# Addressing Microbial Threats through Genomics, Metagenomics, and Bioinformatics

Dissertation zur Erlangung des Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Mathematik und Informatik der
Universität des Saarlandes

vorgelegt von

Georges Pierre Schmartz

Saarbrücken,
2024

|  |  |
|---|---|
| **Dean of the faculty** | Prof. Dr. Roland Speicher |
| **Date of the colloquium** | Nov. 11th 2024 |
| **Chair of the commitee** | Prof. Dr. Volkhard Helms |
| **Reviewers** | Prof. Dr. Andreas Keller<br>Prof. Dr. Jilles Vreeken<br>Prof. Dr. Paul Wilmes |
| **Academic Assistant** | Dr. Alexander Gress |

*Dedicated to my fiancee.*

# Abstract

Bacteria evolved for millions of years with pathogens plaguing humankind throughout all of its history. By comparison, the scientific community could only prove that individual bacterial pathogens cause infectious diseases around 200 years ago. With the market release of the first antibiotics, effective treatment against a wide range of bacterial infections was accessible. Unfortunately, antibacterial-resistant pathogens emerged, capable of surviving existing antibiotic treatments initiating a high-stakes arms race between pathogens and the scientific community. The rapid dissipation of antibiotic resistances and the slow innovation cycle in antibiotic research, made researchers, governments, and health associations voice their concerns on the sustainability of antibiotic treatments. With an estimated 5 million deaths associated with antimicrobial resistance in 2019, projections predict up to 10 million deaths in 2050.

The advent of next-generation sequencing and the developments in the field of metagenomics propagated the interest in microbial communities from ecology into clinical microbiology. While moving away from 16S amplicon sequencing to complete metagenome shotgun sequencing and now to long-read sequencing, our understanding of the role of the human microbiome is constantly improving. More and more community compositions of different body environments keep being associated with diseases. However, our causal understanding of the human-microbiome interactions is likely still in its infancy as the microbiome displays an enormous diversity of molecules. Among these, metabolites, especially those encoded by biosynthetic gene clusters, have been highlighted as potential avenues to uncover new antimicrobial compounds.

Within this doctoral thesis, we present a total of ten research projects that fight against the bacterial threat distinguishing three actionable pillars. The first pillar focuses on the support of ongoing applied studies in the field of clinical microbiology. In this research context, we study antimicrobial-resistant pathogens, new emerging pathogens as well as two dietary interventions. Within resistant clinical isolates, we were able to identify resistance gene-carrying plasmids. The dietary interventions we assessed did not yield any differentially abundant species after dietary interventions. However we did see significant differences comparing e.g. baseline Parkinson's disease to control patients.

The second pillar aims to achieve our research network's long-term

goals in the area of natural compound research through data analysis of metagenomic sequencing data. In two projects, across several host species, disease cohorts, and biospecimens we search for potentially disease-implicated or protective natural products. We achieve this by genome mining for relevant biosynthetic gene clusters through large amounts of metagenomic assembled short-read sequencing data derived from over two thousand samples.

Lastly, the third pillar enables of other researchers in the field of microbiology through the development and maintenance of accessible online web services and databases. In this context, we updated a popular natural plasmid database as well as the only online metagenomic binning web service. Further, we developed a new online web service for short-read metagenomic sequencing data, while successfully addressing the challenges of large data transmission.

In conclusion, the herein presented works contributed several applied research studies in the field of clinical microbiology enabled other researchers in the same research area, and potentially lay the groundwork for the discovery of new clinically relevant natural products.

# *Zusammenfassung*

Bakterien evolvieren seit Millionen von Jahren, wobei einige Pathogene die Menschheit schon immer plagten. Erst vor etwa 200 Jahren gelang es jedoch zu zeigen, dass Bakterien tatsächlich für die Entstehung von Infektionskrankheiten verantwortlich sind. Mit der späteren Vermarktung erster Antibiotika wurde eine effektive Behandlung gegen eine Vielzahl von Bakterien möglich. Nach dieser Erfolgsgeschichte wurden jedoch leider resistente Pathogene festgestellt, die eine Behandlung mit bekannten Antibiotika überlebten. Somit begann ein Wettlauf zwischen Bakterien und der wissenschaftlichen Gemeinschaft. Die schnelle Verbreitung von antibakteriellen Resistenzen und die langsame Entwicklung neuer Antibiotika führen zu Zweifeln in Forschungspolitik und Gesundheitsorganisationen, ob nachhaltige Antibiotikaforschung noch möglich ist. Mit geschätzten 5 Millionen Toten im Jahr 2019, die mit Antibiotikaresistenzen in Verbindung gebracht wurden, liegen mögliche Vorhersagen für 2050 bei bis zu 10 Millionen.

Die Erfindung der Sequenzierung der nächsten Generation und der Metagenomik hat das Interesse an mikrobiellen Gemeinschaften über die Grenzen der Ökologie hinweg bis in die klinische Mikrobiologie erweitert. Durch den Übergang von der 16S-Amplikon-Sequenzierung zur Metagenom-Shotgun-Sequenzierung und schließlich zur Long-Read-Sequenzierung verbessert sich unser Verständnis des menschlichen Mikrobioms stetig. Somit werden immer mehr Assoziationen zwischen Mikrobiomen und Krankheiten hergestellt. Unser Verständnis über die Kausalität dieser Zusammenhänge steckt jedoch wahrscheinlich noch in den Kinderschuhen, da die Diversität der von den Mitgliedern des Mikrobioms ausgeschiedenen Moleküle enorm ist. Besonders jene Metabolite, deren Kodierung in Biosyntheseclustern liegt, gelten auch als potenzielle Kandidaten neuer antimikrobieller Wirkstoffe.

Im Rahmen dieser Doktorarbeit präsentieren wir zehn Forschungsprojekte, die sich auf drei Kernpunkte im Kampf gegen bakterielle Pathogene konzentrieren. Der erste Punkt betrifft die Unterstützung laufender angewandter Studien der klinischen Mikrobiologie. In diesem Kontext untersuchen wir antimikrobiell resistente Krankheitserreger, neu auftretende Erreger sowie zwei Diätinterventionen. Bei den resistenten klinischen Isolaten konnten wir Resistenzgene in Plasmiden identifizieren. Während wir keine unterschiedlich abundanten Spezies in den Interventionen beobachten konnten, sahen wir signifi-

kante Unterschiede im Vergleich zwischen z. B. Parkinson-Patienten und Kontrollpatienten.

Der zweite Punkt bemüht sich um das Erreichen der langfristigen Ziele unseres Forschungsnetzwerks auf dem Gebiet der Erforschung von Naturstoffen mittels metagenomischer Sequenzierungsdaten. In zwei Projekten suchen wir über mehrere Wirtsspezies, Krankheitskohorten und Bioproben hinweg nach krankheitsauslösenden oder vorbeugenden Naturstoffen. Dieses Ziel erreichen wir durch gezielte Suche nach relevanten biosynthetischen Genclustern in großen Mengen metagenomisch assemblierter Short-Read-Sequenzierungsdaten von über zweitausend Proben.

Zuletzt unterstützen wir andere Wissenschaftler im Feld der Mikrobiologie durch Entwicklung und Instandhaltung von Online-Webdiensten und Datenbanken. In diesem Zusammenhang haben wir eine Datenbank für natürlich vorkommende Plasmide und ein metagenomisches Binning-Tool aktualisiert. Außerdem haben wir einen neuen Online-Webdienst für Short-Read-Metagenom-Sequenzierungsdaten entwickelt und dabei erfolgreich die Herausforderungen der Übertragung großer Datenmengen bewältigt.

Zusammenfassend lässt sich sagen, dass die hier vorgestellten Arbeiten zu mehreren angewandten Forschungsstudien im Bereich der klinischen Mikrobiologie beigetragen haben, Kollegen im selben Forschungsbereich unterstützt haben und hoffentlich zur Entdeckung neuer relevanter Naturstoffe beigetragen haben.

# *Scientific publications*

The result section of the thesis includes the following peer-reviewed journal papers: [1–7]. Further, the following papers are not yet published at the time of writing: [8–10] Equal first and last author contributions are highlighted with the dagger symbol ($^\dagger$).

1. Jacqueline Rehner$^\dagger$, **Georges Pierre Schmartz**$^\dagger$, Laura Groeger$^\dagger$, Jan Dastbaz, Nicole Ludwig, et al. Systematic cross-biospecimen evaluation of dna extraction kits for long-and short-read multi-metagenomic sequencing studies. *Genomics, Proteomics and Bioinformatics*, 20(2):405–417, 2022.

2. Anouck Becker$^\dagger$, **Georges Pierre Schmartz**$^\dagger$, Laura Gröger$^\dagger$, Nadja Grammes, Valentina Galata, et al. Effects of resistant starch on symptoms, fecal markers, and gut microbiota in parkinson's disease—the resista-pd trial. *Genomics, Proteomics & Bioinformatics*, 20 (2):274–287, 2022.

3. Jacqueline Rehner, **Georges P Schmartz**, Tabea Kramer, Verena Keller, Andreas Keller, and Sören L Becker. The effect of a planetary health diet on the human gut microbiome: A descriptive analysis. *Nutrients*, 15(8):1924, 2023.

4. Fabian K Berger, **Georges P Schmartz**$^\dagger$, Tobias Fritz, Nils Veith, Farah Alhussein, et al. Occurrence, resistance patterns, and management of carbapenemase-producing bacteria in war-wounded refugees from ukraine. *International Journal of Infectious Diseases*, 132:89–92, 2023.

5. Sophie Roth, Maximilian Linxweiler, Jacqueline Rehner, **Georges-Pierre Schmartz**, Sören L Becker, and Jan Philipp Kühn. Auritidibacter ignavus, an emerging pathogen associated with chronic ear infections. *Emerging Infectious Diseases*, 30(1):8, 2024.

6. **Georges P Schmartz**, Pascal Hirsch, Jérémy Amand, Jan Dastbaz, Tobias Fehlmann, Fabian Kern, Rolf Müller, and Andreas Keller. Busybee web: towards comprehensive and differential composition-based metagenomic binning. *Nucleic Acids Research*, 50(W1):W132–W137, 2022.

7. **Georges P Schmartz**, Anna Hartung, Pascal Hirsch, Fabian Kern, Tobias Fehlmann, Rolf Müller, and Andreas Keller. PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Research*, 50(D1):D273–D278, 2022.

8. Pascal Hirsch, Alejandra Leidy Gonzales, Ernesto Aparicio-Puerta, Annika Engel, Jens Zentgraf, Sven Rahmann, Matthias Hannig, Rolf Müller, Fabian Kern, Andreas Keller, and **Georges P. Schmartz**. Mibianto: ultra-efficient online microbiome analysis through k-mer based metagenomics

9. **Georges P. Schmartz**[†], Jacqueline Rehner[†], Miriam J. Schuff, Sören L. Becker, Marcin Krawczyk, et al. Between cages and wild: Unraveling the impact of captivity on animal microbiomes and antimicrobial resistance.

10. **Georges P. Schmartz**[†], Jacqueline Rehner[†], Madline Gund, Stefan Rupf, Matthias Hannig, et al. Decoding the diagnostic and therapeutic potential of microbiota using pan-body pan-disease microbiomics.

# *Contents*

# *List of Figures*

# *List of Tables*

# *Abbreviations*

*A*

***AMR*** antimicrobial resistance 39, 44, 45, 53, 56, 57

*B*

***BGC*** biosynthetic gene cluster 35–37, 45, 46, 57, 197–199

***BLAST*** Basic Local Alignment Search Tool 20

***bps*** base pairs 19, 29

*C*

***CAMI*** Critical Assessment of Metagenome Interpretation 50

***CCB*** Chair for Clinical Bioinformatics 259

*D*

***DNA*** deoxyribonucleic acid 24, 29–32, 35, 36, 39, 40, 42, 49–51, 53, 55

*E*

***EU*** European Union 39, 44

*G*

***GTDB*** Genome Taxonomy Database 51

*I*

***IBD*** inflammatory bowel disease 47, 48

*M*

***MIC*** minimum inhibitory concentration test 41

***miRNA*** microRNA

***mRNA*** messenger-RNA 29–31, 35, 42

*N*

***NMDS*** non-metric multidimensional scaling 23, 52

*P*

***PCoA*** principal coordinate analysis 52

***PCR*** polymerase chain reaction 49

***PLSDB*** Plasmid Database 56, 57, 198

*R*

***RNA*** ribonucleic acid

*S*

***SGB*** species-level genome bin 51

*T*

***tRNA*** transfer RNA 31, 42

*U*

***UMAP*** Uniform Manifold Approximation and Projection 23

*W*

***WHO*** World Health Organization 44, 45

***WMGS*** whole metagenome shotgun sequencing 11, 49–51, 53

# 1 Introduction

A first version of the complete human genome was released in the year 2000 [11; 12]. At the time it was considered a milestone achievement that was attributed to the collaboration between the fields of molecular biology and bioinformatics. Following over twenty years of research and tremendous efforts from the scientific community many insights on these approximately $3 \times 10^9$ base pairs (bps) and 20,000 protein-coding genes have been gained [13]. However, already well before completing the human genome project, it was well-understood that not all diseases are due to genetic factors. Instead, humans live in and interact with their environment which considerably impacts their health. Microbial pathogens have been demonstrated to be a causative agents of diseases ranging from diseases such as typhus to tuberculosis. Apart from individual pathogenic species, a vast array of different microorganisms resides in and on the human body defining the microbiome. This microbiome is estimated to harbor around 2 million genes i.e. 100 fold the number of human genes [14]. Due to this complexity, researching the impact of the microbiome on its host's health does not only require a deep knowledge of the underlying biological system but also necessitates dedicated bioinformatic tools and data analysis.

In this thesis, we will investigate the clinical importance of microbes through bioinformatic analysis of genomic sequencing data. The general structure of this thesis is split into four separate chapters, beginning with this introduction. In the introduction, we begin in section 1.1 by presenting overarching concepts in the field of bioinformatics that are not bound to a specific type of data (Figure 1.1). We then elaborate on different concepts that are rather associated with the discipline of molecular biology. At the end of each of these sections presenting a concept, we introduce some of the challenges and best practices that are associated with bioinformatics within these general topics. After the introduction we concisely present the goal of this thesis in chapter 2, while also providing a brief digestible outlook on the various projects. The results are then showcased in full detail in chapter 3. Lastly, the results are discussed, in chapter 4 providing further perspective on future work.

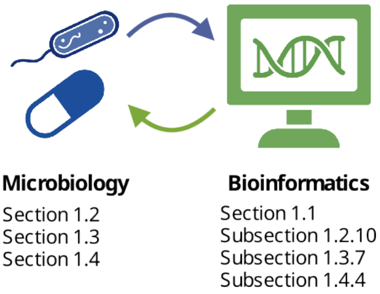


Figure 1.1: General structure of the introduction discussing bioinformatic and microbiology concepts. Created with BioRender.com.

## 1.1 Bioinformatics

The field of bioinformatics can naively be summarized as the processing of biological data with computational tools. Located at the midsection between computer science and biology, it is a relatively young field of research with its beginnings arguably dating back to the middle of the last century when software was still written onto punch cards [15; 16]. With the advent of high throughput sequencing data, the relevance of bioinformatics in biomedical research became evident. These days, the rapid incremental innovations in informatics and new discoveries in molecular biology make bioinformatics a dynamic interdisciplinary research field. Yet also apart from its relevance for research, more and more bioinformatics services find their way into the healthcare system [17].

As bioinformatics is closely intertwined with other fields of research, new discoveries, and technologies are often quickly integrated. Unfortunately, through this relationship, bioinformatics not only inherits many opportunities but also several challenges. On the one hand, data derived from biological experiments is often noisy, due to measurement inaccuracies as well as many confounding factors acting upon the measured biological systems [18]. Further, depending on the experiment, sample sizes may be very limited e.g. due to experimental costs or low number of disease occurrences [19]. On the other hand, the rise of personal computing led to software development having to account for many different hardware environments and software dependencies making it often more difficult to replicate results across systems [20]. Apart from inherited problems, bioinformatics faces also many self-made challenges. Bioinformaticians often must interact with researchers and medical experts who lack programming expertise. Due to these circumstances, dedicated effort is required for accessible software design to accommodate the interdisciplinary nature of bioinformatics. These increased demands for qualitative software are in direct contradiction with development teams showing high fluctuations due to many developers being graduate students on short contracts who also often lack formal experience in software design [21; 22].

### 1.1.1 Algorithms

Algorithmic bioinformatics focuses on solving a specific task on biological data using a set of predefined rules. Especially algorithms that operate on strings, i.e. sequences of letters, see frequent application as most high throughput sequencing outputs have string representations. Classical examples include alignment algorithms, k-mer-based algorithms, and assembly algorithms. Alignment algorithms such as Needlemann-Wunsch, Smith-Waterman, and Basic Local Alignment Search Tool (BLAST) generally aim to compare sequences against each other [23–25]. K-mer based algorithms are not restricted to only sequence alignment and are suitable for large amounts of data. K-mers are defined as substrings of length *k* and are frequently hashed in associated algorithms i.e. a number is derived from the substring [26]. Assembly algorithms try to concatenate individual sequence fragments into longer sequences. Among these algorithms, several subcategories may be distinguished relying on greedy, overlap-layout-consensus, or de Bruijn graph strategies [27]. SPAdes, a popular short-read assembler, relies on de Bruijn graphs. Naively, reads are aggregated into the network representation called a *de Bruijn graph*, where edges represent overlap information of different sequences and nodes represent k-mers [28]. Then, a wide range of dedicated algorithms and heuristics are applied to simplify this graph using e.g. coverage and paired-end information to aggregate nodes [29–31]. In the end, longer fragments are extracted from the simplified graph.

There are also many algorithms in bioinformatics that are not necessarily focused on strings such as the Viterbi algorithm, the Baum-Welch algorithm, a plethora of algorithms tied to data science, etc. [32; 33].

Independent of the selected theoretical data structure and algorithms to solve a problem, the implementation of an algorithm will also have a major impact on the quality of a solution. Frequently, successful implementations are programmed in low-level programming languages for improved performance [34]. Tools such as SPAdes, BWA, or Bowtie2 accumulating many citations and widely across bioinformatics research, are frequently programmed in C++ [28; 35; 36].

### *1.1.2 Workflow management*

True to the Unix philosophy, most bioinformatic software is centered around solving one very specific task for a given input. The output of this software should then be forwarded to the next piece of software, repeating this pattern until the desired outcome is achieved. This procedure is supported by the most popular tools by adhering to well-defined filetypes specific to the field, where examples include formats such as *.bam*, *.vcf*, *.fastq*, etc. [37]. By design, many bioinformatic tools can only be used via the command line. With sample counts in the thousands, and often at least five different tools required per sample, a larger study may easily require the execution of five thousand commands. Manual input is not only a lot of work but also error-prone. Here, modern workflow management tools such as Nextflow or Snakemake try to automate large portions of this work [38; 39]. Instead of executing every command by hand, additional code is supplied to a workflow manager describing a recipe on how to invoke a tool, receive its inputs, and forward its output. This leads to modular pipelines able to perform entire analysis with minimal interaction. Conveniently for users, these workflow management tools are also able to handle many other aspects important for software execution. For example, they can orchestrate parallel execution on large computing clusters, support error logging, manage resource scheduling such as memory, etc.

