between the participants, with a sharp peak ~ 0.08. Interestingly, in the last Phageome sample, there was an increase in the proportion of the scaffolds with high p-value ~ 0.6. Therefore they resembled more the Microbiome distribution, which is consistent with the ACLAME classification patterns.

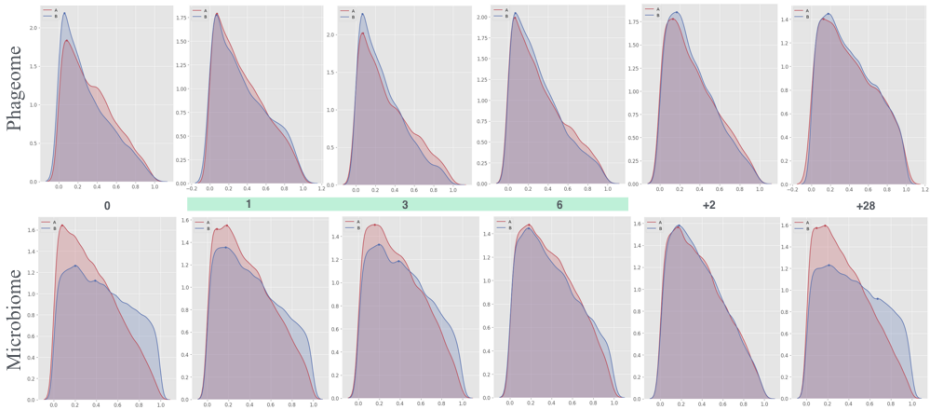


Figure 4.14: Distribution of the VirFinder p-values.

Each of the phage-detection methods resulted in selecting a varying number of scaffolds, and what was worse, those sets rarely overlapped (Table 4.3). Each of the phage identification methods relied on a different aspect of the phage genome. However, each method's ability to discover phages was limited by the scope of the underlying database. None of the methods was sufficiently sensitive to detect the majority of the phages. This lead to the idea of employing a Random Forest classifier based on the scaffold features, with a ground truth being a low p-value from the VirFinder.

Table 4.3: Numbers of scaffolds identified as phage by the four methods, along with the number of co-discovered scaffolds.

	Phageome		Microbiome	
Dataset	A	B	A	B
VirFinder	5,009	5,265	15,214	21,740
CRISPR spacer	2,625	2,676	6,730	9,369
Phage gene	510	406	661	1,173
ACLAME MGE classification: viral	424	296	353	668
Scaffolds co-discovered with two methods				
VirFinder and CRISPR	524	459	792	1,057
CRISPR and Phage gene	180	166	139	199
Phage gene and ACLAME	210	109	70	93
CRISPR and RF	357	395	1,017	1,444

Depth-controlled RFs performed better than those with the default parameters in respect to overfitting (Fig. E.6), as measured by the difference in accuracy and proportion of phage scaffolds. The average OOB-accuracy of the RF runs with the best parameters (Table E.5) reached 69% for participant A and 68% for participant B in both datasets.

The RF also provided a feature importance ranking. It differed between the Phage and Microbiome datasets (Table 4.4). In all runs, the number of functional genes, CRISPR spacers, and GC content were among the most influential features. On the one hand, contrary to the Microbiome runs, in the case of the Phageome, the ACLAME classifications and number of viral genes had non-zero importance. On the other hand, in the Phageome dataset contrary to the Microbiome runs, predicated gene coverage, number or predicted genes, number of reverses oriented ORFs and proportion of overlapping genes had no importance. This supports the assumption that the Phageome dataset contained phage scaffolds solely, as the primary genetic structure features were non-discriminating.

Feature selection, i.e., removing features with a low mean decrease in accuracy, resulted in decreasing OOB accuracy. Therefore, all features were incorporated into the RF model. Finally, RF mining resulted in selecting scaffolds across all ranges of p-values from VirFinder, therefore those scaffolds could not be selected based on a more relaxed p-value cutoff (Fig. E.7).

Although the accuracy of the individual RF classifiers was not convincing, the ensemble approach resulted in selecting scaffolds that would not have been chosen by using a more liberal p-value cutoff. Therefore, RF was an efficient way to combine the phage detection methods. As a result, a large number of phage scaffolds were selected and analyzed. The RF identified 1,218 and 1,317 scaffolds in the Phageome, and 7,944 and 12,708 in the Microbiome for participants A and B respectively.

Phage detection methods seemed highly unreliable. However, I am convinced the selected sets were enriched with MGE sequences. Finally, the phages accounted for 10-12% of all Microbiome scaffolds, which agrees with the previous research reporting

Table 4.4: Mean decrease in accuracy for RF. Red and orange denote the highest and second highest values respectively within the variant.

	Phageome		Microbiome	
Feature	A	B	A	B
Number of MyRast functional genes	26.5 ±2	26.2 ±2	26.9 ±1	15.9 ±1
Scaffolds' GC-content	15.0 ±2	23.3 ±2	16.9 ±2	29.6 ±1
Scaffolds' length	11.9 ±1	14.3 ±1	15.8 ± 1	20.7 ±1
Number of CRISPR spacers	13.4 ±2	12.6 ±1	10.5 ±1	4.5 ±1
Number of viral genes (MyRast)	10.0 ±2	6.0 ±1	2.4 ±1	5.3 ±1
Portion of genes with unknown function	5.0 ±1	6.2 ±1	4.7 ±1	2.2
Number of viral genes (Phaster)	6.4 ±1	4.4 ±1	0.0	0.0
ACLAME classification (vir)	4.3 ±1	1.0	0.0	0.0
ACLAME classification (proph)	1.8 ±1	0.0	0.0	0.0
ACLAME classification (plasmid)	0.9	0.4	0.0	0.0
Predicted gene coverage	0.0	0.0	6.3 ±1	5.0 ±1
Number of predicted genes	0.0	0.0	5.5 ±1	5.9 ±1
Portion of reverse oriented ORFs	0.0	0.0	1.9 ±1	2.9 ±1
Portion of the overlapping genes	0.0	0.0	2.2	1.05

that there are 4-22% phage reads within the standard human gut metagenomic sequencing. This also supports the hypotheses that the majority of bacteria have integrated prophages, which may occupy significant proportions of the bacterial chromosomes [284]. In the Phageome dataset, the percentage varied: reaching maximum (18%) on the 3^{rd} day for participant A and steadily decreasing in participant B from 20 to 8%. Finally, the Microbiome dataset consisted foremost of bacterial ($\sim$ 90%), then phage ($\sim$ 10%), plasmid ($\sim$ 2%), transposon ($\sim$ 0.05%) and integron ($\sim$ 0.3%) scaffolds (Fig. E.8).

4.3.4 Abundance trajectory analysis

High-level analysis

The coverage values had a large variance, but for the vast majority of the scaffolds, especially in the Phageome dataset, the coverage was near-zero (Fig. E.9). However, the idea of *feature abundance* utilized those differences. *Feature abundance analysis* progressed from the most global to the most detailed. The first level included all scaffolds. The global *feature abundance* trajectories for both participants in the Microbiome set decreased on the 3^{rd}, 6^{th} and $+2^{nd}$ days and increased back up to the initial levels, on the $+28^{th}$ day (Fig. 4.15).

This pattern was not present in the Phageome trajectories. The last sample in the Phageome set participant's B was characterized by unusually high abundance. This sample showed unusual patterns in a number of the previous steps, so had to be kept in mind for the downstream analysis.

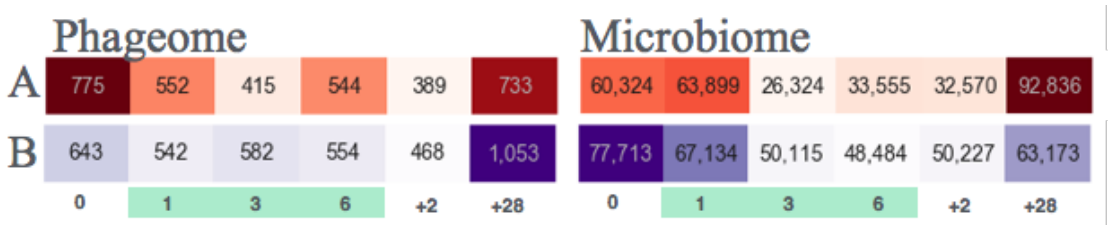


Figure 4.15: Global abundance and number-of-scaffolds trajectories for all scaffolds in both participants and both sets.

Next, *feature abundance trajectory* analysis level focused on the changes of relative abundances of the *Firmicutes* and *Bacteroidetes*. The abundance ratio of those phyla is one of the most widely used global descriptors of the gut microbiome's evolution. The studies agree the fluoroquinolone antibiotics cause an increase of the ratio [285]. However, the trajectories of the *Bacteroides* and *Firmicutes* differed between the participants (Fig. 4.16).

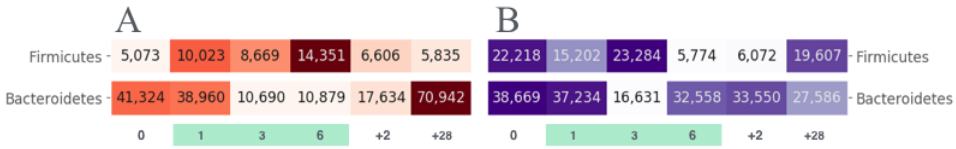


Figure 4.16: *Abundance trajectories* for the *Firmicutes* and *Bacteroides* phyla in the Microbiome dataset.

Behavior of the participant A's Microbiome agreed with the previous research. The *Firmicutes's* abundance started increasing before the *Bacteroides* got entirely suppressed as if the bacterio-static effect was enough to free ecological space for the *Firmicutes*. In participant B there was no direction within the trajectories. Both phyla had high abundance in the first two days. Next, the abundance of the *Bacteroides* bacteria decreased where *Firmicutes* was high but only for one day, and then the relationship reversed, *Firmicutes* had low abundance where *Bacteroides* were high. Lastly, the *Firmicutes* restored their abundance whereas the abundance of the *Bacteroides* ended lower.

Researchers discovered that the abundance of *Firmicutes* genera *Faecalibacterium and Ruminococcus* and *Bacteroides* genus *Alistipes* decreased under pressure of the ciprofloxacin therapy but the *Bacteroides* genus *Bacteroides* and the *Lachnospiraceae* family increased [151]. Exactly this pattern was observed in the participant A's Microbiome (Fig. 4.17). However, in the participant B's Microbiome *Bacteroidetes* abundance started increasing too early, on the 6^{th} day.

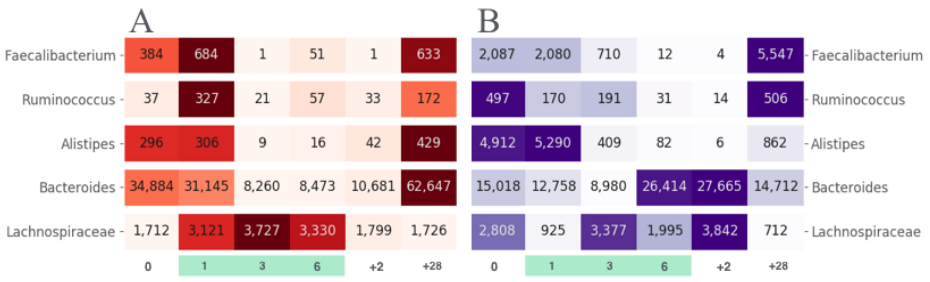


Figure 4.17: *Abundance trajectories* for the genera and families that are the most impacted by ciprofloxacin for the Microbiome set.

Mobile genetic elements, and resistance trajectory analysis

Fig. 4.18 shows *abundance trajectories* for MGEs, ARGs grouped by the antibiotic and MGEs with the ARGs for the Microbiome and Phageome datasets.

The signal in the Phageome set was weaker than for the equivalent selections in the Microbiome set. The patterns also differed between the two participants. However, overall, increase in abundance for MGEs and ARGs coincided with the antibiotics therapy. There were three therapy-related trajectory patterns: trajectory characterized by a relatively low abundance on the first two days, a subsequent increase from the 3^{rd} to $+2^{nd}$ days, and a decrease on the last day, and trajectory characterized by an increase from the 6^{th} to $+2^{nd}$ or until $+28^{th}$ days of sampling. The first trajectory type was more prevalent for participant A and the latter two types for participant B.

Phage *abundance trajectories* universally dropped on the 3^{rd} day and then continued slowly rebuilding in the second part of sampling. In Microbiome participant B, MGEs had therapy-related trajectories. In participant A's Microbiome, the patterns were less unified. The integron scaffolds were the most abundant on the 3^{rd} day when all other MGEs reached minimal abundance. Phageome trajectories for MGEs were weak - which agrees with the hypotheses that the Phageome dataset contained mostly phage scaffolds.

Among the antibiotic-grouped ARG trajectories, fluoroquinolone resistance genes were the most abundant in the participant A's Microbiome, the second most abundant, after β-lactams in the B's Microbiome, and in both cases characterized by therapy-related trajectories. The resistance against aminoglycosides decreased between the 6^{th} and $+2^{nd}$ days. Therefore its pattern was similar to the global trajectory. Phageome trajectories of the resistomes were shifted as the increase started on a later date than in the Microbiome trajectories. A similar pattern was observed for the Phageome. However, their values were much lower. Interestingly, it appeared a large portion of the ARGs could also be attributed to the MGEs. As in the Microbiome dataset, both participants MGEs with ARGs and AR-HMMs had almost uniformly strong therapy-related trajectories.

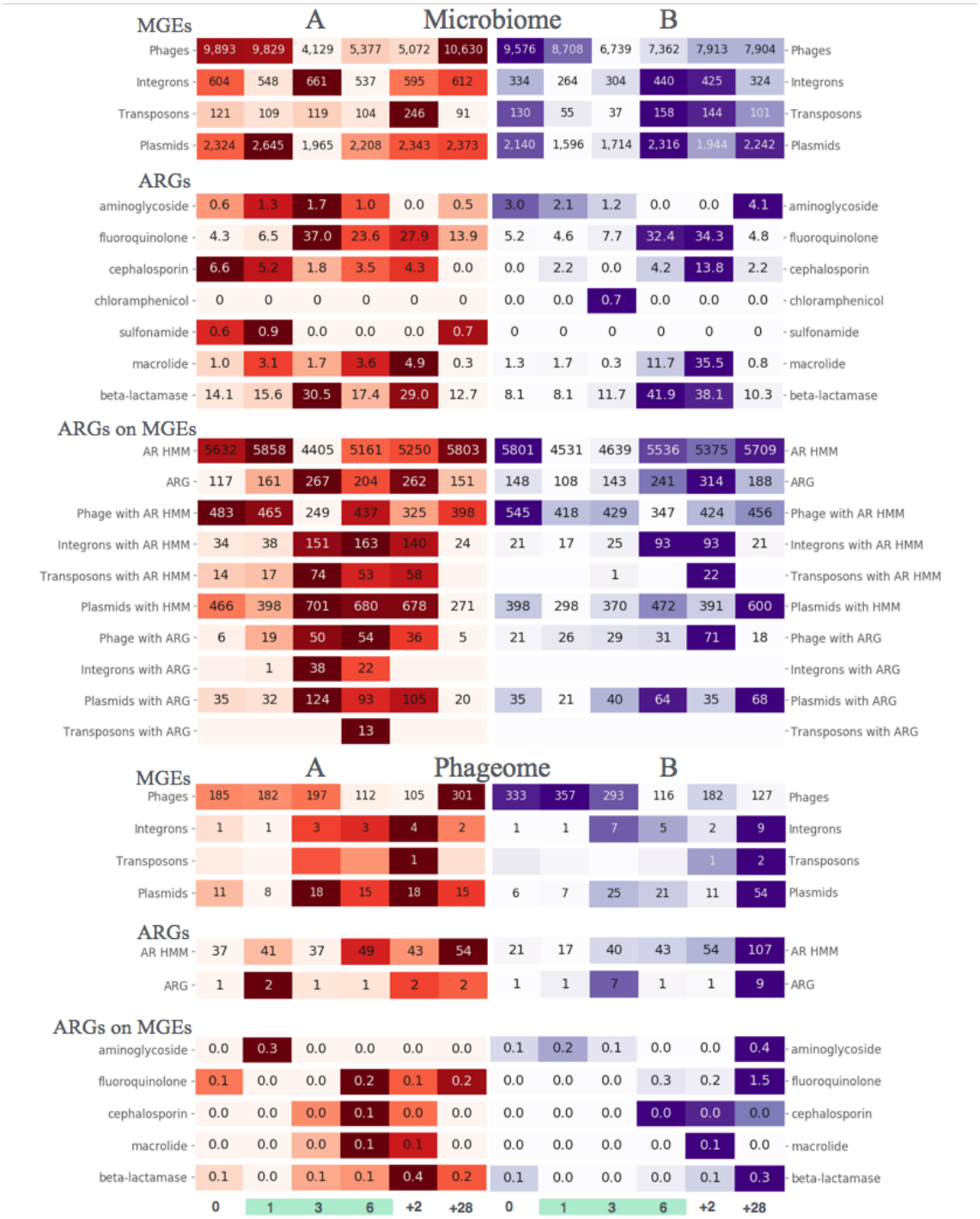


Figure 4.18: *Feature abundance trajectories* of MGEs.

Gene-level analysis

Although the high-level abundance trajectories show clear patterns, the holy grail is to analyze the dynamics within microbiome with gene-level resolution. A unique trajectory characterized every gene. Fig. 4.19 presents the most abundant *average scaled trajectories* for the lowest functional unit in the MyRast, CARD, and AR HMM classifications. Trajectory clustering, average computation, and scaling enabled identifying genes with similar behaviors. Even then, there were hundreds of trajectories observed.

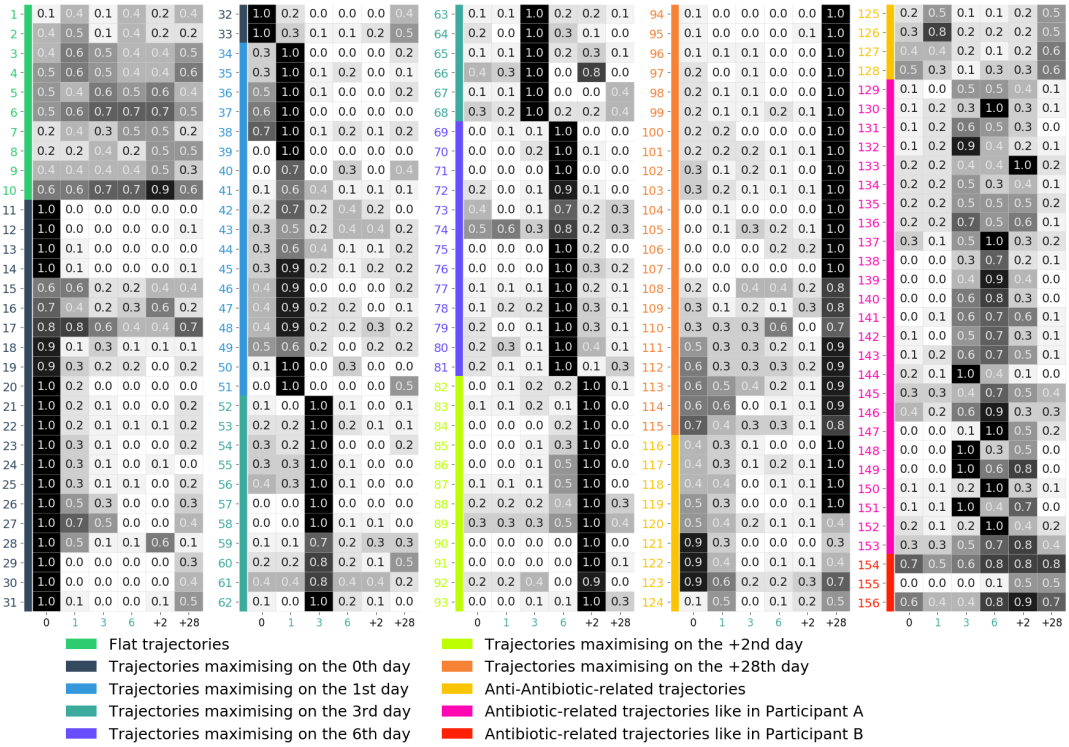
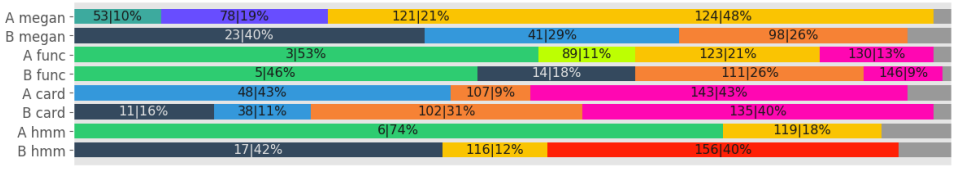
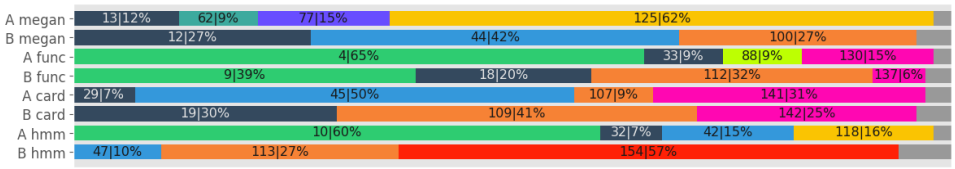


Figure 4.19: *Average scaled trajectories* for all features.

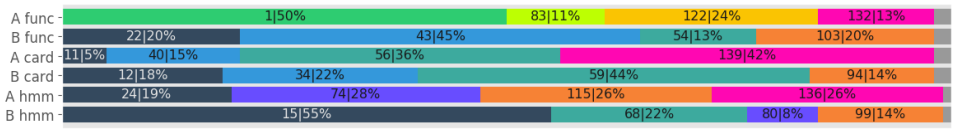
Although particular trajectories do not repeat across the underlying scaffold selections and annotation systems, they fall into ten classes: a flat trajectory with no clear direction, six patterns of trajectories with a certain maximum on one of the sampling days, therapy-related trajectories, and anti-therapy-related pattern. The trajectory types are color-coded in the Figures 4.19 and 4.20. The profiles were divided depending on the MGEs and feature scaffold selections: i.e., Bacteria, Phages, Integrons, Transposons, Plasmids, and functional genes, taxonomic annotations, ARGs, and HMM-profiles. The results described below are based on simultaneous analysis of the three resources: *average scaled trajectories* in Fig. 4.19, *trajectory profiles* in Fig. 4.20 and lists of features assigned to trajectories in the tables that were too large to be included.



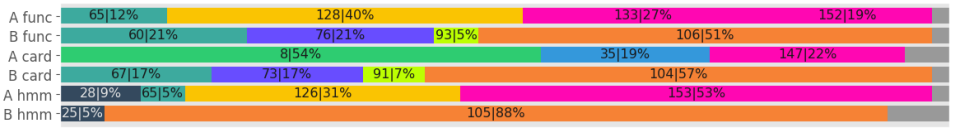
(a) Microbiome|All scaffolds



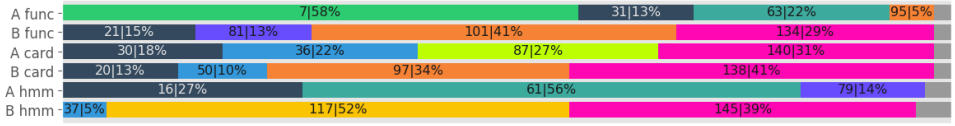
(b) Microbiome|Bacterial scaffolds



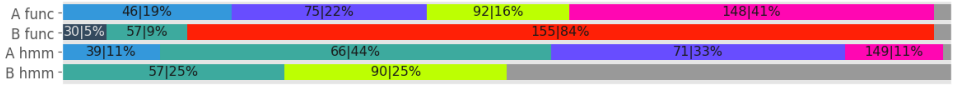
(c) Microbiome|Phage scaffolds



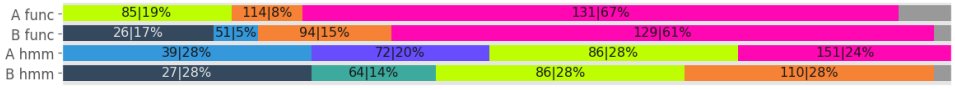
(d) Phageome|All scaffolds



(e) Microbiome|Plasmids



(f) Microbiome|Transposons



(g) Microbiome|Integrons

Figure 4.20: *Feature-abundance trajectory profiles.* Each horizontal plot denotes a *trajectory profile* of the genes, indicated by the label in the y-axis and the scaffold selection, e.g., the first bar presents that 10% of the taxa in the Microbiome of participant A had trajectory number 53, 19% trajectory num 78 and so on. Numbers denote trajectories (Fig. 4.19) and the percentage. The gray bars present portion of genes with low-count trajectories (<5%).

The trajectory profile for taxonomic assignments differed between the two participants. A significant portion of the taxa had an anti-antibiotic trajectory, i.e., they corresponded to those bacteria that were susceptible to ciprofloxacin. The Gram-positive bacteria were prevalent in the trajectories with an abundance increasing during the antibiotic-related therapy such as the trajectory number 78, and of Gram-negative bacteria within those decreasing during the therapy such as the trajectory number 121 (Fig. 4.20(a)).

Phage scaffolds were represented into both profiles represented in Fig. 4.20(c) and 4.20(d). The first selection included the free phages and the prophages incorporated in the bacterial genomes. The latter comprised solely free phages. In participant A and both phage selections, a significant proportion of functional genes had a therapy-related trajectory. Both of those sets for both participants included numerous phage-related proteins such as phage tail, phage terminase, portal proteins and a large proportion of the proteins with unknown function.

A significant proportion of the functional genes had flat trajectories. House-keeping genes are present in all bacteria, and their abundance does not depend on the administration of antibiotics. The functional assignments included also efflux pumps, mobile genetic elements, toxins, and other HGT-related proteins, which had therapy-related trajectories, e.g., site-specific recombinases and conjugative transposon proteins. In the same time, the phage-related and unknown genes were less often found in the Microbiome bacterial scaffolds (Fig. 4.20(b)).

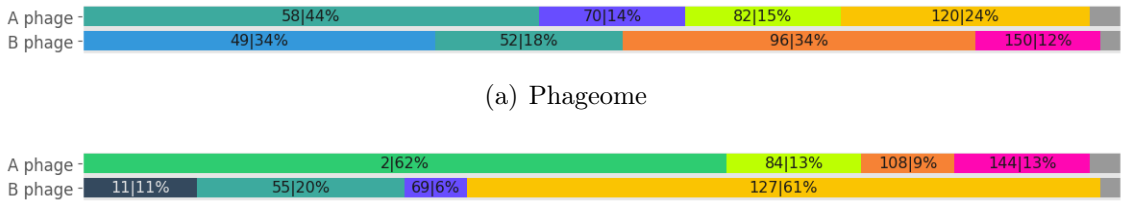
Only a small proportion of the bacterial scaffolds carried full ARGs. As it was reported before [286], this number was even smaller for Phageome. However, among the ARGs between 22% and 43% were characterized by the therapy-related trajectories almost independently of the selection of the scaffolds, also including MGEs.

The majority of the identified ARGs were not ciprofloxacin-specific, and included ABC transporters, efflux pumps and extended-spectrum beta-lactamases, such as OXA-347, which was found on bacterial and plasmid scaffolds of the participant A, confirming both the original study and the read-based analysis. Mostly, the fluoroquinolone-resistance genes found were gyrA genes, coding for mutated resistant gyrases.

There were many more scaffolds with an AR HMM profile, than with the ARGs. A significant portion of the AR HMM profiles on the bacterial scaffolds in participant B follow the participant B therapy-related trajectory. In the case of transposons, those include the multiple-antibiotic resistance protein marC. This suggests that full ARGs were rare within the metagenomics assembly, but partial or potentially further related ARGs are quite common. AR HMM profile alignment was a more sensitive method, however, potentially prone to false positive errors.

Finally, truly a tiny fraction of scaffolds had an alignment to any viral proteins, enabling their taxonomic annotation. The majority of the annotated phages were related to the human gut microbiome, including *Bacilli*, *Clostridia*, *Staphylococcus*.

There was a weak correspondence between the same-patient samples of the two sets. Unculturable crAssphage was found in samples of both participants. The phages had a range of different trajectories, including those related to antibiotic therapy (Fig. 4.21).



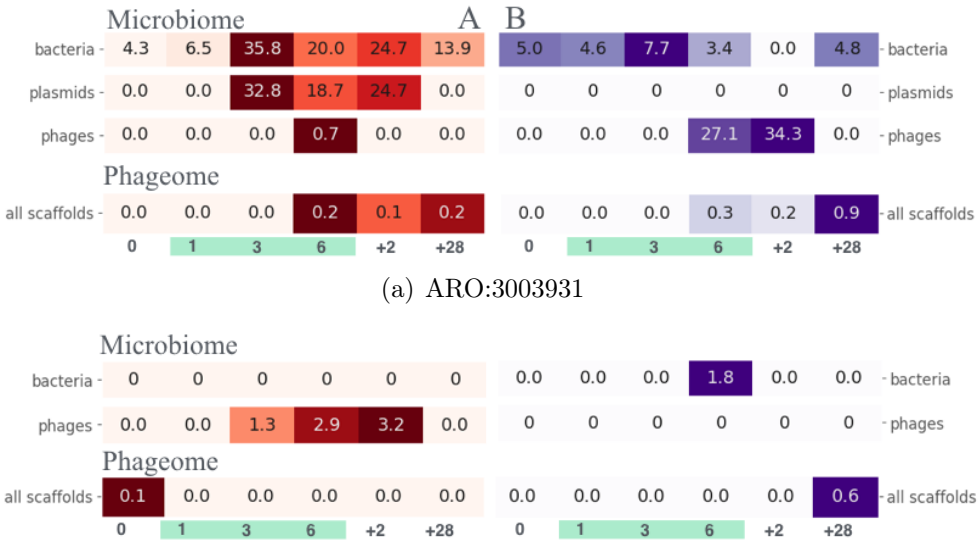
(a) Phageome

(b) Microbiome

Figure 4.21: *Feature abundance trajectories* for phage taxonomic annotations.

4.3.5 Ciprofloxacin resistance

Two genes conferring resistance to fluoroquinolones including ciprofloxacin were found in the assembly. Both genes, ARO3003831 and ARO3003995, encode mutated gyrases *gyrA* to which the fluoroquinolone antibiotics cannot bind. In the Phageome dataset, the *abundance trajectory* of ARO3003831 increased in the second half of sampling consistently for both participants (Fig. 4.22).



(a) ARO:3003931

(b) ARO:3003995

Figure 4.22: *Feature abundance trajectories* for gyrases.

In participant A's Microbiome, gyrase trajectories on the bacterial and phage scaffolds had the therapy-related trajectories. In the Microbiome dataset of

participant B the gyrase was present on the bacterial chromosome scaffolds throughout sampling. However, their abundance increased on the phage scaffolds on the 6^{th} and $+2^{nd}$ days. The abundance of ARO:3003995 was much lower than that of ARO:3003931. ARO:3003995 had therapy-related trajectories on phages in participant A's Microbiome, however they were not present in bacteria.

In both participants, the gyrases were found on bacterial chromosomes before the therapy, but prompted by the antibiotic they transferred to plasmids, pro-phages and the free phages in participant A. In participant B the transfer omitted the plasmids, but gyrases also shifted first to prophages and later to the free VLPs.