Bioinformatics data analysis has been highlighted to contribute to the replication crisis in science [40; 41]. Reasons include unclear documentation of computational methods that were used, missing code, missing data etc. An important contributor to decreasing reproducibility are also software libraries to which newly developed code is dependent on [42]. To solve these dependency issues several solutions such as in virtual machines or the usage of various levels of containerization can be leveraged. One key functionality of more modern workflow management tool is the support for various containerization methods such as Conda or Docker environments. From the over 150 workflow management tools available, we selected Snakemake for most of this thesis [39; 43].

### *1.1.3 Web services*

Bioinformatic web services on the internet, not to be confused with cloud computing, come with various functionalities, advantages, and drawbacks. For example, databases aggregate data from the community and redistribute it through the web. The contents usually consist of sequence information, three-dimensional structure, functional annotations, etc. [44–46]. On the one hand, large databases such as Sequencing Read Archive may store over 20 petabytes of data, underlining how integral data is for a data-centric science [47]. They often play a key role in adhering to good data practices such as the *FAIR* concept, where findability, accessibility, interoperability, and reusability should be maximized for the benefit of the scientific community [48]. Further, uploaded data may not be stashed locally anymore freeing resources for teams with sparse storage infrastructure. On the other hand, uploading data for users is often tedious as they have to supply metadata and match predefined formats. Moreover, large resources frequently suffer from quality issues as manual curation is unfeasible, and automated detection is usually not perfect [49–51].
Apart from databases, web servers capable of analyzing data through the web have seen popularity in the bioinformatics community. Identical to locally executed software their main task is to accept user input, process or transform the data, and present results. The advantages of distributing software in the form of web services are plentiful. First and foremost, they are easily accessible as there is no need for installation or expensive computing power on the user side. The only thing required is a browser, internet access, and data, which most researchers have. Nevertheless, data must be uploaded to the resource which may take time or be limited in size. Similarly, in case sensitive clinical data is treated, a convoluted array of data protection laws may apply that prohibit the redistribution of data [52]. Relevant for larger studies, computational resources for each user are usually limited to guarantee smooth operation for other users [53]. Lastly, most servers are constrained in terms of workflow customization. While individual projects such as Galaxy provide a wide plethora of tools and configurations for users, most web servers present only a restricted set of possible analyses and parameters to explore [54].

Like any other software, also web services require maintenance and are occasionally abandoned. As such, about 50% of web services are no longer reachable after 10 years [55]. Despite the mentioned challenges, the conveniences appear to outweigh, as to this day, there are still many web services published. For example, the scientific journal *Nucleic Acid Research* dedicates every year two issues to such web services accumulating several hundred submissions yearly [56–58]. One issue is dedicated only to databases. The other one targets web servers, where we also contributed to the community by supporting the editor as software tester since 2020 [57; 59; 60].

### *1.1.4 Data Science*

While many colleagues insist on highlighting all the intricate differences between artificial intelligence, machine learning, statistical learning, etc. we are here, for the sake of simplicity, going to summarize all of these crafts under the term of data science. At its core, bioinformatics has the goal of processing and interpreting biological data and is therefore deeply connected to data science. Data science applications in bioinformatics include statistical modeling of genome-scale metabolic networks, dimensionality reductions for efficient visualization of higher dimensional data, etc. [61]. The main application of data science in this thesis will be for the mining of biomarker sequences in genome sequencing experiments. More specifically, in metagenomic experiments, further elaborated in subsection 1.4.3, we will derive a multivariate feature vector for each sample where the dimensionality is equal to the number of different microorganisms we consider in the experiment [62; 63]. This data may be interpreted as compositional data, meaning that the true information of an entry mostly derives through the ratio to other values in the same sample [64]. Given feature vectors representing individual samples, the primary objective is to identify features that are significantly associated with specific conditions, which may vary based on the study design. While many publications use methods designed for differential expression analysis from the field of transcriptomics for this purpose, such an approach is discouraged in dedicated literature [65]. Consequently, we opt for a restricted set of specialized models tailored for conducting differential abundance analysis. Apart from differential abundance analysis, other statistical methods we rely on are mostly centered around data visualization. Here dimensionality reduction methods such as *Uniform Manifold Approximation and Projection (UMAP)* or *non-metric multidimensional scaling (NMDS)* are leveraged to display similarity among data points [66; 67]. To this end, we first compute dissimilarities or distances among all samples and reduce them then to a two-dimensional space in a way that preserves clustering behavior. The dissimilarity among points may be computed based on various distance functions such as the Bray-Curtis distance which takes abundances of microorganisms into account, or the Jaccard distance applied directly on hashed k-mers of the sequencing data as e.g. done in Mash or sourmash [68–70]. Another frequent tool we utilize for data visualization is the center log-ratio transform for data normalization. As previously stated, metagenomics data exhibits compositional nature and feature abundances should therefore be interpreted in relation to the other features. The center log ratio transformation is frequently applied in metagenomics data analysis, due to a wide range of advantageous properties. First, it preserves the relative ordering between feature counts [71–73]. Further, it quickly allows to estimate whether the feature abundance is below average as the transformation is performed with respect to the geometric mean of each sample independently.

## 1.2 Bacteria

The bacterial domain describes a group of usually single-celled organisms that are traditionally characterized by a lack of organelles, internal membranes, and nuclei [74; 75]. Hardly possible to investigate by eye, bacteria are of microscopic scale ranging from 0.29 µm to 750 µm in diameter [76; 77]. Following the central dogma of molecular biology, the deoxyribonucleic acid (DNA) found within the bacterial cytoplasm is transcribed into ribonucleic acid (RNA) which may be translated into proteins [78]. Proteins perform enzymatic reactions and partake in large pathways to produce a plethora of various metabolites [79]. Genes and their products, predispose bacterial cells to be a complex dynamic biochemical machinery capable of interacting with the outside world and able to react to external conditions [80]. This flexibility combined with the possibility to replicate and mutate over billions of years, enabled bacteria to adapt to a wide range of different environments, even those considered extreme by measures of pH, pressure, or temperature [81]. Accordingly, bacterial organisms have been found in the depths of our oceans, deep caves, and geysers [82; 83]. Therefore, bacteria are often said to be ubiquitous [84–86]. For example, an estimated $1.3 \times 10^{29}$ bacteria are expected to be found in in the oceanic subsurface alone [87; 88].

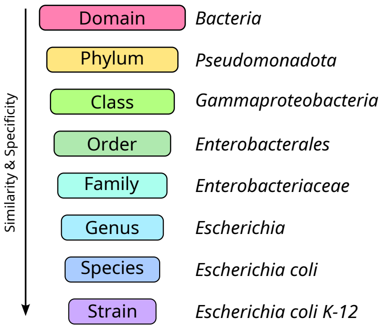


Figure 1.2: Selection of various taxonomic ranks displaying the classification of the *Escherichia coli K-12* strain.

Bacteria are known to reside in and on higher eukaryotes including humans. Especially in case of a perturbation of a host e.g. by a wound, pathogenic species can infect a host [89]. Depending on the bacterial infection, a host may suffer detrimental consequences ranging from minor symptoms such as fever to amputation of limbs, and in the worst cases, death [90]. Accordingly, it has been of historic importance to understand, characterize, and, in some cases, combat the bacterial domain. In the last centuries, tremendous progress in these endeavors could be celebrated. Nevertheless, many questions remain unanswered in the field of bacteriology while the threat of microbial infections may become ever more relevant in the future.

### 1.2.1 *Classification*

Starting with Aristoteles' History of Animals, a major goal in biology has always been the classification and grouping of similarly composed and behaving organisms [91]. This grouping allows the transfer of knowledge gained about one organism to other identically classified species. Already early classification systems, such as the Systema Naturae by Carl Linnaeus in the year 1735 built on a hierarchical structure to organize different organisms at various levels of similarity [92]. Hundreds of years later, after incorporating additional molecular knowledge and phylogenetics into classifications, hierarchical structures subdividing taxonomies prevailed [93–95]. In the most prevalent taxonomic classification systems, in the early branching structures of their hierarchy, an acknowledged split is separating eukaryotic and procaryotic organisms, e.g. based on information on the presence or absence of a nucleus [96; 97]. However, it is currently assumed that prokaryotes, lacking a nucleus, separated already earlier in evolution into what we now define as archaea and bacteria [98; 99]. Thus, three domains of life are traditionally distinguished with each of these domains is then further subdivided into several taxonomic ranks, such as order or genus (Figure 1.2).

Species diversity estimates within the three domains of life are highly variable ranging e.g. from the lower millions up to trillions of different species for the bacterial domain [100–107]. A large portion of microorganisms found in and on humans were successfully put into culture and isolated [108–111]. However, for other biomes, some researchers claim that only 1% of microorganisms are culturable on standard agar, complicating documentation [112–114]. Accordingly, only about 20,000 bacterial species were successfully isolated and described [115]. Genome analysis of uncultured bacteria in molecular experiments motivated the revision of early estimates [116]. The genome taxonomy database, for example, currently holds around 80,000 different bacterial species which are defined based on genome similarity [117]. To provide context, there are an estimated 8.7 million eukaryotic species on Earth, of which about 1.6 million have been described (Figure 1.3) [102; 118; 119].

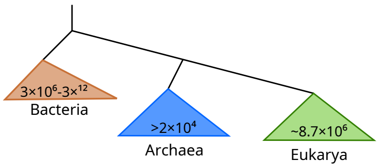


Figure 1.3: Estimated number of species grouped into the three domains of life. Note, the estimates are highly volatile and were taken from literature [102; 107; 120].

### 1.2.2 *Pathogens & Infections*

On the one hand, there are many commensal bacteria that live on a host organism without causing harm. Some even maintain mutually beneficial relationships with their host. Similarly, for the majority of bacterial species, contact is safe for humans. On the other hand, some bacterial species have been described as causal disease agents which defines them as pathogens. As such, bacterial infections are among the leading causes of death worldwide. Only 33 selected bacterial pathogenic species, were globally associated with an estimated 7.7 million deaths in 2019 alone, which constitutes approximately every eighth death [121]. As of today, around 1500 bacterial pathogens are described [122]. Frequently the difference between obligate and opportunistic pathogens is distinguished, with opportunistic pathogens requiring an immunocompromised host for successful infection [123]. We present a selection of the most frequent or relevant bacterial pathogens in Table 1.1.

Until the nineteenth century scientific consensus on infectious diseases was that pathogens emerged and propagated from a miasma which was linked to the idea of bad air [124]. These days, it has been shown that bacterial cells or spores are transmitted to hosts where they colonize in different organs and on surfaces [125]. Hereby, specific modes of transmission can vary among species [126]. Whereas e.g. *Yesinia pesits* is transmitted via flea bites, *Salmonella enterica* is taken up with food and water, and *Bacillus* spores may transmit through inhalation [127–132]. Despite this diversity in transmission modes, many bacterial infections have in common that they can be traced back to a lack of hygiene [133; 134]. In clinics or hospitals, where a more vulnerable and diseased population coexist closely together and operations may be performed, nosocomial infections can pose a serious threat to staff and patients [135–138]. Accordingly, hygienic measures like disinfecting hands and materials, play an important role within healthcare facilities and considerably impact budgets [139–142]. In community-acquired infections, outside of healthcare facilities, diseases like cholera are associated with poor wastewater treatment infrastructure [143].

| **Bacterial Species** | **Associated Diseases** |
|---|---:|
| Bacillus anthracis | Anthrax |
| Borrelia burgdorferi | Lyme disease |
| Campylobacter jejuni | Gastroenteritis |
| Chlamydia trachomatis | Chlamydia infection |
| Clostridium botulinum | Botulism |
| Clostridium difficile | Diarrheal infections |
| Clostridium perfringens | Gas gangrene |
| Enterococcus faecalis | Urinary tract infections |
| Escherichia coli | Gastroenteritis |
| Gardnerella vaginalis | Bacterial vaginosis |
| Haemophilus influenzae | Respiratory infections |
| Helicobacter hepaticus | Liver inflammation |
| Helicobacter pylori | Peptic ulcers |
| Klebsiella pneumoniae | Pneumonia, UTI |
| Legionella pneumophila | Legionnaires' disease |
| Listeria monocytogenes | Listeriosis |
| Mycobacterium tuberculosis | Tuberculosis |
| Mycoplasma pneumoniae | Atypical pneumonia |
| Neisseria gonorrhoeae | Gonorrhea |
| Salmonella spp. | Food Poisoning |
| Shigella spp. | Shigellosis |
| Staphylococcus aureus | Pneumonia |
| Streptococcus pneumoniae | Pneumonia |
| Treponema pallidum | Syphilis |
| Vibrio cholerae | Cholera |
| Yersinia pestis | Plague |

Table 1.1: Assortment of relevant pathogens selected based on historic and clinical relevance.

### 1.2.3 Historic Relevance of Bacteria

Estimates suggest that bacteria inhabit our planet for up to 3.9 billion years [144]. Various fossil findings arguably prove that cyanobacteria existed already 3.5 billion years ago [145–147]. Accordingly, bacteria inhabited this planet long before humans did, where first fossils only date back as far as around 300,000 years ago [148]. Interestingly, early signs of osteomyelitis hint that already dinosaurs have been subject to bacterial infections [149]. Similarly, paleontological findings from South Africa suggest that *Australopithecus africanus*, an ancient *homoinis* species, was subject to *Brucella* infections 2.4 million years ago [150]. In *Homo sapiens*, findings from the bronze age dating back almost four thousand years hint towards plague infections [151; 152]. On a similar note, ancient human mummies from Egypt and Peru provide the first signs of tuberculosis infections in humans dating back over two thousand years ago [153–155]. Since methods to identify and characterize bacterial pathogens as causal disease agents are rather novel by comparison, written proof of bacterial infections only existed since 1876 [156]. However, the earliest conserved written documentations in human history on what were likely bacterial infections are already described in various ancient texts dating back over three thousand years ago [157; 158]. Written records of bacterial pandemics are scattered all over human history too with the worst examples usually being the bubonic plagues caused by *Yesinia pesits* infections.

Including the *Plague of Justinian*, *Black Death* and *the third plague pandemic* it is estimated that over 200 million people have succumbed to this pathogen throughout history [159]. Apart from pandemics decimating entire populations in one sweep, common bacterial infections, such as tuberculosis constituted a major threat on a daily basis.

### 1.2.4 Golden age of Bacteriology

With bacterial pathogens historically claiming so many lives, it comes as no surprise that breakthrough research on microbial organisms has a tremendous impact on our society. Frequently considered the earliest microbiologist, Anton van Leeuwenhoek was 1676 the first person to ever observe and describe the microbial world with the help of his advancements in microscopy [160]. Ignaz Semmelweis achieved a pivotal milestone in combating childbed fever through the implementation of infection control measures in 1847 [161]. Lacking a profound comprehension of the disease's intricacies and relying on evidence-based medicine, he introduced a hand-sanitization protocol using a chlorine solution before delivering healthcare during childbirth. This initiative ultimately resulted in a reduction of bacterial infections and a decline in childbirth-related deaths [162]. The early beginnings of modern bacteriology as a research field are frequently attributed to the contributions of Ferdinand Cohn, Louis Pasteur, and Robert Koch in the 19th century. Ferdinand Cohn elaborated on the endospores of *Bacillus subtilis* and proposed early taxonomies of bacteria in 1875 [163]. Louis Pasteur published a wide range of contributions across several fields. Some of his most influential works include his work on pasteurization in 1864 [164], the germ theory of disease during the 1860s [165], and his work on the rabies vaccine in 1885 [166]. He thus did not only provide a first explanation linking the cause of many diseases to microorganisms but he also provided a trivial way to make food and beverages safer for consumption [167]. Around the same time, Robert Koch published his famous postulates, identifying pathogens as causal disease agents, and his works on the identification of three different bacterial pathogens namely, *Bacillus anthracis*, *Mycobacterium tuberculosis*, and *Vibrio cholerae* [156; 168–171].

### 1.2.5 Bacterial Structure

Due to the large diversity within the bacterial domain, it is difficult to derive a single description for the structure of a bacterial cell, encompassing all members of the domain. Nevertheless, we here try to characterize a majority of noteworthy properties and mechanisms of known bacteria. Some of these will later be relevant as they serve as antibiotic targets.

*Cell envelope* On the outside, bacterial cells possess a rigid cell wall, which provides structural support and protects the cell [172]. Bacteria may be categorized into Gram-positive and Gram-negative based on their susceptibility to Gram staining [173]. In Gram-positive bacteria, the cell wall is mostly composed of multiple layers of peptidoglycan [174]. This peptidoglycan can be deconstructed into linear chains of alternating sugars namely N-acetylglucosamine and N-acetylmuramic acid. These chains are then cross-linked by a D-alanyl-D-alanine-cleaving-peptidase using D-alanyl-D-alanine dipeptides as a substrate to form a mesh-like peptidoglycan layer [175]. In Gram-negative bacteria, while peptidoglycan may be included in the cell wall, the overall number of layers is reduced [176]. During Gram staining, e.g. hexamethyl pararosaniline chloride, or crystal violet, is used to color the peptidoglycan layer of the cell wall [177]. Depending on the thickness of this layer, staining results will vary allowing this histological staining technique to be used during species identification. Gram-positive cell walls also include large amounts of anionic polymers and a wide range of different proteins [174]. In contrast, Gram-negative bacteria typically possess lipopolysaccharides as an outer membrane surrounding the optional peptidoglycan layer i.e. cell wall [178]. This membrane is composed of a lipid bilayer where the inner layers are phospholipids and the outer layers are glycolipids.

The cell wall surrounds the plasma membrane of the cell. The cell membrane, consisting of a phospholipid bilayer, plays an important role in regulating the exchange of signal molecules, nutrients, and waste products with the environment [179]. Apart from a cell wall and cell membrane, some bacterial species own a third cell envelope layer called a capsule or slime layer [180]. The capsule surrounds the cell wall and can play key roles in immune evasion and protection against environmental stress factors [181].

Despite sophisticated, metabolic pathways and the capability to synthesize a plethora of different molecules, bacterial cells are in constant molecule exchange with their environment to survive. While this encompasses the uptake of nutrients for cell growth, it also includes a wide range of metabolites that are excreted into the environment [182]. To orchestrate these fluxes, a wide range of mechanisms evolved exploiting e.g. passive diffusion, and active specialized transport systems. Some of these cellular transport systems, such as pumps are anchored into the cell envelope [183].

Several bacterial species own one or multiple flagella which are characteristic whip-like appendages that allow a cell to traverse its environment [184; 185]. The flagella's motility mechanism differs across bacterial species [186]. Other optional external structures that can be used to differentiate bacterial species are pili or fimbriae, which are hair-like extensions on the cell surface. They can be used to attach to other cells and surfaces [187–189] A more exotic structure that may rarely be observed in bacteria is a magnetosome [190]. These membrane-bound iron particles or crystals enable magnetotactic bacteria to orient themselves along magnetic field lines [191; 192].

*Bacterial DNA* The inner of the bacterial cell is filled with cytoplasm which is a fluid with increased viscosity [193]. Here, a wide range of different macromolecules and metabolites may be found that are either synthesized or imported into the cell [193]. Organelles in the classical sense, such as a nucleus or other structures enclosed by a phospholipid membrane are typically not found inside bacterial cells. Nevertheless, other subcellular structures with defined boundaries have been described [75]. Within the cytoplasm, one cyclic chromosome may be found in a region called nucleoid [194]. Due to the multiploidy of some bacterial species, several chromosomes may exceptionally be observed [195]. Similarly, linear chromosomes have been described [196]. Just like eukaryotic DNA, bacterial DNA is arranged into a double helix, harboring the same four nucleotides. Further, a wide range of functional genomic regions, such as promoters, mobile genetic elements, etc., despite being of different sequences, are also found in bacteria [197]. Operons, traditionally associated with prokaryotes and occasionally observed in eukaryotes, are functional units of DNA [198]. These units consist of clustered genes that are co-regulated and transcribed as a single entity. The transcription of operons may produce polycistronic messenger-RNA (mRNA), encoding multiple proteins [199].

Apart from chromosomal DNA, Prokaryotes can harbor extrachromosomal DNA in the form of plasmids. Generally also described as cyclic, natural plasmids can appear in multiple copies in the same cell ranging from 746 bps to several Mb [200; 201]. While multiple different plasmids may coexist, incompatibility groups have been defined due to shared replication or partitioning mechanisms [202; 203]. As further discussed in subsection 1.3.2, plasmids can be transferred horizontally among bacteria.

*Bacterial Gene Expression* The bacterial genome encodes for about 500-7500 genes with lengths up to around 100kb [204; 205]. Hereby, a gene may be defined as the necessary DNA that contains the required information to generate a functioning RNA or protein. DNA information is transcribed into RNA by RNA polymerase during gene transcription. Resulting mRNA fragments can then be further translated into proteins by ribosomes. Finally, these proteins can partake in complex metabolic pathways or compose the cell structure. In order to remain responsive to outside stimuli and environmental changes, cells regulate their genes yielding various levels of gene expression [206]. Gene regulation may happen on many different levels. The methods of bacterial gene regulation share many similarities to the eukaryotic domain, yet may generally be described to be simpler [207; 208].

Whereas epigenetic gene regulation is important in eukaryotic cells e.g. for cell differentiation, this type of gene regulation is less studied in bacteria [209]. Histone modifications, i.e. modified versions of the DNA packaging proteins, play a major role in eukaryotes. In contrast, bacteria were long thought to lack histones and therefore also to lack histone modifications [210; 211]. However, recently histones were also described to be prevalent across the bacterial domain [212; 213]. Thus, new histone modifications may be found in the future. Certainly, bacteria have been described to perform DNA methylation for epigenetic gene regulation [214–217]. Using DNA methyltransferases bacteria can attach methyl groups to adenine or cytosine bases in their DNA [218]. Based on the presence or absence of these methylation patterns, subpopulations with different phenotypes may emerge [219].

A well-studied level of regulation is transcriptional gene regulation dictating how quickly and frequently a gene will be transcribed by the DNA-dependant-RNA polymerase complex. Here, transcription start can be regulated based on the recognition of the promoter site, RNA polymerase activity, and its holoenzyme formation [220; 221]. Promoter recognition behavior may be altered by activator or repressor proteins [222; 223]. The lac operon in *Escherichia coli* is the classical textbook example of such acting inducers, activators, and repressors, where based on the presence or absence of lactose and glucose, different regulation patterns are observed [224]. An alternative method for promoter-based regulation would be the modification of the promoter e.g. by sequence inversion [225; 226]. Further, the formation of the holoenzyme, composed of the RNA polymerase and a required sigma factor subunit, can be regulated [227]. With different sigma factors being more susceptible to specific promoters, expression can be adapted based on the concentration of a selected sigma factor [228]. Moreover, bacteria may alter the activity of their already active holoenzyme [229]. Here, another well-explained mechanism of transcriptional regulation is attenuation, where the textbook example would be the tryptophan operon in *Escherichia coli*. Through altered stem-loop conformation of the mRNA dependent on the tryptophan concentration in the cell, transcription may end prematurely [230]. Lastly, 20-30% of transcripts, are terminated via rho factor concentrations, providing another avenue for regulation [231].

Since bacteria do not possess a nucleus, the transcribed mRNA does not pass a membrane before translation. Similarly, most bacterial transcripts do not undergo splicing [232]. Accordingly, translation can already start before transcription is completed [233]. Once a functional mRNA has been transcribed regulation may still occur at the post-transcriptional level. So-called riboswitches have been described to play an important role in bacteria gene regulation. By binding ligands, riboswitches alter the conformation of a transcript prohibiting translation [234]. Frequently, riboswitches regulate the same mRNA fragment they are located on, called cis-acting. However, also trans-acting riboswitches have been described which regulate a different mRNA molecule [235]. Further, small RNAs and antisense RNAs have been described to bind to mRNA and influence translation rates [236; 237]. In this context, Hfq, a protein that mediates small RNA binding, is frequently mentioned alongside [238; 239]. Lastly, the mRNA may be degraded by ribonucleases, such as the RNase E [240; 241].

The bacterial translation machinery consists of two ribosomal subunits namely, 50S and 30S that slide along a transcribed mRNA [242; 243]. Several initiation factors are required as well as the Shine-Dalgarno sequence on the mRNA opening alleyways for translational regulation [244]. After initiation, several factors can influence efficiency, e.g. codon usage and transfer RNA (tRNA) availability [245–247]. Furthermore, transfer-messenger-RNA have been shown to regulate translation, as these RNAs are capable of re-initiating translation of stalled ribosomes [248].

After successful translation into a protein, additional mechanisms exist to regulate protein activity. Methylation, acetylations, glycosylation, lipidation, and other protein modifications may activate, deactivate, or alter enzymatic efficiency in proteins [249]. Another example of post-translational regulation involves the cleavage of peptide bonds as exemplified in the cholera toxin activation where the toxic-active A-subunit is cleaved by a protease into two fragments named CTA1 and CTA2 which are then linked via a disulfide bond [250; 251]. Further, allosteric regulation is observed in bacteria. Here, effector molecules binding outside of a protein's active site induce conformational changes, altering protein activity [252; 253]. Lastly, protein degradation rates impact protein expression levels. In contrast to eukaryotic post-transcriptional regulation, ubiquitin degradation tags are less relevant in bacteria [254]. Instead, protein degradation is mostly regulated by specific recognition sequences or signals that mark proteins for degradation called degrons. One common mechanism in bacteria involves the tagging of proteins with small peptides, known as ssrA tags or tmRNA tags respectively [255].

*Metabolism* Bacterial cells have developed a large union of chemical reactions to orchestrate survival, growth, proliferation, energy conversion, waste product transformation, and cell component synthesis to stay alive. These reactions required to sustain life are usually referred to as *metabolism* [256]. To improve our understanding of metabolism, researchers group the reactions into structured pathways consisting of a sequence of chemical and enzymatic reactions. The pathways are then described to take one or many inputs and provide several outputs while being labeled according to the main function of the pathway [257]. An example of such a well-studied pathway is the folate biosynthesis pathway. Here para-aminobenzoic acid and pteridine derivatives are converted into dihydrofolate and ultimately tetrahydrofolate which is vital for purine and thymidine synthesis, both key components of DNA [258].

Historically bacteria were characterized based on the different kinds of energy metabolisms characterized by the capability of bacteria to grow in the presence or absence of oxygen. Obligate aerobe bacteria, such as *Mycobacterium tuberculosis*, require oxygen for respiration. For obligate anaerobes, such as *Clostridium difficile*, oxygen is toxic. They may rely on fermentation for energy conversion [259; 260]. Further groups, such as facultative aerobic, aerotolerant anaerobe, etc. may be distinguished [261].

Significant in the context of Earth's history is also the capability of various bacteria to leverage photosynthesis in their metabolism.

Cyanobacter as well as some purple sulfur bacteria can transform carbon dioxide into carbon-rich organic compounds or oxygen using light [262; 263]. These cyanobacteria are discussed to have played a vital role in oxygenating our planet [264]. In the wake of ever-increasing greenhouse gases in our atmosphere, these photosynthetic bacteria are also explored as a means to reduce e.g. $CO_2$ concentrations [265]. The chemolithotrophy performed by various bacterial species is similarly important to our environment. Here, bacteria harvest inorganic compounds as energy sources. Examples include sulfur-oxidizing bacteria, such as *Beggiatoa* and *Thiomargerita* [266]. Other bacteria have the ability to detoxify pollutants, making them valuable assets in bioremediation efforts [267]. Nitrogen-fixing bacteria convert atmospheric nitrogen into ammonia, making it available for other organisms [267]. Examples include *Rhizobium* in plant root nodules [268].

*Reproduction* Bacterial reproduction is asexual and happens most often through binary fission [269]. To prepare for cell division, the bacterium duplicates its chromosomal DNA through DNA replication. Typically replication begins at a unique origin of replication by cooperative binding of the initiator protein DnaA to multiple recognition sites [270]. This triggers DNA separation and allows the replisomes to enter in between strands [271]. The replisome is a protein complex that is dragged along the DNA double helix during replication. On each strand, DNA information is copied into a new strand resulting in two new double strands [272]. While replisomes are reading along the initial double strand, DNA starts supercoiling. Here, topoisomerases introduce DNA breaks and heal them to avoid negative and positive supercoiling of the DNA strand [273]. Once the DNA is duplicated, the cell elongates and initiates septum formation. The septum is a wall in the middle of the cell that divides it into two compartments [274]. It continues to grow during cytokinesis until it fully splits the cell [275]. As a consequence, one larger cell divides into two cells of almost identical genetic information. Note that through the distribution of plasmids during fission and due to errors made during DNA replication genetic information does not need to perfectly coincide between offspring.

*Specialized Adaptations and Functions* There is a wide range of other functionalities and mechanisms found in bacteria that are maybe less common within the domain yet interesting nevertheless. Sporulation, for example, allows some bacteria to form highly resistant dormant structures to protect their DNA allowing the cell to survive higher pH or temperature gradients [276]. Other bacteria are capable of chemotaxis, i.e. navigation based on a chemical gradient in a cell's environment [277]. While most described properties so far were attributed to individual bacterial cells, bacteria also evolved to display interesting properties acting in communities. For example, some communities can form biofilms [278; 279]. Additionally, several species developed intercellular communication. Via a mechanism called quorum sensing, bioluminescence, production of virulence factors, or coordinated cell death may be organized among cells [280].

### 1.2.6 Host Interaction Mechanisms

The study of bacterial species in isolation and their impact on community dynamics provides deep insights into the fundamental mechanisms governing bacterial life. However, in a clinical context, understanding the interaction between bacteria and their host organism is often of central importance. Bacteria can interact in complex ways establishing a parasitic, symbiotic, or mutualistic relationship depending on the derived benefits for bacteria and host [281]. Especially the parasitic relationships which are detrimental to the host, have enjoyed special attention in research, and many interaction mechanisms have been uncovered.

Before consistent interaction with the host is possible, a bacterial cell must somehow enter, establish contact, or adhere to the host tissue. Hereby, bacteria may enter the host via simple surface contact with a wound, with food intake, or via aerosols. Attachment to host surfaces may be facilitated via adhesins or pili as seen in *Neisseria gonorrhoeae* or *Escherichia coli* respectively [282–284]. In rare cases, invasion of host cells may happen as done by *Salmonella enterica* [285–287]. Once inside or on the host the bacteria has to counteract the host's immune response. A wide range of immune response evasion mechanisms can improve the bacteria's odds of successfully inhabiting a host [288]. *Mycobacterium tuberculosis*, for example, can disrupt phagosome maturation which is a vesicle structure formed by macrophages to engulf pathogens [289]. Apart from host system evasion, bacteria may also try to improve their own growth condition by competing with host cells for nutrients. For example, *Staphylococcus aureus* produces proteins that scavenge iron from host proteins [290]. The most spectacular pathogenic interaction is probably displayed in the secretion of toxins. These virulence factors are capable of directly damaging host tissue. Examples include the diphtheria toxin by *Corynebacterium diphtheriae*, tetanospasmin by *Clostridium tetani*, or the most toxic bacterial toxin botulinustoxin by *Clostridium botulinum* [291; 292].

### 1.2.7 Combating Bacterial infections

To avoid infection with bacterial pathogens in the first place, a wide range of prevention measures may be taken. The overwhelming majority of these actions are centered around the improvement of hygiene and contact reduction. First, food hygiene is a key aspect to avoid e.g. infection with *Salmonella enterica* and *Clostridium botulinum* [293–295]. This includes thorough cooking of meat, refrigerating of perishable items, and proper canning for long-term storage [296; 297]. Also drinking water needs to be clean and proper sewage infrastructure should be in place to avoid e.g. *Legionella pneumophila* and *Campylobacter jejuni* infections respectively [298–300]. Respiratory hygiene by following cough etiquette or wearing a mask can reduce aerosol-based infections, such as *Pseudomonas aeruginosa* [301–303]. Similarly, as discussed in subsection 1.2.4, regular hygiene and washing of hands with soap and water can contribute to the prevention of contact-transmitted infections with e.g. *Staphylococcus aureus* [304]. Especially in the healthcare environment, disinfection of surgical instruments, sheets, and further utilities plays a central role in avoiding germ spread [141]. Contact avoidance measures can be helpful in infection prevention too. Apart from the isolation of e.g. infected patients in isolation precautions, general close contact avoidance can already show major results on a population scale [305]. Regarding close contact, safe sex practiced with condoms can already decrease *Chlamydia trachomatis* and *Treponema pallidum* transmission rates by as much as 90% [306]. Further, proper wound care by keeping the perturbation site clean can avoid opportunistic pathogens, such as *Pseudomonas aeruginosa*. Lastly, another key preventive measure is vaccination [307]. Vaccination ahead of infection against e.g. *Clostridium tetani*; *Heamophilus Influenza B*, or *Corynebacterium diphtheriae*, enables the adaptive host immune system to fight infections successfully [308].

With preventive measures likely never being feasible to avoid all pathogens, methods to combat ongoing infections will remain important. In fact, there is a lot of active research on new approaches, investigating concepts, such as phage therapy, CRISPR-based antimicrobials, nanoparticles, biofilm disruptors, quorum sensing inhibition, immune stimulation, etc. [309–313]. While highly intriguing to look at each of these concepts potentially capable of disrupting the field in the future, we will instead limit our scope here to methods that are frequently and reliably applied in state-of-the-art healthcare facilities. Offering supportive care through hydration, rest, pain relief, and, if necessary, fever-reducing medications often enables the natural immune response of the host to eliminate various infections without the necessity for excessive treatment. In case these measures do not suffice, a more direct way to combat bacteria is with the prescription of antibiotics, which is further discussed in detail in section 1.3. However, antibiotics are not always a treatment option. Potentially none of them provide an improvement to the situation. Then, surgical intervention may be necessary to remove infected tissue before sepsis can occur. In rare cases, this may imply limb amputation [314; 315].

### *1.2.8 Bacterial Isolates*

The analysis of a cultured bacterial strain in isolation remains one of the pillars of microbiology. In clinical microbiology this state-of-the-art analysis, starts with the incubation of the complex native sample harboring the bacterial community in a growth medium. Once incubated, selected colonies are transferred to new plates and incubated again. The results are individual strains that may then be classified for diagnostic purposes using microscopy, or histological staining [316]. One prominent method for microbial identification in clinical microbiology is the usage of mass spectrometry due to its high troughput, cost-effectiveness and high accuracy. In *Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry*, a laser ionizes particles. Next the ions are accelerated using an electrical field and the time which the different particles need to pass the field is measured generating a spectrum. This spectrum is then compared to a database for identification. As not all bacteria can proliferate in isolation and in the same growth conditions, only a fraction of all species in a community is often isolated and diagnosed [317]. Clinical isolates often find vast application e.g. in pharmaceutical and biotechnological research. Here, often molecular profiling of macromolecules such as e.g. lipids, mRNAs, or proteins, may be performed to gain additional insights [318].

### *1.2.9 Biotechnological value of bacteria*

Many pathogenic microorganisms are capable of significantly harming humans. Nevertheless, the positive effects of a few selected species did not remain unnoticed by our earliest ancestors. Despite only shallow scientific understanding at the time, successful taming of microorganisms may date back over 10,000 years [319; 320]. Here, beer and bread production with the help of yeast presents one of the oldest and most well-known examples of this success. In the bacterial domain, domestication has likely happened more recently with evidence dating back over 8,000 years [321]. The food industrial application of bacteria is found in the milk fermentation for yogurt, cheese, etc. [322]. Many years and biotechnological advancements later, genetic engineering opened up many more ways to exploit bacteria in other industries [323; 324]. As such bacteria are now considered for the synthesis of biofuel, biodegradability plastic, and biopesticides [325–327]. Bacteria are leveraged for environmental and agricultural purposes bioremediation for detoxification, wastewater treatment, and nitrogen fixation [328–330]. Newer exploratory methods aim to harness bacteria for biosensors and biocomputing [331]. Several prominent applications remain in the medical and biotechnological field where bacteria are, among others, used to research new vaccines, probiotics, and antibiotics. Undoubtedly, one of the most widely presented breakthroughs in bacteriology in recent years was the CRISPR-Cas9 system published in 2012 [332; 333].

*Biosynthetic Gene Clusters* are groups of collocated genes in the DNA that code for biosynthesis pathways of specialized metabolites [334]. Widely documented in fungi, plants, and bacteria, they provide blueprints for a wide array of secondary metabolites, such as pigments, antioxidants, and toxins [335–337]. Due to their physical proximity in the genome, gene expression of a biosynthetic gene cluster (BGC) can be tightly orchestrated by the organism. The genes within a BGC are often categorized by the role they play during the synthesis of the compounds. Core biosynthetic genes, for example, are coding for enzymes relevant for the synthesis of the precursor molecule, whereas additional biosynthetic genes catalyze subsequent reactions modifying the precursors. BGCs themselves may be grouped based on gene similarity which often goes hand in hand with the similarity of the final compound [338–341].

BGCs have received a lot of attention within the biotechnology community due to their capability to produce bioactive compounds [343–345]. Streptomyces are renowned for their antibiotic production, with streptomycin constituting a landmark discovery in the fight against tuberculosis (Figure 1.4) [346]. The pathway involved in the synthesis is encoded in a BGC. In the long evolutionary fight, bacteria developed a whole range of such protective mechanisms and toxins to get an advantage over other community members. Following this reasoning, researchers started to systematically search within the genomes of microbial organisms for BGC that may encode for new compounds. This procedure is referred to as genome mining [347].

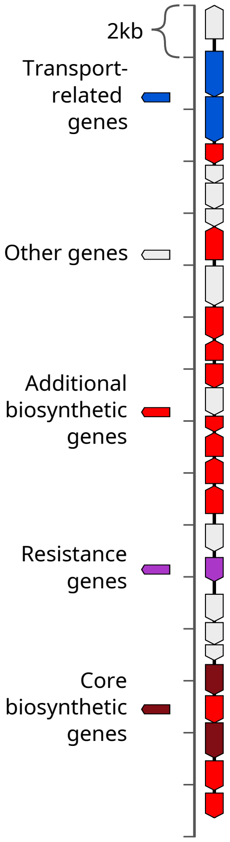


Figure 1.4: Streptomycin BGC of *Streptomyces griseus*. Visualization taken and adapted from the MIBiG database [342].