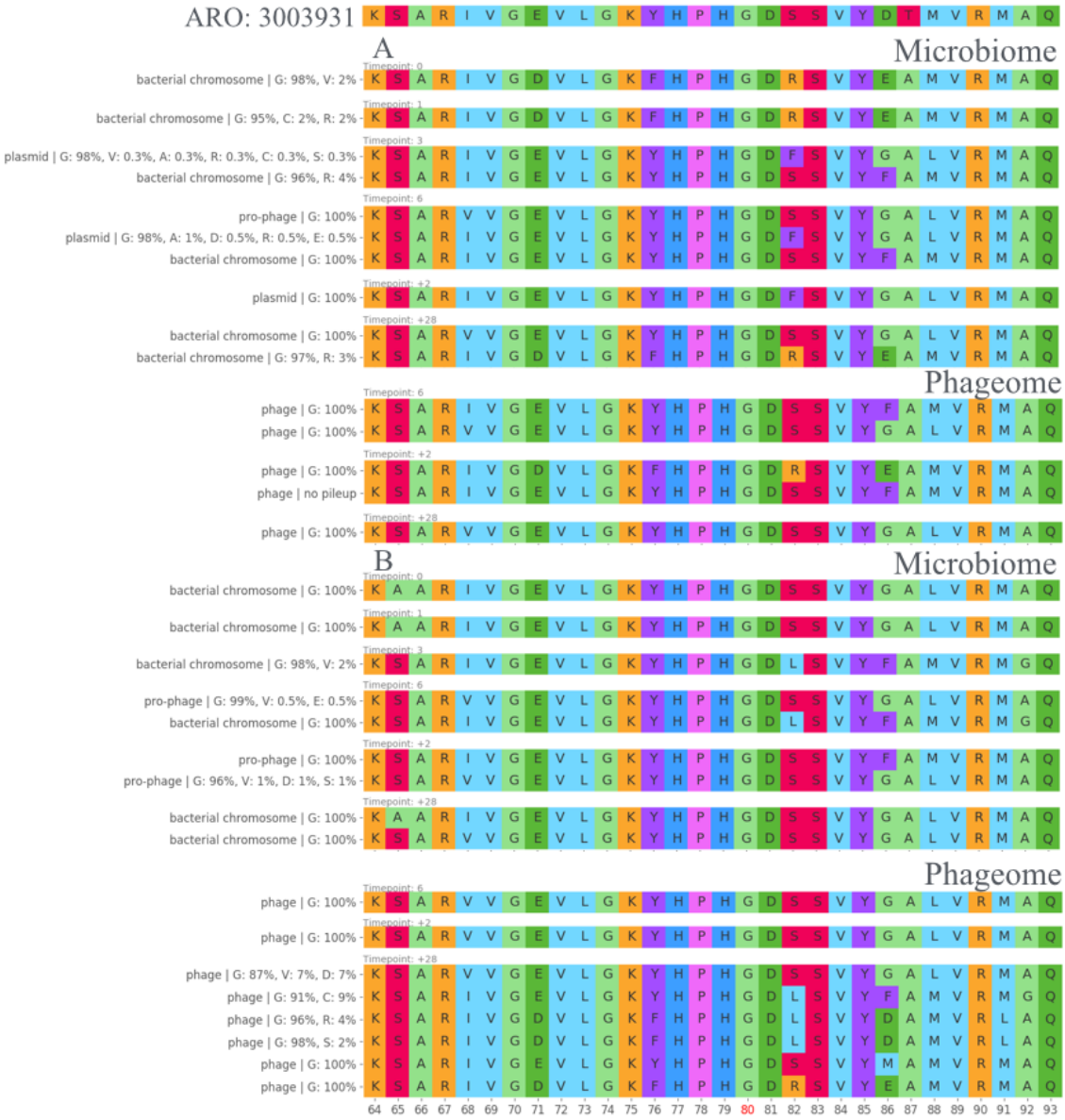


Figure 4.23: Protein sequence alignment for ARO:3003931 gyrase.

As the resistance of gyrases relies on the point mutations, their detection in the metagenomic analysis could be subject to false positive. However, in the case of both analyzed gyrases, both mutations were highly supported by the reads mapping to the mutation sites. The great majority of the mapped reads carried the mutated codon (Fig. 4.23). The amino acids in the 82 position differed across the alignment of the gyraze sequences isolated from various scaffolds. However the mutations in this position cluster according to the MGEs within the set and participant.

4.3.6 Global diversity trajectories

To avoid database bias, the diversity was measured based on the clustered scaffolds. First, scaffolds longer than 200 bp for all samples within a variant were put together. Second, they were clustered (90% sequence similarity) using CD-HIT [287]. Third, the longest sequences within a cluster were taken. Subsequently, the reads for each sample separately were mapped onto the cluster representatives. Finally, cluster representative was treated as the unit of diversity, and the proportion of reads mapped onto them as an abundance. In the very end the Shannon index was computed (Fig. 4.24).

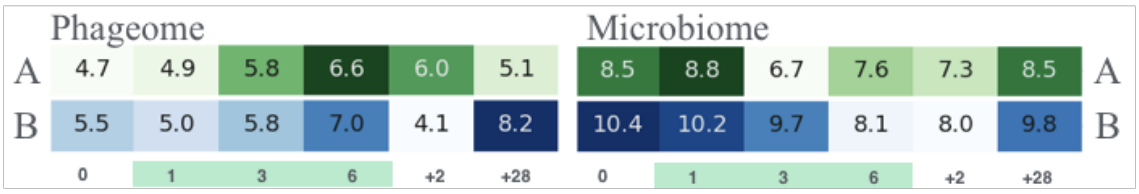
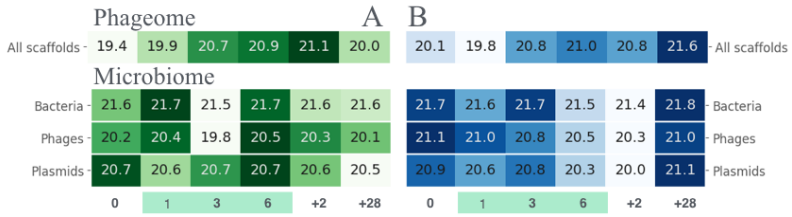


Figure 4.24: Shannon index computed in a read-recruiting manner.

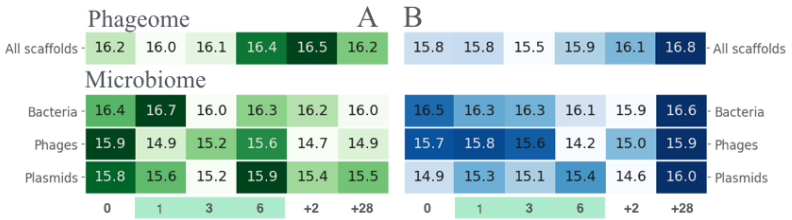
The global diversity in the Microbiome set decreased in response to antibiotic therapy. The decrease was more rapid in A patient than B, but in both, it happened on the 3^{rd} day of antibiotic therapy and restored on the last day of sampling. This was not the case in the Phageome set, where the diversity maximized on the 6^{th} day.

4.3.7 Diversity trajectories

The functional diversity of the Phageome increased in both participants. In the Microbiome, it decreased on the 6^{th} day of therapy in participant B regardless of the MGE selection. In the Microbiome of participant A, bacteria and phage functional diversity dropped on the 3^{rd} day but it increased back already on the 6^{th} day, whereas the plasmid diversity remained constant.



(a) Functional



(b) ARGs

Figure 4.25: *Diversity trajectories* for functional genes and ARGs.

The *resistome diversity* increased in Phageome for both participants in the second half of sampling (Fig. 4.25(a)). It remained elevated even on the last day of sampling, especially in the case of participant B. Microbiome diversity trajectories differed between the MGE annotations. Participant A's ARGs diversity decreased on the 3^{rd} day and restored already on the 6^{th} day, independently of the MGE selection. In the case of participant B, the decrease happened later, on the 6^{th} day and the restoration was slower, as it happened on the last day. Plasmid *resistome diversity trajectories* were elevated during the therapy. Overall the *diversity trajectories* were harder to interpret the *abundance trajectories*, but they confirmed the ARGs diversity increased on phages after the therapy.

4.3.8 Phage/host dynamics

Not only did the abundance of MGEs increase under the antibiotic pressure but also so did the frequency of the occurrences of the integration of phages into the bacterial genomes. Fig. 4.26 shows the increase had an antibiotic-related pattern, between 1^{st} and 3^{rd}, or 3^{rd} to 6^{th} days, for participants A and B, respectively. The phages that incorporated were enriched with the AR HMMs. Overall approximately 5% of the Microbiome phage scaffolds had an alignment to the AR HMM profile but, among those that were considered integrated the majority had an AR HMM profile.

However, each of the phage/host trajectories was unique. I tried to find a phage/host relationship pattern that suggested a transduction event of a gene increasing the fitness of the bacterial host, such as an ARG. Such a pattern was observed in *Flavobacterium psychrophilum* and *Sphingobacterium mizutani* in participant A, and *Bacillus cereus*, *Chryseobacterium piperi* in participant B

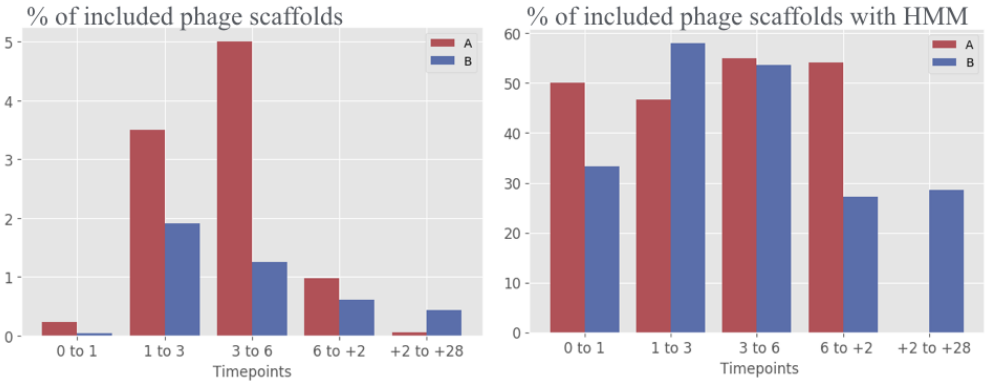


Figure 4.26: The proportion of phage scaffolds that have integrated into the bacterial chromosomes within the following timestep, and proportion of those with an alignment to the AR HMM profile.

(Fig. E.10). The corresponding difference trajectories were relatively constant for some time and followed by a rapid decrease, indicating an increase of the phage abundance, and then a rapid increase, indicating an increase in the host abundance.

4.4 Summary and conclusions

The antibiotics exert environmental pressure on bacteria and phages. In the Microbiome, antibiotics prompted an increase in abundance of the overall MGEs and ARGs, specific to fluoroquinolones and beta-lactamases. Moreover, I was able to track the transfer of the ciprofloxacin resistance gene, from the bacterial chromosomes, onto plasmids, pro-phages, and free phages in the last place. The resistance to ciprofloxacin is conferred by the mutated gyrazes, where the mutation decreases their affinity to the antibiotic - I showed the sequence of the protein carried the mutation on the relevant position and that changed when transferring to MGEs.

From the bioinformatics perspective, this study presented a unique setup as it incorporated the time-series samples sequencing of both the total gut microbiome and the free-phage fraction. The ground step of the pipeline: metagenomic assembly, causes data loss and introduces bias. However, it enables the association between the various features, such as taxonomic and functional assignments, time steps and abundance. This together with the metadata provided by the databases, enables a unique trajectory-based high-resolution analysis of the individual elements of the microbiome. The trajectory-oriented approach to the metagenome analysis is scalable regarding samples and applicable in the analysis of other time-series dataset.

Identification of the MGEs within the whole metagenomic sequencing data is a difficult task, as all of the elements: phages, plasmids, transposons, integrons have similar genetic traits. Machine learning was used to separate the metagenomic assembly - this class of methods is able to utilize multiple various features at once.

With all certain errors in the classification of the individual scaffolds, I hope the final sets were greatly enriched with them, and therefore, enabled the *feature trajectory analysis*. Nevertheless, the separation of the MGEs, especially in the time-series scheme, facilitates a more detailed analysis of the dynamics within the microbiome.

The taxonomic analysis of the MGEs makes little sense. They are weakly represented in current databases, carry unknown genes, or genes misclassified as other MGEs [282]. Their biology does not support the tree-like taxonomy, as they could be associated with multiple bacterial hosts. Therefore the separation of the MGEs before the bacterial taxonomic analysis could improve the latter. However, the association between the MGEs and their bacterial hosts could be discovered by other methods as shown in this chapter. The result should be a network rather than a tree.

I showed that antibiotics resistance emergence is mediated by HGT, in time as short as a month. This phenomenon can be studied with WGS of metagenomes. However, it requires a novel step before classical metagenomics analysis, to separate the dataset into the MGEs and bacterial sequences. The analysis could be improved with the long read sequencing technologies, which would enable skipping the assembly step.

Chapter 5

SATURN project

5.1 Introduction

Antibiotics facilitate colonization by the multi-drug resistant strains such as MRSA or ESBL-carrying *Enterobacteriaceae* (ESBLs). During the SATURN Work-Package 4 (WP4) for three years in three clinical centers in three countries, scientists collected detailed data about admitted patients and repeatedly tested for the presence of MRSA and ESBLs. The question posed by the SATURN project was to stratify antibiotics therapies by their impact on the probability of acquiring MRSA and ESBL, to improve the guidelines for prescribing antibiotics.

I aimed to construct a machine learning (ML) model that would enable the discovery of relationships between patients characteristics, antibiotics therapy, and colonization with MRSA or ESBLs, without the expert involvement. The second aim was to provide a predictive tool to select better antibiotics therapies from presented alternatives. Prof Evelina Tacconnelli coordinated the SATURN WP4. The SATURN project team collected the data. Members of Prof. Tacconelli's research group worked on collecting and cleaning the data. Dr. Primrose Beryl Gladstone performed a parallel analysis using logistic regression. Prof. Bernhard Schölkopf from Max Planck for Intelligent Systems advised on how to construct the data vector initially for SVM. The RF part was discussed with Prof. Michael Cummings, University of Maryland.

The data were carefully collected in the medical and surgical wards of the Italian Università Cattolica Sacro Cuore (UCSC), the Romanian Institute for Infectious Diseases Matei Bals (IDMB) and the Clinical Centre of Serbia (CCS). All of the institutions followed the same protocols and questionnaires designed by the SATURN consortium. The staff was trained to ensure the data were comparable. ESBL microbiological tests were based on the fecal swabs and MRSA tests on the nasal swabs. Samples were analyzed at the local laboratories in the hospitals. Negative or unclear results were confirmed by microbiological tests on bacteria plated and after overnight growth in an ESBL-selective medium. Samples were

plated on ESBL medium and incubated at $37°C$ in the air for 18 to 24 hours. If ESBL-positive colonies were observed, a single colony of each morphology was subcultured and ESBL presence confirmed. For each patient, extensive metadata were collected, regarding their demographics, age, gender, comorbidities such as surgeries and diseases, they also contained a detailed description of the patients' treatment and stay in the hospital.

First, on admission, all of the background data were collected: patient's medical history, reasons for admission, whether they were taking antibiotics at the moment of admission, and tests for the presence of MRSA and ESBLs were performed (Fig. 5.1). Next, the patients underwent treatment, which potentially included antibiotics therapy. The details of the antibiotics therapy were collected: antibiotics, dates, and dosage. A patient could have received more than one antibiotic. On the day of discharge, the tests for MRSA or ESBL were performed again. If the patient was treated with antibiotics, they entered the follow-up study. The follow-up MRSA/ESBL and blood samples were taken on days 0^{th}, 3^{rd}, 7^{th}, 15^{th} and 30^{th} of antibiotic therapy, or after it. Each time the patient was tested for MRSA and ESBL the basic blood parameters were measured. Hence, the follow-up could be concluded during the hospital stay, but could also be continued after the discharge.

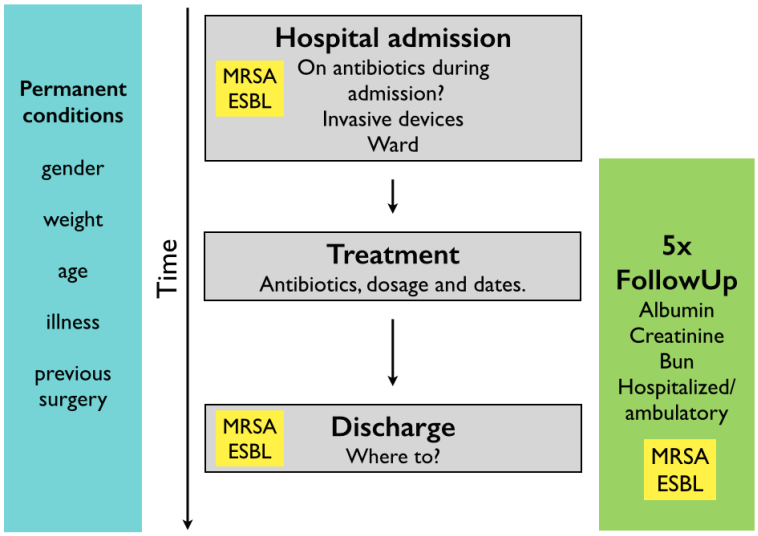


Figure 5.1: Scheme of data collection for SATURN WP4.

This chapter describes my contributions to the analysis and machine learning for the WP4 dataset. The chapter does not adhere to the classical Methods and Results division. The significant portion of the machine learning especially feature-engineering depends on the data itself. Therefore, in the first place I describe the data and its encoding, and later the machine learning, feature importance, and finally the `AskSaturn` website.

5.2 Data processing

Before machine learning can be applied data had to be prepared. First, the data needed to be migrated from the original files into the database. Then, they had to be encoded. Namely, the real-world values had to be transformed into the features. Next, the data were analyzed, and missing data were inputted. Finally, the data could be scaled and balanced (Fig. 5.2).

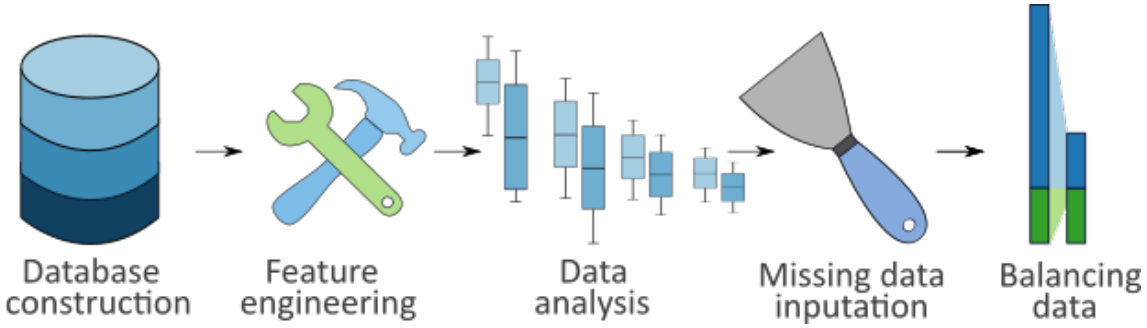


Figure 5.2: Data processing pipeline.

5.2.1 Database construction

The dataset was first provided in the form of a six `Excel` sheets. Three per every country: main, antibiotics therapies and follow up, which corresponded to the questionnaires patients and stuff filled in the hospitals. The main part consisted of 146 columns containing information collected at the moment of admission. The antibiotics sheet had only 16 columns and contained all details about the antibiotics therapies. Finally, the follow-up sheet had 416 columns with data of all five follow-up samples, MRSA and ESBL positivity and blood parameters. Most of the columns held binary data, and the majority of the table is empty.

Every patient was identified by a unique patient id (PID). If they were admitted to the hospital multiple times during the three years period of this study, each time they were given another PID. The PID was a primary key enabling relating the tables. The Excel database was not normalized as most of the cells were empty. Therefore, this data storage method had many disadvantages: it was slow to query, and burdened with the platform- and version- incompatibility.

Therefore, the data needed to be migrated into a structure that would enable easy and fast access. Because of the mentioned problems with format conversion, I wanted to avoid using a type-controlled structure not to introduce errors. Dataset was also not too large, therefore the size was not a concern. Additionally, it was static and was not to be further altered. Knowing all that, I decided to use a document-based database in opposition to SQL-based one, avoiding laborious construction of the relation-database, type control, and normalization.

Among the many available engines, `MongoDB` [288] was selected, as it is an open-source project with a well documented `Python API` and supported by a large community. A record in `MongoDB` is a document, a dictionary, which can be embedded as the values can be of any type. The `MongoDB` is organized in sets of documents called collections. The `MongoDB Python API` translates `Python` dictionaries into database documents and vice versa. Navigating this database is intuitive, as every field is accessible by specifying a sequence of the keys. Most importantly, `Mongo` enables building SQL-like queries from the `Python` level.

The final database for SATURN data was split into three collections - one for every country. Every patient was described with one nested document. If the field is empty, it is not added to the document. Therefore every record had a different size. The database was queried to construct input for machine learning.

5.2.2 Feature engineering

The input for ML is a data matrix (X), where columns are features and rows are observations. In this study, rows are feature vectors, each describes a single patient. The last Y column contained class labels describing MRSA/ESBL positivity and negativity. *Encoding* is the process of creating a feature vector from a database record. *Encoding* uses a functional programming style in Python. The `CodePatient` class has a database record and functions generating all of the *feature* values, e.g. `code_BMI`. To create a feature vector computer iterates through all of the functions in the `CodePatient` class, and calls those functions which name starts with `code`. In this step, the functions can be filtered so that various feature vectors can be constructed. The results of those functions are added to the dictionary encoding a single patient. To ensure all feature vectors are sorted in the same way, the `GroupCoder` iterated over all of the `CodePatient` dictionaries using a static list of keys. Finally, `GroupCoder` holds the X array and the Y vector, and list of PIDs.

The data features can be classified depending on their data type or the real-world semantics. The majority of the features are binary, but some are categorical or numerical. Encoding of the binary features was straightforward. Categorical data needed to be expanded into multiple binary columns and numerical had to be scaled. The diseases are grouped and simplified for encoding. As each disease is described with specific features in the data, each of them required a separate function. However, overall the diseases are encoded with binary features, as a rule not taking into account their details. Table 5.1 describes how the features were encoded.

Table 5.1: Features

Features	Description
Positivity	See Paragraph 5.2.2
Age, BMI, Female, Male	Age was encoded directly, body mass index (BMI) was computed from patients weight and height. Gender was encoded with two exclusive binary columns.
Ward type	Surgical or Medical ward, encoded with exclusive binary columns.
Previous antibiotics	Positive if patient took antibiotics within the month prior to admission.
Previous hospitalization, ICU	Positive if a patient had been hospitalized or in ICU within the last three months respectivelly.
Admission from LCTF	Positive if the patient was admitted from another care facility such as acute care, extended care facility, institutions.
Domiciliary assistance	Positive when patient used domiciliary nursing or medical assistance.
Heart disease	Positive if patient had any of the diseases: myocardial infarction, congestive heart failure, peripheral vascular disease
Cardiovascular disease	Positive if patient had Cardiovascular disease
Malignancy	Positive if patient had malignancy in the past, or active independently of its type.
Neurological disease	Positive if patient had hemiplegia, dementia, cerebrovascular disease or any other major neurological disease.
Chronic skin lesions, renal, lung, connective tissue disease, respiratory, liver diseases, diabetes, organ damage, ulcer, HIV, infectious and parasitic diseases	Binary features, positive if patient had any of the diseases.
Immunodeficiency	Positive if patient was immunodeficient, independelty of the cause such as HIV, radio or chemo-therapy, post-transplant etc.
Surgeries, Transfers	Number of surgeries, number of times the patient was transferred within the hospital
ESBL, MRSA roommates	Binary feature, positive if patient had a roommate carrying ESBL or MRSA
Dialysis, invasive devices	Binary features denoting if patient underwent dialysis or used other invasive devices.
Bedridden, diarrhea, wounds	Binary features describing patients state while in hospital
Season	Four binary columns denoting the seasons of the hospitalisation
Antibiotic therapies	see Paragraph 5.2.2
Length of hospitalization	see Paragrah 5.2.2
Colonization pressure	see Paragraph 5.2.2

Positivity A patient was labeled positive if at least one sample of the follow-ups: 3^{rd}, 7^{th}, 15^{th} or 30^{th} or discharge was positive. The patient was labeled negative when all of the collected samples were negative and if there was at least one more sample tested apart from the admission and 0^{th} follow-up.

Length of hospitalisation The number of days between the dates of admission and discharge, or the date of the first positive sample if it was earlier than admission. Therefore, only the time of hospitalization before the positive sample was taken into account.

Colonization pressure The probability of colonization depends on the frequency of the patient's interactions with the MDR carriers. The *colonization pressure* is an average proportion of the positive patients within the hospital population, over the days of the patient's recent hospitalization history. The period depends on the sampling timing. In the case of this dataset, it was four days. Feature vectors included an average of the colonization pressure from admission to the hospital to discharge or the first positive sample.

Antibiotic therapies First, to decrease the complexity of the data Prof. Tacconelli grouped the antibiotics so that from the original 68 classes, the 17 super-classes were created: Aminoglycosides, Anti/Anaerobes, Carbapenems, Cephalosporins, Clindamycin, Colistin, Cotrimoxazole, Daptomycin, Glycopeptides, Linezolid, Macrolides, Metro, Penicillins, Quinolone, Tetracyclines, Tigecycline, Piperacillin. This step was necessary because otherwise, almost all of the records would be unique and it would be impossible to generalize across the patients.

The antibiotic therapy was encoded up to the first positive sample, or the discharge day if the patient remained negative at all times. First, we encoded the number of days each separate antibiotic was taken, the number of days pairwise combinations of antibiotics were prescribed and finally the permutations, describing the sequence of prescribed antibiotics. The permutations were encoded with the binary features.

Although using the antibiotic superclasses, the therapy of a single patient can be very complicated. In Fig. 5.3 the antibiotic therapy for one of the patients in the dataset is shown, and in Table 5.2 the encoding for that therapy is presented. For some patients, the antibiotic therapy before admission was also recorded in detail. However, as it was not recorded for all of the patients, the details of the therapy before admission were ignored and denoted only with a binary feature. Finally, the data vector had 468 features.

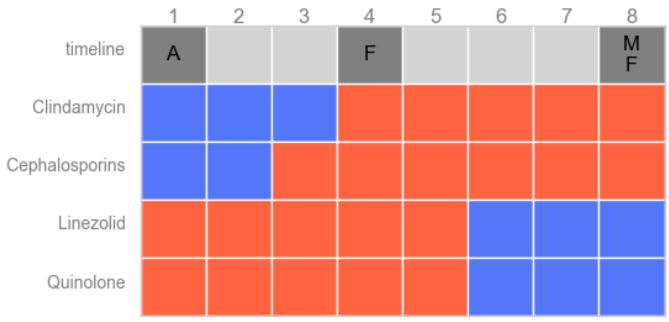


Figure 5.3: Antibiotic therapy for one of the patients. Columns represent days of therapy and rows prescribed antibiotics. Red fields indicate the days the antibiotics were taken, and blue when they were not. In the timeline row, A stands for admission, F for follow-up samples and E/M for samples positive for ESBL/MRSA.

Single antibiotic therapy	
Clindamycin	5
Cephalosporins	5
Linezolid	6
Quinolone	5
Combinations	
Clindamycin and Cephalosporins	5
Clindamycin and Linezolid	3
Clindamycin and Quinolone	2
Linezolid and Quinolone	2
Cephalosporins and Linezolid	3
Cephalosporins and Quinolone	2
Linezolid and Quinolone	5
Permutations (order)	
Clindamycin before Linezolid	1
Clindamycin before Quinolone	1
Cephalosporins before Linezolid	1
Cephalosporins before Quinolone	1
Linezolid before Quinolone	1

Table 5.2: Encoding.

5.2.3 Missing data

Records of 300 patients lacked information about BMI (Fig. 5.4). The missing data were inputted by randomly choosing a value from the distribution model fitted to the distribution of the known data. In 98% of the cases, both height and weight values were missing. Therefore, rather than modeling the two-dimensional distribution, which would assume one of the values is known, after ensuring the BMI was not correlated to main features such as age, I fitted a normal distribution of the BMI.

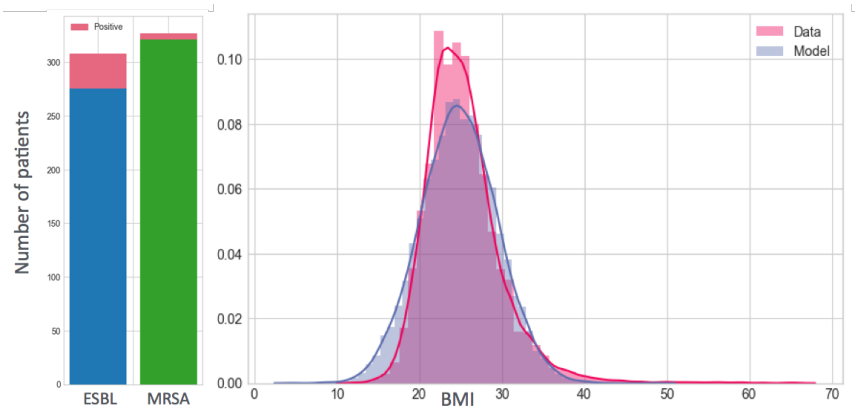
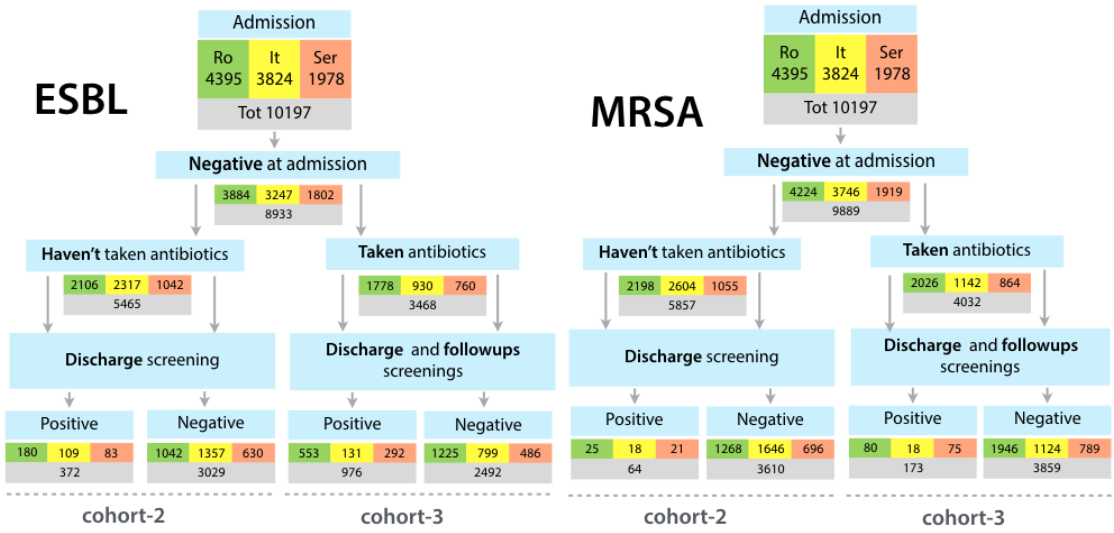


Figure 5.4: Number of patients with missing BMI separated by positivity/negativity for MRSA and ESBL. Distribution of the known BMI values along with the density plot for the probabilistic model used for BMI imputation.

Other missing feature values could not have been modeled. In the majority of cases, one of the admission or discharge dates were missing. These dates were especially important as they provided a frame for the hospitalization and therapy lengths, and antibiotic usage. Therefore, those records were removed from the analysis. 93 records were removed from the MRSA dataset and 85 from the ESBL dataset.

5.2.4 Patients and cohorts

The data had been collected for three years in the three hospitals located in three countries: Serbia, Romania, and Italy. In total 10,197 patients were recruited for the WP4 study. Among them 8,933 patients were ESBL-negative, and 9,889 were MRSA-negative at admission. Therefore, they were eligible for the study. Although this was a single dataset, it was treated as two parallel projects, with two unrelated outcomes: colonization with MRSA and ESBL. Roughly 40% of patients were treated with antibiotics and underwent follow-up MRSA/ESBL screenings. All of the patients were tested at hospital discharge (Fig. 5.5).



(a) ESBL (b) MRSA

Figure 5.5: Numbers of patients at admission and discharge in each of the three centers. Cohort-2 denotes patients who were not treated with antibiotics and cohort-3 denotes patients who were treated with antibiotics.

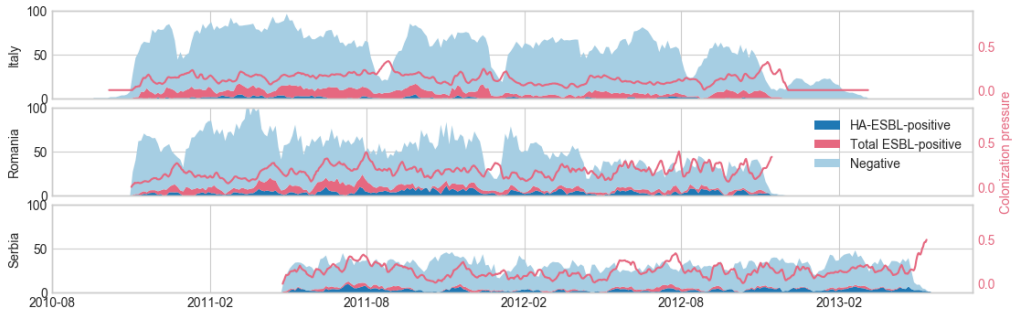
Patients who did not take antibiotics were tested only twice: at admission and discharge. The patients who took antibiotics could have been tested up to five times depending on how long they stayed in the hospital. Regarding all that, we considered three groups of patients: cohort-1: patients positive at the admission

screening, cohort-2: patients negative at admission, not taking antibiotics, cohort-3: patients negative at admission, taking antibiotics.

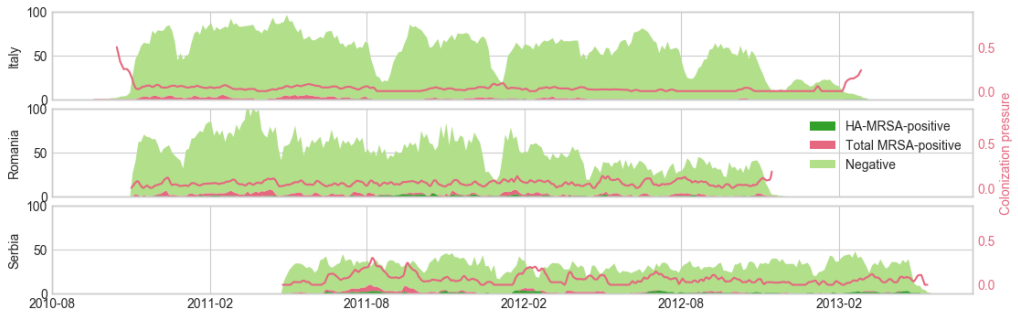
The machine learning analysis used only two groups of patients: cohort-2 consisting of patients taking antibiotics and undergoing a followup, and all patients who were negative at admission namely the cohorts-2 and -3 together. Therefore, four patient groups were analyzed: cohort-2 and cohorts-2 and -3 for both ESBL and MRSA positivity.

5.2.5 Features

Features were divided into three groups depending on the data type and real-world meaning. There were numerical demographic features, binary demographic features including comorbidities and features encoding antibiotic therapy. The distribution of the demographic characteristics and comorbidities did not differ between the cohorts, and overall the features did not show any anomalies (Fig. 5.7). Colonization rate, and as a consequence colonization pressure, was larger for ESBL than for MRSA. It fluctuated heavily across the sampling time (Fig. 5.6).



(a) ESBL



(b) MRSA

Figure 5.6: Number of patients, colonized patients and colonization pressure for both MRSA and ESBL and three countries, across three sampling years. HA stands for Hospital Acquired.

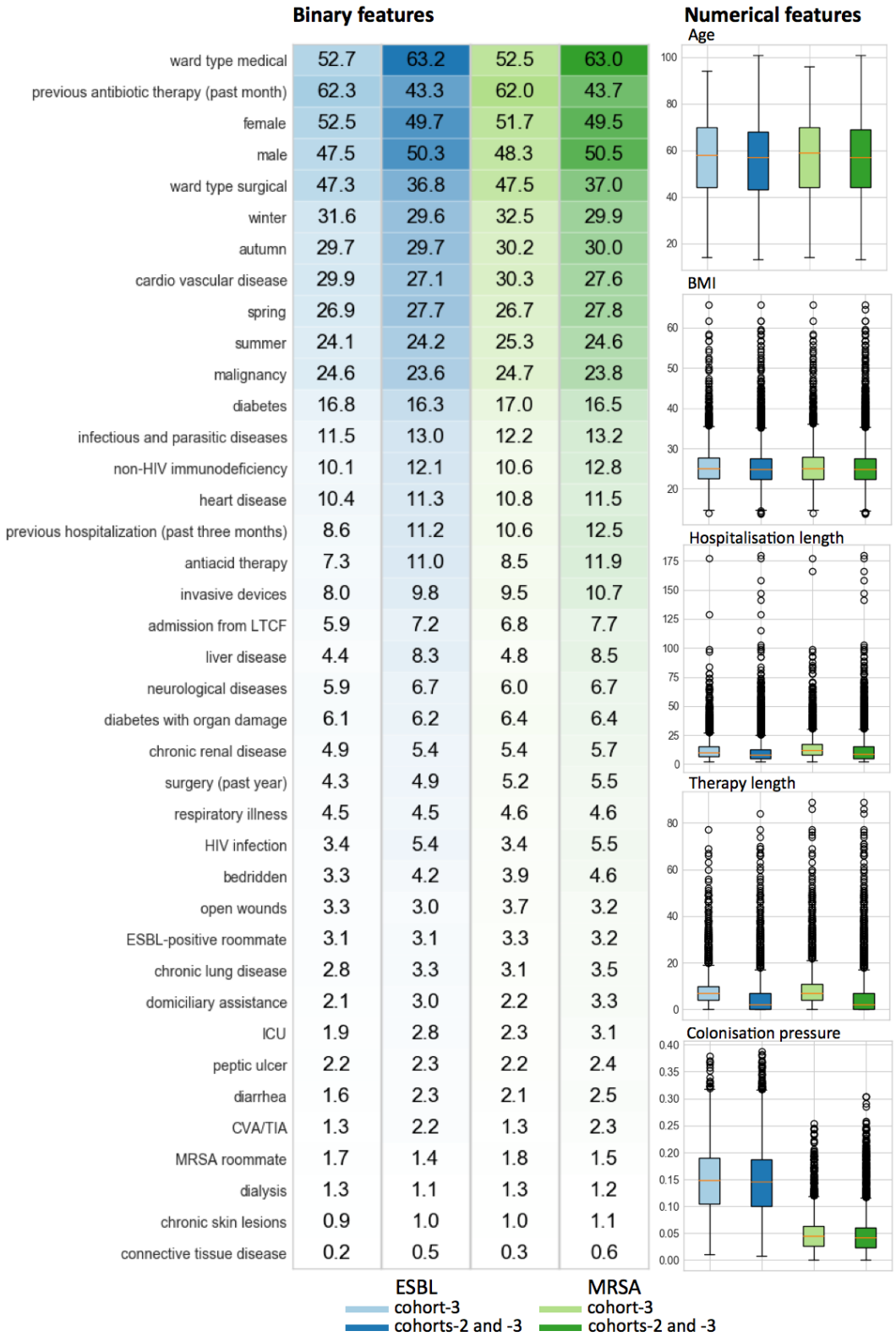


Figure 5.7: Distribution of features across all four patient groups.

The rate and structure of the colonization pressure differed between the hospitals. In Italy, ESBL colonization pressure could be mostly attributed to the positive patients admitted to the hospital rather than ESBL emergence inside the hospital. This seemed to be an opposite case to Romania, where the colonization pressure curve reflected the variation of the number of patient with the hospital-acquired ESBL (HA-ESBL).

On average patients were 56 years old, and were characterized by a healthy BMI $\sim$ 25. Average antibiotic therapy was nine days long, although characterized by a significant variance, as standard deviation was seven days. Patients stayed in the hospital on average for twelve days $\pm$ ten days, before they became positive, or got discharged. However, the average hospitalization length was three days longer in the antibiotic-taking patient group than for all of the patients.

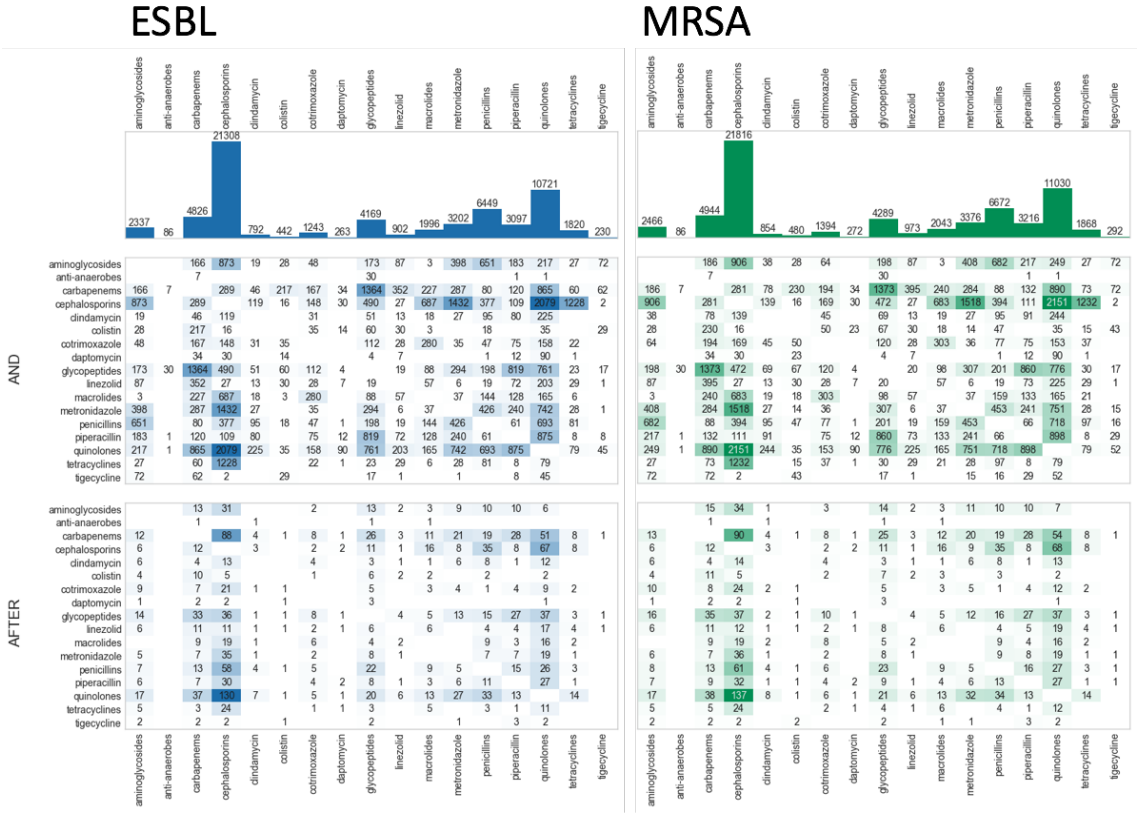


Figure 5.8: Antibiotics usage in cohorts-2 and 3. The top bar charts present the number of days each antibiotic was prescribed, as a single-therapy and in combinations. The middle heatmaps present the number of days the combinations of antibiotics were prescribed. The bottom heatmaps present number of instances of the transitions from one antibiotic (y-axis) to the other (x-axis).

As expected the antibiotic usage in the data was not uniformly distributed. Cephalosporins that constitute a large group of broad-spectrum antibiotics was by far the most commonly prescribed class of antibiotics (Fig. 5.8). Consequently, combination therapies with cephalosporins were also the most popular. The second most used were quinolones, and the next penicillins.

5.2.6 Undersampling

The SATURN dataset was unbalanced. In the ESBL cohorts-2 and -3 only 28%, and in the MRSA cohort just 4% of the patients were positive. Therefore, the dataset needed to be undersampled or weighted. The dataset was of high dimensionality, and also unbalanced regarding antibiotic therapies and comorbidities, i.e., not all feature values were equally represented. Therefore, I preferred to undersample the dataset rather than weight the samples.

I tested multiple domain-based approaches to undersampling, such as limiting the dataset only to those patients who took only antibiotics of a single class, or conversely, at least two antibiotics, to the group of patients with no antibiotic before admission. However, neither of this domain-based selection methods yielded a balanced dataset.

Therefore, undersampling of the majority class was used. The undersampled dataset comprised of the of the samples of the positive class and an equal number of the randomly chosen negative classes. The undersampling was parametrized by the size of the proportion of the minority class. Consequently, each time the undersampling was performed the dataset undergoing machine learning differs - mostly in the negative class.

5.3 Machine learning

The undersampled dataset constituted direct input into the ML. Training the ML classifier often follows the basic pipeline as presented in Fig. 5.9. In our case, the process of feature selection, the ML algorithm and parameter selection were repeated multiple times. Since, on the one hand, the ML methods were used for feature selection, and on the other hand, feature selection could influence the method and parameter selection.

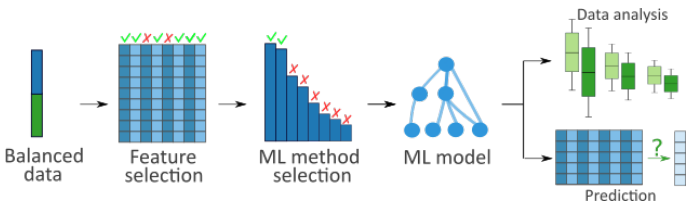


Figure 5.9: Pipeline for developing a ML model.

5.3.1 Algorithm selection

There are several extensive toolkits for application of various machine learning tools. One of the most extensive is Python's `sklearn`. Multiple steps were applied before the final classifier could be developed. The first step entailed finding the best-suited classifiers among those programmed in the `sklearn` package [250]. To avoid parameter bias, in this initial step I used a `parameter grid search`, which allows the specification of a list of parameters and their values so that the computer can test a classifier with all possible combinations of parameters under a five-fold cross-validation regime for the scaled and balanced datasets.

The selection of the ML algorithm relied on three measurements: accuracy, overfitting and permutation significance. Fig. 5.10 presents both preferences computed across all classifiers and their parameter selection. The objective was to select the classifier, dataset, and parameters maximizing the accuracy, minimizing overfitting and permutation significance p-value.

The majority of the models built for ESBL cohorts were successful and managed to generalize, as evidenced by low overfitting and permutation significance p-value. Further, the runs of ESBL cohorts-2 and -3 performed better than the ESBL cohort-3. The ensemble ML methods such as RF or Extremely Randomized RF performed better regarding accuracy than linear models such as Logistic Regression or SVM. Unfortunately, regarding overfitting, the ML methods ranking appeared worse. However, the ensemble tree-based method with the lowest average overfitting was RF. Random Forests and Neural Networks were the two algorithms picked for further analysis of the cohorts 2 and -3 dataset.

Models trained for the MRSA dataset performed quite bad. This was understandable as the proportion of the positive patients in the dataset was very low. Therefore, from this point onwards, I focused on the analysis of the ESBL dataset.

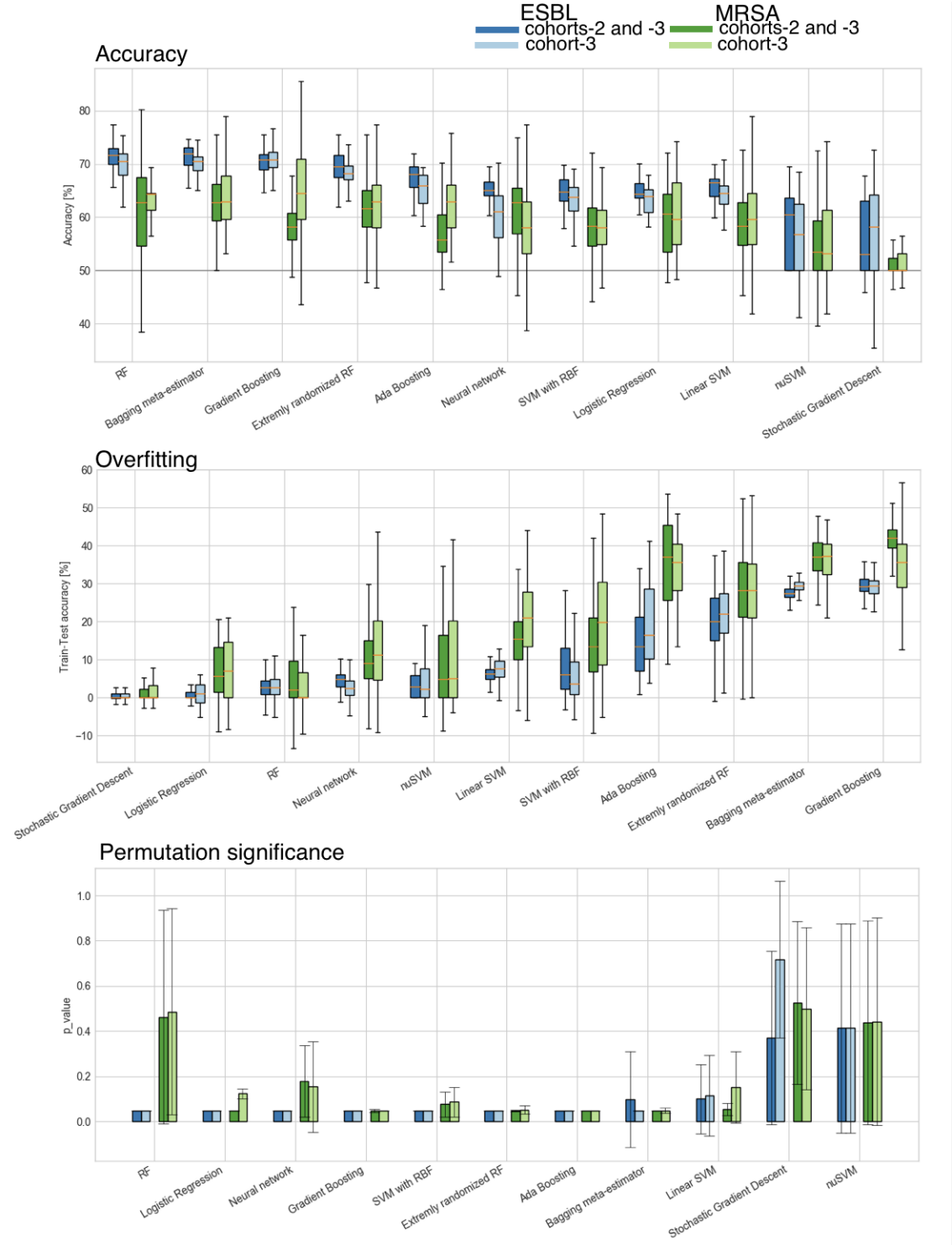
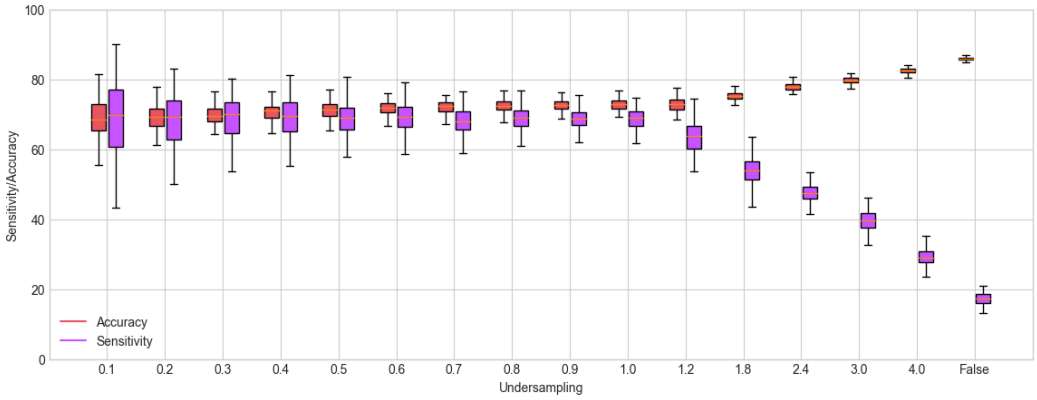


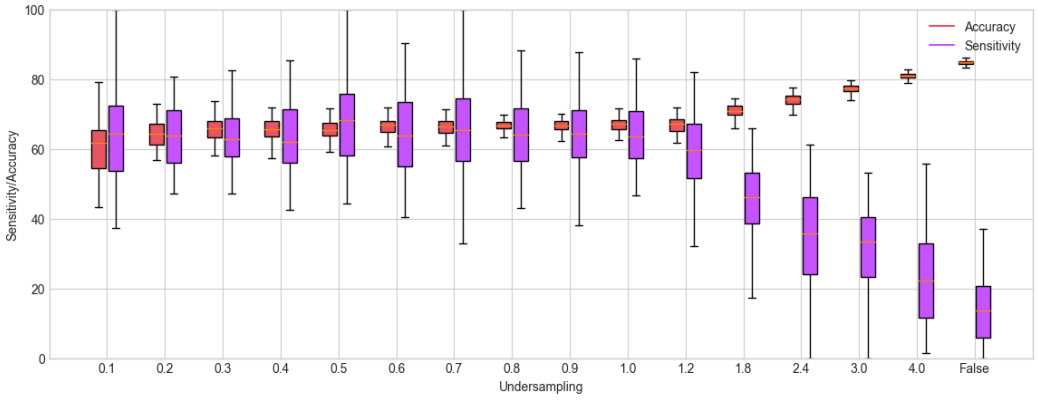
Figure 5.10: Performance of machine learning algorithms.

5.3.2 Undersampling parameter selection

An unbalanced dataset causes a decrease in sensitivity and an increase in the accuracy of the fitted models. Precisely this was observed for RF and NN runs of the ESBL-cohorts 2 and -3 (Fig. 5.11). However, the decrease in sensitivity was steeper and more substantial than the accompanying increase in accuracy. The decrease begins right when the negative class started outnumbering the positive class. The pattern was confirmed with Neural Networks (NN). NN achieved similar sensitivity with lower accuracy. The parameters for both algorithms were selected based on the previous step, i.e., were the best regarding accuracy and overfitting.



(a) RF



(b) NN

Figure 5.11: Accuracy and sensitivity of NN and RF classifiers relative the undersampling parameter. False denotes no undersampling.

Finally, undersampling at a level of 0.9 was used. Therefore, at this point, two decisions had been made. The ESBL-cohorts-2 and -3 dataset was used with 0.9 undersampling. The 0.9 undersampling also emphasized the need to repeat all tests several times so that all of the data points were used.

5.3.3 Feature selection

The feature selection objective was increasing the accuracy of the trained classifier by improving the ratio of the vector size to the number of data points. The first step of feature selection was to remove features with no positive values at all, which reduced the number of features from 469 features to 345. The next step was removing features with a deficient proportion of positive values, namely those that characterized $\leqslant 0.2\%$ of the overall patients. This step further reduced the number of features to 118. For such a dataset with reduced dimensionality, the machine learning algorithm and parameter selection was repeated (Fig. 5.12). The ranking of the ML algorithms had not changed for the reduced data set in either the accuracy or overfitting. However, it managed to increase the accuracy, especially for the MRSA dataset.

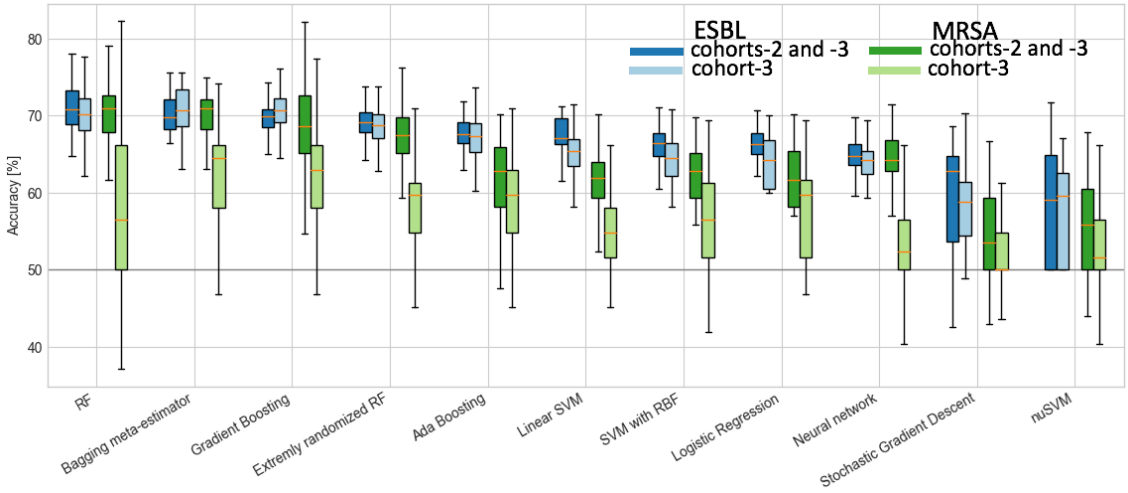


Figure 5.12: Accuracy for different ML tools and 118-feature dataset.

For such prepared dataset, more elaborate feature selection was further pursued. First, I measured the scores of the univariate feature importance. For the MRSA dataset, neither method could be computed. Therefore, it seems the MRSA dataset was weak and little information value with respect to MRSA-colonization. For the majority of the features, the ANOVA and Chi-square measurements did not agree with each other. Those methods were blind towards the complex multi-feature relationship with the classes. Therefore, I employed an ML-based approach to select features.

The classifiers were trained for each of the features separately (*just one*), the features for which accuracy was higher than 50% were regarded as important. Next, the classifier was trained for the increasing number of features starting from those with the highest accuracy in the previous step (*adding*). The more important the feature, the more it improves the accuracy. In the last approach, *singular*, the

classifiers were trained for all but one feature, here, the most important features were those whose removal caused the most significant drop in accuracy (Fig. 5.13).

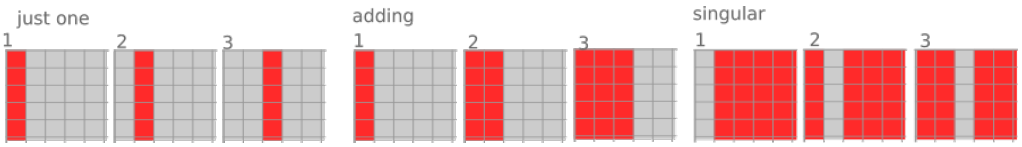


Figure 5.13: Feature selection strategies. Columns represent features and rows data points. Red denotes the features used and gray not used. Numbers denote the sequential steps.

All feature selection strategies showed the majority of the 118 features were uninformative, consequently the features could be further reduced (Fig. F.11). Fig. 5.14 presents comparative scores of the features of the three strategies.

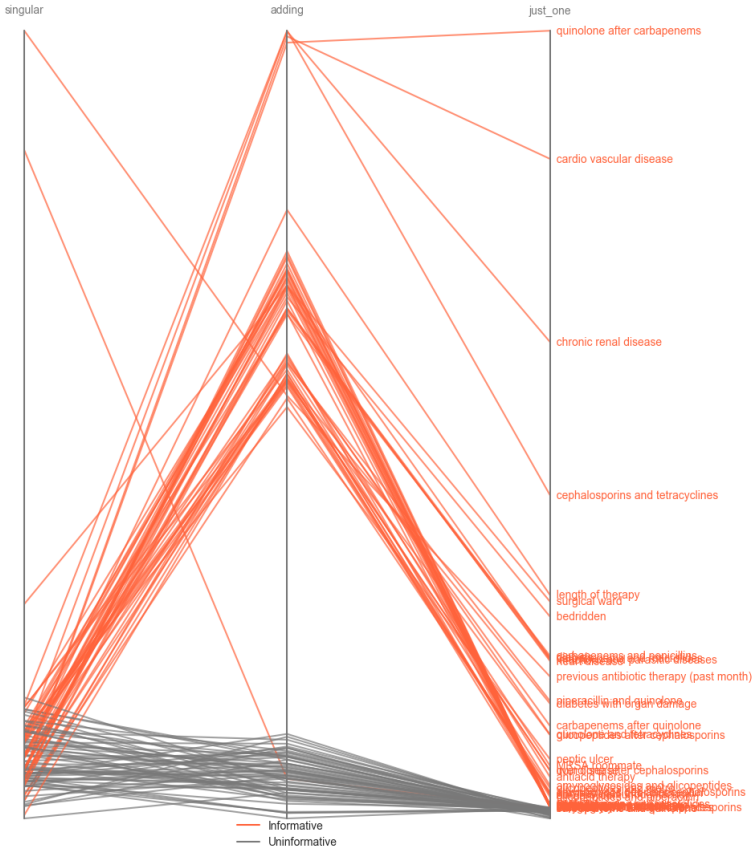


Figure 5.14: Parallel coordinates for feature scores across three feature selection strategies. The higher the position on the axes the more important the feature.

Unimportant features were those that scored low in all three of the measurements, conversely, the features that scored high in at least one of the feature selection strategies comprised a reduced dataset. Finally, 56 features were selected. Among the removed features were such intuitively important ones like the length of hospitalization, admission from LTCF and age. Those features had been selected in the previously attempted univariate approach.

Next, I rerun the parameter and algorithm selection for the 56-feature set and full cohort-2 and cohort-3 datasets. The ranking of methods remained unchanged, with RF coming in the first place. However, comparing to the parameter and algorithm selection for the 118-feature dataset (Fig. 5.12) the accuracy for different methods was better and had smaller variance. This suggests that the feature selection reduced the noise in the dataset. This reduction did not cause a reduction in performance, or even a slight improvement (Fig. 5.15). The smaller complexity of the dataset also improves the computational time.

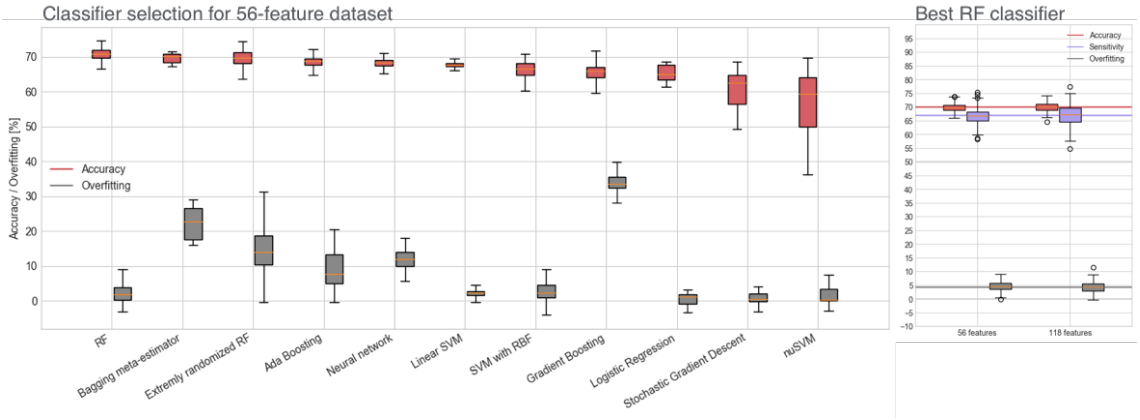


Figure 5.15: Classifier selection ranking for the dataset with 56 features, and the comparison between the best-parameter RF for 56-feature and the 118-feature dataset.

5.4 Results

5.4.1 Final pipeline

In the process of finding the optimal combination of ML algorithm and parameters, feature selection, undersampling parameters, the pipeline has changed (Fig. 5.16) in comparison to the one I assumed in the beginning (Fig. 5.9).

The pipeline started with the ML algorithm selection. This enabled to pick the undersampling parameter based on the best scoring ML algorithms. Then multiple different strategies for feature selection were attempted, starting from the variance-based, univariate and ML-based methods. Finally, the two-step strategy was used. The first step was reducing the features from 470 to 118 based on their

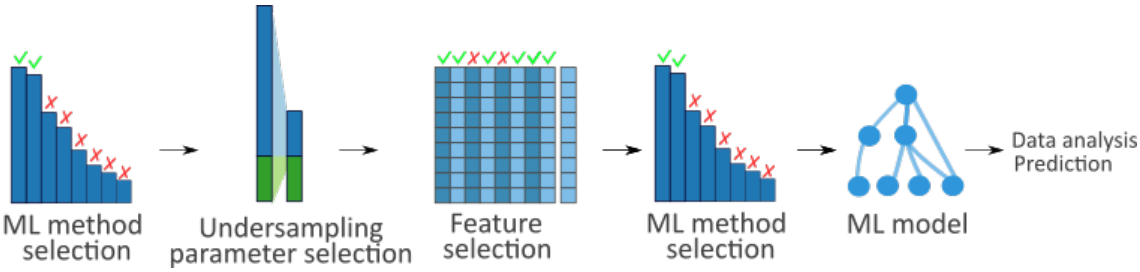


Figure 5.16: Final pipeline.

representation in the data, and the second step was the further reduction to 56 features based on the RF. Every time the ML algorithm selection step was repeated. The accuracy-based ranking of the algorithms did not change. However, the overall performance increased with the feature reduction. Finally, the permutation feature importance can be measured.

5.4.2 Features driving ESBL colonization

For the RF classifier and 56-feature dataset, the permutation feature importance was measured. Overall, non-antibiotic-therapy features proved to be more important for the ESBL colonization (Fig. 5.17) than those describing antibiotic therapy (Fig. 5.18). The most important features were the length of the antibiotic therapy, being treated in a surgical ward, and BMI. Next important feature was the antibiotic therapy before hospitalization. This could also be regarded as contributing to the overall length of antibiotic therapy and supports a conclusion that overall antibiotic usage drives ESBL-colonization.