While sounding highly promising, antimicrobial compound discovery through genome mining faces many challenges. The discovery of antibiotics with a novel mechanism of action is inherently difficult since most BGC detection tools rely in one way or another on sequence similarity [348]. Especially, machine-learning-based in-silico prediction of BGCs will generate many false positives [349]. If BGCs are characterized directly from a bacterial community instead of an isolate, isolation of the desired strain may pose a challenge in the laboratory. Lastly, only because the genes can be found in the genome does not imply that they are expressed [350; 351]. Coupling all these challenges with the time-intensive lab work requiring a major amount of expertise, it comes as no surprise that even this methodically sound approach rarely yields success.

### *1.2.10 Isolate genome analysis*

One method to get a detailed characterization of an isolated bacterial strain is to perform DNA sequencing [352]. The first fully sequenced bacterium was *Haemophilus influenzae* in 1977 [353]. In theory, whole genome sequencing allows to assign each individual nucleotide in the bacterial genome to either adenine, cytosine, guanine, or thymine. In practice, measurement errors arise. Modern next-generation DNA sequencing captures millions of short fragments of the DNA that are traditionally limited to several hundred bases, called reads. Thorough sequencing combined with state-of-the-art bioinformatic analysis allows to assemble the fragments into longer sequences and to identify individual genes within the DNA sequence [205]. This is combined with documented database knowledge and similarity search against existing isolates which allows extrapolating gene functions e.g. for virulence factors [354]. However, while reads generally tend to have good quality and low error rates, the short-read sequencing technology has many limitations in the analysis of long repetitive regions and mobile genetic elements [355]. In recent years, long-read technologies, such as Oxford Nanopore sequencing or PacBio HiFi reads emerged, having overall lower accuracy but therefore having average read lengths of several kilobases [356]. To compensate for this decreased accuracy, hybrid methods leveraging long and short-read technology are currently widely used [357]. However, with future improvements to long-read technology, accuracy is expected to improve likely superseding short-read sequencing.

With the increasing amount of available reference data, tools, and different sequencing technologies, genome sequencing data analysis pipelines are continuously evolving. For the sake of this thesis, we will only limit ourselves to short-read sequencing data analysis for isolate analysis. We will not elaborate on long-read sequencing data analysis in this context. Several workflows may be distinguished for analysis, both relying on methods elaborated in subsection 1.1.1 [358]. On the one hand, a reference-guided analysis may be performed where reads are aligned against a reference genome using aligners such as e.g. Bowtie 2 or bwa [35; 36]. Once aligned, differences between reference and alignment can be highlighted with variant callers such as FreeBayes or GATK HaplotypeCaller [359; 360]. On the other hand, a de-novo assembly may be performed using e.g. SPAdes [205]. De-novo assembly does not use any reference material and only works with the reads generated from the sample. Both methods come with their own disadvantages. As the name suggests, reference-guided analysis requires reference material that may not be available for a newly discovered bacterial species. Further, working with small reads as input, structural variants or large insertions and deletions remain often undetected. The reference-free workflow is often unable to provide a fully closed genome and instead returns several longer fragments which then require additional manual processing [357]. Moreover, de-novo assembly requires a lot of computational power and data.

Once a satisfactory representation of a strain's genomic landscape is attained, a plethora of downstream analyses become available. In case a new genome has been assembled, often functional annotation of the genome is required. Here, pipelines such as PGAP annotate genes based on sequence similarity to other annotated genes in databases [361]. Annotations may be of clinical relevance including virulence genes or antimicrobial resistance genes which are further elaborated in subsection 1.3.7. Similarly, tools such as antiSMASH can mine genomes for BGCs [344]. Another potential analysis is to identify plasmids which may be done with dedicated classifiers or via the comparison to existing databases. Lastly, researchers may want to compare different genomes to each other. In case different strains are compared, multi-locus sequence typing may be performed to classify strains based on a selection of predefined housekeeping genes [362]. Such comparison can also extend beyond the comparison of different strains, e.g. with phylogenetic analysis. Often a phylogenetic tree is constructed representing the evolutionary history of genomes included in the analysis. To construct this tree, homologous genes are compared. Next, either a dissimilarity matrix is derived followed by the popular neighbor-joining algorithm, or statistical approaches like maximum parsimony, Bayesian inference, or maximum likelihood approaches are implemented [363].

## 1.3 Antibiotics

Antibiotics are substances that are capable of treating ongoing bacterial infections via a reduction of growth rate (bacteriostatic) or active decimation of pathogenic cells (bacteriocidal) [364]. Starting with the discovery of arsphenamine in Paul Ehrlich's lab in 1907, the first commercially available antimicrobial compound was released in 1910. In practice, it was used against syphilis infections [365; 366]. This breakthrough was followed by the well-known discovery of penicillin in 1928 [367] and prontosil in 1931 [368]. Initiating this golden age of antibiotics many new antimicrobial compounds were discovered, e.g. streptomycin in 1944 [369], chloramphenicol in 1947 [370], tetracycline in 1948 [371], and many more. Over 150 antibiotics have been discovered since the discovery of penicillin [372].

### 1.3.1 Antibiotic Resistance

With the discovery of new antimicrobial compounds during the golden age of antibiotics, more and more reports also emerged describing bacterial strains that resisted treatment [183; 373–376]. Due to a combination of selective pressure, mutations in the DNA, and horizontal gene transfers the bacterial pathogens acquired genes that improved survival in the presence of antibiotics. Through various mechanisms described in subsection 1.3.4, antibiotic resistances allow bacteria to fully or partially avoid the consequences of selected drugs or entire categories of antibiotics. A famous example is methicillin-resistant *Staphylococcus aureus* which is frequently found in healthcare facilities [377–379]. While methicillin was introduced in 1959, resistance also appeared against more modern antibiotics to the point where, unfortunately, for the majority of available antibiotic drugs, resistance has been documented [372]. Bacteria may also resist several categories of antibiotics e.g. through the acquisition of several resistant genes as exemplified by mobile genetic elements in *Staphylococcus aureus* [380]. Frequently labeled as *superbugs* these pathogens are claimed to constitute a serious health risk. Especially multidrug-resistant pathogens may lead to deadly infections if no more antibiotics display an effect. In this context, multidrug-resistant *Mycobacterium tuberculosis* is frequently mentioned [381]. To address the appearance of new resistances, new antibiotics must be discovered in a continuous arms race between bacteria and research. However, whereas in the the golden age of antibiotics, antibiotic compounds with new mechanisms of action entered the market relatively quickly, these days they are rather a rare sight [382]. This decrease in new

discoveries is linked to a combination of economic, political, and scientific reasons [383].
In 2015 and 2016 estimates of the antimicrobial burden in Europe evaluated that around 6 deaths per 100,000 population can be attributed to antimicrobial-resistant bacteria with overall almost 700,000 total infections in the European Union (EU) [384; 385]. Globally antimicrobial resistance (AMR)-associated deaths were estimated to be around 700,000 per year [386]. Annually, global economic costs due to AMR may reach up to 100 trillion dollars US dollar and 10 million deaths may have to be reported by 2050 [387]. Updated estimates predicted around 4.95 million deaths to be associated with bacterial AMR in 2019 [388]. Accordingly, antibiotic resistance remains a global health threat.

### 1.3.2 Antimicrobial Resistance Dissipation

Due to DNA duplication during proliferation, offspring of resistant bacteria usually inherit the resistant genes from their parent cell if no mutation happens in the concerned genes and all extrachromosomal DNA was shared. However, apart from this vertical inheritance of resistances, bacteria can also transfer genetic material horizontally, even between different species [389]. Here, genes can be passed laterally between organisms. This phenomenon promotes the rapid spread of desirable evolutionary traits among bacteria. Unfortunately, as exemplified by the dissemination of the New Delhi metallo-beta-lactamase (NDM-1) gene in *Escherichia coli* and *Klebsiella pneumoniae*, also resistance genes may spread [390; 391]. Three canonical mechanisms are described for the horizontal dissipation of genetic material, all happening at different frequencies [392; 393].

*Conjugation* is the most important and prevalent method in nature for bacteria to transfer genetic material horizontally [394; 395]. During conjugation, a conjugation pilus is formed which physically connects the donor to the recipient cells. Then, a secretion channel is established and chromosomal mobile genetic elements or plasmid DNA may be transferred to the recipient [396]. While conjugation has been described to cross the inter-species boundary, it has even been highlighted to occur between cells from different kingdoms [397].

*Transformation* describes the uptake of foreign DNA from the environment into a cell which is then said to be *competent*. During transformation, only one strand of the exogenous double-stranded DNA passes through a protein channel across the membrane into the cytosol [398]. In biotechnology, this horizontal gene transfer mechanism is frequently exploited to insert prepared genetic information into colonies. Here, competence may be induced or increased e.g. with chemical agents or polarization of the cell membrane [399].

*Transduction* describes the gene transfer through viruses. Phages enter bacterial cells and integrate their DNA into the bacterial genome. In its lytic cycle, a virus multiplies and assembles. During this process, a phage can accidentally integrate host DNA into the capsule exclusively or incorporate its own DNA along with host genome fragments. [400]. Upon infection of the next host cell, the acquired bacterial genes will then be passed along to the next new host cell dissipating potential resistance genes. Such phage DNA with resistance genes has already been detected e.g. in urban sewer water [401].

### 1.3.3 Consequences of Antibiotic Resistance on Therapy

Antibiotic therapy aims to treat bacterial infections through the use of antibiotics. Due to the large variety of infected host sites, the diversity in pathogens that may be targeted, as well as the antimicrobial resistances that can be encountered, there is no one-size-fits-all solution when it comes to prescription. Similarly, depending on the selected antibiotic compound, dosage and duration of therapy must be considered in order to maximize efficiency and minimize the risk of resistance. Accordingly, many guidelines emerged to consult physicians [402].

*AWaRe* State-of-the-art treatment of bacterial infections includes the prescription of antibiotics. However, conscious of the global threat that the increased prevalence of antibiotic resistance poses to humankind, the World Health Organization (WHO) proposed the AWaRe classification in 2017 to promote antibiotic stewardship [403]. Following the 2021 extension, a total of 258 drugs have been classified into three groups: *Access*, *Watch*, and *Reserve* [404]. According to this recommendation, physicians should follow a tiered approach when selecting antibiotics. Access antibiotics, often named first-line antibiotics are usually antibiotics with a lower resistance potential [405]. They are widely available, accessible at lower costs, and yield fewer side effects. They also have a narrow activity spectrum. Examples include amoxicillin, clindamycin, and cefalexin. Watch antibiotics, such as azithromycin, ciprofloxacin, vancomycin have a broader spectrum but usually also higher costs [406]. They select more aggressively for resistant strains and should generally be prescribed to patients where the pathogen is expected to be resistant to antibiotics of the previous category. Reserve Antibiotics, or last-resort antibiotics, are only to be used when empirical evidence is provided that the drugs of both other categories failed. Accordingly, they should only be prescribed for severe multidrug-resistant infections. By closely monitoring the use of reserve antibiotics, such as linezolid, polymyxin b, or meropenem, their effectiveness is hoped to be preserved.

*Antimicrobial Susceptibility Tests* may be done upfront before antibiotic prescription or after their intake did not show any improvements. In a clinical setting, the infected wound from a patient may be sampled and deposited into agar plates. The clinical isolate can then be tested quantitatively or qualitatively for antimicrobial resistance. As a qualitative test, the Kirby-Bauer disk diffusion test is usually cheap [407]. Here, a paper disk soaked in a selected antibiotic drug is placed into the plate and incubated together with the bacteria [408]. In case there is no bacteria around the paper disc after incubation, the isolate is susceptible. In case a quantitative assessment on how high the required concentration of a drug must be in order to overcome a potential resistance, the minimum inhibitory concentration test (MIC) may be performed [409]. For the MIC test, several agar plates are prepared with varying concentrations of antimicrobial compounds. The isolated bacterium is added to each plate and incubated. Finally, the MIC score can be deduced depending on the minimum concentration where no growth was observed.

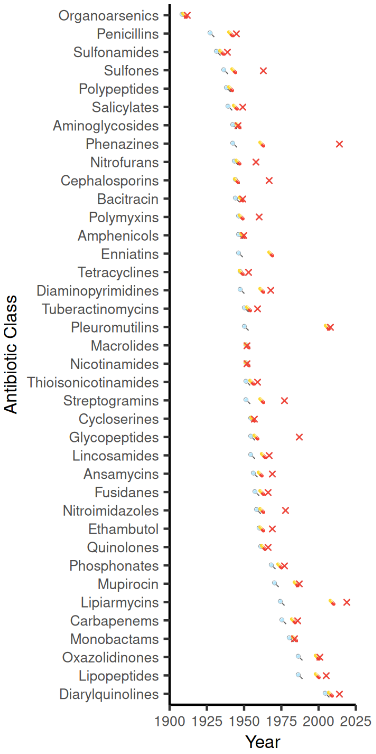


Figure 1.5: Overview of compound classes and their resistances over the years. Indication is given when a compound class was first discovered to have antibiotic activity, when a first antibiotic of this class was in clinical use, and when the first resistant clinical isolate was observed. Data taken from Stennett et al. [410].

### 1.3.4 *Antibiotics Mechanisms*

With over 250 existing antibiotics of various compound classes, fully elaborating each molecular mechanism would leave the scope of this work [405]. Nevertheless, elaborating on the mechanisms behind the most prominent classes of antibiotics provides some insights into their effectiveness and potential shortcomings. We thus showcase seven antibiotic drug classes.

*β-Lactam Antibiotics* have a chemical structure that includes an intramolecular amide bond (carbonyl group with a nitrogen molecule) who causes a cyclization of the molecule. Differences among members of this drug category lie in the neighboring structures to the beta-lactam ring [411]. For example, penicillins display a pentagon whereas a hexagon can be found in cephalosporins. These drugs generally function by inhibiting the synthesis of the bacterial cell wall's peptidoglycan layers by inhibiting the D-alanyl-D-alanine-cleaving-peptidase activity, which is required to crosslink D-alanyl-D-alanine dipeptide with the linear polysaccharide chains into a three-dimensional mesh. Inhibition happens by the beta-lactam ring irreversibly covalently binding to the active site of the D-alanyl-D-alanine-cleaving-peptidase [412]. Since Gram-negative bacterial cell walls are surrounded by a thinner layer of peptidoglycan and possess a lipopolysaccharide layer that inhibits antibiotic entry (See subsection 1.2.5), *β*-lactam antibiotics are more effective on Gram-positive bacteria. Probably the most famous class of *β*-lactam antibiotics are the penicillins. Active molecules include penicillin, amoxicillin, and ampicillin. However, also cephalosporins, carbapenems, and monobactams follow this general mechanism of action.

*Glycopeptides* function similarly to *β*-lactam antibiotics in that they inhibit cell-wall synthesis. However, a key difference to *β*-lactam antibiotics is that the glycopeptides bind to the acyl-D-alanyl-D-alanine in an intermediate product of the peptidoglycan biosynthesis pathway which is called lipid II [413]. Once bound, the glycopeptides prevent the extension of new units to the peptidoglycan. Vancomycin falls into this category of drugs.

*Tetracyclines* are characterized by their chemical structure backbone consisting of four hydrocarbon rings where different chemical groups are attached. Tetracyclines achieve bacteriostatic activity by inhibiting elongation during protein translation. By binding to the 30S ribosomal subunit which is part of the mRNA-ribosome complex, new aminoacyl-tRNAs cannot bind at the acceptor site prohibiting the continuation of translation [414]. Examples of tetracyclines are doxycyclin and minocyclin.

*Lincosamides* follow a similar mechanism of action to tetracyclines. However, instead of targeting the 30S ribosomal subunit, they target the 23S portion of the 50S ribosomal subunit. As a consequence, the peptidyl transferase activity of the ribosome, which is responsible for catalyzing the formation of peptide bonds between amino acids, is inhibited [415]. Apart from halting protein synthesis, lincosamides have also been documented to lead to premature release of incomplete peptide chains from the ribosome [416]. Examples include clindamycin.

*Macrolides* bind, similar to Lincosamides, to the 50S ribosomal subunit to inhibit ribosomal peptidyl transferase activity. Chemically they all share a 12- to 16-membered macrocyclic lactone ring with a $\beta$-glycosidically bound sugar [417; 418]. The first and lead substance is erythromycin. Due to its acid-labile properties, erythromycin is usually only used externally. More developed macrolides, such as azithromycin and clarithromycin, do not have this weakness due to chemical changes and are therefore more orally bioavailable [419].

*Fluoroquinolones* inhibit an important enzyme for DNA replication, repair, and transcription, called topoisomerase. These are present in humans as well as in bacteria, but they differ enough structurally to serve as an antibiotic target in bacteria. The specific topoisomerase attacked in bacteria is called gyrase, hence the general name of the group as gyrase inhibitor [420]. From a chemical point of view, these drugs have a quinolone with a keto group in the C4 position and a carboxylic acid in the C3 position, as well as a fluorine atom in the neighboring ring. Due to the increased number of side effects on the muscular system and nervous system, this group should only be prescribed with caution [421]. Examples include ciprofloxacin, moxifloxacin, and levofloxacin.

*Sulfonamides* have a common feature: the sulfanilamide structure. They imitate p-aminobenzoic acid and thus inhibit the enzyme dihydropteroate synthase [422]. This is important for the synthesis of folic acid. For humans, this is an essential vitamin and can only be absorbed through food. Bacteria are dependent on the synthesis pathway. Sulfonamides are rarely used and when they are, it is primarily sulfamethoxazole in combination with trimethoprim as so-called cotrimoxazole. Trimethoprim specifically inhibits another enzyme involved in folic acid synthesis, dihydrofolate reductase.

### 1.3.5 *Antibiotic Resistance Mechanisms*

Antibiotic resistance mechanisms may vary among bacterial species and will also depend on the specific antibiotic. A wide range of different mechanisms have been described some on individual cell level, some acting on a community level [423–428]. Here, we elaborate on four major categories [429].

*Spatial exclusion* A first category of antibiotic resistance mechanisms is to try to avoid the accumulation of antibiotic compounds in the cell. This can be achieved in two directions. On the one hand bacterial cells have several possibilities to limit compound intake, such as down regulation of porin expression. On the other hand, efflux pumps may be used to reduce the amount of compound that already entered the cell to nonlethal concentrations [430].

*Compound modifications* Enzymatic reactions enable bacteria to deconstruct antibiotics. An important example are $\beta$-lactamase enzymes which convey resistance against beta-lactam antibiotics. Here, the four amid bonds within the $\beta$-lactam ring are hydrolyzed, disabling the antibiotic property of the compound [431]. On a similar note, for aminoglycosides, a whole array of enzymes have been described as able to modify the drugs and alter their binding affinity as well as effectiveness, without completely deconstructing the compound [432].

*Target modification* is another mechanism used by bacteria to avoid antibiotics [433]. Acquired mutations of the genes forming the drug targets can lead to alternated target molecule structures, reducing the binding affinity of the antibiotic compounds [434]. Another target modification mechanism is exemplified by the many *van* genes in glycopeptide resistances, where termini of peptidoglycan precursors are enzymatically modified yielding the same effect of decreased drug binding affinity [435]. Accordingly, bacterial cells can also actively protect a target from antibiotic compounds.

*Bypass* mechanisms to avoid drugs are diverse. Some bacteria can tolerate the antibiotic at the cost of a reduced metabolism or growth rate. Persister cells are dormant cells within a strain that can tolerate antibiotics exactly trough this mechanism [436]. When the drug concentration in the environment reduces again, these persister cells may increase their metabolism again, proliferate, and cause a relapse of an infection [437]. Another bypass mechanism is the overexpression of the target molecule to overload the compound has been described as a functioning bypass [438]. Lastly, bacteria may activate alternative pathways or genes that are unaffected by the drug and still perform a similar role as the drug target. The *sul* gene conveying resistance against sulfonamides serves here as an example [439].