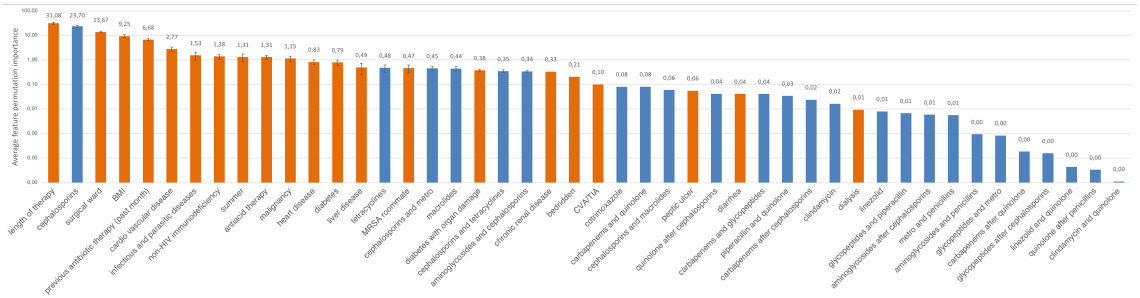


Figure 5.17: Average permutation feature importance for colonization with ESBL. Blue color indicates antibiotic-related features.

As presented in Fig. 5.8, antibiotic-therapy features were unevenly represented in the database. Therefore, the important features could be influenced by the amount of information in the dataset. The features were ranked from the most to the least common (Fig. 5.18), and from the most important to least important. On the one hand, combination and permutation therapies such as quinolone after

cephalosporins, carbapenems after cephalosporins had an increasing ranking. On the other hand, other features mostly connected to linezolid, carbapenems and linezolid, linezolid and quinolone had lower importance than the data ranking.

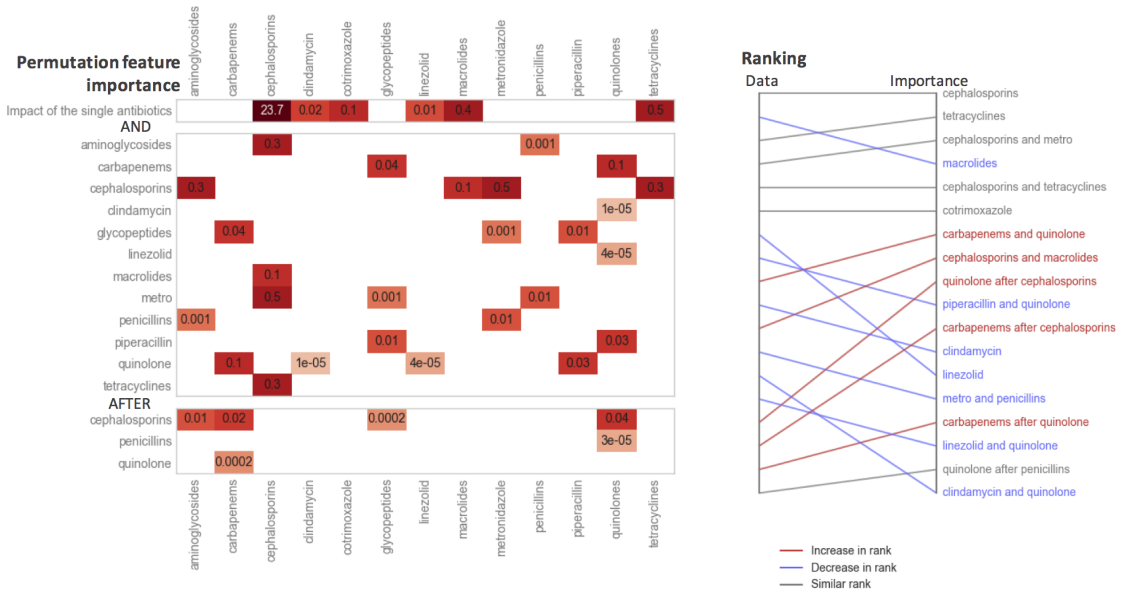


Figure 5.18: Average RF-based permutation importance for antibiotic-therapy features. Right-hand side parallel coordinates plot shows differences in rank between the data and feature importance.

In RF antibiotic therapy features (not shown), the feature ranking for therapy features corresponded to the antibiotic usage. This suggests the RF could have failed to learn from the diverse antibiotic usage.

5.4.3 AskSaturn website

The second aim of the project was enabling doctors to score antibiotic therapies for their patients regarding the probability of the colonization with ESBLs. I designed a website that first collects the demographic features of the patient, then comorbidities and antibiotic therapies, and predicts the probability of ESBL colonization for the two provided therapies. However, doctors do not have access to the full list of features, e.g., roommates positivity, therefore, the features had to be further reduced. For simplicity, I also removed antibiotics with low importance. The website classifier should be small to enable fast querying, therefore finally it had only 20 features. The minimal RF classifier performed with similar accuracy, however slightly worse regarding overfitting.

The website consists of a form gathering basic-patient information and two panels enabling encoding of the two alternative antibiotics therapies to be compared (Fig. 5.19). In the results, the probability of ESBL-colonization depending on the therapy

is shown. The user can quickly move between the panels and adjust the antibiotic therapies. The results are updated every time the compute button is hit.

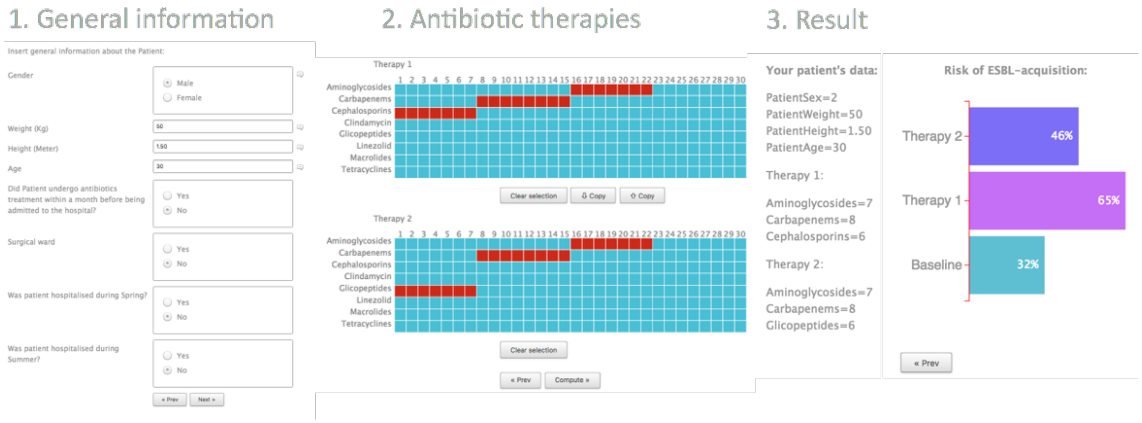


Figure 5.19: Panels from the AskSaturn website available under the address http://asksaturn.informatik.uni-tuebingen.de.

5.5 Summary and conclusions

The result showed the antibiotic therapy on its own drives colonization with ESBL-producing bacteria. However, the antibiotics usage could be stratified to minimize the probability of colonization with MDR bacteria as the probability differs depending on the antibiotic therapy. This relationship is quite complex and identifying clear guidelines is difficult. Therefore, the website enabling doctors to compare antibiotic therapies regarding the probability of ESBL-colonization is especially useful, as the ML-classifiers can handle complex therapies.

In this Chapter, I focused on the successful approaches. However, on each step of this project relied on choosing one of the multiple available approaches. Therefore, here, more than in the other projects, I have a feeling so much more could be done and tested, regarding feature encoding, algorithm/parameter selection and feature selection steps. The SATURN dataset was complex and sparse - namely there rarely were patients differing only in some of the features, and not all of them. However, this is extremely hard to achieve for this class of data, as the prescription of antibiotic therapy depends on the patient's demographics and comorbidities.

The solution could be separating the dataset to be able to answer the two questions separately: stratify by the patients' demographic and comorbidities to discover the space of the antibiotic therapies and conversely stratification by antibiotic therapy, to discover the impact of the comorbidities. This leads to the popular conclusion that more data is needed. In the same time, machine learning methods constitute a powerful toolkit enabling implicitly reasoning from multiple various features simultaneously.

Chapter 6

Tübiom project

6.1 Introduction

Antibiotic therapy causes an increase in abundance and diversity of the resistome in the patient's gut, and as a result, it impairs future antibiotic therapies. Therefore, the effect of an individual therapy on a particular patient's health also depends on the state of the person's microbiome. This dependency remains uncharacterized.

Characterization of the gut-health relationship requires a detailed analysis of the thousands of gut microbiome profiles. However, since the microbiome analysis is quite susceptible to the processing protocols and sequencing itself, those needed to be performed in a controlled way for all of the samples. Therefore, researchers undertook large-scale microbiome projects such as American Gut. However, they do not have access to *our* population.

Therefore, a broad collaboration of research groups and the CeMeT company located in Tübingen brought to life the Tübiom project. In the first phase, the goal was to collect 10,000 samples of fecal swabs and perform 16S rRNA sequencing to compute gut microbiome profiles [289]. For the second phase, the hope was to identify interesting phenotypic groups of participants to study further, with the deep whole-genome sequencing. Currently, the project is paused, with $\sim$ 3,500 samples collected. We encouraged the participants to submit multiple samples, especially if they underwent antibiotics therapies or went on a trip to a distant country. Therefore, Tübiom had the potential to provide a coherent dataset to measure the variability of the microbiome in the local population.

Our group was entrusted with setting up both the website and the backend. The primary role of the website was to show the results to the participant. The backend included the databases, for samples and patient information, and the bioinformatics pipeline for computing the taxonomic profiles from the 16S rRNA sequencing samples. Sina Beier wrote the analysis pipeline, the website and databases were set up by Patrick Groupp (MSc student), Theresa Harbig and I designed and programmed the data visualization, and Prof. Huson supervised the entire process.

Since a team performed the vast majority of work in this project, I will use a pronoun *we* throughout this chapter.

We presented the informatics and bioinformatics infrastructure created for the Tübiom project during the German Bioinformatics Conference 2016 [290]. In this chapter, I shortly described the Tübiom setup, the visualizations that we have developed and the first collected data, focusing on my responsibilities.

6.2 Tübiom setup

In the first place, a participant had to register on the website, where they had to fill in a questionnaire and order a sampling kit (Fig. 6.1). Next, a kit and use instructions were sent to their house. The participant could also receive a kit first, and then they would register it on the website later using the kit's identifying number. Once CeMeT received the sample back, the DNA was extracted, and the sample was sequenced.

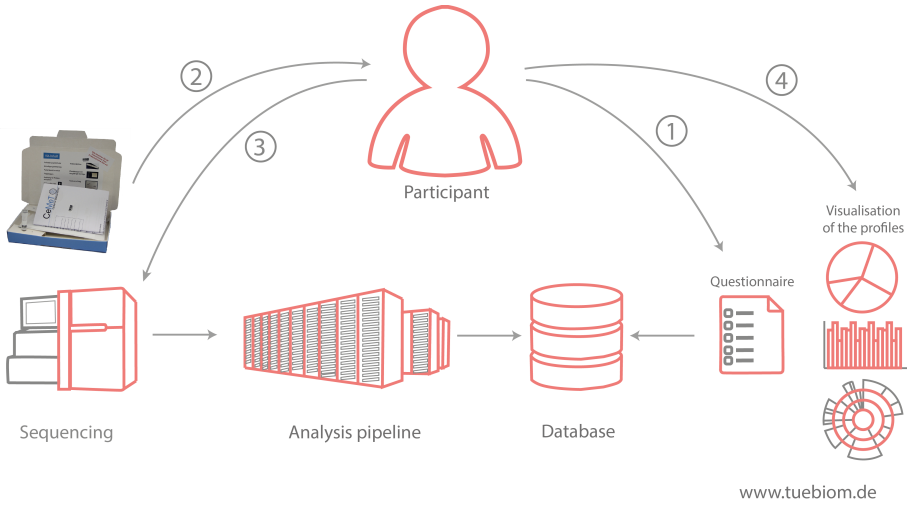


Figure 6.1: Setup of the Tübiom project.

The analysis pipeline, presented in Fig. 6.2 ran automatically after the files with the sequencing reads appeared in the designated folder. The computed taxonomic profile was stored in the profile table in the database. Once the sample was sequenced, taxonomic profile computed and the metadata inputted, the user could view and analyze the taxonomic profile of their samples on the website. The participant's online profile was connected to the samples identified by the kit's numbers. Each taxonomic profile had their metadata since the samples might differ according to circumstances, such as recent antibiotic usage or a trip to a distant country.

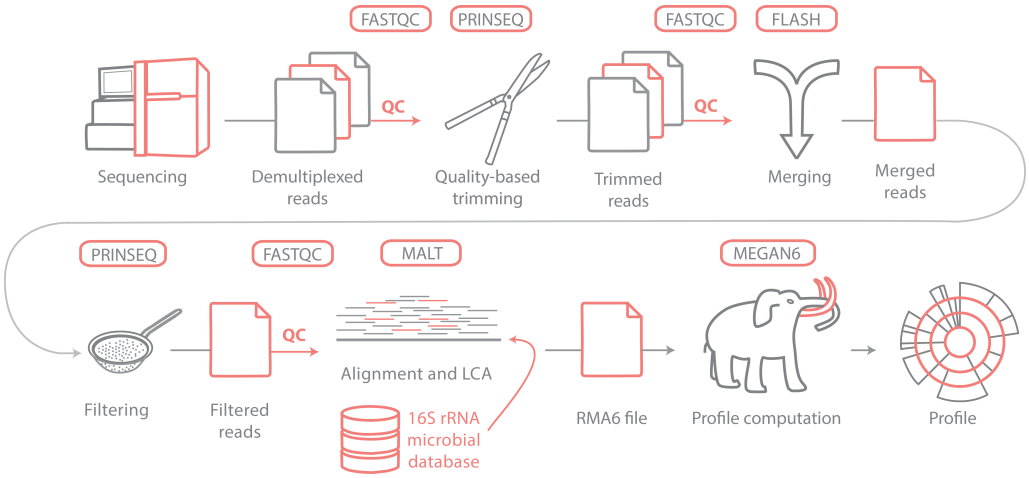


Figure 6.2: Pipeline for analysis of the Tübiom 16S rRNA sequencing data developed by Sina Beier.

6.3 Data visualization

Visualization of the microbiome profiles for this project was a different task, then when it serves the scientific publication. In scientific reports and publications, the audience consists of scientists, mostly of experts in the field. Also usually the publication plots are static figures presenting the author's message. In Tübiom the audience consisted of the general public interested in health and science. The visualization needed to be designed to serve the future data. Consequently, to provide context we assumed the Tübiom participants, as non-experts were mostly interested in a comparison of their samples to other profiles.

Therefore, we computed the overall average profile for the samples with normal BMI as the primary baseline. After there were enough samples in the database, the average profiles of the other large phenotypic groups such as vegetarians, young people, or those regularly exercising, were computed. The computation of the metadata-defined average profiles was fully automatized. The phenotypic group could be defined by any SQL expression using the metadata columns. We intended to update the list of the average profiles often, along with the growing number of samples in the database.

A taxonomic profile is on its own a complex dataset. In Tübiom the taxonomic profile consisted of the five taxonomic levels: phylum, class, order, family, and genus, since on the one hand, domain and kingdom levels are not informative for the human gut microbiome analysis, and on the other hand, the species level is too specific for the 16S rRNA analysis. With each taxonomic level, there are more taxa, starting with 8 phyla to $\sim 1,800$ genera. Consequently, the metagenomic datasets constitute challenging data to visualize.

In the vast majority, visualizations of the metagenomic datasets that include multiple taxonomic levels have two forms: taxonomic trees, like in MEGAN, or sunburst plots, popular in web-based visualization tools [291]. Out of them, only the phylogenetic tree supports visualization of multiple data series, and consequently, a comparison of the profiles. In such visualization, leaves of the taxonomic tree can contain bar charts with multiple series. However, we thought those were too complex and difficult to interpret to be used in this setting.

Consequently, we decided to rely mostly on the most straight-forward bar charts and stack charts for the main visualization (Fig. 6.4). Before the participant had to make any decisions in the panel (2), we presented a bar chart with the values of the three main phyla compared to the average profile for healthy participants in the panel (1) (Fig. 6.3). The bar chart was our overview. This way we implemented the first rule of the interactive visualization, the famous Shneiderman's mantra *zoom and filter, then details on demand* [292].

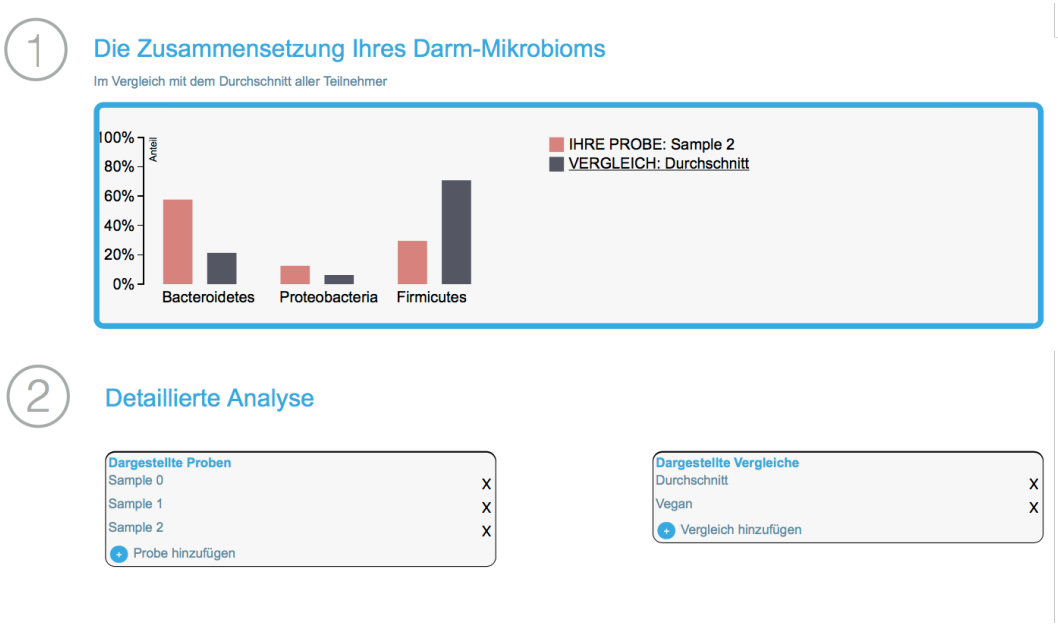


Figure 6.3: The main view of the participant's samples and selected comparison profiles.

The next panel contains the detailed plots for comparison of the selected profiles, on the level set before in the panel (3). Bar charts in the panel (4) enabled direct analysis of the proportions of taxa selected in the left and right panels. The color distinguished between the participant's samples and the phenotype group profiles. The drop-down menus contained all of the taxa of the level. Stack charts in the next panel (5) showed the comparison of the profiles on the selected taxonomic level. The final horizontal bar chart (6) presented the distance between the profile selected in the above drop-down menu and the rest of the samples.

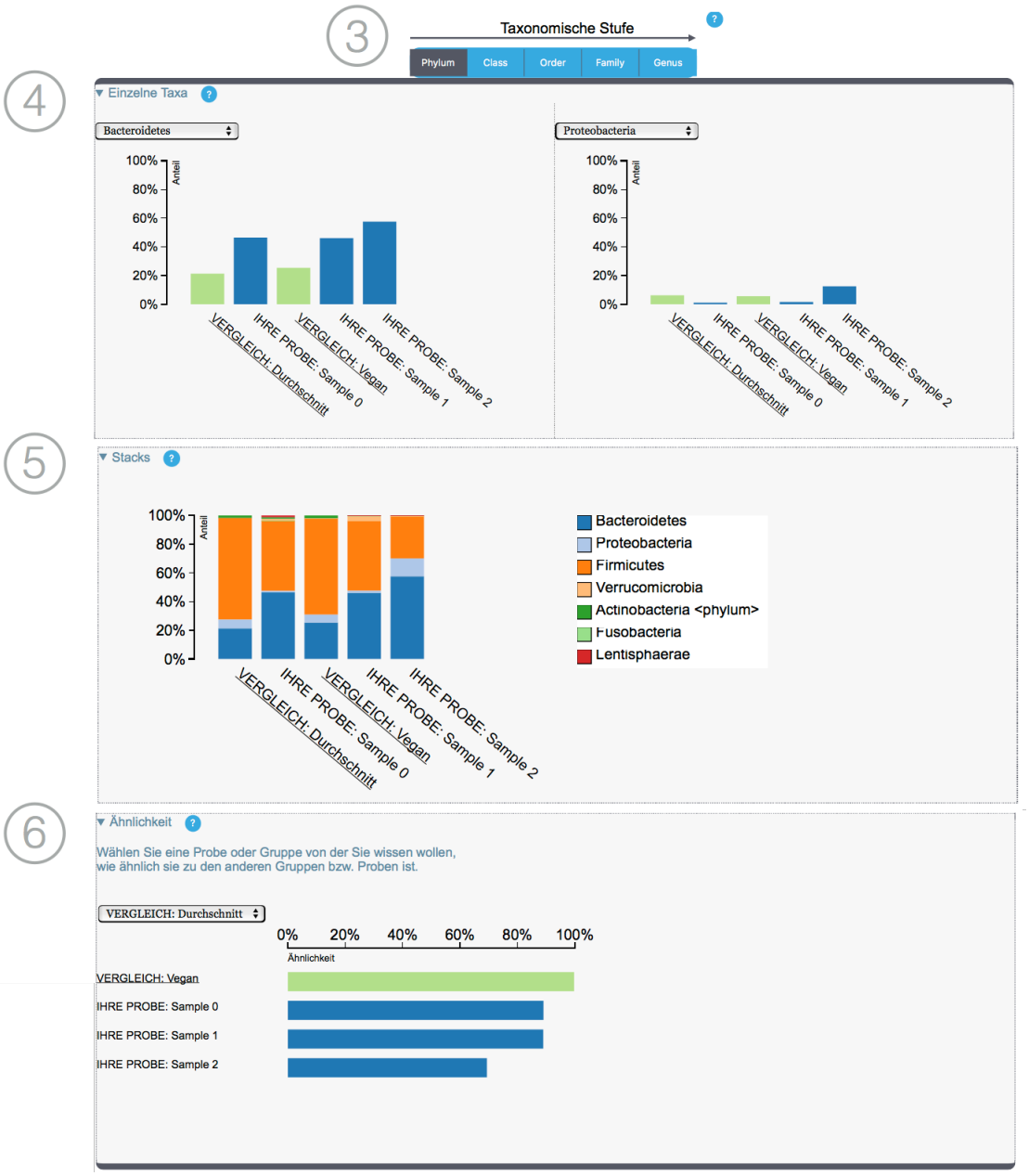


Figure 6.4: The main view of the participant's samples and selected comparison profiles.

The panels presented in Fig. G.12 were not included in the default view. They were only accessible under an advanced analysis button. Panel (7) showed the taxonomic profile again, with values encoded with color instead of bar height. In such heatmap, it is easier to analyze the low abundant taxa, and to read, when analyzing the low taxonomic levels, where there are much more taxa. So we thought

it would be suitable for users with more expertise. The final panel (8) included sunbursts. They presented all taxonomic levels at once. The colors corresponded to those in the stack chart in the panel (5). All of the plots provided hints - when hovered on different elements of the plots, the adequate values were displayed.

6.4 Metadata for collected samples

One of the motivations for Tübiom was gathering data for the local population that differs from the populations sampled by the American or British gut projects. However, as the project relied on volunteers, the sampled population was not fully representative. It turned out that women were more willing to take part in the Tübiom project, consequently, among the first $\sim 3,500$ participants 71% were female (Fig. 6.5). The distribution of the metadata suggested that the participants were both healthy and health-aware. The majority of participants had normal BMI, rarely consumed alcohol, did not smoke and regularly exercised. The age distribution was quite broad including a handful of infants and young children.

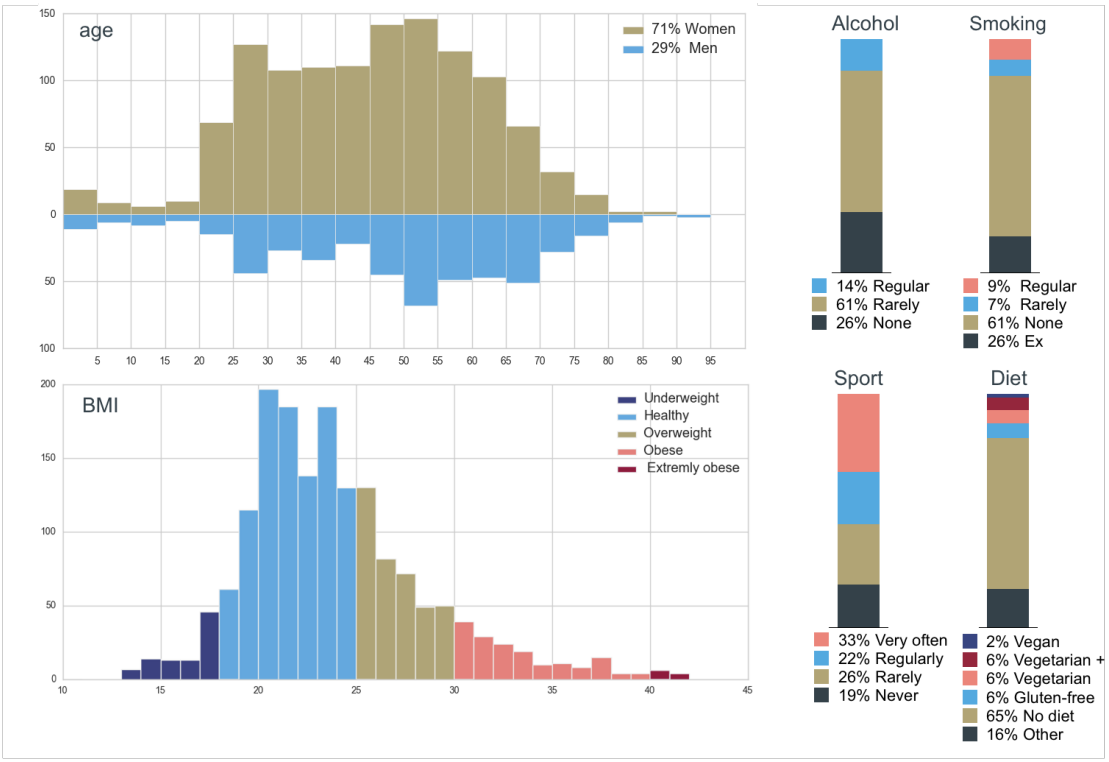


Figure 6.5: Distribution of the main metadata features for 3,491 samples.

6.5 Analysis of the preliminary data

So far we know that the relationship of the gut microbiome with the host's body is bidirectional and complicated. Accordingly, we did not expect the Tübiom data will right away reveal a strong correlation with any of the metadata features. Instead, we hoped that the dataset of such a large size and high quality would enable the discovery of those more elusive relationships. Our approach relied on stratifying the samples based on the metadata to reduce the complexity and training ML classifiers to discover the relationships between the remaining metadata and the taxonomic profiles.

Sample collecting and sequencing took longer than planned. Consequently, we were only able to perform a preliminary analysis of the full data points (metadata and taxonomic profile) for the first 1,252 samples. The metadata for each sample consisted of the 109 features. The microbiome taxonomic profiles were projected onto a genus level, and the values were normalized, in the end, the profiles consisted of the 769 taxa. The dimensionality, meaning the number of samples in relation to the number of features, of this small dataset was not sufficient for the ML.

However, it provides an overview of the diversity and quality of the data the Tübiom would provide. Fig. 6.6 presents the vectorized metadata of the initial dataset. Some metadata features were encoded with binary values, some were numerical and some categorical. The more heterogeneous the metadata, the more ways a data vector can be encoded, as some columns could be expanded, and others collapsed.

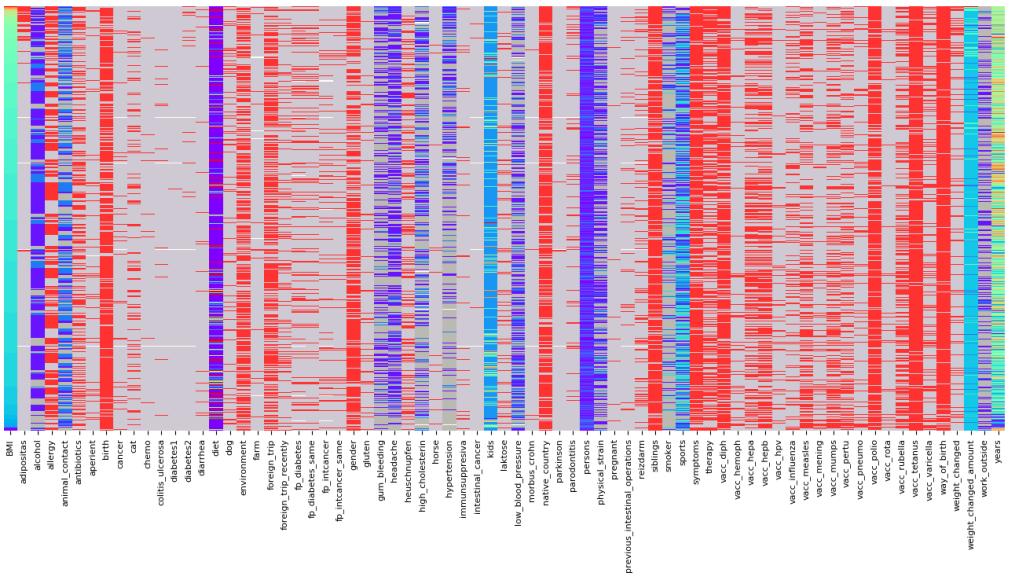


Figure 6.6: Metadata for preliminary set. Rows represent participants (data items), and columns the features. Red represents positive values (1.0) and grau negative (-1.0), white: value not known. In the color columns the values were

Even this small dataset was not entirely correct: some values were missing (white fields), and some samples had to be filtered out as they lacked too many values (not shown). On the one hand, it meant our set up for collecting data needed work, as it did not ensure full correctness. On the other hand, this also suggested we needed to expand the analysis pipeline to ensure missing data imputation.

However, the sequencing and bioinformatic analysis appeared to be working correctly. The taxonomic profiles were quite variable (Fig. 6.7), as some had high diversity (Fig. 6.8), and a single taxon dominated others, what reached even 82% like *Bacteroides*.

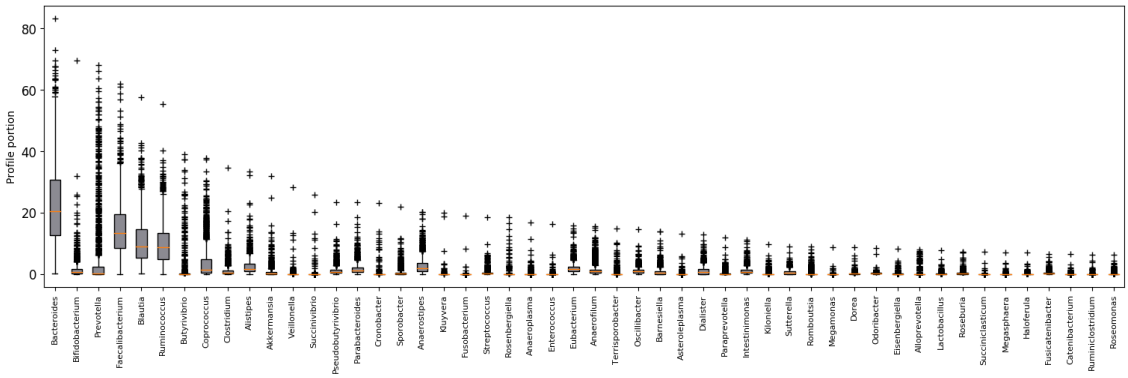


Figure 6.7: Distribution of the values of the taxonomic profiles for the first 50 most abundant genera.

6.5.1 Antibiotic therapies and antibiotic-affected microbiome profiles

I planned to analyze Tübiom dataset focusing on a comparison of the untreated profiles, to those affected by an antibiotic therapy. Among the first 1,252 samples, 413 participants had taken antibiotics within a month before collecting the sample. However, according to the dataset, the majority of the antibiotic therapies were only one day long (Fig. 6.8). Consequently, either the data were incorrect, a large group of participants sent a sample within the first day of the therapy, or the participants did not follow the orders of their doctors. However, all explanations appear to be unlikely.