### *1.3.6 Fighting Antibiotic Resistance*

Independent studies sporadically display local geographic reduction concerning the presence of AMR genes for various antibiotic-pathogen combinations [440; 441]. However, with prophylactic prescription of antibiotics during the COVID-19 pandemic, the rise of military conflicts increasing the need for infectious treatment, and the combination of increasing inequality and socioeconomic factors linking to AMR, invite us to speculate on negative developments during the last few years [442; 443]. Of course, the best method to combat AMR would be to avoid all bacterial infections. Unfortunately, despite all the hygienic protocols enlisted in subsection 1.2.7, perfect global avoidance of bacterial infections appears currently quite distant [444]. Accordingly, complementary measures apart from disease prevention are required.

*Antibiotic Stewardship* Due to the diversity of resistance mechanisms, the ever-increasing amount of antimicrobial resistance observed in patients, and the slow development of new antibiotics, scientists, health organizations as well as many governments have recognized the need to intervene [445; 446]. Accordingly, promoting antibiotic stewardship has become a central element in the combat against antimicrobial resistance [447]. Already presented in subsection 1.3.3, the AWaRe guideline of the WHO is a famous example. The overall goal is to expose microorganisms to as few antimicrobial drugs as possible. As a consequence, less selection is made on strains with resistance genes, reducing their overall presence.

*One health* describes the idea that population health is closely interconnected with the health of animals and the environment. The concept may be considered broadly fetched as it includes also e.g. the impacts of nutrition or environmental pollution on human health [448]. Nevertheless, it also captures the idea of AMR dissipation through animals and the environment [407; 449]. For example, extensive use of antimicrobial drugs in livestock has been linked to an increase presence of resistance genes in animal guts. Presumably, via the food chain, these resistances may then end up in humans [450]. Accordingly, the previously mentioned antibiotic stewardship expands even further than just the prescription of drugs to humans. Already passed EU directives, such as regulation (EU) 2019/4 on Medicated Feed and (EU) 2019/6 on Veterinary Medicinal Products, banning all forms of routine antibiotic use in farming, are clear examples that also governing bodies have recognized the One Health concept as well as the antimicrobial resistance threat [451]. Interestingly, human-to-animal transfer of resistant pathogens has also been documented [452].

*Monitoring* of AMR in the environment and on a population level is of key importance. On the one hand, it is important to inform policymakers and governments on the AMR situation by presenting reliable numbers [453]. In this context, the WHO initiated 2015 the *Global Antimicrobial Resistance and Use Surveillance System*, with the aim of providing standardized approaches to the data acquisition and analysis on AMR [405]. Similarly, in 2023 the WHO published its *Global research agenda for antimicrobial resistance in human health* which defines 40 research priorities to generate evidence for the fight against AMR by 2030 [454]. On the other hand, monitoring is also important to update prescription recommendations. With resistance against a few antibiotics on the rise, some antibiotics lose in relevance and should be substituted. Methicillin is a fitting example for this circumstance.

*New Therapies* must be developed in order to combat multidrug-resistant pathogens. Here, alternatives to antibiotics drugs which were previously already mentioned in subsection 1.2.7 may present a compelling solution as they circumvent AMR. Unfortunately, many of these methods require additional research or have inherent limitations [455]. Accordingly, many academic research efforts remain channeled into finding new antimicrobial compounds. The newest marketed antibiotic compound in Germany at the time of writing is Eravacycline in August 2022 [456]. However, this is not a new compound class as it belongs to the drug class of tetracyclines [457]. Thus, the capabilities of the drug to fight resistant bacteria are limited [458; 459]. The newest, antibiotic compound class that eventually released on the market was discovered in 1987. Only five new classes of antibiotics have seen successful market introduction since 1987 [460]. With only a few candidates in the clinical trial pipeline, high dropout rates, lack of incentives for the pharmaceutical companies to partake in the early stages of drug development, lack of funding in academia, and slow development cycles, antibacterial drug research requires structural interventions to remain sustainable [383; 461; 462].

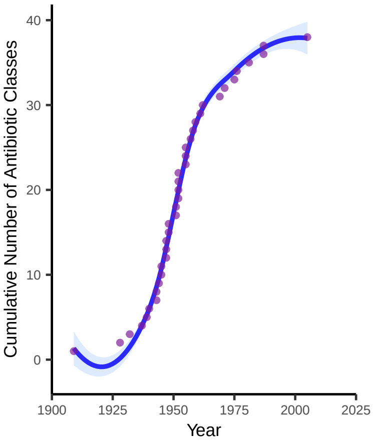


Figure 1.6: Cumulative number of antibiotic drug classes discovered. Data taken from Stennett et al. [410].

### 1.3.7 *Bioinformatics in Antibiotics research*

The applications of bioinformatics in antibiotics research are plentiful. First of all, bioinformatics plays an important supportive role during the design of new antibiotics. Bioinformatic approaches can e.g. be used to identify new potential drug targets based on metabolic pathway analysis [463]. The three-dimensional structure of an identified target's protein may be predicted in silico [464]. Similarly, docking algorithms can predict interactions of drug candidates with the target protein [465].

As mentioned in subsection 1.2.8 and subsection 1.2.10, BGCs can harbor genetic information for the construction of diverse natural compounds and more specifically antibiotic compounds [466; 467]. Here, bioinformatic algorithms for genome mining contribute to the discovery of novel antibiotic compounds [468–471]. Tools such as antiSMASH and PRISM often leverage profile-hidden Markov models for gene classification and then rely on predefined rules on the absence and presence of genes to predict BGCs [467]. As an alternative e.g. DeepBGC relies on neural networks to predict BGCs [470].

Apart from finding new treatments, bioinformatic methods can also support the monitoring and understanding of dissipating antimicrobial resistances [472]. Based on well-documented resistance genes detected in sequencing experiments, antimicrobial resistance can be deduced. Again, different approaches may be distinguished for this analysis. On the one hand, longer fragments resulting from assembly or long-read sequencing capturing the entire resistance gene may be compared against databases such as CARD using e.g. AMRFinderPlus [473–477]. On the other hand, resistance genes may be profiled directly from short-reads using tools such as SRST2 or ARIBA, often preserving adequate accuracy [476–480].

## 1.4 *Community analysis*

Under natural conditions, bacterial species rarely occur in isolation. Instead, they coexist in a mixture with other microbiological entities, such as archaea, viruses, and lower eukaryotes, that are interacting competitively as well as symbiotically in an ecological space, the microbiome [481–483]. Such microbial ecosystems also exist on body surfaces of higher eukaryotic host organisms [484; 485]. From the host's perspective, the symbiotic or parasitic interaction with microorganisms happens through a wide range of different molecules that are emitted from the microbiome. Early microbiological research focused heavily on bacterial pathogens that induced strong symptoms upon infection while closely following Koch's postulates to establish causal relationships. However, by observing e.g. protective effect of commensal microorganisms by conferring a colonization resistance towards pathogens, the one pathogen - one disease model required revision [486; 487]. As such, modern approaches try to understand the role of pathogens existing within their microbial communities while being subject to the community's dynamics and interaction with its surroundings [488]. To analyze these host-associated communities, medical microbiology leverages many established experimental methodologies from the field of ecology which are used e.g. to analyze microbial communities of glacier, soil, or water samples [489–492].

To fully understand the underlying workings of microbial communities, a wide range of macro-molecules can be experimentally quantified and qualified depending on the research question [493]. Unfortunately, each of the existing protocols comes with its own drawbacks and blind spots. To this end, recent trends advocate in favor of multi-omics experiments that integrate data e.g. from metagenomic, metatranscriptomic, and metabolomic experiments to get a holistic view of the communities' inner workings [494; 495]. However, not only is this type of analysis limited by costs and throughput, but it is also a methodological challenging task. E.g. heterogeneous composition of communities combined with uneven sampling across experiments already reflects itself in the data during integration. Nevertheless, based on the gained insight of these multi-omics approaches, interaction networks may be built to simulate e.g. in and output fluxes of entire communities, predict consequences of interventions, or search for new probiotic therapies [496; 497].

### 1.4.1 Human Microbiome

The microbiome of a male adult is estimated to harbor around $4 \times 10^{13}$ bacterial cells with the gut housing 500 to 1,000 different species [14; 498; 499]. While microorganisms have been shown to transmit from person to person, it is generally agreed upon that the microbiome of a person is unique [14; 500]. Community composition has been shown to considerably differ across body sites, such as skin, or gut [501–504]. Even further, it has been demonstrated that collocated microbiota, e.g. within the oral cavity, differ depending on the exact sampling site, due to differences in pH, oxygen, and temperature gradients which can impact the growth and survival conditions of microorganisms [505; 506]. Thus, the concept of a human microbiome has to capture the ensemble of different local communities which are each including smaller eukaryotes, viruses, archaea, and bacteria.

The human microbiome project is frequently considered a key important deep dive into the various human microbiota, such as the vagina, gut, skin, etc. moving beyond culture-based approaches [507]. Following these early initiatives, a plethora of studies emerged assessing the human microbiome accumulating over 2,500 studies in total [508]. With this enormous quantity of data, the human microbiome can be considered one of the best-described host microbiomes. Nevertheless, large portions of sequenced, metagenomic DNA, remain unexplained to the point that even in recent studies many reads remain unmapped [509]. Accordingly, to this day there are still many ongoing efforts to catalog and sequence all the different aspects of the human microbiome such as individual strains and rare community members [510–519].

The human microbiome has been described as being dynamic and changing over time [520]. Here especially the early development of the gut microbiome within infants has received a lot of attention from the scientific community. Apart from changing over time, the microbiome has also been associated with a wide range of demographic factors such as gender, ethnicity, and diet. Lastly, many microbiome-disease associations have been described. For example, a follow-up project to the human microbiome project called the integrative human microbiome project, focused on associations between inflammatory bowel disease (IBD) and diabetes with the microbiome [521]. However, since correlation is not causation the scientific community often remains doubtful if new claims of disease-microbiome associations are truly causal [522]. Especially since within the microbiome research community, a wide range of misconceptions persisted [111; 523].

### 1.4.2 *Dysbiosis*

Microbial communities find themselves in permanent interaction with their surrounding. As such they take up and metabolize nutrients from their environment. Similarly, the microbiome releases a diverse array of macromolecules and metabolites into its environment, known as the expobiome [524; 525]. For example, microbiota in the human gut have been described to remove carcinogens and toxins, synthesize vitamins, and support the decomposition of dietary components for uptake [526]. Further, a healthy microbiome can prevent pathogens from colonization and train the host immune system [527]. However, e.g. consumption of drugs and antibiotics can alter microbiome composition and activity [528–531].

Examining the adverse effects of microbiome imbalance beyond the healthy state of homeostasis reveals an extensive bulk of literature. Dysbiosis, characterized by an abnormal distribution or reduced diversity of microorganisms, can negatively affect host health through a large array of mechanisms [532]. The gut microbiome has long been claimed to play a causal role in IBD including Crohn's disease and Ulcerative Colitis. Here, several microbial immunomodulatory molecules are involved in the pathomechanism of IBD such as a reduction of short-chain fatty acids concentrations, explicitly butyrate, which has been associated with immune modulatory capacities via the induction of regulatory T-cells [533–535]. Nonetheless, also genetic components were associated to IBD [536]. While the relevance of the gut microbiome in a gut disease may appear obvious due to the close proximity of events, there are also less obvious, remote-acting mechanisms described. The gut microbiome is e.g. suspected to impact several neurological and mental disorders including Parkinson's disease, autism, and depression [537–541]. Here the vagus nerve is involved in the bidirectional signal transmission between brain and gut, defining what is referred to as the gut-brain axis [542–545]. Bacteria in the microbiome can release neuroactive compounds, such as dopamine, which the vagus nerve then transmits to the brain [546].

The oral cavity is a heterologous environment revealing strong differences in community compositions between e.g. interdental plaque and tongue dorsum [547]. Within this environment, several dental diseases such as caries and periodontitis are associated with dysbiosis [548; 549]. In caries, acidophilic and acid-producing species such as *Streptococcus mutans* were associated with the disease while nitrate-reducing bacteria like *Rothia* species were discussed to play a protective role [550]. Similarly, to the gut microbiome, the oral microbiome has been linked to diseases affecting apparently remote organs, for example by the dislocation of oral microorganisms in periodontal disease [551]. Apart from the oral cavity and the gut, local as well as remote associations have also been described for microbiota of the skin, vagina, and respiratory tract [503; 552–554].

### 1.4.3 *Metagenomics*

Early research on microbiota relied mostly on isolate analysis [555]. With the description of the 16S ribosomal RNA gene as a phylogenetic marker and advances in primer design as well as sequencing technology, first protocols emerged that avoided culturing bias [556]. These amplicon sequencing protocols such as 16S, 18S, and 23S target specific ribosomal genes of phylogenetic relevance, amplify these regions via polymerase chain reaction (PCR), and almost exclusively sequence these genes [557]. Thus, while they allow an insight into the microbial community without the need for isolate culturing, they are limited to revealing only a small selection of the total genes present in a sample. Therefore, amplicon sequencing does not explain the whole physiological potential of a genome or community. Further, the amplification step introduces PCR-bias which impacts relative measurements on community composition [558].

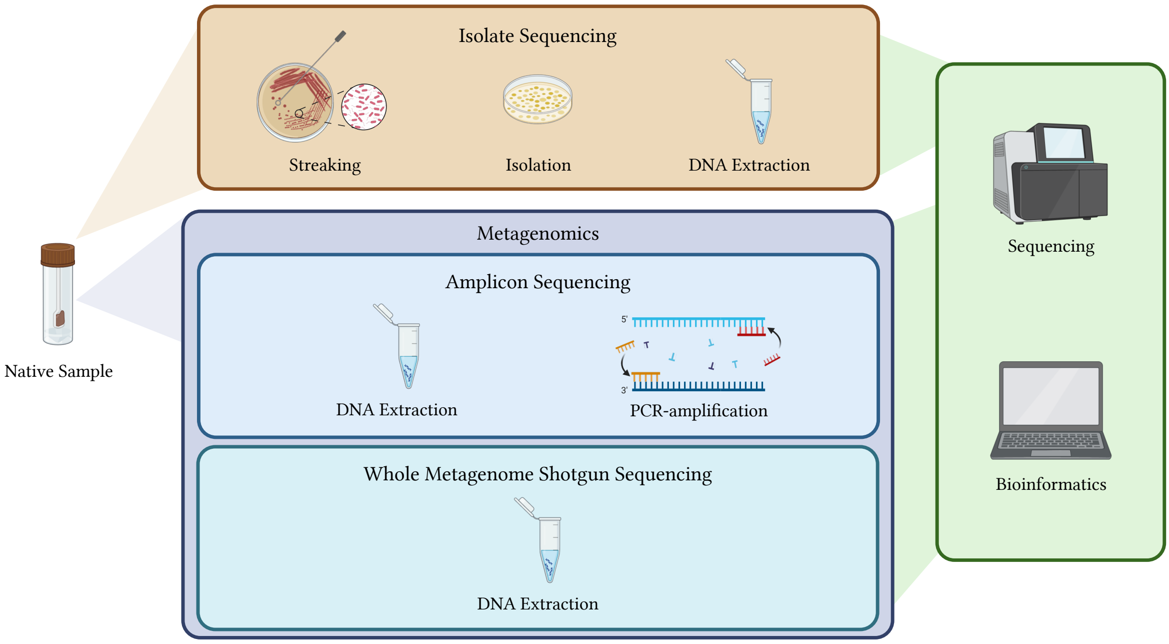


Figure 1.7: Protocols of the various genome sequencing workflows. Whereas isolate sequencing focuses solely on one species, the metagenomic methods measure several species at the same time. Note, the bioinformatic analyses must always be adapted according to the selected protocol. Created with BioRender.com.

A competitor to the amplicon sequencing method is whole metagenome shotgun sequencing (WMGS). Here, all DNA is extracted from the sample and sequenced without selectively targeting specific marker sequences [559]. During the humble beginnings of this technology, only short 40kb fragments of an entire sample could be assessed [560]. However, due to the rapid drop in sequencing costs coupled with protocol refinements, entire communities can now be deeply characterized, to the point where also the rare biosphere can be captured, i.e. species with a low relative abundance in the community are traceable

[561]. With this gain in resolution over amplicon sequencing more genes are available for better taxonomic assignments [562]. Further, entire chromosomes and plasmids can be assembled and analyzed directly from the community allowing one to understand phenotypes in detail [563]. With the further advances in long-read sequencing technology, assessment even up to strain level is discussed elaborating on phenotypical differences attributed to within-host evolution [564]. Lastly, some viral DNA can be detected without the need for explicit amplification [565]. However, despite sequencing costs continuously decreasing, WMGS studies remain considerably costlier in comparison to amplicon sequencing [566]. Moreover, the amount of data that needs to be generated, stored, and analyzed as well as the complexity of these analyses are increased for WMGS. [567]

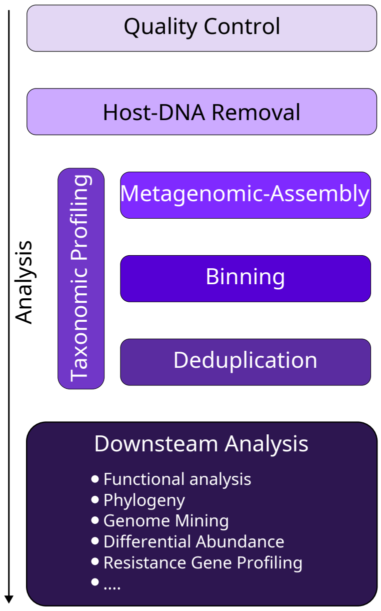


Figure 1.8: Core components of a WMGS data analysis pipeline. Taxonomic profiling may be performed directly on reads instead of requiring an assembly.

### 1.4.4 *Whole metagenome shotgun sequencing data analysis*

The exact workflow used to evaluate WMGS data, as in every bioinformatics analysis, depends on the scientific question that is targeted, the protocol that is employed, and the sequencing technology that is selected [568]. With the bacterial domain being so diverse and ubiquitous it comes as no surprise that within the field of metagenomics, a plethora of tools, pipelines, and data repositories emerged leading to a situation that has been described as the *Wild West*, where no real gold standards exist for data analysis [569]. Unfortunately, there exists plenty of empirical evidence that the results of WMGS studies are dependent on the exact tools, databases, parameters, and statistical tests that were used during analysis [570–572]. This volatility of results frequently prohibits straightforward cross-study comparisons [573]. With the many benchmarking efforts such as the Critical Assessment of Metagenome Interpretation (CAMI) challenges trying to objectively compare tool performances, efforts were made by the community to improve on the situation [574–577]. Users are being incentivized to migrate to the best-performing tools. Nevertheless, a large variety of workflows persisted, as the field remains far away from unified [578]. Despite the large heterogeneity in the field, we present in Figure 1.8 a generally agreed-upon core of a data analysis workflow as also presented in several reviews [579–582]. In accordance with most sequencing data analysis workflow, WMGS data analysis usually starts with quality control of reads. Here a plethora of tools is available [583–585]. For host-associated microbiomes, an optional host-removal step is recommended where sequencing reads that align to the host genome or other contaminating sequences are removed [586; 587]. Afterward, roughly two main workflows may be distinguished in the early phases of the analysis. On the one hand, there is the read-based approach which heavily relies on databases to be complete enough for fragment comparison. On the other hand, there is the assembly-based approach which relies mostly on the data of the experiment [588]. The assembly-based data analysis is considered superior if rarely analyzed microbiota are analyzed or if functional analysis is desired.

However, the overall complexity and amount of computational power is higher in comparison to the read-based analysis. Further downstream analysis following these pipelines is closely tailored to the biological research question.

*Assembly-based* WMGS data analysis, as the name suggests, is centered around the step of metagenomic assembly, that is similar to genome assembly elaborated in subsection 1.1.1[31]. Due to the nature of the experimental setup, the output of metagenomic assemblies contains several thousand long fragments deriving from a wide range of different species. To group these fragments by taxonomy, binning is performed. Metagenomic binning uses various features such as k-mer frequency, GC content, and coverage information to group assembled fragments [589–592]. Depending on the quality of the output bins, individual groups may be interpreted differently. Frequently, bins are interpreted as a single species being termed species-level genome bin (SGB). To assess the quality of these bins, so-called unique marker genes may be used which are expected to occur exactly once per bin [593; 594]. Multiple occurrences hint towards contamination. The lack of a gene indicates deficiencies in completeness. In the context of a larger study where species are expected to occur in multiple samples, SGB would be dereplicated to select a bin that best represents the genomes [595; 596]. Once a set of high-quality SGB has been defined, bins are frequently classified to a certain taxonomic depth based on e.g. average nucleotide identity with e.g. the Genome Taxonomy Database (GTDB) serving as reference material [597; 598]. Here, new genomes worth cataloging may be identified.

*Read-based* analysis avoids the computationally expensive step of metagenomic assembly. Instead, sequencing reads are directly compared against a reference database of marker sequences [599–601]. This method of taxonomic profiling frequently yields only information on the composition of the community. The quality of the marker sequence library is central to this method. Due to shared DNA sequences, reads may not always uniquely be attributed to a species. Similarly, if new species are being assessed that have not yet been documented in the marker gene library, the assignment will fail. Exceptionally, within the biobakery suite around the group of *C. Huttenhower*, tools such as HUMAnN and StrainPhlAn yield results that go beyond community composition without relying on assembled fragments [602; 603]. Instead, phylogenetic similarity as well as functional profiles can be assessed. Similarly, there exist several read-based antimicrobial resistance gene profiler [604; 605].

*Further analysis* The microorganism abundance information resulting from the previous workflows is often used for many state of-the-art compositional analysis from the field of ecology. One such assessment is to quantify the taxonomic diversity of an ecosystem, expressed by the so called alpha-diversity. A wide range of formulas and measures including the Shannon index, which is identical to the Shannon entropy from information theory may be used to describe alpha-diversity [606]. Similarly, beta-diversity captures the difference between ecosystems and is often coupled to ordination analysis where those differences are represented in a two-dimensional embedding. Again several dissimilarity measures as well as embedding methods exist. Frequent examples from the literature include Bray-Curtis or weighted UniFrac distance, and principal coordinate analysis (PCoA) or NMDS, for distance measures and embedding respectively [607]. As elaborated in subsection 1.1.4 differential abundance analysis is also performed on abundance data and can be used to highlight relevant operational taxonomic units that explain potential differences in cohort behavior. These may then serve as future potential biomarkers [608].

A plethora of additional information may be uncovered from metagenomic sequencing data that goes beyond abundance information. Many of these downstream analysis require metagenomic assembly beforehand as they rely on longer sequences. Depending on the quality of assembled genomes all analysis for bacterial isolates elaborated in subsection 1.2.10 may be applied here, including phylogenetic trees, resistances gene profiling, and genome mining. Similarly, plasmids can also be detected in metagenomic experiments [609–616]. Further, a wide range of functional annotations are possible, with examples including virulence factor analysis [354]. However, more specific to metagenomic sequencing, functional analysis spanning several species is frequently performed. Popular targets for this analysis include e.g. carbohydrate-active enzymes that are present across several species [617]. Lastly, community network analysis may be performed. While some approaches try to understand abundance dynamics in bacterial communities, more elaborate approaches try e.g. to simulate entire molecule fluxes which requires a deep functional understanding of the community members [618].

# 2
# Goals of the PhD thesis

The central goal of this thesis is to combat the bacterial threat at two fronts in close collaboration with biotechnologists, microbiologists, and clinicians by providing the bioinformatic expertise needed to work on bacterial genome data. The first front targeted the diagnostic aspect of bacterial diseases. Here, we analyzed new emerging pathogens, we assessed the potential of dysbiosis in the context of dietary intervention as well as diseases, and we developed web tools to enable other researchers with less programming knowledge. The second front was the fight against AMR. Within this scope, we accompanied the earliest stages of drug development, we monitored emerging AMR in conflict regions, and we extended a database focusing on the key mechanism for resistance gene dissipation. Following this goal, seven manuscripts that target at least one of the two fronts fully matured into published scientific articles and are herein included. At the time of thesis submission, an additional three manuscripts are still ongoing. The submitted versions are included in the thesis. We will shortly present all works following the grouping depicted in Figure 2.1. Subsequently, the completed manuscripts are presented in chapter 3.

## 2.1 Metagenomics Projects

① As described in subsection 1.4.3, amplicon sequencing has long been the method of choice in metagenomics. Transitioning to WMGS is a costly commitment especially when large-scaled studies at a higher sequencing depth are considered. Therefore, one central basic research project focused on establishing a reliable and robust experimental metagenomics workflow. Here, we compared three DNA extraction protocols across six specimens based on sequencing results encompassing next-generation sequencing (NGS) as well as Oxford Nanopore sequencing [1]. Many quality control measures were compared, visualizing e.g. the amount of DNA, reads, assembly quality etc. From this study we had several conclusions to take away. First, we decided on a DNA extraction kit that was capable of extracting decent amounts of prokaryotic DNA even in low-input samples such as conjunctiva samples. Further, we observed the overall robustness of larger input samples in cheaper kits. Thus, in the long run, this study

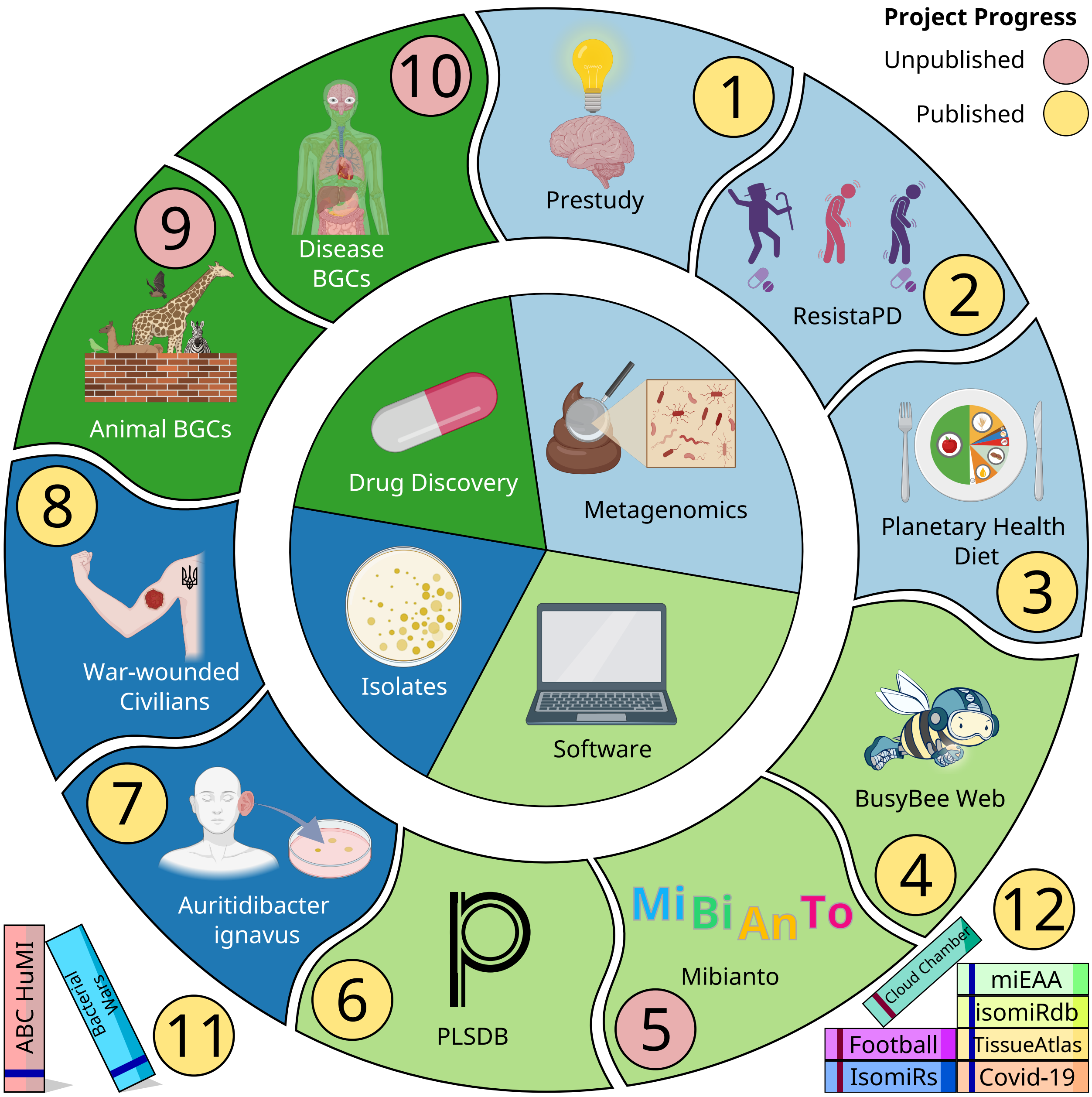


Figure 2.1: The ten main projects included in this thesis. Seven are published in scientific journals. Three manuscripts are submitted as of the time of writing. Two published manuscripts, not included in this thesis, fall within the same general research area. Created with BioRender.com.

allowed us to confidently analyze microbial communities deriving from different specimens at scale while optimizing for costs.

② Parkinson's disease is a neurodegenerative disease that killed around 329,000 people in 2019 with an increasing tendency over the last 25 years [619]. Frequent symptoms include limb tremors and muscular rigidity. On a microscopic level, Lewy bodies which consist of aggregated proteins inside nerve cells are characteristic of the disease [620]. The microbiome acting through the gut–brain axis has been extensively described as a potential avenue for pathogenesis in Parinson's disease [621]. In this context, we analyzed the potential of resistant starch as a prebiotic for Parkinson's disease patients using metagenomics and metabolomics methods [2]. Additionally, two control cohorts were defined. The main conclusion from the study was that the prebiotic intervention partially restored short-chain-fatty concentrations which have consistently been reported to be reduced in Parkinson's disease across multiple studies [621].

③ In 2019 the EAT-Lancet commission proposed a diet plan that supposedly is able to sustainably feed a population of 10 billion people without health deficits [622]. To assess whether a dietary adjustment to this planetary diet would lead to dysbiosis, we performed a dietary intervention on sixteen individuals and compared them to a western and vegetarian/vegan diet using metagenomics [3]. Focusing mostly on compositional changes, we did not observe any significant differences in abundance which were due to the intervention. Accordingly, adopting the planetary health diet does not appear to yield a risk for dysbiosis, at least for the twelve week time period we assessed.

## 2.2 *Software Development Projects*

All software we distributed to the community was provided in the form of web services. ④ In the field of metagenomics, we released and published another version of Busybee Web which, to the best of our knowledge, has remained the only online metagenomic binning tool [6]. As clarified in subsection 1.4.4 Busybee Web allows researchers to group longer DNA fragments derived from metagenomic experiments by their taxonomy. Hereby, it is especially interesting for long read sequences, as no assembly is required, and the overall amount of data is smaller which is favorable for network traffic. Apart from binning, Busybee performs a wide range of annotations including taxonomy and comparisons to plasmids. Lastly, the tool also allows to compare cohorts by co-embedding two datasets at the same time enabling a reference-free differential abundance analysis.

⑤ Aware of the data limitations of Busybee Web, we also released Mibianto which is a web server focused on short-read metagenomic data analysis, that is optimized for larger amounts of data [8]. The tool does not require longer fragments but instead accepts short NGS reads as input. We reduce the amount of network traffic by only sending a selection of hashes computed on the data to our server. Specifically, we compute so-called FracMinHashes on the user side using the sourmash tool which we compiled into web assembly [69]. While this enables us to handle large amounts of data, it comes with several disadvantages. Functional analysis is impossible as the server does not have complete gene information. Similarly, quality control can not be performed on the server side. Lastly, for a web server, it requires above average computational power on the user side as the hashes need to be computed before sending. Once arrived on the server, the data undergoes the proposed workflow for taxonomic profiling with hashes, which involves the solving of the set cover problem using a greedy algorithm. Finally, we provide a wide range of state-of-the-art downstream analysis for compositional analysis such as differential abundance analysis, and provide proxies for quality control. At the time of thesis submission, the paper has not yet been published but is included.

⑥ In the context of combating AMR, the first central piece of work we published was the update of the Plasmid Database (PLSDB) [7]. Since conjugation remains one of the most frequent mechanisms for horizontal gene transfer in nature, plasmids with AMR genes pose a serious threat at the population scale as described in subsection 1.3.2. PLSDB aims to aggregate naturally occurring plasmids from multiple data repositories to provide a valuable resource for the research community to compare their own detected plasmids against. In PLSDB we add a wide range of filtering, data cleaning, and new annotations in order to provide data with an adequate quality. With the over two-fold increase in entries, an update of the web server as well as the data collection pipeline was due. In a larger scope, this allows monitoring of potential emerging resistances.

## 2.3 *Bacterial Isolates Projects*

Apart from metagenomic samples, we also investigated clinical isolates that derived from infected patients. ⑦ A published example of this work discusses the species *Auritidibacter ignavus*. In the clinical microbiology community, this species has recently been suspected to be a potential pathogen implicated in ear infections [623]. With only a few dedicated isolates previously analyzed, the reference material on this organism is rather sparse. In our study, three bacterial isolates from three different patients with otorrhea were extracted and cultivated [5]. Following NGS sequencing, the first goal was to establish relatedness among the three isolates. Using a reference-based approach we did not see any signs that the isolates were related. Next, we assessed the isolates for resistance genes, yet we did not uncover any genes explaining the resistant phenotype.

⑧ 24 February 2022 marks the day of the invasion of Russian troops into the Ukraine. Whereas in the early phases of the war, it appeared that the Russian troops would be able to take in Kyiv, the Ukrainian troops were able to stand their ground and push back the invaders. After over two years, as of the time of writing, the conflict is still ongoing with major portions of Ukranian land remaining under Russian occupation and a front line that is currently almost static. Most member states of the EU, despite not actively participating in the conflict, support the Ukraine. In this context, after receiving primary care in Ukraine, several war-wounded civilians were treated at the *Saarland University Medical Center* for antibiotic infections. After observing antibiotic resistance, pathogens were isolated and whole genome sequencing was performed. Using the techniques mentioned in subsection 1.2.10, we were able to capture a wide array of resistance genes. At one point, we found ourselves in the position of using PLSDB when we investigated multiple multiresistant *Klebsiella pneumoniae* strains[4]. By leveraging PLSDB we identified identical plasmids across multiple patients that carried resistance genes explaining similar phenotypical behavior.

## 2.4 *Drug Discovery Projects*

Whereas two of the previously elaborated published manuscripts focused on AMR monitoring, we further tried to contribute to the sustainable discovery of novel antibiotic compounds in two projects that are currently in submission. These endeavors are based on the potential of metagenomic assembled genomes to harbor BGCs capable of synthesizing natural products with antimicrobial properties as presented in subsection 1.2.9. ⑨ In one of these projects, we looked into the BGC landscape of animal oral and gut microbiota by analyzing metagenomes across a total of 45 host species from the Saarbrücker Zoo [9]. As exotic animal metagenomes are strongly underrepresented in data collections and more heterogeneous than human-only datasets, we documented many novel metagenome-assembled genomes. This diversity in the landscape of microorganisms reflected itself in the large diversity in BGCs that we observed during genome mining.

⑩ The other project in this category, our largest applied research metagenomics project, is a metagenomics study compromising over three thousand initial samples and over six hundred participants [10]. Here, we sampled up to eight biospecimens from the same patients while extensively documenting clinical information. The participants consisted of healthy controls but also a wide array of diseased patients with e.g. oral diseases such as caries but also lung cancer or IBD. We assessed this data pool for potential biomarkers not only on the bacterial species level but also on a BGC level.

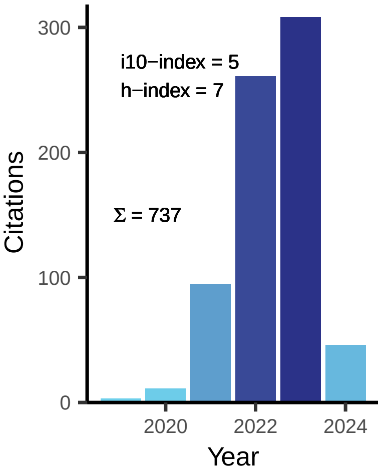


Figure 2.2: Personal citation data taken from Google Scholar. Data accessed on 25/2/2024.

## 2.5 *Other Projects*

⑪ & ⑫ Apart from the scientific work mentioned above discussing the common theme of bacterial research a mixture of scientific curiosity and courtesy towards colleagues led to a total of nine additional publications where seven may not fall into this general research area [624–632]. Examples, listing only (co-)first or last authorship positions, were published in the fields of miRNAs, sports medicine, and experimental physics [626; 630; 632]. None of these works are included in the thesis.

# 3 Results

This thesis includes seven peer-reviewed publications and three submitted manuscripts. The published and submitted versions of the manuscripts are herein included.