Importantly, the taxonomic profiles for antibiotic-influenced samples had on average lower taxonomic diversity, especially in the case of ciprofloxacin usage (Fig. 6.8), which agreed with previous research. The variability between the values meant that the response to the antibiotic therapies varied, suggesting it dependent on the other features.

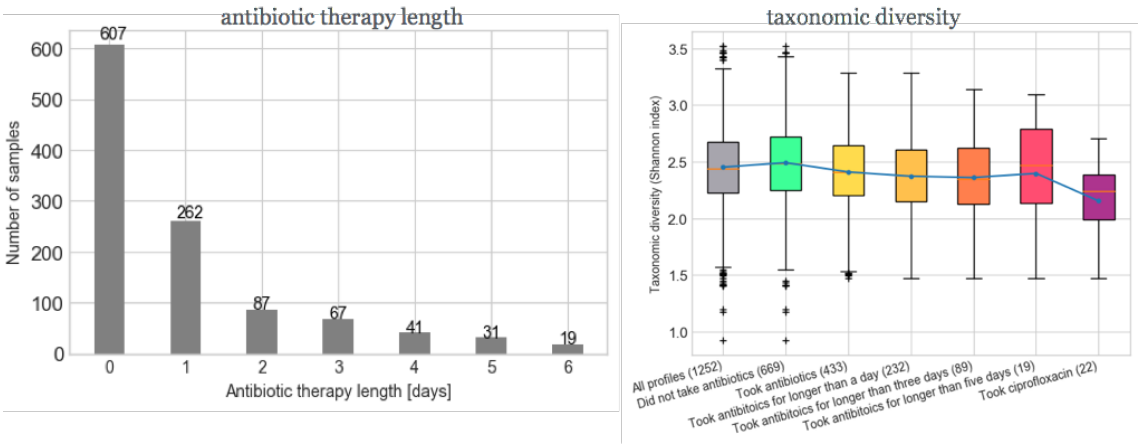


Figure 6.8: Distribution of the length of antibiotic therapy and taxonomic diversity depending on the antibiotic therapy. The numbers in the boxplot labels denote the number of samples used for the distribution.

6.6 Summary and conclusions

My contribution to the Tübiom project was data visualization and web development, neither of which falls strictly into the realm of bioinformatics. Nevertheless, the Tübiom project can be treated as a complement to the SATURN project, since it investigated the healthy population outside of the hospital. Although the Tübiom participants lead healthy lifestyles, a large proportion ($\sim$ 30%) of samples was affected with recent antibiotic therapy. However, their response varied depending on the length of therapy, antibiotic and between the participants. The preliminary analysis of the first 1,252 data points showed decreased Shannon diversity in the samples belonging to participants who underwent antibiotic therapies, especially to those who took ciprofloxacin. Hence, the full dataset would have been a great resource to study the influence of antibiotic therapies on the gut microbiomes, in relation to other meta-data features such as diet or lifestyle.

Chapter 7

Other projects

7.1 Analysis of the molecular dynamics trajectories of RNA

The bacterial ribosome or precisely its two main components, the two rRNA molecules are targets for eight antibiotic classes [20]. Previously, I worked on designing a sequence of a short anti-sense artificial nucleic acid to inhibit bacterial translation through binding to the 16S rRNA [293]. The design process relied on finding rRNA regions with the best score, which encapsulated features speaking to the functional importance and physical characteristics of the region. Accessibility of the rRNA regions was measured based on the behavior of an RNA in the molecular dynamics trajectory.

Molecular dynamics (MD) is a widely used method for investigating dynamical properties of biomolecules. It employs a purely macro-scale model [294], where each atom is a sphere of a certain size connected with other atoms by harmonic bonds and subjected to electrostatic forces. All of the energy terms are input into a Newtonian equation of motion with which positions of atoms are computed. As all *in silico* modeling methods, MD is criticized for over-simplification of both representation of molecules and laws of physics [295, 296, 297, 298]. However, it has great advantages. MD enables observation of molecular behavior in the atom-detailed scale. MD was successfully used to simulate riboswitches [299], protein-RNA complexes [300] and even the entire ribosome [301]. Although, when simulating RNA molecules especially delicate protocols need to be employed, MD is quite useful, as it enables observing formation and breaking of the secondary and tertiary contacts in RNA molecules [302]. To that end, I created tool MINT. It measures all types of nucleotide/nucleotide interactions in the RNA molecules during the MD trajectories [303].

During the Ph.D., I developed a novel approach for the analysis and visualization of RNA dynamics. Based on the output from MINT, a graph of interactions

is computed: the nodes represent nucleotides and edges correspond to the Watson-Crick, Hoogsteen, Sugar edge or stacking interactions. The edge weights indicate strength of the interaction expressed by a proportion of trajectory in which the interaction was observed. The graph encapsulates entire molecular dynamics into a distinct mathematical entity. The graphs enabled comparison of the dynamics of the similar structures, e.g., mutational variants. After the two analysis are performed, their contact graphs are deducted. The deducted contact graph represents differences in the interaction patterns.

Retrotransposons are transposon elements that undergo transcription. The tertiary structure of the transcribed RNA molecule is crucial for retrotransposition. Collaborators from the Max Planck Institute for Developmental Biology dr. Oliver Weichenrieder and Steffen Schmidt who investigate retrotransposing RNAs, suggested that I perform an MD-based analysis of the structures of the two retrotransposition elements. Namely, the small but crucial fragment on the 3' tail of the eel's UnaL2 LINE element (PDB: 2FDT) and the human Alu SRP14/16 complex (PDB: 5AOX). For both elements, the retrotransposition frequency (RFr) of several mutation variants were measured. I investigated the impact those mutations have on the structure of the RNA molecule.

The native and each of the mutated structures underwent the same extremely cautious MD protocol. First, the molecules are protonated and solvated with a water box. Next, ions are added (Mg^{2+}, Na^{+}, Cl^{-}). All components of the solvent are minimized. The solution is fixed, and the minimization steps are performed: water minimization, water and ions minimization. Next, the system is thermalized - with a fixed solution the temperature is being gradually raised. Equilibration follows the thermalization step. The harmonic energy constraints were put on the solute to prevent it from moving too quickly, and being destroyed at the beginning of the simulation. Initially, the structures are almost completely restrained, next the constraints are slowly loosened to the moment the structure is equilibrated and the unconstrained simulation can be performed. Finally, for the production trajectories, the MINT analysis, and the networks are computed. Computations were performed using NAMD program [294] mostly on the Paderborn computational cluster OCuLUS.

7.1.1 Analysis of the structure of 3' end of the UnaL2 LINE element

The first structure was a 36-nucleotide fragment of the 3' tail of the eel's Long Interspersed Nuclear Element (LINE) [304]. It consists of two helices, one bulge and a hairpin loop (Fig. 7.1). This fragment is crucial for retrotransposition. The researchers mutated the cysteine located in the middle bulge. Replacing it with Adenine (mutation: C8A) resulted in the increase of the RFr to a 120%, and mutating it to Guanine (C8G) resulted in RFr decrease to 4%. The MD trajectories for native

and mutated structures were performed, the 210 ns production trajectories were analyzed with MINT, and the deducted contact graphs were computed.

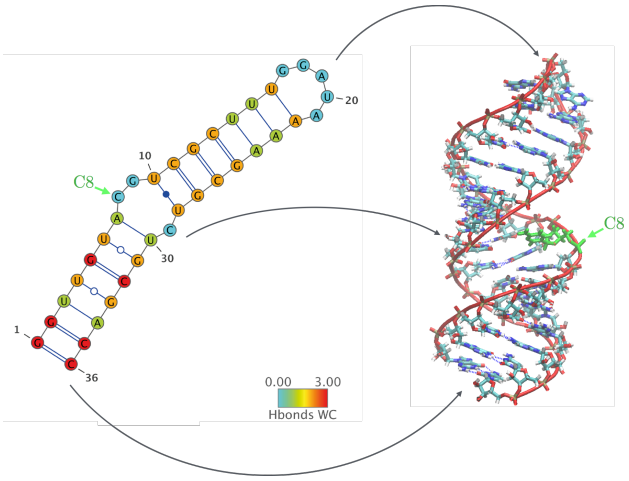


Figure 7.1: Secondary and tertiary structures of the native 3' end of the UnaL2 LINE element RNA molecule.

The deducted contact graphs revealed that mutations of the nucleotide on the 8^{th} position caused changes in the interactions in the lower part of the molecule (Fig. 7.2). Both mutations caused loss of the tertiary hydrogen bonds between the 8^{th} nucleotide and the C26 and G27 located on the other side of the helix, and the stacking interaction with the U28. However, C8A mutation enabled binding of the G9 to the U30, and the C8G mutation promoted hydrogen binding between C8 and G25. Changes introduced by C8G mutation caused breakage of the WC-pairing at the beginning of the helix.

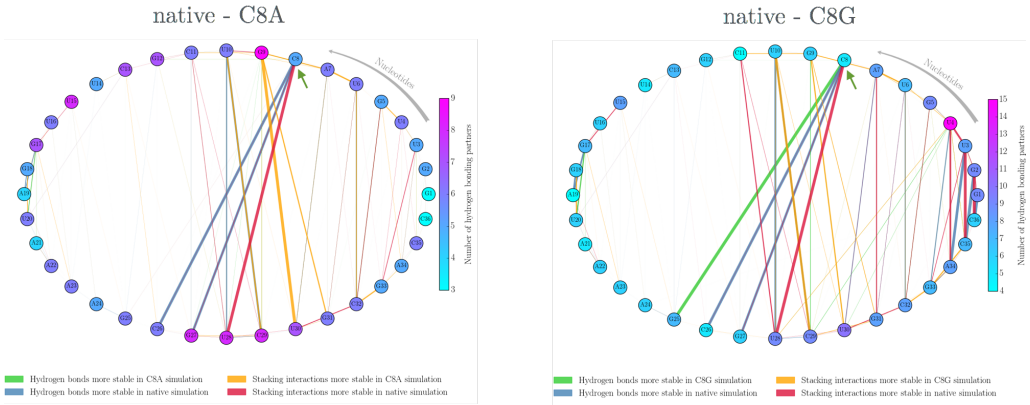


Figure 7.2: Deducted contact graphs for native structure and the two mutated molecules.

A possible explanation is that the strong hydrogen binding in the middle of the molecule between the G8 and G25 nucleotides impeds the reverse transcriptase unwind the molecule during retrotransposition, which results in the low RFr of the molecule with the C8G mutation. This agrees with the results for the analysis of C8A-mutated structure, where the hydrogen bonds were removed, and replaced by a weaker stacking interaction, which resulted in an RFr increase up to 120%.

7.1.2 Analysis of the structure of Alu SRP9/14 complex

In the cytoplasm, Alu RNA forms a complex with two proteins SRP 9 and 14. All of those elements are crucial for retrotransposition. The G25C mutation in the Alu RNA causes $\sim 50\%$ decrease in RFr [305]. I compared simulations of the native and mutated structures of the complex. The mutated nucleotide on the 25^{th} position was a center of the interaction between the RNA and the protein. The 25^{th} nucleotide is located almost directly opposite from the pseudo-knot, which was proven to be crucial for retrotransposition (Fig. 7.3).

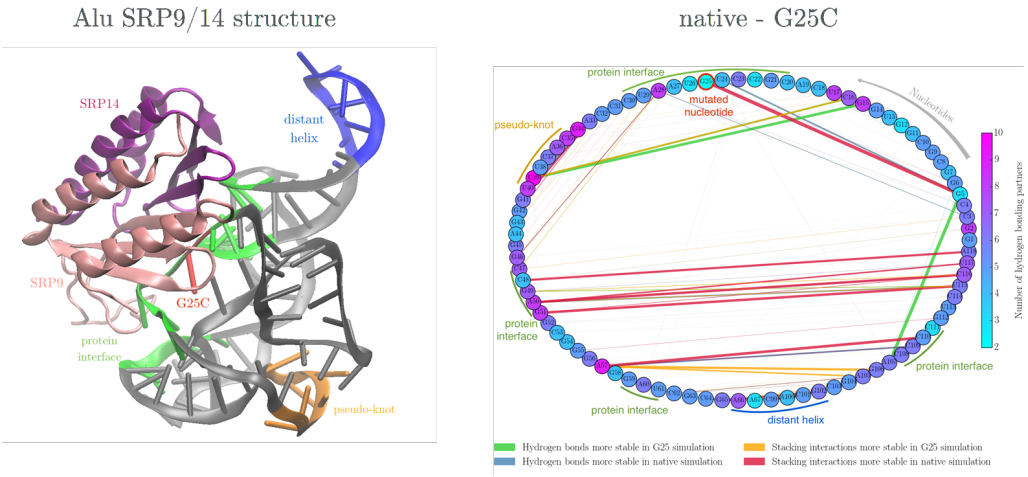


Figure 7.3: Structure of the complex and the contact difference graph (native-G35C).

The G25C mutation caused loss of the strong binding between G25 and G5. However, the G5 formed a strong hydrogen bond with the A107. The mutation also caused loosening of the stacking interactions on the site of the protein interface. Interestingly another strong hydrogen bond appeared between the U39 and G15. The U39 is typically a part of the pseudo-knot that is crucial to the biological function of Alu. This could be a cause for the decrease in the retrotransposition frequency.

7.1.3 Summary and conclusions

I established a reliable molecular dynamics protocol to perform simulations of RNA molecules and RNA/protein complexes. The novel approach to the analysis of the trajectories provided explanations of the mutation-caused changes of the structure. In the first structure, most of the differences were located around the mutated nucleotide or its interacting partner located on the other side of the helix. However, in the second, more complex structure, the analysis revealed a whole sequence of changes. The mutation of the nucleotide located in the center of interaction with the two proteins resulted in the destruction of the pseudo-knot on the opposite side of the structure. In the sequence-sense that is a long distance interaction, however, it was revealed by my MD/MINT analysis.

In the future, the molecular dynamics of the native structure could enable a precise prediction of the structural changes, and biological function for the various mutated sequences. This will enable clustering of retrotransposon sequences, e.g., those found in the human genome, into active and inactive.

7.2 Phase the turtle!

7.2.1 Introduction

Cells of diploid organisms have two copies of each chromosome. Although they carry the same genes, their sequence may vary on the nucleotide level. During the crossing-over, those small differences are mixed. Sequence fragments are called allele, and a collection of alleles inherited together is a haplotype. Each chromosome has multiple haplotype blocks. A probability of the two genes to be in the separate haplotypes grows proportionally to the distance separating their loci. Therefore the probability of finding two alleles in the same individual differs depending on their location. This phenomenon is called linkage disequilibrium.

The majority of the genome sequences of diploid organisms deposited in databases are consensuses of the two haplotypes [306]. Knowledge about haplotypes and their frequency in the population is especially crucial for medicine [307, 308], as the single nucleotide polymorphisms (SNPs) can confer a diseased phenotype. Separating the haplotype sequences, or *phasing* uncovers the exact sequence of the genome.

Computational approaches to haplotype phasing fall into two groups depending on the data. Either, they consist of the large numbers of genomes of unrelated individuals, or of the genomes of the members of a single family. In the first variant, all of the methods are based on statistical modeling of haplotype frequencies within a given population [309]. In the second variant, the parents and offspring genomes, enable direct determination of the haplotype blocks. Those methods are developed for human genomes and are not always applicable to other organisms.

This project aimed at obtaining a full, phased genome of a *Diamondback terrapin*. The University of Maryland researchers formed a collaboration between the Center for Bioinformatics and Computational Biology (CBCB) and the local Biology department. They sequenced the genome of a female turtle and fourteen of her progeny. Complicated reproductive biology of the turtle made it impossible to determine the father of the progeny. Therefore, this case falls into neither existing haplotype-phasing approaches and requires developing a novel approach (Fig. 7.4).

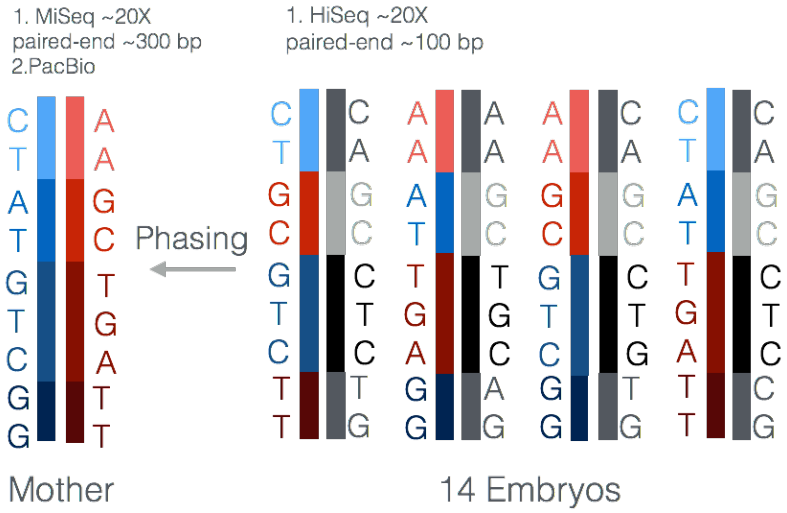


Figure 7.4: Project scheme. The colorful blocks represent haplotypes coming from the mother, and the gray colors those coming from father, whose sequence we did not know.

The turtle's genome was sequenced using two technologies: PacBio generating long scaffolds but characterized with relatively high error rate and Illumina MiSeq producing short, high-quality paired-end reads 300bp long reads, with 20x coverage. The genomes of the 14 embryos were sequenced with HiSeq Illumina technology, generating 100bp long paired-end reads, with 20x coverage.

Prof. Pop's group performed initial steps of analysis. The k-mers correlation analysis revealed that among the 14 progeny, groups of 3 and 11 embryos shared a father. My role in the project was to phase the haplotypes for the turtle's genome. The work started during the 2^{nd} BEST summer school and continued during my internship in the Prof. Pop's lab at the University of Maryland. Since all of the analysis was done in close collaboration with Victoria Cepeda and under the supervision of the professors from CBCB, I used pronoun *we* throughout this section.

7.2.2 Phasing pipeline

We assumed each PacBio read belonged to a single LD block, but its sequence was a consensus of two haplotypes. MiSeq reads are homozygous since they come from one of the DNA strands. The short reads were first mapped with Bowtie2 [230] against the PacBio scaffolds. Next, Pilon [310] was used to correct the PacBio sequencing errors. Our goal was sorting MiSeq reads, into two haplotypes, in the context of a single PacBio read at the time (Fig. 7.5).

Figure 7.5: Reads mapped to the PacBio scaffolds, were sorted into two sets for each haplotype. Next, for each side separately Pilon introduced the haplotype changes.

In order to separate the mapped reads according to the haplotype, first, we counted k-mers (size 19) in all of the short reads using jellyfish and jellyfish-matrix [259, 311]. A k-mer corresponded to a single allele. Next, each mother's k-mer was assigned a binary segregation profile encoding its presence in the progeny sequencing data (Fig. 7.6). An ideally homozygous profile consisted of solely 1'ns, and an ideally heterozygous profile was encoded by a complementary pair, e.g.: `10110011001` and `01001100110`.

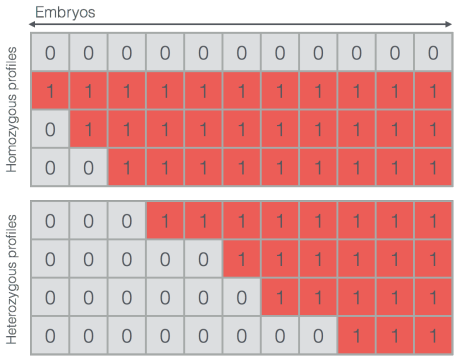


Figure 7.6: Each k-mer extracted from the mothers sequencing, was assigned a binary profile vector based on the presence or absence in the embryos.

At this point, each scaffold was linked to a set of reads, and each read to a number of k-mers, each with a segregation profile. Most of the aligned reads had a homozygous profile since most of the sequence in both haplotypes was identical.

If the scaffold was heterozygous, the profiles of the mapped reads should include complementary profile pairs (Fig. 7.7). However, the profiles were not ideal due to the sequencing errors, k-mer frequency sampling and the fact that a k-mer could be present in the father's genome. If the k-mer came from the homozygous locus in the father, their profile was also homozygous, independently whether the locus was heterozygous in the mother's genome. Therefore, the `0` in the profile denoting an absence of a k-mer in an embryo was more informative than its presence. The profiles of the reads mapped to a single scaffold were clustered. Subsequently, the clusters were used to sort reads into the haplotypes.

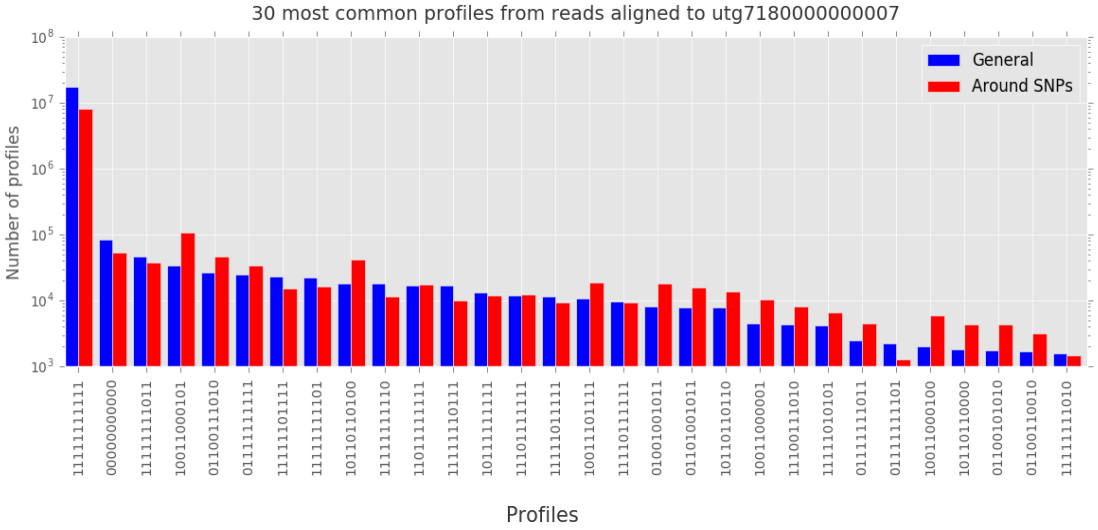


Figure 7.7: Histogram of profiles found in reads aligned to one of the scaffolds. As expected the most abundant are homozygous and empty profiles that emerged from the k-mer frequency cutoff. The third most frequent profile is also erroneous since it is doubtful that an allele omits a single progeny. On the fourth and fifth positions, there are reciprocal profiles responsible indicating the two haplotypes. There is little difference in the distribution of profiles between those located close to the SNPs and others.

For each scaffold, four clusters were computed: homozygous and empty profiles, A, B haplotype profiles and erroneous profiles. The profiles with almost only zeros or ones were uninformative. The first cluster contained the most abundant informative profile (A) and all of the profiles which share at least one zero with its representative. Analogously cluster B contained the most abundant informative profile that did not share zeros with the cluster A representative. The clustering algorithm was inspired by the CD-Hit [287] algorithm (Fig. H.13).

Using the profile clusters, we separated the reads aligned to a single scaffold. In the case, their distribution was dominated by one of the clusters the read was

assigned to the adequate group. Otherwise, if both A and B clusters were equally represented the read was assigned to the mixed category, which might be used later to resolve the rare situations when the LD break was located within a scaffold. The last step of the pipeline (Fig. 7.8) was running Pilon for the scaffolds with heterogenous sites and the two sets of reads.

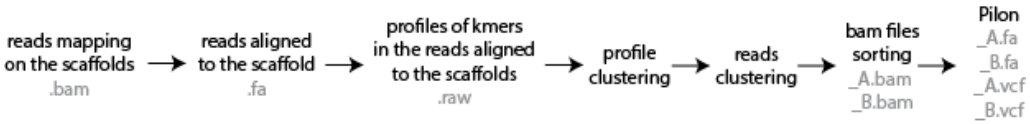


Figure 7.8: Steps of the pipeline. All of the clustering, sorting and file manipulations were performed with Python.

7.2.3 Validation

The haplotype separation of the reads was validated by the mapping of the embryos' reads onto the phased scaffolds. The representative profiles enoded the segregation pattern. Therefore, we should be able to find reads solely of the embryos indicated by an adequate representative profile (Fig. 7.9).

emb	3	4	5	7	8	9	10	11	12	13	15
A	1	0	0	1	1	1	0	1	0	1	1
10011101011	4 %	0	0	10 %	9 %	8 %	0	7 %	0	1 %	9 %
B	1	1	1	1	0	0	1	0	1	1	1
11110010111	10 %	6 %	6 %	4 %	0	0	5 %	0	5 %	0	5 %

Figure 7.9: Comparison the number of aligned reads per embryo to the representative profiles for an exemplary scaffold.

7.2.4 Summary and conclusions

We managed to solve the phasing issue in an atypical setup, with the unique, k-mer based approach. The entire pipeline was set up on the University of Maryland computational cluster. It computed for the majority of the 21,684 PacBio scaffolds. However, it requires more work. In some cases, we ran into computational problems, as the files became extremely large. Also, each of the steps in the pipeline could be parametrized. Especially we should check if the k-mer frequency cutoffs influence the results. Next steps should be rerunning pipeline, including validation steps to find out the most successful set of parameters.

Chapter 8

Discussion and outlook

Antibiotics are primarily natural substances used by microorganisms in their struggle for ecological space. We weaponized them to fight bacterial infections. It came with great success but also with a high toll. Scientists observed emergence of the dangerous multi-drug resistant bacteria and rise of resistance levels, so high that it threatens a continuous safe use of antibiotics. Antibiotics were introduced into the therapy before their impact was fully understood. Nowadays, the sequencing, bioinformatics tools, and data mining methods enable insight into the mechanisms of resistance emergence on all of the levels of biological organization.

The resistance starts in the bacterial cell. First, I investigated how the genomes of Methicillin-resistant *Staphylococcus Aureus* (MRSA) change throughout antibiotic therapy. The observed genetic differences between the MRSA strains concerned mostly the virulence factors, whose function is fighting the environment that is the human body, often, in the presence of high antibiotic concentrations. It was not possible to directly prove that the antibiotic was driving the micro-evolution of those strains as the isolates came from different states of infection and there were too few patients, with too complicated and individual therapies in the dataset. However, the observed variability between the isolates could in a large portion be attributed to mobile genetic elements. MRSAs are multi-drug resistant bacteria packed with virulence and resistance factors, and even more with transposons, phages, and plasmids. This emphasizes the crucial role of horizontal gene transfer (HGT) in resistance emergence.

The second project focused on resistance emergence mediated by HGT. I investigated the dynamics of the gut microbiome under the antibiotic-excreted ecological pressure. I observed the expected shifts in the microbiome composition as well as an increase of the phage integration events. The resistance appeared firstly in the antibiotic's direct pressure point, the bacterial chromosomes, and secondly, in the mobile genetic elements: the plasmids, prophages, and lastly in the free phages.

The genetic information travels through the microbiome as if it was a network with two types of nodes: bacteria and phages, connected by HGT events. It seems the phages work as a temporary reservoir, like a retention pool on the river, storing genetic information, as they do not respond to the ecological pressure of an antibiotic. Antibiotic therapies remove some of those nodes, leaving ecological space for the resistant bacteria. The connections between bacteria are strengthened and thickened as HGT is prompted. Such an analysis was possible only for the dataset of the deep sequencing enabling metagenomic assembly for time-series samples. The network of each person's gut microbiome is different, and so is its response to antibiotic therapy. This was also shown by the analysis of the preliminary data from the Tübiom project. The gut microbiomes' responses vary depending on the lifestyle, past treatments, and the therapy itself.

During the third project, I investigated the impact of different antibiotic therapies on the probability of colonization with MDRs. The data, coming from this observational study, were quite large, diverse and unbalanced regarding antibiotic usage and colonization. We showed that MRSAs pose less threat than other MDR types. The length of the antibiotic therapy drove colonization with extended spectrum beta-lactamase-producing *Enteriobacteriaccae* (ESBL). However, using an extensive machine learning pipeline, we were able to observe that the impact of some antibiotics or their combinations was lower than others. To help doctors navigate those complicated dependencies, we launched the `AskSaturn` website, that compares the antibiotic therapies with respect to the predicted probability of getting colonized by MDR.

Choosing a therapy based on the patient's features, treatment history and the current state of health is a definition of *personalized medicine* or *medicine P4* [312]. The hope is that in the future the medical practice will be fully personalized with information about the patient's genome and full medical history available to their doctors. Currently, the idea of personalized medicine is being implemented by the identification of the clusters of patients for whom the therapy outcomes are similar. This boils down to the analysis of the large and diverse datasets of patients. From a data perspective, each patient and therapy denotes a long data vector of mixed data types, much like those discussed in the SATURN project. Therefore, on the one hand, the main workhorse in the analysis of those datasets are machine learning methods. On the other hand, a large portion of the features will in the future rely heavily on sequencing data and bioinformatics analysis.

Results of all of the projects presented in this thesis showed that the adverse effects of antibiotic therapy depend on the intricate inner workings of the gut microbiome. The SATURN project confirmed that antibiotics were the main reason driving the colonization with MDR. No other patient characteristic or comorbidities turned out to be as important. Knowing that the gut response to antibiotic differs between patients and that it depends on their lifestyle and past treatments - personalization of antibiotic therapy needs more basic research into the gut

microbiome. It also requires communication of the results in the form of research papers and tools providing doctors meaningful, actionable access to the study results, of which a modest example is our `AskSaturn` website.

Analysis of the metagenomic sequencing resembles pooling since it is not possible to sequence each bacterium of the gut microbiome. Researchers analyze small samples hoping they are representative for the entire gut microbiome. For this reason research into the gut microbiome needs large datasets of sequencing samples along with the metadata prepared with unified and carefully controlled methods. However, that is not enough. Researchers have also stressed the importance of time-series analysis, as only such data enables discovery of the interactions [155]. Characterization of the MGEs within the gut, to be able to track resistance accurately, requires scaffolds. However, metagenomic assembly is quite error-prone. Here, long-read sequencing technologies paired with the new algorithms enable functional and taxonomical annotations and comparisons of the large datasets [263].

During my Ph.D. I had the opportunity to work on the projects dealing with various aspects of antibiotics resistance. In the methodological aspects, those projects were quite diverse. However, each of them corresponded to one of the crucial parts of a pipeline to provide the gut microbiome information to the routine medical practice. Such a pipeline will enable better antibiotic usage and control of resistance emergence.

Bibliography

[1] J O'Neill. Tackling a Crisis for the Health and Wealth of Nations. Technical Report February 2015, Wellcome trust, 2015.

[2] Eileen R Choffnes, David a Relman, and Alison Mack. *Antibiotic Resistance: Implications for Global Health and Novel Intervention Strategies: Workshop*. The national academies press, 2010.

[3] WHO. Antimicrobial resistance. Global Report on Surveillance. *Bulletin of the World Health Organization*, 61(3):383–94, 2014.

[4] Richard J Fair and Yitzhak Tor. Antibiotics and Bacterial Resistance in the 21st Century. *Perspectives in Medicinal Chemistry*, pages 25–64, 2014.

[5] John G. Bartlett, David N. Gilbert, and Brad Spellberg. Seven ways to preserve the Miracle of antibiotics. *Clinical Infectious Diseases*, 56(10):1445–1450, 2013.

[6] B. Brismar, C. Edlund, A. S. Malmborg, and C. E. Nord. Ciprofloxacin concentrations and impact of the colon microflora in patients undergoing colorectal surgery. *Antimicrobial Agents and Chemotherapy*, 34(3):481–483, 1990.

[7] D. Van Der Waaij, J. M. Berghuis-de Vries, and J. E C Lekkerkerk-Van Der Wees. Colonization resistance of the digestive tract in conventional and antibiotic-treated mice. *Journal of Hygiene*, 69(3):405–411, 1971.

[8] E J Vollaard ' and H A L Clasener2. MINIREVIEW Colonization Resistance. *Antimicrobial Agents and Chemotherapy*, 38(3):409–414, 1994.

[9] Ebimieowei Etebu and Ibemologi Arikekpar. Antibiotics: Classification and mechanisms of action with emphasis on molecular perspectives. *International Journal of Applied Microbiology and Biotechnology Research*, 4:90–101, 2016.

[10] Allen S. Johnson. Medicine's responsibility in the propagation of poor protoplasm. *New England Journal of Medicine*, 238(22), 1948.

[11] Walter Sneader. 22 Antibiotics. In *Drug Discovery. A History.*, chapter 22, pages 287–313. John Wiley & Sons Ltd, 2005.

[12] K H Schleifer and O Kandler. Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriological Reviews*, 36(4):407–477, 1972.

[13] G S Wilson and A A Miles. *Topley and Wilson's Principles of Bacteriology and Immunity*, volume 1. The Williams & Wilkins Company, Baltimore, 1946.

[14] G. A. Pankey and L. D. Sabath. Clinical Relevance of Bateriostatic versus Bactericidal Mechanisms of Action in the Treatment of Gram-Positive Bacterial Infections. *Clinical Infectious Diseases*, 38(6):864–870, 2004.

[15] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. How cells readthe genome: from DNA to protein. In *Molecular biology of the cell*, pages 299–374. Garland Science, 2002.

[16] Katsuhiko S. Murakami. Structural biology of bacterial RNA polymerase. *Biomolecules*, 5(2):848–864, 2015.

[17] Katsuhiko S. Murakami and Seth A. Darst. Bacterial RNA polymerases: The wholo story. *Current Opinion in Structural Biology*, 13(1):31–39, 2003.

[18] Jayanta Mukhopadhyay, Kalyan Das, Sajida Ismail, et al. The RNA Polymerase "Switch Region" Is a Target for Inhibitors. *Cell*, 135(2):295–307, 2008.

[19] Cong Ma, Xiao Yang, and Peter J. Lewis. Bacterial Transcription as a Target for Antibacterial Drug Development. *Microbiology and Molecular Biology Reviews*, 80(1):139–160, 2016.

[20] Daniel N. Wilson. Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nature Reviews Microbiology*, 12(1):35–48, 2014.

[21] T. Martin Schmeing and V. Ramakrishnan. What recent ribosome structures have revealed about the mechanism of translation. *Nature*, 461(7268):1234–1242, 2009.

[22] Shinichiro Shoji, Sarah E. Walker, and Kurt Fredrick. Ribosomal translocation: One step closer to the molecular mechanism. *ACS Chemical Biology*, 4(2):93–107, 2009.

[23] Lisa S McCoy, Yun Xie, and Yitzhak Tor. Antibiotics that target protein synthesis. *Wiley interdisciplinary reviews. RNA*, 2(2):209–32, 2011.

[24] Norbert Polacek and Alexander S. Mankin. The Ribosomal Peptidyl Transferase Center: Structure, Function, Evolution, Inhibition. *Critical Reviews in Biochemistry and Molecular Biology*, 40(5):285–311, 2005.

[25] Marie Paule Mingeot-Leclercq, Youri Glupczynski, and Paul M. Tulkens. Aminoglycosides: Activity and resistance. *Antimicrobial Agents and Chemotherapy*, 43(4):727–737, 1999.

[26] Ian Chopra and Marilyn Roberts. Tetracycline Antibiotics : Mode of Action, Applications , Molecular Biology, and Epidemiology of Bacterial Resistance. *Microbiology and Molecular Biology Reviews*, 65(2):232–260, 2001.

[27] G. V. R. Born. Virginiamycin as an Antibiotic for Poultry Feeds. *Nature*, 196:952–953, 1962.

[28] Mark Casewell, Christian Friis, Enric Marco, Paul McMullin, and Ian Phillips. The European ban on growth-promoting antibiotics and emerging consequences for human and animal health. *Journal of Antimicrobial Chemotherapy*, 52(2):159–161, 2003.

[29] S M Swaney, H Aoki, and M C Ganoza. The oxazolidinone linezolid inhibts initiation of protein synthesis in bacteria. *Antimicrobial Agents and Chemotherapy*, 42(12):3251–3255, 1998.

[30] Zohar Eyal, Donna Matzov, Miri Krupkin, et al. Structural insights into species-specific features of the ribosome from the pathogen *Staphylococcus aureus*. *Proceedings of the National Academy of Sciences*, 112(43):E5805–E5814, 2015.

[31] Patricia Kloss, Liqun Xiong, Dean L. Shinabarger, and Alexander S. Mankin. Resistance mutations in 23 S rRNA identify the site of action of the protein synthesis inhibitor linezolid in the ribosomal peptidyl transferase center. *Journal of Molecular Biology*, 294(1):93–101, 1999.

[32] Michael R. Barbachyn and Charles W. Ford. Oxazolidinone structure-activity relationships leading to linezolid. *Angewandte Chemie - International Edition*, 42(18):2010–2023, 2003.

[33] P Courvalin R Leclercq. Bacterial resistance to macrolide, lincosamide, and streptogramin antibiotics by target modification. *Antimicrobial Agents and Chemotherapy*, 35(7):1267, 1991.

[34] J Feng, a Lupien, H Gingras, et al. Genome sequencing of linezolid-resistant Streptococcus pneumoniae mutants reveals novel mechanisms of resistance. *Genome research*, pages 1214–1223, 2009.

[35] Vickers Burdett. Tet(M)-promoted release of tetracycline from ribosomes is GTP dependent. *Journal of Bacteriology*, 178(11):3246–3251, 1996.

[36] Sean R. Connell, Dobryan M. Tracz, Knud H. Nierhaus, and Diane E. Taylor. Ribosomal Protection Proteins and Their Mechanism of Tetracycline Resistance. *Antimicrobial Agents and Chemotherapy*, 47(12):3675–3681, 2003.

[37] J. L. Burns, L. A. Hedin, and D. M. Lien. Chloramphenicol resistance in pseudomonas cepacia because of decreased permeability. *Antimicrobial Agents and Chemotherapy*, 33(2):136–141, 1989.

[38] B. S. Speer and A. A. Salyers. Novel aerobic tetracycline resistance gene that chemically modifies tetracycline. *Journal of Bacteriology*, 171(1):148–153, 1989.

[39] T J Silhavy, D Kahne, and S Walker. The bacterial cell envelope. *Cold Spring Harb Perspect Biol*, 2(5):a000414, 2010.

[40] Manhong Wu, Elke Maier, Roland Benz, and Robert E. W. Hancock. Mechanism of Interaction of Different Classes of Cationic Antimicrobial Peptides with Planar Bilayers and with the Cytoplasmic Membrane of *Escherichia coli*. *Biochemistry*, 38(22):7235–7242, 1999.

[41] Judith N. Steenbergen, Jeff Alder, Grace M. Thorne, and Francis P. Tally. Daptomycin: A lipopeptide antibiotic for the treatment of serious Gram-positive infections. *Journal of Antimicrobial Chemotherapy*, 55(3):283–288, 2005.

[42] Jared a Silverman, Nancy G Perlmutter, M Howard, and Howard M Shapiro. Correlation of daptomycin bactericidal activity and membrane depolarization in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 47(8):2538–2544, 2003.

[43] Tianhua Zhang, Jawad K. Muraih, Ben MacCormick, Jared Silverman, and Michael Palmer. Daptomycin forms cation- and size-selective pores in model membranes. *Biochimica et Biophysica Acta - Biomembranes*, 1838(10):2425–2430, 2014.

[44] Sina Jordan, Matthew I. Hutchings, and Thorsten Mascher. Cell envelope stress response in Gram-positive bacteria. *FEMS Microbiology Reviews*, 32(1):107–146, 2008.

[45] Anton Y. Peleg, Spiros Miyakis, Doyle V. Ward, et al. Whole genome characterization of the mechanisms of daptomycin resistance in clinical and laboratory derived isolates of staphylococcus aureus. *PLoS ONE*, 7(1), 2012.

[46] Truc T. Tran, Jose M. Munita, and Cesar A. Arias. Mechanisms of drug resistance: Daptomycin resistance. *Annals of the New York Academy of Sciences*, 1354(1):32–53, 2015.

[47] Brunello Oliva, Gloria Gordon, Paul Mcnicholas, George Ellestad, and I a N Choprat. Evidence that Tetracycline Analogs Whose Primary Target Is Not the Bacterial Ribosome Cause Lysis of Eschierichia coli. *Antimicrobial Agents and Chemotherapy*, 36(5):913–919, 1992.

[48] Jennifer J. Stepanek, Tadeja Lukežič, Ines Teichert, Hrvoje Petković, and Julia E. Bandow. Dual mechanism of action of the atypical tetracycline chelocardin. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1864(6):645–654, 2016.

[49] David W Green. The bacterial cell wall as a source of antibacterial targets. *Expert opinion on therapeutic targets*, 6(1):1–19, 2002.

[50] Lynn L. Silver. Novel inhibitors of bacterial cell wall synthesis. *Current Opinion in Microbiology*, 6(5):431–438, 2003.

[51] D J Tipper. Mode of action of beta-lactam antibiotics. *Pharmacology & therapeutics*, 27:1–35, 1985.

[52] B. Korat, H. Mottl, and W. Keck. Penicillin-binding protein 4 of Escherichia coli: molecular cloning of the dacB gene, controlled overexpression, and alteriations in murein composition. *Molecular Microbiology*, 5(3):675–684, 1991.

[53] Hiroyuki Kishida, Satoru Unzai, David I. Roper, Adrian Lloyd, Sam Yong Park, and Jeremy R H Tame. Crystal structure of penicillin binding protein 4 (dacB) from Escherichia coli, both in the native form and covalently linked to various antibiotics. *Biochemistry*, 45(3):783–792, 2006.

[54] R. P. Elander. Industrial production of β-lactam antibiotics. *Applied Microbiology and Biotechnology*, 61(5-6):385–392, 2003.

[55] PA Bradford. Extended spectrum betalactamase in the 21 century: characterization, epidemiology, and detection of this important resistant threat. *Clinical Microbiol Rev*, 14(4):933–951, 2001.

[56] Karen Bush and George A. Jacoby. Updated functional classification of β-lactamases. *Antimicrobial Agents and Chemotherapy*, 54(3):969–976, 2010.

[57] Naasm T, , and Nordman Poiller P. Minor extended-spectrum B-lactamases. *Expert Review of Anti-Infective Therapy*, 8(11):1251–1258, 2010.

[58] M. Gniadkowski. Evolution and epidemiology of extended-spectrum β-lactamases (ESBLs) and ESBL-producing microorganisms. *Clinical Microbiology and Infection*, 7(11):597–608, 2001.

[59] K. Poole. Resistance to β-lactam antibiotics. *Cellular and Molecular Life Sciences*, 61(17):2200–2223, 2004.

[60] Françoise Van Bambeke. Glycopeptides and glycodepsipeptides in clinical development: a comparative review of their antibacterial spectrum, pharmacokinetics and clinical efficacy., 2006.

[61] P. E. Reynolds. Structure, biochemistry and mechanism of action of glycopeptide antibiotics. *European Journal of Clinical Microbiology & Infectious Diseases*, 8(11):943–950, 1989.

[62] P. Courvalin. Vancomycin Resistance in Gram-Positive Cocci. *Clinical Infectious Diseases*, 42(Supplement 1):S25–S34, 2006.

[63] Guy R Pupp, William J Brown, D Ph, Denise Cardo, and Scott K Fridkin. Infection with Vancomycin-Resistant Staphylococcus aureus Containing the vanA Resistance Gene. *The New England Jornal of Meficine*, 348(14):1342–1347, 2003.

[64] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. DNA replication, repair and recombination. In *Molecular biology of the cell*, pages 235–298. Garland Science, 2002.

[65] Michael O'Donnell, Lance Langston, and Bruce Stillman. Principles and concepts of DNA replication in bacteria, archaea, and eukarya. *Cold Spring Harbor Perspectives in Biology*, 5(7):1–13, 2013.

[66] Stephan Heeb, Matthew P. Fletcher, Siri Ram Chhabra, Stephen P. Diggle, Paul Williams, and Miguel Cámara. Quinolones: From antibiotics to autoinducers. *FEMS Microbiology Reviews*, 35(2):247–274, 2011.

[67] João H. Morais Cabral, Andrew P. Jackson, Clare V. Smith, Nita Shikotra, Anthony Maxwell, and Robert C. Liddington. Crystal structure of the breakage-reunion domain of DNA gyrase. *Nature*, 388(6645):903–906, 1997.

[68] Anna Fàbrega, Sergi Madurga, Ernest Giralt, and Jordi Vila. Mechanism of action of and resistance to quinolones. *Microbial Biotechnology*, 2(1):40–61, 2009.

[69] M. I. Andersson. Development of the quinolones. *Journal of Antimicrobial Chemotherapy*, 51(90001):1–11, 2003.

[70] Katie J. Aldred, Robert J. Kerns, and Neil Osheroff. Mechanism of quinolone action and resistance. *Biochemistry*, 53(10):1565–1574, 2014.

[71] Helen C. Davison, Mark E.J. Woolhouse, and J. Chris Low. What is antibiotic resistance and how can we measure it? *Trends in Microbiology*, 8(12):554–559, 2000.

[72] Santiago Sandoval-Motta and Maximino Aldana. Adaptive resistance to antibiotics in bacteria: A systems biology perspective. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 8(3):253–267, 2016.

[73] Jose M Munita, Cesar A Arias, Antimicrobial Resistance Unit, and Alemana De Santiago. Mechanisms of Antibiotics Resistance. *Mechanisms of Antibiotic Resistance*, 4(2):1–37, 2016.

[74] Mark A Toleman, Peter M Bennett, Timothy R Walsh, et al. Aminoglucoside Modifing Enzymes. *Sciences-New York*, 13(6):151–171, 2011.

[75] Sylvie Garneau-tsodikova and Kristin J Labby. Mechanisms of Resistance to Aminoglycoside Antibiotics: Overview and Perspectives. *MedChemComm*, 7(1):11–27, 2016.

[76] Stefan Schwarz, Corinna Kehrenberg, Benoît Doublet, and Axel Cloeckaert. Molecular basis of bacterial resistance to chloramphenicol and florfenicol. *FEMS Microbiology Reviews*, 28(5):519–542, 2004.

[77] Karen Joy Shaw and Michael R. Barbachyn. The oxazolidinones: Past, present, and future. *Annals of the New York Academy of Sciences*, 1241(1):48–70, 2011.

[78] Yang Wang, Yuan Lv, Jiachang Cai, et al. A novel gene, optrA, that confers transferable resistance to oxazolidinones and phenicols and its presence in Enterococcus faecalis and Enterococcus faecium of human and animal origin. *Journal of Antimicrobial Chemotherapy*, 70(8):2182–2190, 2015.

[79] Katherine S. Long and Birte Vester. Resistance to linezolid caused by modifications at its binding site on the ribosome. *Antimicrobial Agents and Chemotherapy*, 56(2):603–612, 2012.

[80] Nicole J Johnston, Tariq a Mukhtar, and Gerard D Wright. Streptogramin antibiotics: mode of action and resistance. *Current drug targets*, 3(4):335–344, 2002.

[81] Roland Leclercq and Patrice Courvalin. Resistance to macrolides and related antibiotics in Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy*, 46(9):2727–2734, 2002.

[82] Corey Fyfe, Trudy H. Grossman, Kathy Kerstein, and Joyce Sutcliffe. Resistance to macrolide antibiotics in public health pathogens. *Cold Spring Harbor Perspectives in Medicine*, 6(10):1–38, 2016.

[83] Peter Spanogiannopoulos, Maulik Thaker, Kalinka Koteva, Nicholas Waglechner, and Gerard D. Wright. Characterization of a rifampin-inactivating glycosyltransferase from a screen of environmental actinomycetes. *Antimicrobial Agents and Chemotherapy*, 56(10):5061–5069, 2012.

[84] Yasutaka Hoshino, Shoko Fujii, Hideki Shinonaga, et al. Monooxygenation of rifampicin catalyzed by the rox gene product of Nocardia farcinica: Structure elucidation, gene identification and role in drug resistance. *Journal of Antibiotics*, 63(1):23–28, 2010.

[85] L P Kotra and S Mobashery. B-Lactam antibiotics, B-lactamases and bacterial resistance. *Bull. lnst. Pasteur Paris*, 96:139–150, 1998.

[86] George A Jacoby. Mechanisms of resistance to quinolones. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 41 Suppl 2(Supplement_2):S120–6, 2005.

[87] S. S. Hegde. A Fluoroquinolone Resistance Protein from Mycobacterium tuberculosis That Mimics DNA. *Science*, 308(5727):1480–1483, 2005.

[88] Cesar A. Arias, Diana Panesso, Danielle M. McGrath, et al. Genetic Basis for In Vivo Daptomycin Resistance in Enterococci. *New England Journal of Medicine*, 365(10):892–900, 2011.

[89] Christopher M. Thomas and Kaare M. Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology*, 3(9):711–721, 2005.

[90] D Dubnau. DNA uptake in bacteria. *Annual Review of Microbiology*, 53(1):217, 1999.

[91] G J Stewart and C A Carlson. The biology of natural transformation. *Ann. Rev. Microbiol.*, 40:211–235, 1986.

[92] T. A. Brown. Mapping genomes. In *Genomes*, pages 125–162. Garland Science, 2002.

[93] Benjamin Lewin. The replicon. In *Genes VII*, pages 349–384. Oxford University Press, 2000.

[94] LM Proctor, A Okubo, and Jed A. Fuhrman. Calibrating estimates of phage-induced mortality in marine bacteria: Ultrastructural studies of marine bacteriophage development from one-step growth experiments. *Microbial Ecology*, 25(2):161–82, 1993.

[95] Martha R.J. Clokie, Andrew D. Millard, Andrey V. Letarov, and Shaun Heaphy. Phages in nature. *Bacteriophage*, 1(1):31–45, 2011.

[96] John L. Mokili, Forest Rohwer, and Bas E. Dutilh. Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*, 2(1):63–77, 2012.

[97] Harald Brüssow and Roger W. Hendrix. Phage Genomics: Small is beautiful. *Cell*, 108(1):13–16, 2002.

[98] John Carter and Valentina Saunders. Bacterial viruses. In *Virology. Principles and applications*, pages 229–255. Willey, 2009.

[99] Gipsi Lima-Mendez, Jacques Van Helden, Ariane Toussaint, and Raphaël Leplae. Reticulate representation of evolutionary and functional relationships between phage genomes. *Molecular Biology and Evolution*, 25(4):762–777, 2008.

[100] Hélène Deveau, Josiane E Garneau, and Sylvain Moineau. CRISPR/Cas system and its role in phage-bacteria interactions. *Annual review of microbiology*, 64:475–493, 2010.

[101] L A Marraffini and E J Sontheimer. CRISPR interference limits horizontal gene transfer in *staphylococci* by targeting DNA. *Science*, 322(5909):1843–1845, 2008.

[102] Philippe Horvath and Rodolphe Barrangou. CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science*, 327(5962):167–170, 2010.

[103] Benjamin Levin. Transposons. In *Genes VII*, pages 457–484. Oxford University Press, 2000.

[104] Ben Langmead. Aligning short sequecning reads with Bowtie. *Current Protocols Bioinformatics*, pages 1–24, 2011.

[105] Elise Darmon and David R F Leach. Bacterial genome instability. *Microbiology and molecular biology reviews : MMBR*, 78(1):1–39, 2014.

[106] Michael R. Gillings. Integrons: Past, Present, and Future. *Microbiology and Molecular Biology Reviews*, 78(2):257–277, 2014.

[107] Hiroshi Nikaido. Multidrug Resistance in Bacteria. *Annu Rev Biochem.*, 78(2):119–146, 2009.

[108] A. Magiorakos, A Srinivasan, R B Carey, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clinical Microbiology and Infection*, 18(3):268–281, 2011.

[109] Evelina Tacconelli, Giulia De angelis, Maria A. Cataldo, Emanuela Pozzi, and Roberto Cauda. Does antibiotic exposure increase the risk of methicillin-resistant Staphylococcus aureus (MRSA) isolation? A systematic review and meta-analysis. *Journal of Antimicrobial Chemotherapy*, 61(1):26–38, 2008.

[110] J. Davies and D. Davies. Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews*, 74(3):417–433, 2010.

[111] S. Defres, C. Marwick, and D. Nathwani. MRSA as a cause of lung infection including airway infection, community-acquired pneumonia and hospital-acquired pneumonia. *European Respiratory Journal*, 34(6):1470–1476, 2009.

[112] Fd Lowy. Antimicrobial resistance: the example of *Staphylococcus aureus*. *Journal of Clinical Investigation*, 111(9):1265–1273, 2003.

[113] Hare Krishna Tiwari and Malay Ranjan Sen. Emergence of vancomycin resistant Staphylococcus aureus (VRSA) from a tertiary care hospital from northern part of India. *BMC Infectious Diseases*, 6:1–6, 2006.

[114] Mary T. Bessesen, Cassandra Vogel Kotter, Brandie D. Wagner, et al. MRSA colonization and the nasal microbiome in adults at high risk of colonization and infection. *Journal of Infection*, 71(6):649–657, 2015.

[115] Mission Report. MISSION REPORT ECDC country visit to Romania to discuss antimicrobial resistance issues. (March), 2017.

[116] K Hiramatsu, Y Katayama, H Yuzawa, and T Ito. Molecular genetics of methicillin-resistant Staphylococcus aureus. *Int.J.Med.Microbiol.*, 292(2):67–74, 2002.

[117] Sahreena Lakhundi and Kunyan Zhang. Methicillin-Resistant Staphylococcus aureus: Molecular Characterization, Evolution, and Epidemiology. *Clinical Microbiology Reviews*, 31(4):1–103, 2018.

[118] Rishi H.P. Dhillon and John Clark. ESBLs: A clear and present danger? *Critical Care Research and Practice*, 2012, 2012.

[119] Marta Tacão, Alexandra Moura, António Correia, and Isabel Henriques. Co-resistance to different classes of antibiotics among ESBL-producers from aquatic systems. *Water Research*, 48(1):100–107, 2014.

[120] Alma Brolund. Overview of ESBL-producing *Enterobacteriaceae* from a Nordic perspective. *Infection Ecology & Epidemiology*, 4(1):24555, 2014.

[121] Elizabeth K. Costello, Christian L. Lauber, Micah Hamady, Noah Fierer, Jeffrey I. Gordon, and Rob Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, 2009.

[122] Inna Sekirov, Shannon L Russell, L Caetano M Antunes, and B Brett Finlay. Gut Microbiota in Health and Disease. *Physiol Rev*, pages 859–904, 2010.

[123] F Backhed. Host-Bacterial Mutualism in the Human Intestine. *Science (New York, NY)*, 307(5717):1915–1920, 2005.

[124] G L Simon and S L Gorbach. The human intestinal microflora. *Dig Dis Sci*, 31(9):147S–162S, 1986.

[125] Ruth E. Ley, Daniel A. Peterson, and Jeffrey I. Gordon. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124(4):837–848, 2006.

[126] Daniel N Frank, Allison L St Amand, Robert A Feldman, Edgar C Boedeker, Noam Harpaz, and Norman R Pace. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13780–5, 2007.

[127] Xing Yang, Lu Xie, Yixue Li, and Chaochun Wei. More than 9,000,000 unique genes in human gut bacterial community: Estimating gene numbers inside a human body. *PLoS ONE*, 4(6):0–7, 2009.

[128] Paul B Eckburg, Elisabeth M Bik, Charles N Bernstein, et al. Diversity of the Human Intestinal Microbial Flora. *Science*, 308(2005):1635–1639, 2005.

[129] Grace Y Chen. Characterizing the Role of the Gut Microbiome in Colorectal Cancer. *Clinics in Colon and Rectal Surgery*, 31(3):192–198, 2014.

[130] Xochitl C Morgan, Timothy L Tickle, Harry Sokol, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):R79, 2012.

[131] Xiaokai Wang, Xiaoqiang Xu, and Yan Xia. Further analysis reveals new gut microbiome markers of type 2 diabetes mellitus. *Antonie van Leeuwenhoek*, 110(3):445–453, 2017.

[132] Fredrik H. Karlsson, Frida Fåk, Intawat Nookaew, et al. Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nature Communications*, 3, 2012.

[133] Cesare Mancuso and Rosaria Santangelo. Alzheimer's disease and gut microbiota modifications: The long way between preclinical studies and clinical evidence. *Pharmacological Research*, 129:329–336, 2018.

[134] Robin N. Groen, Nicolien C. de Clercq, Max Nieuwdorp, H. J. Rogier Hoenders, and Albert K. Groen. Gut microbiota, metabolism and psychopathology: A critical review and novel perspectives. *Critical Reviews in Clinical Laboratory Sciences*, 55(4):1–11, 2018.

[135] R E Ley, F Backhed, P Turnbaugh, C A Lozupone, R D Knight, and J I Gordon. Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A*, 102(31):11070–11075, 2005.

[136] Peter J Turnbaugh, Ruth E Ley, Michael a Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–31, 2006.

[137] Alexander Koliada, Ganna Syzenko, Vladislav Moseiko, et al. Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. *BMC Microbiology*, 17(1):4–9, 2017.

[138] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

[139] Paul I. Costea, Falk Hildebrand, Arumugam Manimozhiyan, et al. Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, 3(1):8–16, 2018.

[140] Paul W. O'Toole and Ian B. Jeffery. Microbiome–health interactions in older people. *Cellular and Molecular Life Sciences*, 75(1):119–128, 2018.

[141] Carlotta De Filippo, Duccio Cavalieri, Monica Di Paola, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 107(33):14691–6, 2010.

[142] Jason Lloyd-price, Galeb Abu-ali, and Curtis Huttenhower. The healthy human microbiome. *Genome Medicine*, pages 1–11, 2016.

[143] Dachao Liang, Ross Ka-Kit Leung, Wenda Guan, and William W. Au. Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities. *Gut Pathogens*, 10(1):3, 2018.

[144] Susan Westfall, Nikita Lomis, Imen Kahouli, Si Yuan Dia, Surya Pratap Singh, and Satya Prakash. Microbiome, probiotics and neurodegenerative diseases: deciphering the gut brain axis. *Cellular and Molecular Life Sciences*, 74(20):3769–3787, 2017.

[145] S M Finegold, H R Attebery, and V L Sutter. Effect of diet on human fecal flora: comparison of Japanese and American diets. *The American journal of clinical nutrition*, 27(12):1456–1469, 1974.

[146] Marcus J. Claesson, Qiong Wang, Orla O'Sullivan, Rachel Greene-Diniz, James R. Cole, R. Paul Ross, and Paul W. O'Toole. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*, 38(22), 2010.

[147] Juan Jovel, Jordan Patterson, Weiwei Wang, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*, 7(APR):1–17, 2016.

[148] Daniel H. Huson, Sina Beier, Isabell Flade, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Comput Biol*, 12(6):4–12, 2016.

[149] Hilary P. Browne, Samuel C. Forster, Blessing O. Anonye, et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604):543–546, 2016.

[150] Ana Elena Pérez-Cobas, Alejandro Artacho, Henrik Knecht, et al. Differential effects of antibiotic therapy on the structure and function of human gut microbiota. *PLoS ONE*, 8(11), 2013.

[151] A. J. Stewardson, N. Gaïa, P. François, et al. Collateral damage from oral ciprofloxacin versus nitrofurantoin in outpatients with urinary tract infections: A culture-free analysis of gut microbiota. *Clinical Microbiology and Infection*, 21(4):344.e1–344.e11, 2015.

[152] Mamun Ur Rashid, Andrej Weintraub, and Carl Erik Nord. Effect of new antimicrobial agents on the ecological balance of human microflora. *Anaerobe*, 18(2):249–253, 2012.

[153] Kathleen Lange, Martin Buerger, Andreas Stallmach, and Tony Bruns. Effects of Antibiotics on Gut Microbiota. *Digestive Diseases*, 34(3):260–268, 2016.

[154] Suchita Panda, Ismail El Khader, Francesc Casellas, et al. Short-term effect of antibiotics on human gut microbiota. *PLoS ONE*, 9(4), 2014.

[155] Les Dethlefsen and David A Relman. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(Suppl 1):4554–61, 2011.

[156] M F De La Cochetière, T Durand, P Lepage, A Bourreille, J P Galmiche, J Doré, and M F De La Cochetie. Resilience of the Dominant Human Fecal Microbiota upon Short-Course Antibiotic Challenge. *Journal of Clinical Microbiology*, 43(11):5588, 2005.

[157] Sheetal R. Modi, James J. Collins, and David A. Relman. Antibiotics and the gut microbiota. *Journal of Clinical Investigation*, 124(10):4212–4218, 2014.

[158] Cecilia Jernberg, Sonja Löfmark, Charlotta Edlund, and Janet K. Jansson. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *ISME Journal*, 1(1):56–66, 2007.

[159] Willem Van Schaik. The human gut resistome. *Philosophical Transactions of the Royal Society of London B*, 370:1–9, 2015.

[160] Hedvig E. Jakobsson, Cecilia Jernberg, Anders F. Andersson, Maria Sjolund-Karlsson, Janet K. Jansson, and Lars Engstrand. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS ONE*, 5(3), 2010.

[161] David Fitzpatrick and Fiona Walsh. Antibiotic resistance genes across a wide variety of metagenomes. *FEMS Microbiology Ecology*, 92(2):1–8, 2016.

[162] Kristoffer Forslund, Shinichi Sunagawa, Jens Roat Kultima, Daniel R Mende, Manimozhiyan Arumugam, Athanasios Typas, and Peer Bork. Country-specific antibiotic use practices impact the human gut resistome. *Genome*, 23:1163–1169, 2013.

[163] Yongfei Hu, Xi Yang, Junjie Qin, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nature communications*, 4:2151, 2013.

[164] The Nih and H M P Working. The NIH Human Microbiome Project. *Genome Research*, 19(12):2317–2323, 2009.

[165] Curtis Huttenhower, Dirk Gevers, Rob Knight, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[166] Daniel McDonald, Embriette Hyde, Justine W Debelius, et al. American Gut: an Open Platform for Citizen Science Microbiome Research. *American society for microbiology*, 3(3):1–28, 2018.

[167] Xiu Qiu, Jin Hua Lu, Jian Rong He, et al. The Born in Guangzhou Cohort Study (BIGCS). *European Journal of Epidemiology*, 32(4):337–346, 2017.

[168] Jan Hendrik Hehemann, Gaëlle Correc, Tristan Barbeyron, William Helbert, Mirjam Czjzek, and Gurvan Michel. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature*, 464(7290):908–912, 2010.

[169] Eric Bapteste and Yan Boucher. Horizontal Gene Transfer. *Gene*, 532:55–72, 2009.

[170] Brian V. Jones and Julian R. Marchesi. Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nature Methods*, 4(1):55–61, 2007.

[171] Brian V. Jones, Funing Sun, and Julian R Marchesi. Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome. *BMC genomics 2010*, 11(46), 2010.

[172] Rebecca V Thurber, Matthew Haynes, Mya Breitbart, Linda Wegley, and Forest Rohwer. Laboratory procedures to generate viral metagenomes. *Nature protocols*, 4(4):470–483, 2009.

[173] Sheetal R Modi, Henry H Lee, Catherine S Spina, and James J Collins. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*, 499(7457):219–22, 2013.

[174] Lesley A Ogilvie and Brian V Jones. The human gut virome : a multifaceted majority. *Frontiers in microbiology*, 6(September):1–12, 2015.

[175] Jie Ren, Nathan A. Ahlgren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5(1):69, 2017.

[176] Pawel S Krawczyk, Leszek Lipinski, and Andrzej Dziembowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, 46(6):1–14, 2018.

[177] Alma Brolund, Oscar Franzén, Öjar Melefors, Karin Tegmark-Wisell, and Linus Sandegren. Plasmidome-Analysis of ESBL-Producing Escherichia coli Using Conventional Typing and High-Throughput Sequencing. *PLoS ONE*, 8(6), 2013.

[178] B. Stecher, R. Denzler, L. Maier, et al. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proceedings of the National Academy of Sciences*, 109(4):1269–1274, 2012.

[179] Cory S. Straub, Anthony R. Ives, and Claudio Gratton. Evidence for a Trade-Off between Host-Range Breadth and Host-Use Efficiency in Aphid Parasitoids. *The American Naturalist*, 177(3):389–395, 2011.

[180] Samuel Minot, Rohini Sinha, Jun Chen, et al. The human gut virome: Inter-individual variation and dynamic response to diet The human gut virome : Inter-individual variation and dynamic response to diet. *Genome Research*, 10(21):1616–1625, 2011.

[181] Pilar Manrique, Benjamin Bolduc, Seth T. Walk, John van der Oost, Willem M. de Vos, and Mark J. Young. Healthy human gut phageome. *Proceedings of the National Academy of Sciences*, page 201601060, 2016.