## 3.1 *Systematic Cross-biospecimen Evaluation of DNA Extraction Kits for Long- and Short-read Multi-metagenomic Sequencing Studies*

Genomics Proteomics Bioinformatics 20 (2022) 405–417

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

# Systematic Cross-biospecimen Evaluation of DNA Extraction Kits for Long- and Short-read Multi-metagenomic Sequencing Studies

**Jacqueline Rehner [1,#], Georges Pierre Schmartz [2,#], Laura Groeger [3,#], Jan Dastbaz [4], Nicole Ludwig [3], Matthias Hannig [5], Stefan Rupf [5], Berthold Seitz [6], Elias Flockerzi [6], Tim Berger [6], Matthias Christian Reichert [7], Marcin Krawczyk [7], Eckart Meese [3], Christian Herr [8], Robert Bals [8,#], Sören L. Becker [1,#], Andreas Keller [2,#,*], Rolf Müller [4,#], The IMAGINE Consortium**

[1] *Institute of Medical Microbiology and Hygiene, Saarland University, D-66421 Homburg, Germany*
[2] *Clinical Bioinformatics, Saarland University, D-66123 Saarbrücken, Germany*
[3] *Department of Human Genetics, Saarland University, D-66421 Homburg, Germany*
[4] *Helmholtz Institute for Pharmaceutical Research Saarland, D-66123 Saarbrücken, Germany*
[5] *Clinic of Operative Dentistry, Periodontology and Preventive Dentistry, Saarland University, D-66421 Homburg, Germany*
[6] *Department of Ophthalmology, Saarland University Medical Center, D-66421 Homburg, Germany*
[7] *Department of Medicine II, Saarland University Medical Center, D-66421 Homburg, Germany*
[8] *Department of Internal Medicine V – Pulmonology, Allergology, Intensive Care Medicine, Saarland University, D-66421 Homburg, Germany*

Received 23 August 2021; revised 13 May 2022; accepted 19 May 2022
Available online 6 June 2022

Handled by Fangqing Zhao

KEYWORDS

Whole-genome analysis;
Comparative genomics;
Short-read sequencing;
Long-read sequencing;
DNA extraction;
Metagenomics

**Abstract** High-quality **DNA extraction** is a crucial step in metagenomic studies. Bias by different isolation kits impairs the comparison across datasets. A trending topic is, however, the analysis of multiple metagenomes from the same patients to draw a holistic picture of microbiota associated with diseases. We thus collected bile, stool, saliva, plaque, sputum, and conjunctival swab samples and performed DNA extraction with three commercial kits. For each combination of the specimen type and DNA extraction kit, 20-gigabase (Gb) metagenomic data were generated using **short-read sequencing**. While profiles of the specimen types showed close proximity to each other, we observed notable differences in the alpha diversity and composition of the microbiota depending on the DNA

* Corresponding author.
E-mail: andreas.keller@ccb.uni-saarland.de (Keller A).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.05.006

extraction kits. No kit outperformed all selected kits on every specimen. We reached consistently good results using the Qiagen QiAamp DNA Microbiome Kit. Depending on the specimen, our data indicate that over 10 Gb of sequencing data are required to achieve sufficient resolution, but DNA-based identification is superior to identification by mass spectrometry. Finally, long-read nanopore sequencing confirmed the results (correlation coefficient > 0.98). Our results thus suggest using a strategy with only one kit for studies aiming for a direct comparison of multiple microbiotas from the same patients.

## Introduction

In the past decade, microbiome research has become a trending topic with an exponential increase of available data [1]. Researchers worldwide acknowledge the importance of the human microbiome for health [2] regarding a variety of diseases, with the gut microbiome taking a leading role [3]. Recently, the link between a healthy gut microbiome influenced by a Mediterranean diet and cardiometabolic disease risk has been found [4]. In addition, the gut microbiome of Parkinson's disease patients has also been associated with intestinal inflammation [5]. Next to the gut microbiome, the microbiome of the respiratory tract has been studied extensively. For example, it has been previously shown that certain bacteria are associated with chronic rhinosinusitis. Bachert et al. [6], as well as Olzowy et al. [7], detected overgrowth of *Corynebacterium*, *Curobacteria*, *Pseudomonas*, *Staphylococcus*, and *Haemophilus influenzae* in patients with chronic rhinosinusitis compared to the healthy respiratory microbiota. There is accumulating evidence that microbiome research should also identify commensal bacteria and investigate their potential to protect from diseases. Several species are already known to synthesize compounds that inhibit the growth of pathogenic bacteria, thereby establishing a crucial balance within the microbiome. Besides the intended effects on pathogenic bacteria, antibiotic therapy also affects commensal bacteria, and may facilitate overgrowth of potentially dangerous microorganisms, as it is frequently seen in *Clostridioides difficile* infection, a common intestine complication after previous antibiotic treatment [8]. How is the growth of pathogens suppressed under normal conditions? During a co-infection, *Pseudomonas aeruginosa* produces rhamnolipids, which disperse the biofilms of sulfate-reducing bacteria and, additionally, are effective against the biofilms of opportunistic pathogens such as *Escherichia coli* and *Bacillus subtilis* [9]. *Staphylococcus lugdunensis* has been found to produce lugdunin, which is a recently discovered thiazolidine with antibiotic activity. Lugdunin inhibits the growth of the opportunistic pathogen *Staphylococcus aureus* [10]. Furthermore, certain lactic acid bacteria are known to produce a variety of secondary metabolites which inhibit the growth of other bacteria, such as bacteriocins, hydrogen peroxide, and diacetyl [11].

Bacteria have evolved for 4.3 billion years, and their metabolism and entire biosynthesis have perfectly adapted to their environments. They constantly fight for nutrients and space, trying to inhibit the growth of their competitors which renders them the perfect target for searching novel natural compounds to fight bacteria-associated diseases [12]. Also, in the sustainable development of new antibiotics, microbiota plays an essential role [13].

All these aspects can be discovered by examining the human microbiome of various compartments of the body by extracting the whole-genome DNA of clinical samples while depleting the human DNA. The usage of the extracted DNA for next-generation sequencing (NGS) can then shed light on all microorganisms that are in the native sample. This very precise method can be augmented with microbiological cultivation of the same samples. Which bacteria are cultivatable, and also during routine diagnostics which are only detectable by sequencing the native samples?

Many steps in the process of sample collection, DNA extraction, sequencing, and data analysis can introduce significant bias. One example is the stool collection kits used that already affect the reported microbial compositions [14]. Likewise, in oral microbiomes, bias is known and addressed [15]. The extraction of the whole-genome DNA is a crucial step. It is evident that the topic of comparing different DNA extraction kits is essential and thus has become an evolving field of research. For different specimen types, respective protocols have been compared, *e.g.*, for breast milk [16], stool [17], skin [18], vaginal swabs [19], sputum [20], postmortem eye tissue [21], nasal washes [22], and meconium [23]. As for one specific sample type, the most suitable DNA extraction method has been evaluated over several studies, but an analysis of different DNA extraction kits on their suitability for a variety of sample types has, to our knowledge, not been performed yet. It is interesting, however, to analyze various microbiomes without causing bias due to the use of different extraction protocols, to understand the complexity and connectivity of microbiomes at different body sites in health and disease. Analyses of different biospecimens yield inconsistent results, which renders the selection of the very best protocol challenging. While for studies on single specimen types the best kit for the respective specimen can be selected, multi-microbiome studies potentially suffer from bias if different kits are used.

To understand microbiota in health and disease, multi-metagenomic studies that combine the microbiota from many samples of the same patients are however promising. We thus set out to identify a commercially available DNA extraction kit that is suitable to be used on such diverse biospecimens (**Figure 1**A). Here, we presented the data on the comparative extraction efficiency and sequencing quality obtained by whole-genome sequencing for six types of clinical samples (conjunctival swabs, stool, saliva, interdental plaque, bile, and sputum) after DNA extraction with three commercial kits, including 1) Qiagen DNeasy PowerSoil Pro (QPS; Qiagen, Hilden, Germany), 2) Qiagen QiAamp DNA Microbiome Kit (QMK; Qiagen, Hilden, Germany), and 3) ZymoBIOMICS DNA Miniprep Kit (ZYMO; Zymo Research Corp, Irvine, CA). QMK includes the advantage of host DNA depletion, presumably without causing taxonomic bias, which is a crucial step during DNA extraction for biospecimens such as skin and conjunctival swabs, for which more host material than bacterial mass is expected. This additional processing step might

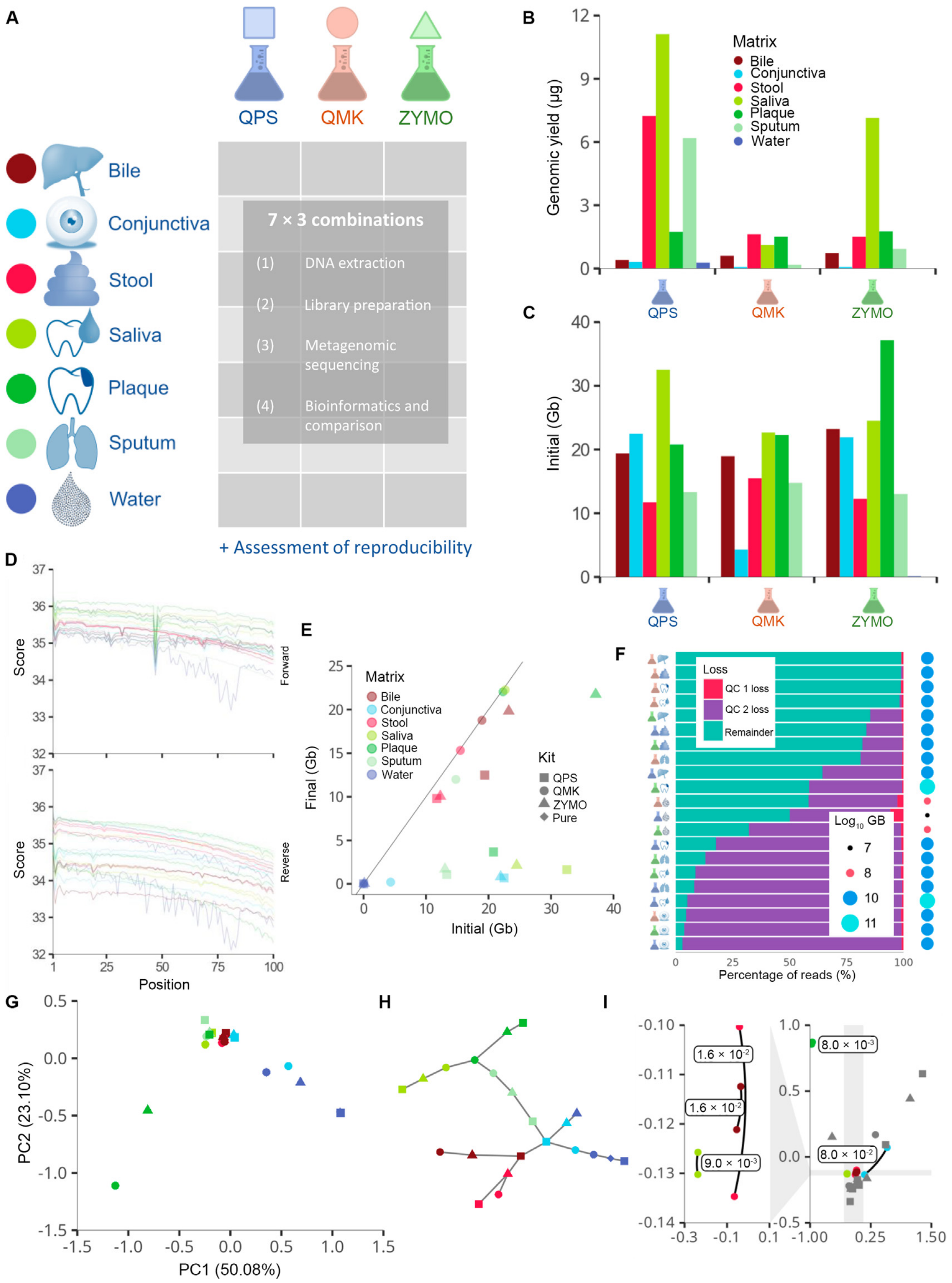


**Figure 1 Study setup and QC**

**A.** For six specimen types and water, we performed DNA extraction using three different commercial kits. Following library preparation and sequencing, the metagenomes were evaluated and compared to each other. **B.** The DNA yield of the different specimen types with different kits is given as a bar diagram. **C.** Comparison of the raw sequencing output in Gb before QC. **D.** Q30 values of the raw sequencing reads. Colors indicate the various biospecimens, in line with (A). **E.** Scatter plot of the raw reads to the reads obtained after QC. Shapes represent the different kits, and colors represent the different biospecimens. **F.** Percentage of reads filtered in the different QC steps and remaining dataset size. QC 1 mostly captures loss attributed to read quality, while QC 2 focuses on contamination by host sequences. **G.** Principal component analysis of the different samples and kits using the Mash distances after QC. Shapes represent the different kits, and colors represent the different biospecimens. **H.** Minimum spanning tree of the Mash distances after QC. Shapes represent the different kits, and colors represent the different biospecimens. **I.** Recomputed embedding displaying Mash distances between replicates. Grey points are without replicates. Colors indicate biospecimens. QC, quality control; Gb, gigabase; PC, principal component; QPS, Qiagen DNeasy PowerSoil Pro; QMK, QiAamp DNA Microbiome Kit; ZYMO, ZymoBIOMICS DNA Miniprep Kit.

be a potential explanation for the increased price of QMK in comparison to the two competitor kits tested in this study. In contrast to the novel QMK, we tested ZYMO and QPS that have both been used frequently in regard to microbiome analysis [24–27]. We followed a staged approach. We first performed a total of 108 DNA extractions and then chose the most promising samples for library preparation and sequencing. After evaluation of the sequencing data, we performed replicates for the best DNA extraction kit to analyze the reproducibility.

## Results

### DNA yield and sequencing quality vary between extraction kits and specimen types

As a first aspect, we compared the DNA yield and sequencing output for the different sample types and DNA extraction kits. The results showed that the DNA amount and concentration varied substantially between the different setups (Figure 1B, Figure S1F). It is known and expected that the different sample types — each with a different human background — lead to varying results in terms of reads and read quality. In line with the yield of DNA, the number of raw reads from the sequencing was likewise diverse (Figure 1C). Here again, the DNA extraction kits had a limited influence as compared to specimen types. However, the read quality in terms of Q30 value matched well, independent of the specimen type and DNA extraction kit, indicating that from all combinations interpretable microbiomes can be extracted (Figure 1D). The number of reads prior to and following quality control filtering generally correlated well for the different kits and sample types. Again, the fraction varied with the different specimen types depending on the expected human background, *e.g.*, introduced by human immune cells and human epithelial cells in saliva and conjunctiva, respectively (Figure 1E). This fact became more evident when considering the lost read fraction in the quality control steps. Again, independent of the kit, the conjunctival swab samples yielded only a fraction of 5% of all reads after quality control, dominated by the mapping of reads to the human genome (Figure 1F). Focusing on the fractions remaining after quality control, the QMK kit retrieved the highest amount of metagenomic information for each specimen type. However, the quantitative aspects were not the only criteria relevant for the selection of a kit, but also the composition of contents. Accordingly, we computed a 2-dimensional embedding using multidimensional scaling based on the Mash distances between samples (Figure 1G). Both, the embedding and the minimum spanning tree of the samples based on the mash distance, confirmed the general considerations: the kit has a limited influence on the output as compared to the difference introduced by the specimen types (Figure 1H). To provide further evidence for this behavior, we carried out technical replicates for five different QMK samples, demonstrating a high reproducibility of metagenomic measurements (Figure 1I).

In the light of the results in this section we might conclude that the variability introduced by the kits is so limited as compared to the difference between sample types and that for each specimen the very best kit might be selected even when multi-microbiome studies are performed. The high-level results, however, also call for a higher resolution analysis of the substantial metagenomic datasets.

### Metagenomes vary strongly between different DNA extraction kits, yet stronger between different specimen types

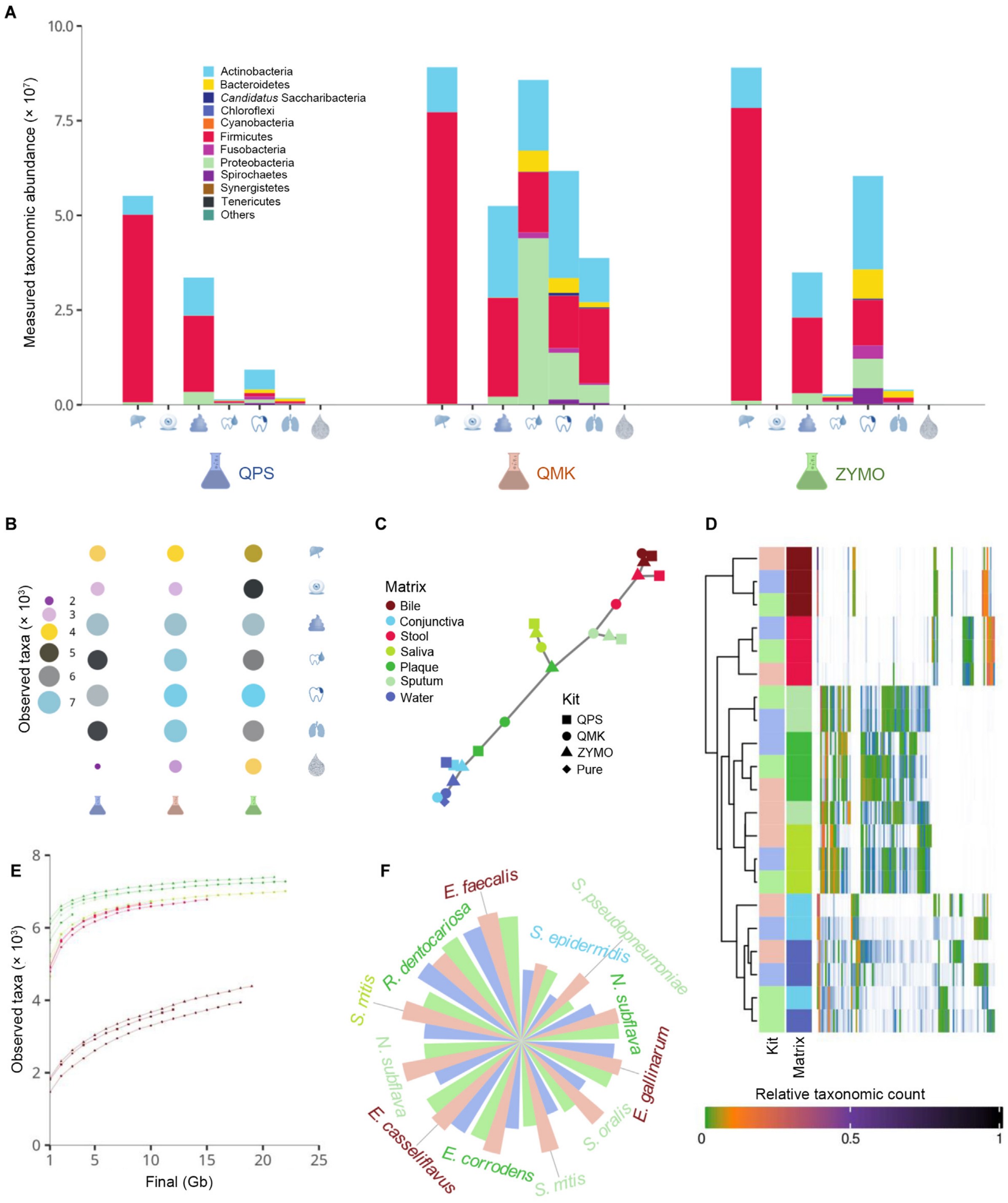
First, we computed the bacterial phyla and families contained in the different samples to get an overview of the taxonomic profile (**Figure 2**A, Figure S1A). Overall, the large quantitative differences in data yield reflect abundance counts. Again, strong differences in relative composition were present for the specimen types. A more detailed consideration revealed a low relative amount of Proteobacteria for several ZYMO- and QPS-extracted samples compared to those extraced by QMK. Especially, the relative portion of Firmicutes decreased in the QMK-extracted samples as compared to those extraced by the other two kits. Mostly Proteobacteria measurements profited from this shift, which was most pronounced in saliva.