[182] Alejandro Reyes, Matthew Haynes, Nicole Hanson, Florent E Angly, Andrew C Heath, Forest Rohwer, and Jeffrey I Gordon. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466(7304):334–338, 2011.

[183] S. Roux, M. Krupovic, D. Debroas, P. Forterre, and F. Enault. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biology*, 3(12):130160–130160, 2013.

[184] Shira R. Abeles, Melissa Ly, Tasha M. Santiago-Rodriguez, and David T. Pride. Effects of long term antibiotic therapy on human oral and fecal viromes. *PLoS ONE*, 10(8):1–18, 2015.

[185] François Enault, Arnaud Briet, Léa Bouteille, Simon Roux, Matthew B Sullivan, and Marie-Agnès Petit. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *The ISME journal*, page 053025, 2016.

[186] Eileen Broaders, Cormac G.M. Gahan, and Julian R. Marchesi. Mobile genetic elements of the human gastrointestinal tract. *Gut Microbes*, 4(4):271–280, 2013.

[187] Britt Koskella and Sean Meaden. Understanding bacteriophage specificity in natural microbial communities. *Viruses*, 5(3):806–823, 2013.

[188] N B Shoemaker, H Vlamakis, K Hayes, and a a Salyers. Evidence for Extensive Resistance Gene Transfer among Bacteroides spp . and among Bacteroides and Other Genera in the Human Colon. *Appl Environ Microbiol*, 67(2):561–8, 2001.

[189] Anamika Gupta, Hera Vlamakis, Nadja Shoemaker, and Abigail A. Salyers. A New Bacteroides Conjugative Transposon that Carries an ermB Gene. *Applied and Environmental Microbiology*, 69(11):6455–6463, 2003.

[190] Arpana Sagwal Chaudhary. A review of global initiatives to fight antibiotic resistance and recent antibiotics' discovery. *Acta Pharmaceutica Sinica B*, 6(6):552–556, 2016.

[191] Idsa Public Policy. The 10x'20 Initiative : Pursuing a Global Commitment to Develop 10 New Antibacterial Drugs by 2020. Technical Report April, Infectious Diseases Society of America, 2010.

[192] Adis Data, Information Bv, and Adis. Fidaxomicin. Technical Report 1, 2010.

[193] Bradley M. Hover, Seong-hwan Hwan Kim, Micah Katz, et al. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nature Microbiology*, 3(4):415–422, 2018.

[194] P Davey, Marwick Ca, Scott Cl, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients (Review). Technical Report 2, Cochrane Database of Systematic Reviews, 2017.

[195] Christophe Van Dijck and Janneke Arnoldine. Antibiotic stewardship interventions in hospitals in low-and middle- income countries : a systematic review. *Bull World Health Organ*, 4(96):266–280, 2018.

[196] EU Publications Office. Community Research and Development Information Service (CORDIS), 2018.

[197] The Saturn Project.

[198] Fraser V Kollef MH Vlasnik J, Sharpless L, Pasque C, Murphy D. Scheduled change of antibiotics classes: a strategy to decrease the incidence of ventilator associated pneumonia. *Am J Respir Crit Care Med*, 156:1040–1048, 1997.

[199] Pleun Joppe van Duijn, Walter Verbrugghe, Philippe Germaine Jorens, et al. The effects of antibiotic cycling and mixing on antibiotic resistance in intensive care units: a cluster-randomised crossover trial. *The Lancet Infectious Diseases*, 18(4):401–409, 2018.

[200] A. J. Stewardson, J. Vervoort, N. Adriaenssens, et al. Effect of outpatient antibiotics for urinary tract infections on antimicrobial resistance among commensal Enterobacteriaceae: A multinational prospective cohort study. *Clinical Microbiology and Infection*, (2018), 2018.

[201] Giulia De Angelis, Giovanni Restuccia, Silvia Venturiello, et al. Nosocomial acquisition of methicillin-resistant Staphyloccocus aureus (MRSA) and extended-spectrum beta-lactamase (ESBL) Enterobacteriaceae in hospitalised patients: a prospective multicenter study. *BMC Infectious Diseases*, 12(1):74, jan 2012.

[202] V. Schechner, T. Kotlovsky, M. Kazma, et al. Asymptomatic rectal carriage of blaKPCproducing carbapenem-resistant Enterobacteriaceae: Who is prone to become clinically infected? *Clinical Microbiology and Infection*, 19(5):451–456, 2013.

[203] Matthew S. Thiese. Observational and interventional study design types; an overview. *Biochemia Medica*, 24(2):199–210, 2014.

[204] Artur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[205] Nello Cristianini and John Shave-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[206] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-Supervised Learning with Deep Generative Models. pages 1–9, 2014.

[207] S H Walker and D B Duncan. Estimation of the probability of an event as a function of several independant variables. *Biometrika*, 54(July):167–179, 1967.

[208] Sandro Sperandei. Understanding logistic regression analysis. *Biochemia Medica*, 24(1):12–18, 2014.

[209] Bernhard Scholkopf, Koji Tsuda, and Jean-Philippe Vert. A primer on Kernel Methods. In Bernhard Scholkopf, Koji Tsuda, and Jean-Philippe Vert, editors, *Kernel Methods in Computational Biology*, pages 35–70. The MIT Press, 2004.

[210] Rosanna Upstill-Goddard, Diana Eccles, Joerg Fliege, and Andrew Collins. Machine learning approaches for the discovery of gene-gene interactions in disease data. *Briefings in Bioinformatics*, 14(2):251–260, 2013.

[211] Vladimir Guyon, Isabelle, Weston, Jason, Barnhill, Stephen, and Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(4):389–422, 2002.

[212] Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, and Muin J Khoury. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC medical informatics and decision making*, 10:16, 2010.

[213] Harris Drucker, Chris J C Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in Neural Information Processing Dystems*, 1:155–161, 1997.

[214] Diego Alejandro Salazar, Jorge Iván Velez, and Juan Carlos Salazar. Comparison between SVM and Logistic Regression: Which One is Better to Discriminate? *Revista Colombiana de Estadistica*, 35(2 SPEC. ISSUE):223–237, 2012.

[215] Lipo Wang. Support Vector Machines : Theory and Applications. *Springer*, 2005.

[216] Tin Kam Ho. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282, 1995.

[217] Stan Hatko. Random Forests. *European Journal of Mathematics*, 45(1):5–32, 2014.

[218] Yoav Freund and Robert F Schapire. A decision-theoretic generalization of on-line learning and application to boosting. *Journal of Computer and System Sciences*, 1(55):119–139, 1997.

[219] Jerome H. Friedman. Greedy Function Approximation: a Gradient Boosting Machine. *The Analysis of Statistics 2001*, 29(1189-1232), 2001.

[220] Michael A. Nielsen. *Neural networks and deep learning*. Determination Press, 2015.

[221] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[222] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning*. Springer, second edi edition, 2009.

[223] L. F. Carvalho, G. Fernandes, M. V O De Assis, J. J P C Rodrigues, and M. Lemes Proen??a. Digital signature of network segment for healthcare environments support. *Irbm*, 35(6):299–309, 2014.

[224] Richard Bellman. *Dynamic Programming*. 1957.

[225] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics*, 8(25), 2007.

[226] S Roweis and L Saul. Nonlinear dimensionality reduction by locally linear embedding. *Sciene*, 290(5500):2323–2326, 2000.

[227] Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11(1):51, 2011.

[228] Paul D Allison. Missing Data. *Quantitative Applications in the Social Sciences*, page 104, 2001.

[229] Paul D. Allison. *Missing data., Inc.; 2001.* Sage Publications, Thousand Oaks, CA, 2001.

[230] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9, 2012.

[231] Min Li, Haokui Zhou, Weiying Hua, et al. Molecular diversity of Bacteroides spp. in human fecal microbiota as determined by group-specific 16S rRNA gene clone library analysis. *Systematic and applied microbiology*, 32(3):193–200, may 2009.

[232] Anton Bankevich, Sergey Nurk, Dmitry Antipov, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

[233] Wenyu Zhang, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, and Bairong Shen. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS one*, 6(3):e17915, 2011.

[234] Keith R Bradnam, Joseph N Fass, Anton Alexandrov, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10, jul 2013.

[235] Yu-Chieh Liao, Shu-Hung Lin, and Hsin-Hung Lin. Completing bacterial genome assemblies: strategy and performance comparisons. *Scientific Reports*, 5:8747, 2015.

[236] Genivaldo Gz Silva, Bas E Dutilh, T David Matthews, Keri Elkins, Robert Schmieder, Elizabeth a Dinsdale, and Robert a Edwards. Combining de novo and reference-guided assembly with scaffold_builder. *Source code for biology and medicine*, 8(1):23, 2013.

[237] Monya Baker. De novo genome assembly: what every biologist should know. *Nature Methods*, 9(4):333–337, 2012.

[238] Aaron C.E. E Darling, Bob Mau, Frederick R Blattner, and Nicole T Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):1394–1403, 2004.

[239] R K Aziz, Daniela Bartels, A A Best, and Matthew DeJongh. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, 9(1):75, jan 2008.

[240] Thomas Brettin, James J. Davis, Terry Disz, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, 5, 2015.

[241] Ross Overbeek, Robert Olson, Gordon D. Pusch, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42(D1):206–214, 2014.

[242] Ea Zankari, Henrik Hasman, Salvatore Cosentino, et al. Identification of acquired antimicrobial resistance genes. *The Journal of Antimicrobial Chemotherapy*, 67(11):2640–4, 2012.

[243] A. Carattoli, E. Zankari, A. Garcia-Fernandez, et al. In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrobial Agents and Chemotherapy*, 58(7):3895–3903, 2014.

[244] Katrine Grimstrup Joensen, Flemming Scheutz, Ole Lund, Henrik Hasman, Rolf S Kaas, Eva M Nielsen, and Frank M Aarestrup. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *Journal of clinical microbiology*, 52(5):1501–10, 2014.

[245] Zhiqun Xie and Haixu Tang. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*, 33(21):3340–3347, 2017.

[246] David Arndt, Jason R. Grant, Ana Marcu, Tanvir Sajed, Allison Pon, Yongjie Liang, and David S. Wishart. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, 44(W1):W16–W21, 2016.

[247] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, page 9, 2012.

[248] Szymon M Kielbasa, Raymond Wan, Kengo Sato, Szymon M Kiebasa, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison Adaptive seeds tame genomic sequence comparison. *Cold Spring Harbor Laboratory Press*, 3(21):487–493, 2011.

[249] Travis E. Oliphant. Python for scientific computing. *Computing in Science and Engineering*, 9(3):10–20, 2007.

[250] F Pedregosa and G Varoquaux. Scikit-learn: Machine Learning in Python. *Journal ofMachine Learning Research*, 12:2825–2830, 2011.

[251] Fatemeh Shahkarami, Ahmad Rashki, and Zahra Rashki Ghalehnoo. Microbial susceptibility and plasmid profiles of methicillin-resistant Staphylococcus aureus and Methicillin-Susceptible S. aureus. *Jundishapur Journal of Microbiology*, 7(7):3–8, 2014.

[252] Young Duck Lee and Jong Hyun Park. Genome analysis of phage SMSAP5 as candidate of biocontrol for Staphylococcus aureus. *Korean Journal for Food Science of Animal Resources*, 35(1):86–90, 2015.

[253] Marian Varga, Roman Pantůček, Vladislava Růžičková, and Jiří Doškař. Molecular characterization of a new efficiently transducing bacteriophage identified in meticillin-resistant Staphylococcus aureus. *Journal of General Virology*, 97(1):258–268, 2016.

[254] Matthias Willmann, Mohamed El-Hadidi, Daniel H. Huson, Monika Schütz, Christopher Weidenmaier, Ingo B. Autenrieth, and Silke Peter. Antibiotic selection pressure determination through sequence-based metagenomics. *Antimicrobial Agents and Chemotherapy*, 59(12):7335–7345, 2015.

[255] H Ochman, J G Lawrence, and E a Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

[256] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):pp. 10–12, 2011.

[257] Alexander Herbig, Frank Maixner, Kirsten I Bos, Albert Zink, Johannes Krause, and Daniel H Huson. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. *bioRxiv*, page 50559, 2016.

[258] Andrew G. McArthur, Nicholas Waglechner, Fazmin Nizam, et al. The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, 57(7):3348–3357, 2013.

[259] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.

[260] Sébastien Boisvert, Frédéric Raymond, Elénie Godzaridis, François Laviolette, and Jacques Corbeil. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome biology*, 13(12):R122, 2012.

[261] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.

[262] Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 2010.

[263] Daniel Huson, Benjamin Albrecht, Caner Bagci, Irina Bessarab, Anna Gorska, Dino Jolic, and Rohan B.H. Williams. MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, 13(6), 2018.

[264] Ruud Jansen, Jan D A Van Embden, Wim Gaastra, and Leo M. Schouls. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, 43(6):1565–1575, 2002.

[265] Anders F Andersson and Jillian F Banfield. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science (New York, N.Y.)*, 320(5879):1047–50, 2008.

[266] Ibtissem Grissa, Gilles Vergnaud, and Christine Pourcel. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC bioinformatics*, 8:172, 2007.

[267] Christine Rousseau, Mathieu Gonnet, Marc Le Romancer, and Jacques Nicolas. CRISPI: A CRISPR interactive database. *Bioinformatics*, 25(24):3317–3318, 2009.

[268] Connor T. Skennerton, Michael Imelfort, and Gene W. Tyson. Crass: Identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Research*, 41(10), 2013.

[269] Junjie Qin, Ruiqiang Li, Jeroen Raes, et al. A human gut microbial gene catalogue established by metagenomic sequencing: Commentary. *Inflammatory Bowel Disease Monitor*, 11(1):28, mar 2010.

[270] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2009.

[271] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire Fraser-liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804–810, 2007.

[272] Stephen F Altschul, Warren Gish, The Pennsylvania, and Universily Park. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[273] Simon Roux, Francois Enault, Bonnie L Hurwitz, and Matthew B Sullivan. VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985, 2015.

[274] Heng Li. Minimap, 2016.

[275] Heng Li, Bob Handsaker, Alec Wysoker, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, aug 2009.

[276] Clovis Galiez. WIsh, 2017.

[277] Germán Bonilla-Rosso, Luis E. Eguiarte, David Romero, Michael Travisano, and Valeria Souza. Understanding microbial community diversity metrics derived from metagenomes: Performance evaluation using simulated data sets. *FEMS Microbiology Ecology*, 82(1):37–49, 2012.

[278] Florian Plaza Onate, Jean-Michel Batto, Catherine Juste, et al. Quality control of microbiota metagenomics by k-mer analysis. *BMC Genomics*, 16(1):183, 2015.

[279] Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(July 1928):379–423, 1948.

[280] Sara Jabbari, John T. Heap, and John R. King. Mathematical modelling of the sporulation-initiation network in Bacillus subtilis revealing the dual role of the putative quorum-sensing signal molecule PhrA. *Bulletin of Mathematical Biology*, 73(1):181–211, 2011.

[281] Falk Hildebrand, Axel Meyer, and Adam Eyre-Walker. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genetics*, 6(9), 2010.

[282] Joanne B. Emerson, Brian C. Thomas, Karen Andrade, Eric E. Allen, Karla B. Heidelberg, and Jillian F. Banfielda. Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Applied and Environmental Microbiology*, 78(17):6309–6320, 2012.

[283] J Mahillon and M Chandler. Insertion sequences. *Microbiology and molecular biology reviews : MMBR*, 62(3):725–74, 1998.

[284] Graham F Hatfull and Roger W Hendrix. Bacteriophages and their Genomes Graham. *Current Opinion in Virology*, 1(4):298–303, 2012.

[285] Ana Elena Pérez-Cobas, María José Gosalbes, Anette Friedrichs, et al. Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut*, 62(11):1–11, nov 2012.

[286] Sajia Akhter, Ramy K. Aziz, and Robert A. Edwards. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Research*, 40(16):1–13, 2012.

[287] Weizhong Li and Adam Godzik. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[288] MongoDB, 2015.

[289] N. R. Peace, D. A. Stahl, D. J. Lane, and G. J. Olsen. The analysis of natural microbial populations by ribosomal RNA sequences. *Advances in Microbial Ecology*, (9):1–55, 1986.

[290] Patrick Grupp, Theresa Anisja Harbig, Sina Beier, Anna Gorska, Isabell Flade, and Daniel H Huson. Bioinformatics support for the Tubiom community gut microbiome project. *PeerJ Preprints*, pages 1–9, 2016.

[291] Brian D. Ondov, Nicholas H. Bergman, and Adam M. Phillippy. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1):385, 2011.

[292] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *The Craft of Information Visualization*, pages 364–371, 2003.

[293] Anna Górska, Agnieszka Markowska-Zagrajek, Marcin Równicki, and Joanna Trylska. Scanning of 16S ribosomal RNA for peptide nucleic acid targets. *J. Phys. Chem. B*, page submitted, 2016.

[294] James C. Phillips, Rosemary Braun, Wei Wang, et al. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, dec 2005.

[295] Thomas E Cheatham and Matthew A Young. Simulation of Nucleic Acids : Successes, Limitations, and Promise. *Biopolymers*, 56(2001):232–256, 2000.

[296] Mark A Ditzler, Michal Otyepka, Jirì Sponer, and Nils G Walter. Molecular dynamics and quantum mechanics of RNA: conformational and chemical change we can believe in. *Accounts of chemical research*, 43(1):40–7, 2010.

[297] Thomas E Cheatham. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Current Opinion in Structural Biology*, 14(3):360–367, 2004.

[298] M Zacharias. Simulation of the structure and dynamics of nonhelical RNA motifs. *Current opinion in structural biology*, 10(3):311–7, jun 2000.

[299] Ryan L. Hayes, Jeffrey K. Noel, Udayan Mohanty, Paul C. Whitford, Scott P. Hennelly, José N. Onuchic, and Karissa Y. Sanbonmatsu. Magnesium fluctuations modulate RNA dynamics in the SAM-I riboswitch. *Journal of the American Chemical Society*, 134(29):12043–12053, 2012.

[300] Miroslav Krepl, Kamila Réblová, Jaroslav Koča, and Jiří Sponer. Bioinformatics and molecular dynamics simulation study of L1 stalk non-canonical rRNA elements: kink-turns, loops, and tetraloops. *The journal of physical chemistry. B*, 117(18):5540–55, may 2013.

[301] Paul Whitford, José Onuchic, and Karissa Sanbonmatsu. Connecting Energy Landscapes with Experimental Rates for Aminoacyl-tRNA Accommodation in the Ribosome. *Journal of the American Chemical Society*, 132(38):13170–13171, 2010.

[302] Neocles B Leontis, Jesse Stombaugh, and Eric Westhof. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic acids research*, 30(16):3497–3531, aug 2002.

[303] Anna Górska, Maciej Jasiński, and Joanna Trylska. MINT: Software to identify motifs and short-range interactions in trajectories of nucleic acids. *Nucleic Acids Research*, 43(17):5–6, 2015.

[304] Y. Nomura, M. Kajikawa, S. Baba, et al. Solution structure and functional importance of a conserved RNA hairpin of eel LINE UnaL2. *Nucleic Acids Research*, 34(18):5184–5193, jan 2006.

[305] Valentina Ahl, Heiko Keller, Steffen Schmidt, and Oliver Weichenrieder. Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation. *Molecular Cell*, 60(5):715–727, 2015.

[306] Gustavo Glusman, Hannah C Cox, and Jared C Roach. Whole-genome haplotyping approaches and genomic medicine. *Genome Medicine*, 73(6):505–506, 2014.

[307] Derek Aguiar, Wendy S W Wong, and Sorin Istrail. Tumor haplotype assembly algorithms for cancer genomics. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 3–14, 2014.

[308] Jacob O Kitzman, Alexandra P Mackenzie, Andrew Adey, et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature biotechnology*, 29(1):59–63, 2011.

[309] Sharon R. Browning and Brian L. Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[310] Bruce J. Walker, Thomas Abeel, Terrance Shea, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, 9(11), 2014.

[311] Andrew Consroe. jellyfish-matrix, 2015.

[312] The Academy of Medical Sciences. Stratified, personalised or P4 medicine: a new direction for placing the patient at the centre of healthcare and health education. *Summary of a joint FORUM meeting held on 12 May 2015*, (May):37, 2015.

Appendices

A List of abbreviations

A

AR	Antibiotic Resistance
ARG	Genes conferring Antibiotic Resistance
AMR	Antibiotic-Resistant Microorganisms

B

BLAST	Basic Local Alignment Search Tool

C

CBCB	Center for Bioinformatics and Computational Biology
CDC	Centers for Disease Control and Prevention
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats

E

ECDC	European Centre for Disease Prevention and Control
ESBL	Extended-Spectrum β-lactamases
EU	European Union

F

FDA	Food and Drug Administration
FP7	7^{th} Framework Program

H

HGT	Horizontal Gene Transfer
HMP	Human Microbiome Project

I

IBD	Inflammatory Bowel Disease
INDELs	Insertion/Deletions
IR	Inverted Repeats
IS	Insertion Sequence

L

LCB	Locally Collinear Blocks
LD	Linkage Disequilibrium
LTCF	Long-Term Care Facility

M

MAR	data Missing At Random
Mb	10^6 base pairs
MCAR	data Missing Completely At Random
MDR	Multi-Drug Resistant bacteria
MGE	Mobile Genetic Elements
MRSA	Methacilin-resistant *Staphylococus Aureus*

N

NMAR	data Not Missing At Random
NN	Neural Networks

P

PC	Principal Component

R

RPPs	Ribosomal Protection Proteins
RNAP	RNA Polymerase

S

SATURN	Specific Antibiotic Therapies on the prevalence of hUman host ResistaNt bacteria
SNPs	Single Nucleotide Polymorphisms
SVM	Support Vector Machines

R

RF	Random Forest
RFr	Retrotransposition frequency

V

VLP	Virus Like Particles

W

WHO	World Health Organisation
WGS	Whole Genome Sequencing
WP	Work packages in the SATURN project

B Contributions

Prof. Daniel Huson advised me during all of the work throughout my Ph.D. I was a member of the International Max Planck Research School and received input and support from my Theses Advisory Committee that included Prof. Eliza Izaurralde and Prof. Richard Neher from the Max Planck for Developmental Biology in Tübingen.

Chapter 3 | MRSA from colonization to infection

Samples were collected within the SATURN project. As all of the selected samples came from Serbia, they were initially processed by the Serbian team from the Clinical Centre of Serbia under the supervision of dr. Biljana Jovanovic. The sequencing was carried out by the group of Prof. Surbhi Malhotra-Kumar from the University of Antwerp. I was responsible for the assembly and analysis of the data.

Chapter 4 | Gut mobileome under antibiotics

Dr. Silke Peter and Prof. Matthias Willmann planned and carried out the collection of the samples and sequencing, for the original ASARI study. Dr. Silke Peter processed and isolated DNA for the phage sequencing, that was carried out by Prof. Robert Schlaberg from Department of Pathology, University of Utah, and Institute for Clinical and Experimental Pathology, UT, Salt Lake City, USA.

Chapter 5 | SATURN

Prof. Evelina Tacconnelli coordinated the work-package 4 of the SATURN project. The data were collected by the SATURN project team and members of Prof. Tacconelli's research group. Dr. Primrose Beryl Gladstone performed an analysis of the data based using the logistic regression. Prof. Bernhard Schölkopf, Max Planck for Intelligent Systems advised on how to construct the data vector initially for SVM. The RF part was discussed with Prof. Michael Cummings form the University of Maryland.

Chapter 6 | Tübiom project

Tübiom project was organized by a collaboration between several research groups in Tübingen and the CeMeT GmBH, which also performed sequencing. Our group was responsible for creating websites, managing databases and microbiome profile computation. Sina Beier wrote the analysis pipeline, the website and the underlying databases were set up by Patrick Groupp (MSc student). Theresa Harbig and I designed and programmed the data visualizations. We heavily relied on the

expertise of Dr. Isabel Flade from CeMeT GmBH, who was also performing sample processing and sequencing. The website, database, and the analysis pipeline were hosted using the CeGaT computational infrastructure.

Chapter 7 | Other projects

Analysis of the MD trajectories of RNA molecules

The molecular dynamics protocols were based on the work done for the MSc projects. Dr. Oliver Weichenrieder and Steffen Schmidt from the Max Planck Institute for Developmental Biology Tübingen suggested the two molecules used for the analysis.

Phase the turtle!

Sequencing of the Diamondback terrapin turtle and her progeny was carried out by the University of Maryland. The initial bioinformatics work including assembly was performed by Ph.D. students in the Center for Bioinformatics and Computational Biology. I came to the project during the 2^{nd} Bioinformatics Exchange Students and Teachers summer school in Maryland. Then, I went on an internship in Center for Bioinformatics and Computational Biology Maryland in the February 2015 where we continued working on the projects with Victoria Cepeda, supervised by the three professors: Mihai Pop, Michael Cummings, and Steve Mount.

C Manuscripts

In preparation:

Tacconelli, E., **Górska, A.**., Angelis, G. De, Lammens, C., Restuccia, G., Huson, D. H., ... Kazma, M. Estimating the Association between Antibiotic Exposure and Colonisation with Antibiotic-resistant Bacteria using Machine-learning Methods

Górska, A., Willmann, M., Schlaberg, R., Huson, D.H., Autenrieth, I., Peter, S. Horizontal gene transfer in the human gut microbiome during the antibiotic therapy.

Published:

Górska, A., Peter, S., Willmann, M., Autenrieth, I., Schlaberg, R., Huson, D.H., 2018. Dynamics of the human gut phageome during antibiotics treatment. Comput. Biol. Chem.

Huson, D., Albrecht, B., Bagci, C., Bessarab, I., **Górska, A.**, Jolic, D., Williams, R.B.H., 2018. MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. Biol. Direct 13.

Grupp, P., Harbig, T.A., Beier, S., **Górska, A.**, A., Flade, I., Huson, D.H., 2016. Bioinformatics support for the Tubiom community gut microbiome project. PeerJ Prepr. 1–9. *Preprint*

Huson, D.H., Beier, S., Flade, I., **Górska, A.**, El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., Tappu, R., 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Comput Biol 12, 4–12.

Talens-Perales, D., **Górska, A.**, Huson, D.H., Polaina, J., Marín-Navarro, J., 2016. Analysis of domain architecture and phylogenetics of family 2 glycoside hydrolases (GH2). PLoS One 11, 1–17.

D Supplementary Material | MRSA from colonization to infection

Table D.1: Initial statistics of the spades-corrected reads. Sample date column contains the dates the samples were taken from the patient. Size was computed via simple summing lengths of all of the reads in the sample, and coverage was estimated via dividing the size by the size of the reference genome (2 902 619). Both files assigned to a single sample should contain a roughly similar number of reads, have similar size and coverage, what is the case for all of the samples. The samples were taken in 2012 (SE1582) and 2013 for all other samples. All samples were sequenced in 2015 dates.

Sequencing date	Sample date	Samples		Number of reads	Amount of sequence	Coverage
SE1582						
25-03	06-11	BAL	R1	301,838	1,668,811,552	575
			R2	312,579	1,668,202,571	575
18-04	19-11	NT4_S4	R1	219,258	1,191,793,143	411
			R2	303,431	1,180,272,344	407
SE1884						
18-04	12-02	BAL_S1	R1	65,109	354,617,126	122
			R2	75,267	353,292,007	122
18-04	22-02	NT4_S2	R1	134,398	766,990,881	264
			R2	233,981	752,845,860	259
15-08	26-02	ND_S4	R1	238,724	1,271,720,312	438
			R2	298,514	1,265,075,493	436
SE1890						
18-04	11-02	BAL_S3	R1	128,274	775,085,477	267
			R2	187,644	767,205,491	264
13-05	14-02	Nt3_S4	R1	236,144	1,028,737,930	354
			R2	313,531	1,019,725,018	351
13-05	27-03	ND_S3	R1	210,451	1,239,657,010	427
			R2	389,925	1,218,075,700	420
SE1895						
15-08	11-02	N_S1	R1	117,121	752,966,412	259
			R2	158,140	747,909,730	258
15-08	12-02	Nt2_S2	R1	89,403	629,749,620	217
			R2	135,856	624,382,700	215
13-05	21-03	ND_S3	R1	176 601	1 315 932 715	453
			R2	291 804	1 301 823 509	448
SE2054						
13-05	05-04	N3_S2	R1	135,279	816,581,530	281
			R2	227,854	804,874,740	277
13-05	12-04	N_1	R1	170,486	897,273,708	309
			R2	251,747	887,178,336	306

Table D.2: Assembly statistics the Spades runs for the MRSA strains isolated from patients from different timepoints across the antibiotics therapy.

Sample	Num.	Scaffolds	Size	Min.	Max.	Med.	Avg.	N50	N90
Assembly statistics for reads that did map to rRNAs.									
SE1582									
BAL_S1	178	11	4,019	204	1,217	240	365	321	208
NT4_S4	140	10	2,972	193	747	243	297	272	211
SE1884									
BAL_S1	267	7	5,529	197	1,976	712	789	1,065	596
NT4_S2	143	7	3,467	205	1,133	295	495	865	234
ND_S4	332	15	5,511	125	831	248	367	650	190
SE1890									
BAL_S3	196	19	5,575	9	878	248	293	408	205
Nt3_S4	335	7	5,884	125	1,593	898	840	977	748
ND_S3	296	11	5,513	125	956	477	501	645	275
SE1895									
N_S1	444	10	6 445	36	2,043	413	644	1,381	260
Nt2_S2	452	18	7,237	54	1,411	185	402	1,068	156
ND_S3	228	15	4,840	70	1,015	259	322	396	166
SE2054									
N3_S2	366	12	6,237	61	2,568	304	519	835	304
N_S1	372	10	6,044	95	2,059	397	604	1,294	231
Assembly statistics for reads that did not map to rRNAs.									
SE1582									
BAL_S1	4,677,829	1,038	3,063,988	36	178,725	224	2,951	49,248	7,082
NT4_S4	3,215,327	655	2,944,558	36	178,687	209	4,495	51,451	13,434
SE1884									
BAL_S1	1,027,738	792	2 919 473	36	133,608	77	3,686	51,367	13,415
NT4_S2	2,070,723	1 393	3 078 254	36	133,720	216	2,209	43,457	4,699
ND_S4	3,874,473	668	2,897,961	36	133,780	80	4,338	51,538	13,544
SE1890									
BAL_S3	2,177,776	863	2,982,277	36	136,189	198	3,455	47,221	12,171
Nt3_S4	3,133,320	449	2,908,787	36	133,745	187	6,478	49,236	13,544
ND_S3	3,734,560	423	2,915,227	36	136,189	200	6,891	49,236	13,415
SE1895									
N_S1	2,299,719	642	2,951,860	36	133,745	207	4,597	51,484	13,415
Nt2_S2	1,972,950	673	2,937,476	36	136,189	144	4,364	49,236	13,415
ND_S3	3,907,472	814	2,927,208	36	133,993	54	3,596	49,248	12,908
SE2054									
N3_S2	2,431,691	677	2,863,318	36	202,540	71	4,229	49,846	15, 180
N_S1	2,696,889	603	2,854,612	36	173,193	71	4,734	49,580	14, 701

Table D.3: Assembly statistics after assembly and ordering.

Sample	Num.	Size	Min.	Max.	Med.	Avg.	N50	N90
SE1582								
BAL_S1	914	3,108,182	36	1,423,316	219	3,400	868,898	541,419
NT4_S4	539	2,938,646	36	1,451,782	202	5,452	602,353	54,206
SE1884								
BAL_S1	655	2,976,587	36	1,524,182	59	4,544	1,524,182	459,024
NT4_S2	1 251	3,048,306	36	1,437,734	212	2,436	872,009	458,706
ND_S4	513	2,816,844	36	1,035,005	64	5,490	785,834	164,148
SE1890								
BAL_S3	728	2,919,187	9	1,438,221	191	4,009	719,175	108,441
Nt3_S4	298	2,783,330	36	944,927	70	9,340	719,758	451,414
ND_S3	294	2,842,405	36	949,031	105	9,668	868,639	440,875
SE1895								
N_S1	503	2,844,719	36	1,449,745	195	5,655	1,449,745	458,199
Nt2_S2	541	2,953,037	36	1,034,270	71	5,458	779,674	448,012
ND_S3	670	2,839,476	36	944,732	48	4,238	870,185	440,799
SE2054								
N3_S2	578	2,756,233	36	873,630	58	4,768	324,945	123,733
N_S1	499	2,744,829	36	880,579	54	5,500	579,290	125,300
Reference		2,821,361						

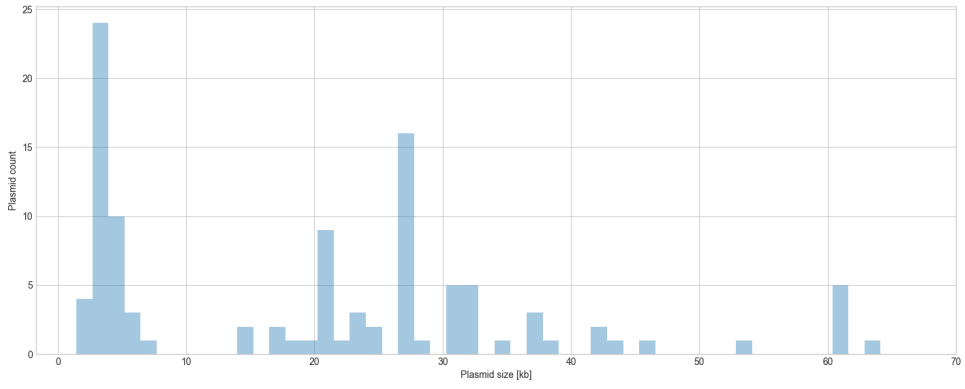


Figure D.1: Distribution of the size of the complete plasmids of the *S. Aureus* found in NCBI (106) as of September 2018.

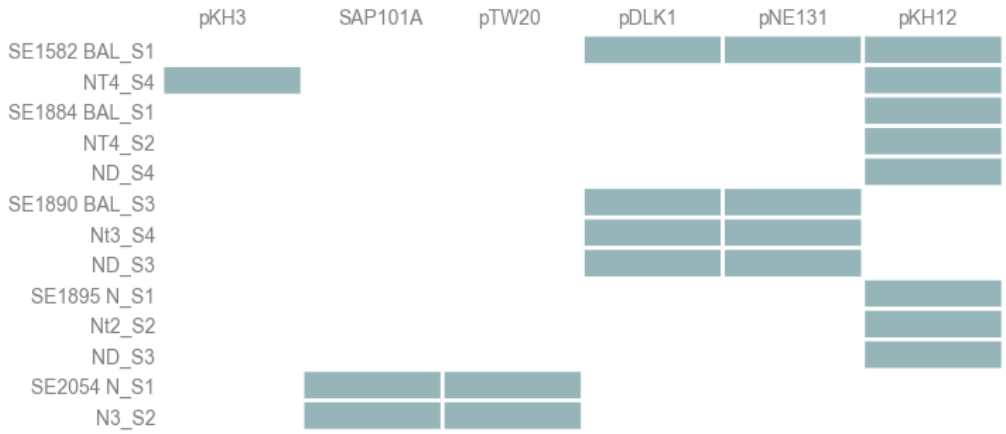


Figure D.2: Presence/absence heatmap for plasmids based on the Plasflow runs for reads.

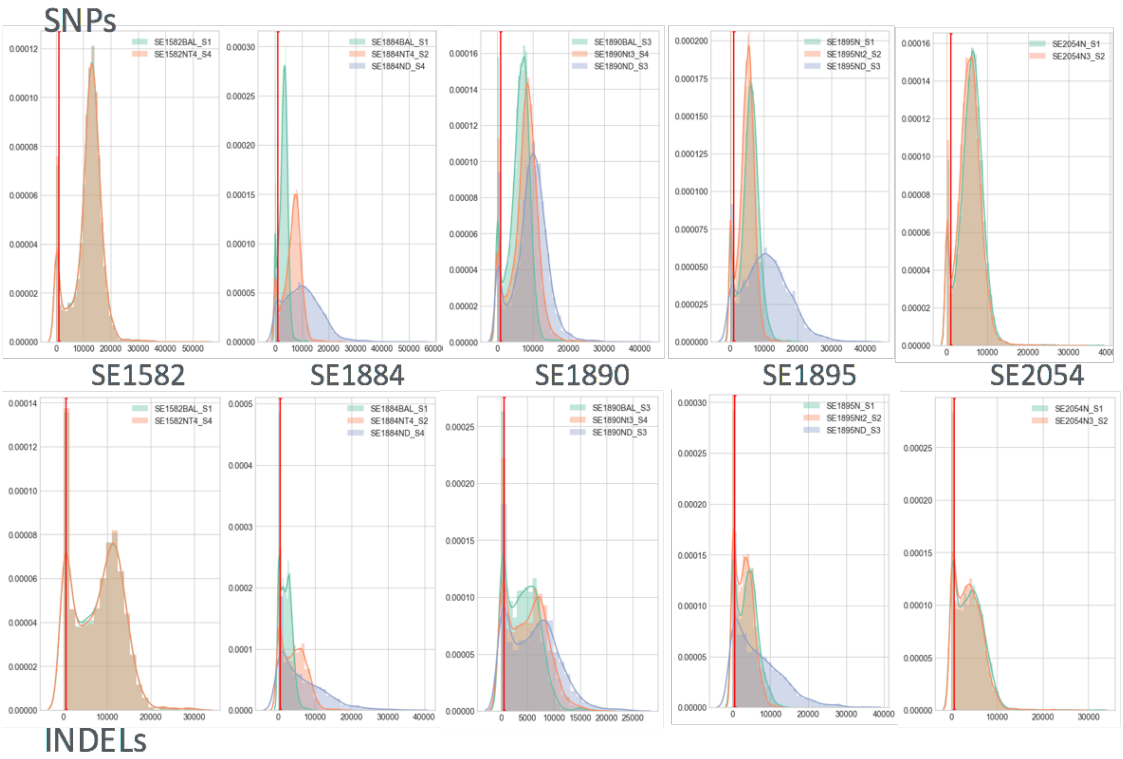
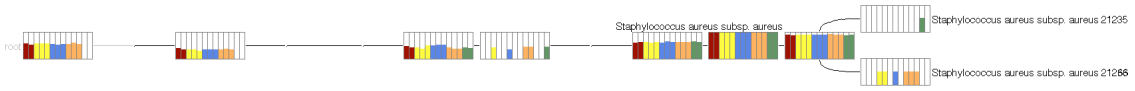
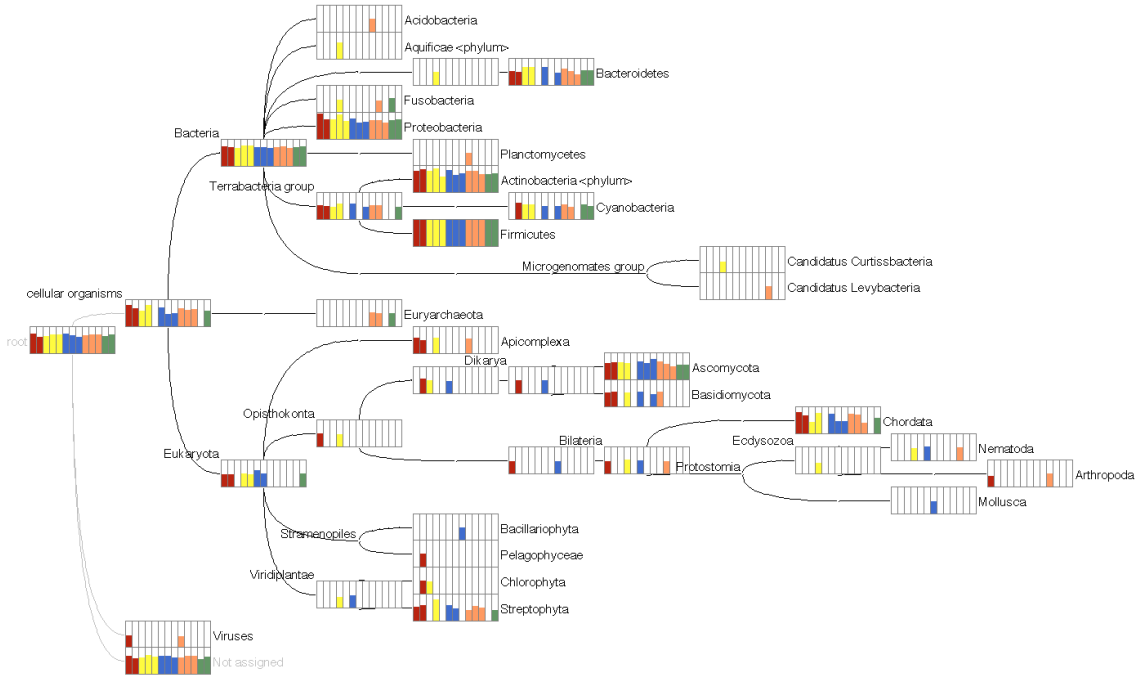


Figure D.3: Histogram of SNP and INDEL qualities from FreeBayes. The quality is defined as $-log_{10}x$ where x is probability the output allele is wrong. Red lines denotes cutoff: 1000 for SNPs and 610 for INDELs - everything lower then cutoff is excluded from analysis.



(a) Genomes



(b) Plasmids

Figure D.4: LAST+MEGAN-LR taxonomic annotation of the genomic and plasmid scaffolds.

E Supplementary Material | Gut mobileome under antibiotics pressure

Table E.4: Assembly quality statistics for the *pooled* and *separated* assembly stategties for k-mer 25.

	K	#Scaff.	Max. scaff.	N50	Tot. length	Predicted genes
Pooled assembly						
Phage | A	23	832,443	511,149	14,057	89,507,509	44,550
	25	860,417	490,755	14,997	89,442,525	876,730
	27	752,382	456,089	15,990	87,633,581	778,730
	29	636,945	453,097	16,698	86,287,015	673,591
Phage | B	23	677,524	283,536	16,113	114,168,468	28,661
	25	714,822	414,336	16,930	112,797,718	775,783
	27	652,966	408,617	18,208	110,874,669	718,350
	29	579,888	424,240	18,616	109,158,037	650,126
Separated assembly						
Phage | A	0	153,500	178,458	1,786	13,521,179	156,536
	1	217,070	240,912	3,287	12,399,639	218,380
	3	408,478	245,922	19,324	25,527,969	400,043
	6	262,850	268,573	18,506	24,635,812	267,716
	+2	321,755	254,477	18,899	33,255,036	327,199
	+28	134,118	129,394	5,704	11,300,454	136,972
	SUM	1,497,771	268,573	-	120,640,089	1,506,846
Phage | B	0	150,953	166,135	2,955	8,415,674	151,092
	1	175,657	166,088	6,914	7,470,370	174,550
	3	208,154	222,378	5,777	14,071,527	210,950
	6	163,288	295,336	29,993	30,182,804	179,575
	+2	117,808	122,256	6,498	23,239,384	128,266
	+28	426,024	204,409	6,565	63,875,494	457,207
	SUM	1,241,884	295,336	-	147,255,253	1,301,640
Microbiome | A	0	300,335	146,934	3,432	70,366,968	344,741
	1	308,781	148,061	2,633	70,100,769	353,686
	3	174,283	117,093	5,233	51,539,491	210,535
	6	222,303	115,288	3,787	68,088,430	268,335
	+2	132,288	91,584	5,934	58,993,467	173,594
	+28	283,884	86,739	1,568	61,614,432	322,699
	SUM	1,421,874	148,061	-	380,703,557	1,673,590
Microbiome | B	0	1,052,066	95,632	1,667	175,505,980	1,139,007
	1	584,744	95,310	1,876	101,185,762	650,853
	3	565,967	93,048	2,469	76,529,178	616,995
	6	181,045	77,360	3,718	58,126,443	220,381
	+2	114,994	70,173	3,631	40,908,611	141,944
	+28	551,686	122,815	2,999	129,162,426	639,726
	SUM	3,050,502	122,815	-	581,418,400	3,408,906

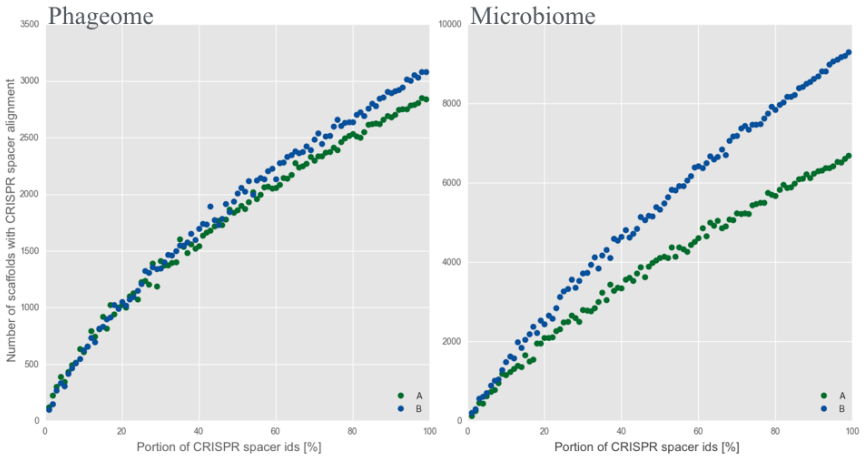


Figure E.5: Rarefaction curves of the number of scaffolds with an alignment to any of the CRISPR spacers vs. portion of whole spacer database.

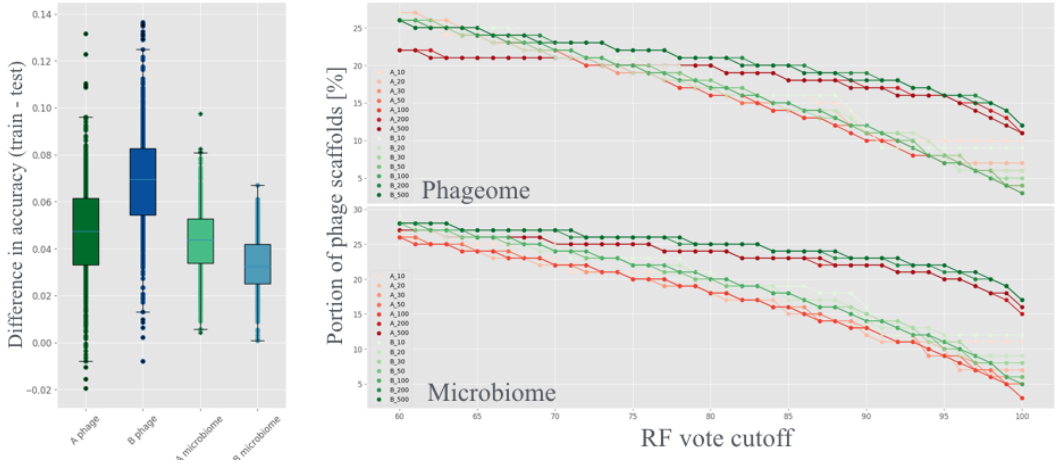


Figure E.6: RF overfitting control. a) Box plot presents distributions of difference in accuracy between the train and the test sets. b) A portion of phage scaffolds across different cutoffs across the number of RF runs and two max_depth parameters: None (10 to 100) and 5 (200 and 500).

Table E.5: Random Forest (RF) parameters.

Parameter	Value	Description
n-trees	10,000	Forest size
max_depth	5	Maximal depth of the trees
max_features	3	Number of features considered for the best split
min_samples_split	2	Minimal number of samples required for splitting
runs	500	Number of RF runs.
vote cutoff	90%	Portion of positive predictions required to denote a scaffold as phage

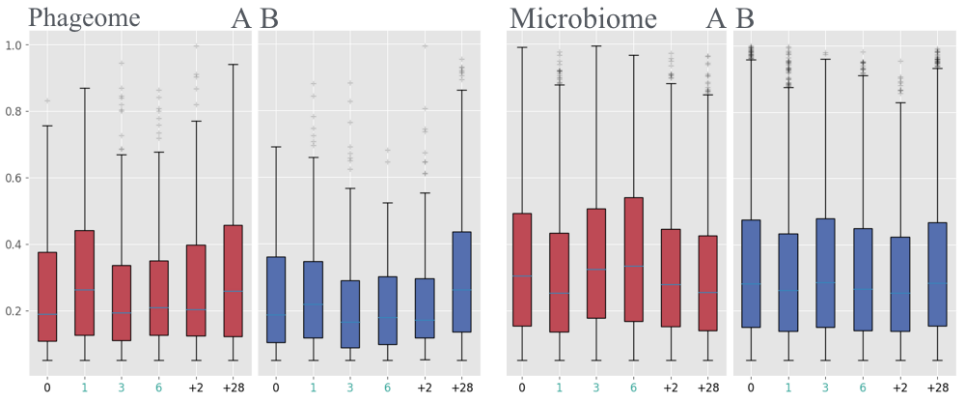


Figure E.7: VirSorter p-values for scaffolds selected by RF.

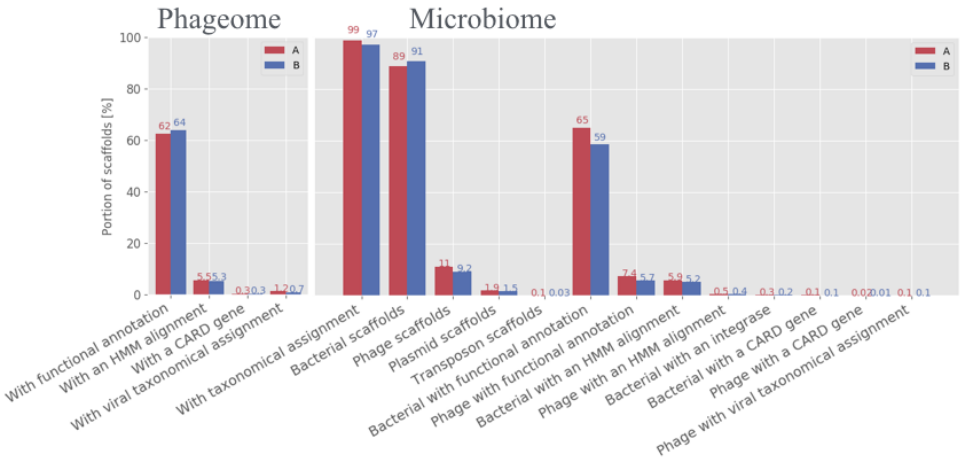


Figure E.8: Proportions of scaffolds of the two sets for all groups undergoing the abundance trajectory analysis.

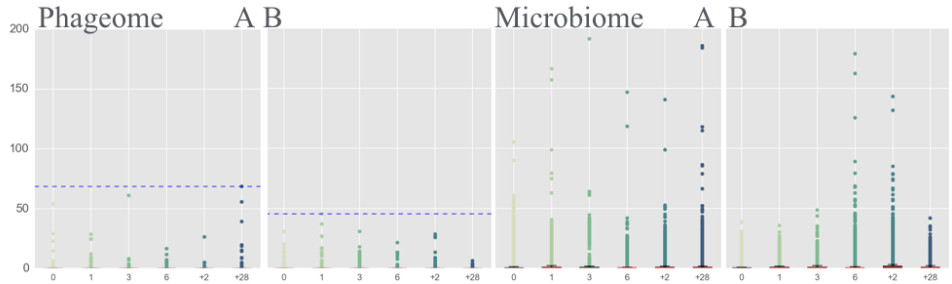


Figure E.9: Average coverage (abundance measurement) per scaffold. The blue line denotes the highest value in the Phageome and the red denote the average.

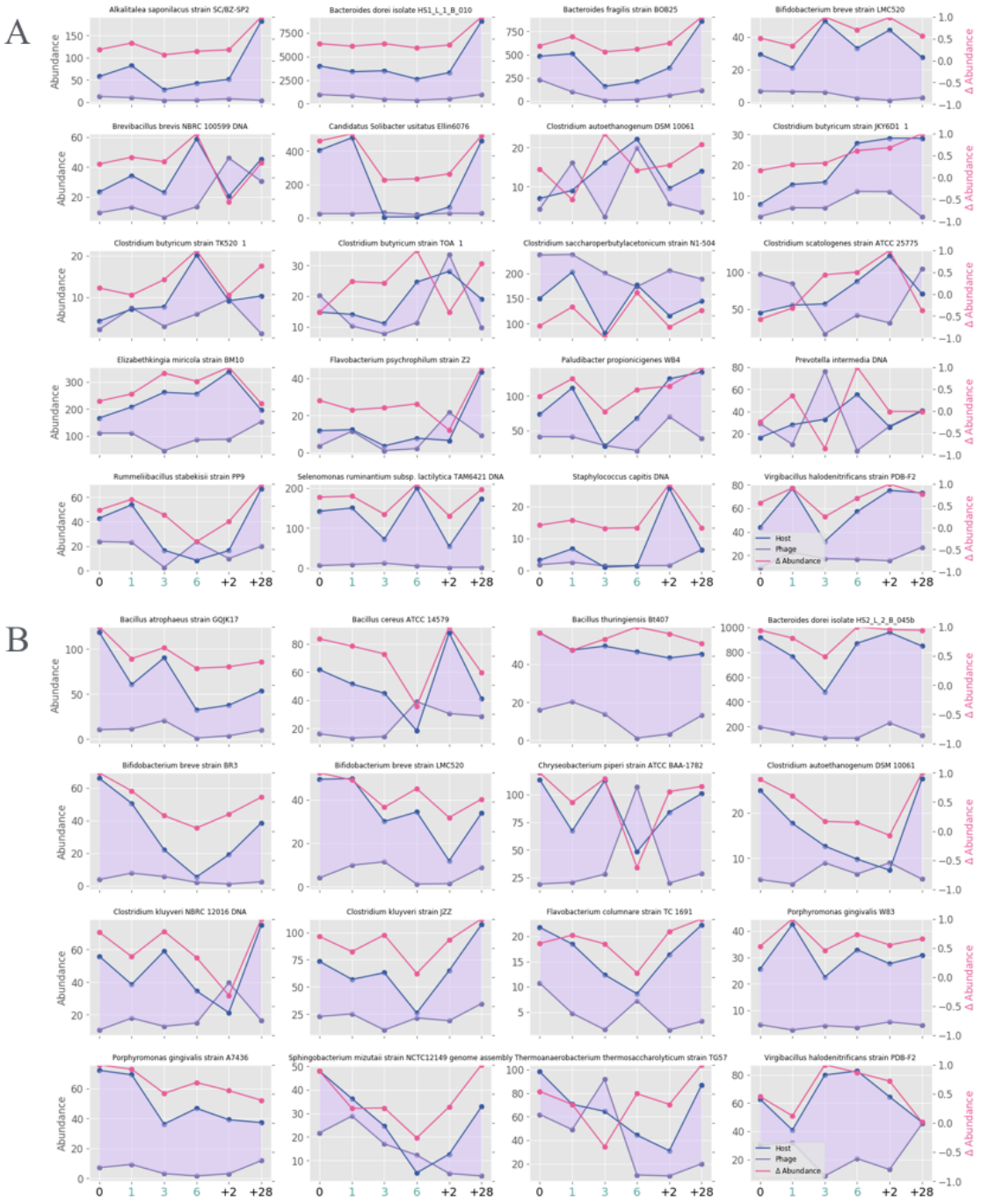


Figure E.10: Abundance of the phages and their bacterial hosts detected by WiSH, along with the trajectories of their relative difference in abundance.

F Supplementary Material | SATRUN

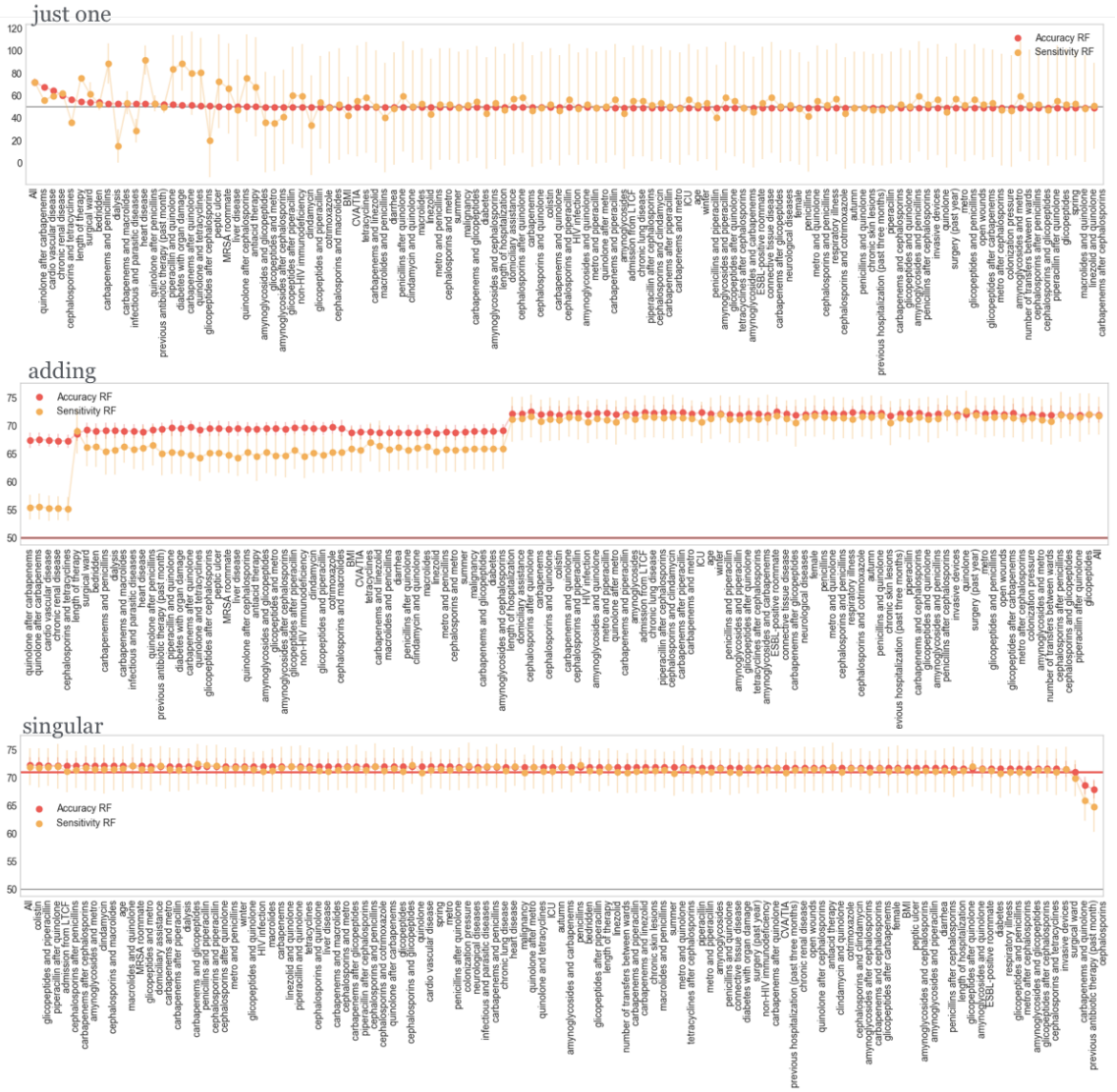


Figure F.11: Average accuracy along with the standard deviation in the three schemes of the feature selection.

G Supplementary Material | Tübiom project

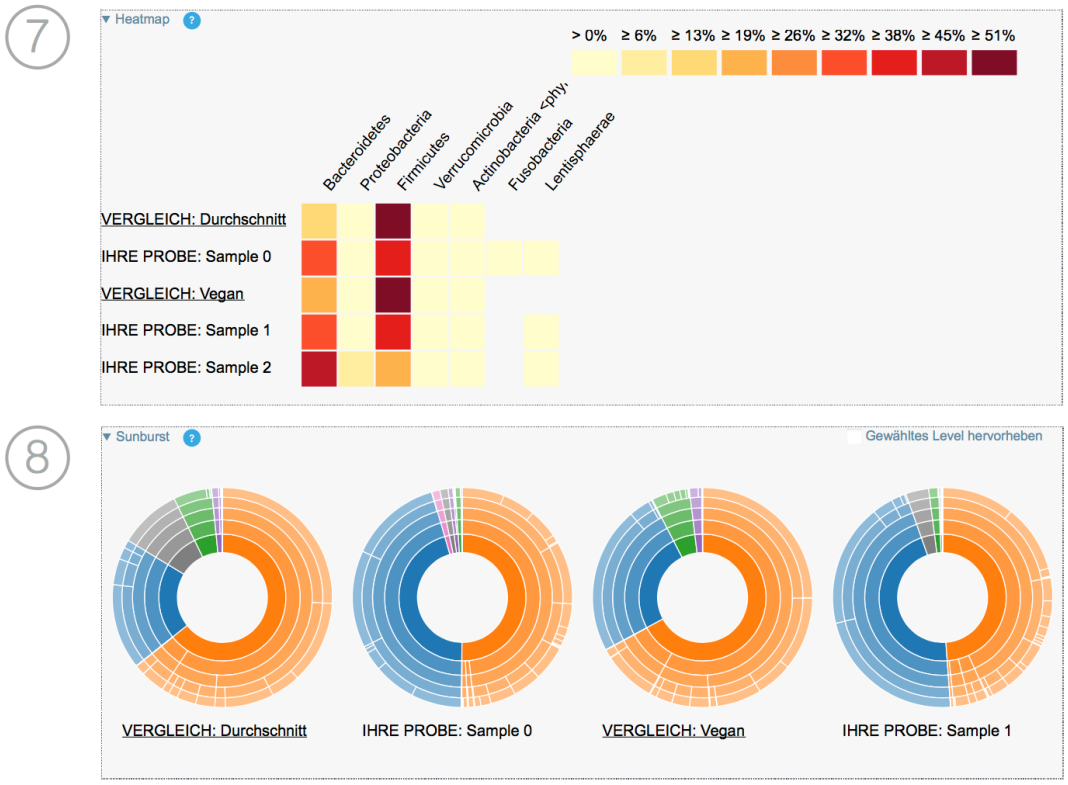


Figure G.12: *Advanced plots* of the Tübiom project. The plots were hidden in the default view. They present profiles of the samples and general profiles the user selected previously.

H Supplementary Material | Phase the turtle!

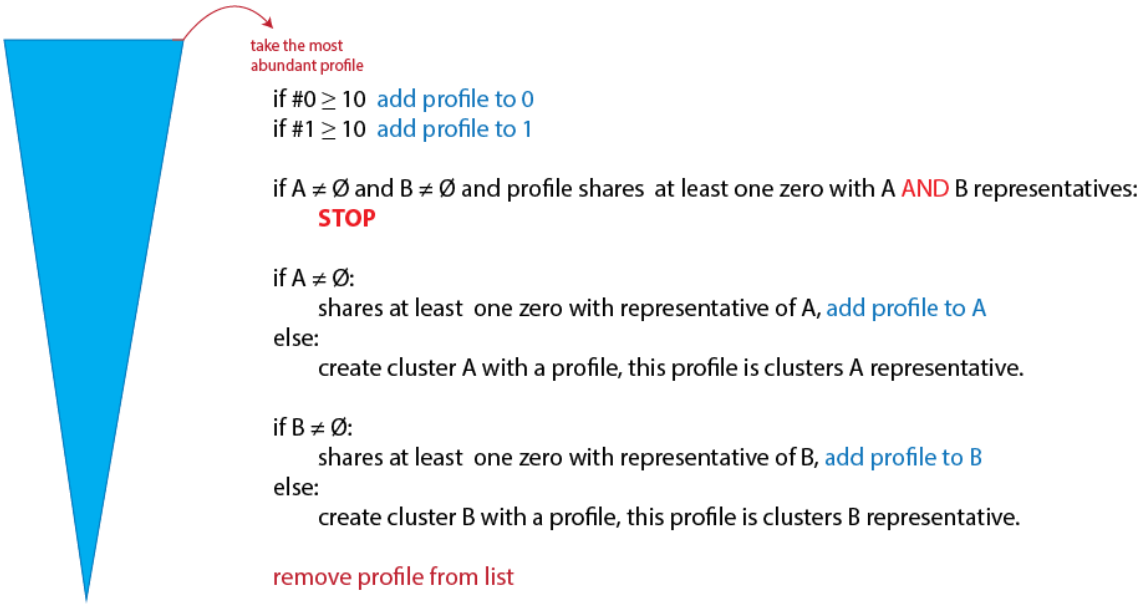


Figure H.13: Profile clustering algorithm. Firstly, the profiles are reversely sorted by abundance. Program classifies one, the most abundant profile, at the time. At each step, it decides if to assign a profile to one of the four clusters or to stop. After processing, the profile is removed from the list. Profiles containing mostly zeros or ones is assigned to an uninformative cluster. The first encountered informative profile is assigned to cluster A, and this profile becomes the representative for the cluster. Analogously the cluster B is created with the first informative profile that did not share zeros with the A's representative. If the profile shares zeros with the representatives of both clusters the program includes that and all further profiles into the erroneous cluster.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Masterarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum Unterschrift