### The alpha diversity crucially depends on the DNA extraction kit and the sequencing depth

Quantitative differences do not necessarily translate to qualitative differences. Accordingly, we investigated the alpha diversity of the various samples. Again, specimen types dominated the overall signal. For bile, conjunctiva, plaque, and water specimens, the ZYMO kit measured the highest number of differing taxonomies (Figure 2B). In case data analysis accounts for signals found in negative controls, the alpha diversity measured in ZYMO began to fall in line with the other two kits (Figure S1B). Despite high fluctuations in alpha diversity and total abundance across kits, bacterial species information recaptured most of the structure previously identified from read information alone, confirming the quality of the taxonomic profiling analysis (Figure 2C, Figure S1C). Investigation of the beta diversity based on dimensionality reduction showed a tendency to group QMK-extracted saliva and sputum samples with plaque samples (Figure S1E). Hereby, the comparably low bacterial abundances of the other two methods may act as a confounding factor. Consistent with the previous minimum spanning trees, we observed a clustering of specimen types into three major categories: 1) the close to sterile water and conjunctiva, 2) the digestive system-focused bile and stool, and 3) the oral cavity specimens saliva and plaque where sputum integrates. Looking closer into the clustering, it is clearly visible that a minority of species contributes a majority of the signal (Figure 2D, Figure S1D). Nevertheless, often rare taxonomies of minimal statistical weight may also be of interest for the analysis due to the potential harboring of *e.g.*, virulence factors. Therefore, to analyze the feasibility of finding rare species in the various environments, we further investigated the number of identified taxonomies changing with sampling depth by doing *in-silico* downsampling (Figure 2E). Hereby, we noted that the kits seemed to converge at a similar rate to their asymptotic behavior. This point of convergence was reliably achieved at around 10 Gb after quality control. The maximal number of taxonomies seemed to differ mostly by specimen types, yet minor differences were also visible for kits, which is consistent with the previous finding. Mass spectrometry (MS)-based identification of 42 colonies indicated for 12 significant results that QMK generated highest counts for all but one confirmed species (Figure 2F). We noted that two species were not detected in our genomic data analysis at all, but were found during MS, which were *Veillonella rogosae* and *Capnocytophaga granulosa* in saliva and sputum, respectively.

**The composition of microbiota considerably varies between DNA extraction kits**

The taxonomic composition of a microbiome is often the key aspect to reveal during a metagenomic experiment. Hereby,



the selected DNA extraction kit may play a crucial role on the qualitative findings of an experiment. While the previously discussed alpha diversity described the general number of different taxonomies captured by an experiment, it failed to discuss the exact nature of these differences. Therefore, we looked at the overlapping sets of detected species across both specimen types and kits (**Figure 3**A). Hereby, we selected a raw abundance count threshold to decide about the presence of a species instead of selecting by relative abundance, to also consider rare species in the analysis whose relative counts may undercut relative thresholds. We first discussed the common species of the individual specimen types. The largest intersecting set is usually the set encompassing all three kits. Only for sputum and water, the consensus was the largest for ZYMO and QMK. For the majority of time, ZYMO built the largest intersections, likely due to frequently constituting the largest stand-alone set. Next, we glanced at potential species that were found independently of input samples for the different kits. Here, the largest intersections were the ones with the largest initial sets. Due to higher measured bacterial abundances, QMK proposed four larger sets including sputum and saliva, whereas ZYMO and QPS only proposed stool and plaque as larger sets. Ignoring the underlying specimen types and aggregating the analysis, ZYMO and QMK had the largest number of species they detected in any specimen type.

Since taxonomic profiling is often limited by the quality and amount of reference organisms available, we further investigated ways to discuss potential differences in taxonomic composition between experiments that remain uncaptured by reference-based analysis. Hereby, we fell back on the core algorithm of BusyBee [28]. Accordingly, for reference-free analysis a Uniform Manifold Approximation and Projection (UMAP) embedding of normalized $k$-mer counts was computed on assembled scaffolds (Figure 3B). Visually, the embedding confirmed several findings of the previous taxonomic profiling. The overall density of the embeddings falls in line with the findings of the alpha diversities. Overall, it appears that ZYMO generates the highest density regions and is spreading all over the two-dimensional plane. While the embedding computed on QPS samples also scatters, there are fewer high-density regions. Last, QMK produces well defined regions of higher density. Moreover, the two clusters found in ZYMO water can be seen in all other samples except for the QPS water and QMK water. However, the left cluster also seems to disappear in QPS conjunctiva.

### Assembly quality depends on the specimen types and the DNA extraction kits

For the previous reference-free analysis, assembly quality was comparably of minor importance due to the decomposition of assembled sequences into short $k$-mers. Yet, depending on further downstream analysis, the quality of metagenomic assembly may play a crucial role. Accordingly, we compared several assembly quality measures across kits and specimen types (**Figure 4**A). Considering length distribution, specimen types were mostly clustered together. However, for the three specimen types of saliva, plaque, and sputum, minor differences were visible with respect to kits, favoring QMK in N50 and N75 measures. Considering the proportion of scaffolds at changing length, QMK was the only kit where no specimen started to dominate after a given length. Last, ZYMO generated the longest assemblies in water and dominated L50 and L75 for water and conjunctiva.

### Taxonomic profiles are consistent across sequencing technologies

With the rising popularity of nanopore sequencing technology and the immense advantages it brings to metagenomics, in terms of assembly quality increase and interpretability, hybrid protocols combining shotgun and nanopore sequencing are continuously gaining in relevance. Correspondingly, the demands to kits supporting both protocols are favored. Since we previously demonstrated clustering behavior into three major clusters, we selected saliva and bile as representatives of the non-sterile specimen types and sequenced the same samples again with nanopore sequencing. Similarly, we removed the ZYMO kit for the experiment, due to our previous findings of a high number of false positives shown across several of our herein presented analyses. Quality control of nanopore reads of all four samples after filtering suggested minor differences between specimen types for both kits (Figure 4B). Pearson correlation between length and PHRED scores was around a low 0.1. Considering both, read length and average read quality scores, Wilcoxon rank sum tests across all reads indicated statistically significant differences between kits, conditioned on the specimen types for each sample ($P < 1 \times 10^{-12}$). After quality control, taxonomic profiling was performed to gauge the effects of interactions among kits, specimen types, and

◀

**Figure 2 Diversity of microbiota**

**A.** Bar plot presenting the composition of bacterial microbiota with respect to different phyla. The color codes represent the phyla. Specimen types corresponding to each DNA extraction kit are in the following order: bile, conjunctival swab, stool, saliva, plaque, sputum, and water (left to right). **B.** For the seven biosamples, the observed unfiltered alpha diversity is presented. The color and bubble size correspond to the alpha diversity. Large and blue bubbles match samples with highest alpha diversity, and the small and purple bubbles match samples with lowest alpha diversity. **C.** Minimum spanning tree on the bacterial species level. Jaccard distance served as distance measure. Shapes represent the different kits, and colors represent the different biospecimens. **D.** Heatmap representing the abundance of different species clustered with respect to the sample type. Only species with a relative abundance above 1% were considered. Colors used to represent different kits are green for ZYMO, blue for QPS, and pink for QMK, in line with (A). Each biospecimen (matrix) is represented by a different color consistent with (C). Relative taxonomic counts are depicted in green for 0, black for 1, and shades of green, orange, and violet with increasing relative taxonomic counts between 0 and 1. **E.** Fraction of observed taxa with respect to the sequencing depth computed on sub-sampled decontaminated reads. Shapes represent the different kits, and colors represent the different biospecimens. **F.** Barplot displaying unnormalized counts in the NGS experiment for the different taxonomies that were detected by mass spectrometry. Font colors indicate specimen types, in line with (C). Bar colors represent the different kits, matching (A). NGS, next-generation sequencing.

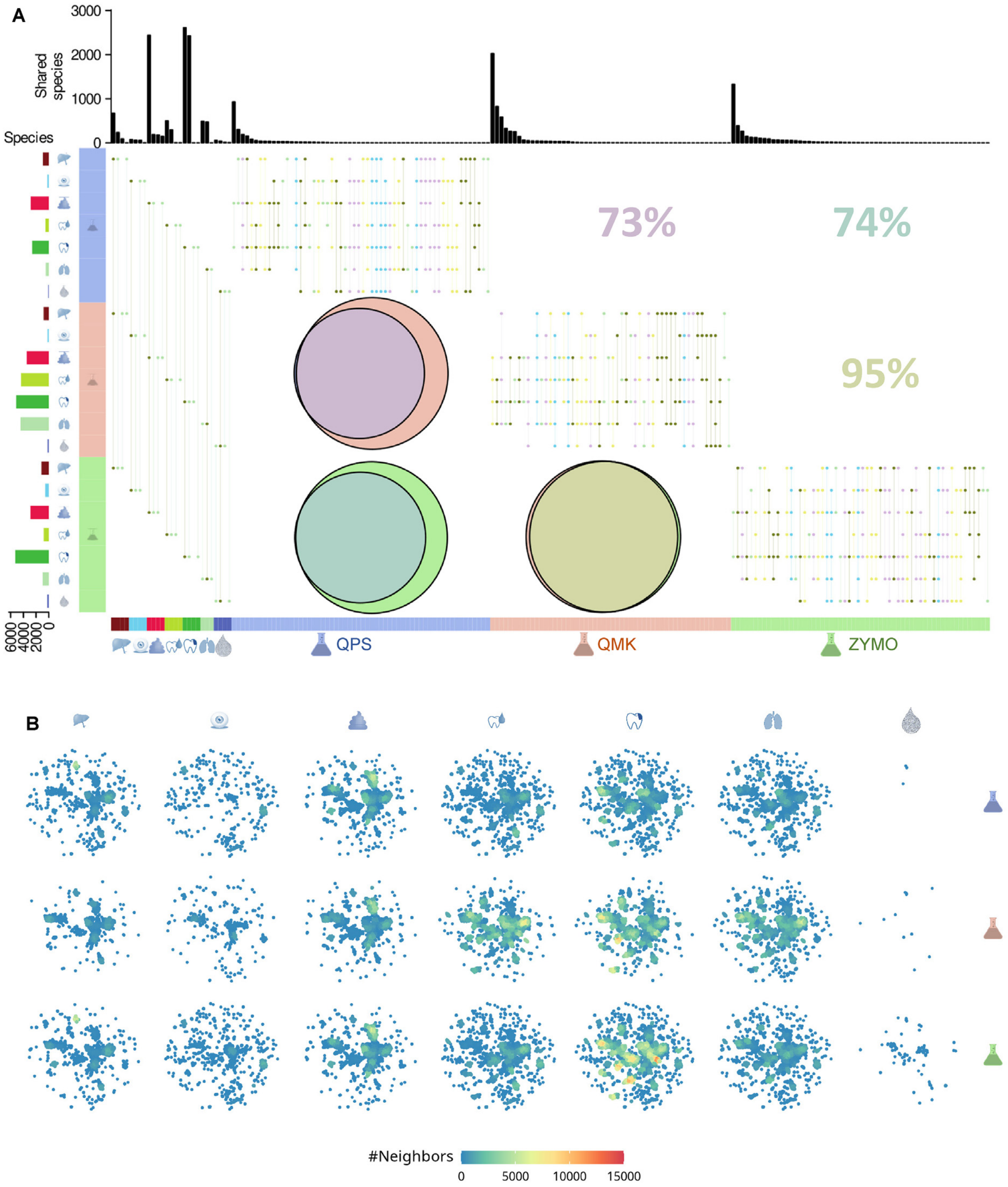


**Figure 3 Similarity of microbiota**

**A.** Combination of ten upset plots each discussing identified species overlap by kit or specimen type. Bottom annotation indicates the aspect the upset plot focuses on, *i.e.*, which kit or specimen type is kept constant. Area proportional Euler diagrams below the diagonal capture the proportion of species identified independently of specimen type. Percentages above the diagonal indicate the overlap numerically. A species is considered identified after surpassing a low count threshold of 20 occurrences. **B.** Embedded microbiota. Each spot represents one scaffold with a length above 3 kb after an embedding of the *k*-mer spectra using UMAP. Colors are indicative of the point density in the respective area. UMAP, Uniform Manifold Approximation and Projection.

sequencing technologies (Figure 4C). As expected, the number of different identified bacterial species and overall abundance were a lot higher for shotgun sequencing due to the generation of false positives during profiling with short reads and the increased sequencing depth. Overall, the ordering of experiments based on uniquely quantified bacterial species remains

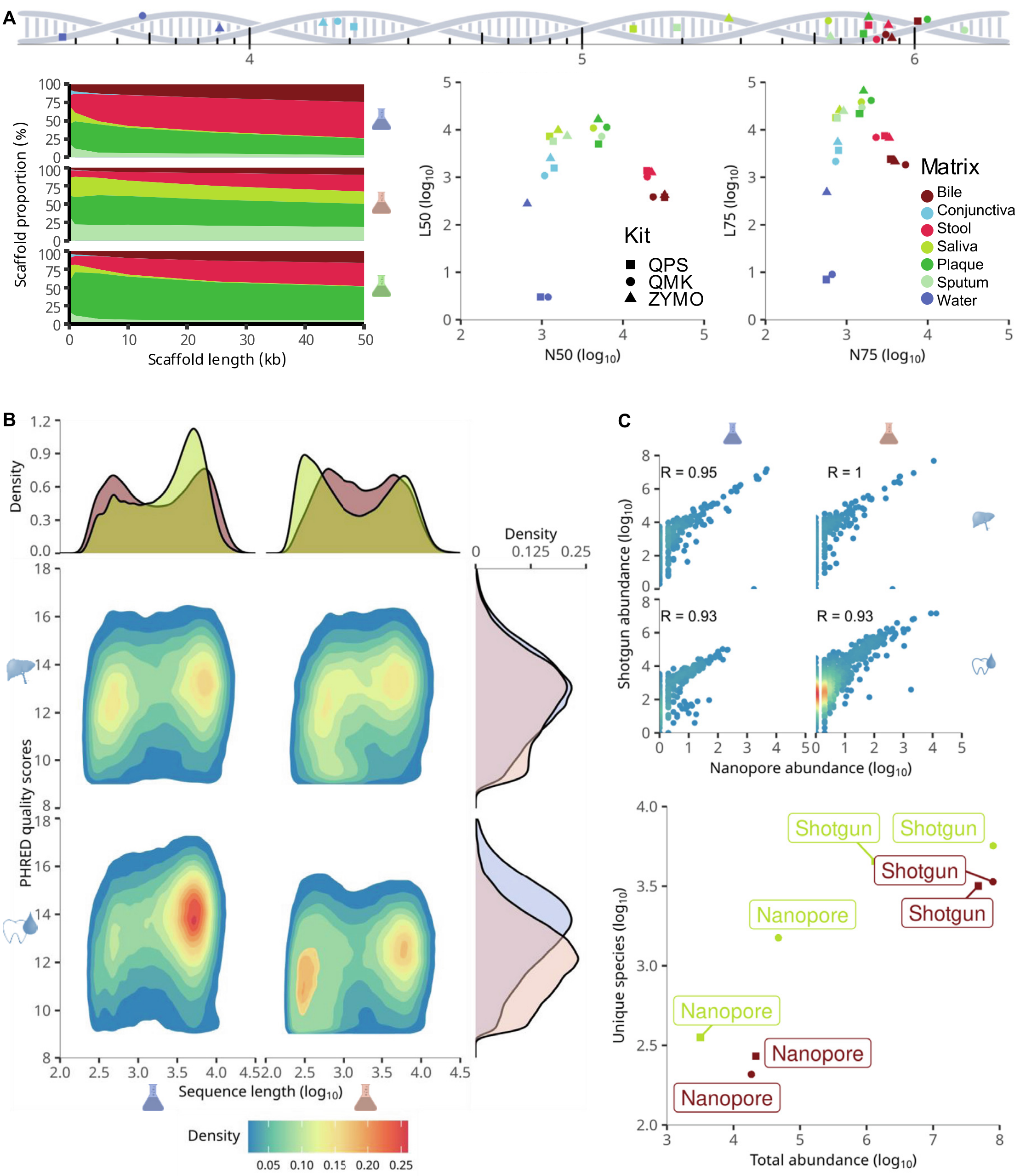


**Figure 4 Assembly and nanopore comparison**

**A.** Assembly quality. One-dimensional line showing the length of the longest scaffold for each assembly. Relative scaffold length distribution by kits together with their N50, N75, L50, and L75 values. **B.** Nanopore QC. Average PHRED scores indicate basecalling quality per read. Sequence length indicates length after basecalling of each read. Density plots on the right and top discuss the conditioned distributions for the different kits and specimen types. Visualized data are representative for data after filtering. **C.** Correlation plot indicating coherence between nanopore and shotgun sequencing taxonomic counts on bacterial species level without thresholding. Numerical values represent the rounded Pearson correlation before log scaling. The total number of measurements and unique different species for the different experiments are shown at the bottom.

unchanged across sequencing technologies. The only difference was that in bile, where QMK detected fewer unique species than QPS for the nanopore technology which may be linked to the difference in sampling depth and correlated total abundance. Glancing at the taxonomic profile on a bacterial species level, we observed strong correlations between nanopore and shotgun sequencing, for all tested kit and specimen combinations.

## Discussion

We evaluated three DNA extraction kits across six different specimen types and water to gauge their suitability for metagenomic experiments. We note that the QMK kit usually yields the highest amount of metagenomic information after host DNA removal. The depletion of human DNA is a significant advantage of QMK compared to ZYMO and QPS. This is consistent with the idea, to lyse all eukaryotic cells in a first step, followed by the degradation of eukaryotic DNA. Therefore, human DNA in particular is depleted during the first part of the DNA extraction with the QMK kit. During the second part, bacterial cells are lysed and the extracted DNA is purified.

Focusing on both, read information and metagenomic data analysis, we showed that the selection of the specimen type dominated the selection of the kit in signal strength. While for the difference between *e.g.*, water and stool, this result was to be expected, the same did not hold true for plaque and saliva samples. Further, we demonstrated the sensitivity of all kits by confirming a selection of taxa using MS. Considering specificity, we demonstrated, using a reference-based and reference-free method, that ZYMO appeared to contain most contamination, going hand in hand with the fact of ZYMO generating the samples with the highest relative amount of human contamination. This effect could be due to unsterile lysis tubes or columns for DNA extraction. However, we noted that no sample remained uncontaminated. The lowest contamination was shown for QPS. Especially in comparison to the QMK water sample, a lower contamination of the QPS sample can be explained by a general lower variety in identified bacteria species. Partly, bacteria in the environment, that contaminate the water samples might be harder to lyse, which the QPS kit might not have to offer. In contrast to the ZYMO and QPS kits, a pre-contamination of columns provided by the QMK kit is highly unlikely, as the special Qiagen ultraclean columns were stored at 4 °C until being used for DNA extraction. Shifting focus away from taxonomic profiling onto assemblies, except for the ZYMO water sample, assemblies were of similar quality at first sight. Here we acknowledge that the scaffold length distribution is not the be-all and end-all of metagenomic assembly quality assessment; however, it is one of the more widely spread [29]. Last, for QPS and QMK we observed that overall, the results after metagenomic analysis are consistent across shotgun and nanopore sequencing. We note that our study is limited by the small sample size and the focus on bacterial microorganisms. Random sampling error may distort our findings. Thus, larger studies, including more replicates, are needed to confirm our results, and similar comparative studies should ideally also assess results for other pathogen classes, such as viruses or parasites [30].

To conclude, we recommend the QMK kit for samples with high eukaryotic host contamination, as it clearly has the least information loss upon host sequence removal. Moreover, if no detection threshold is set, QMK identifies generally more species than QPS, while not showing a strong contamination of sequencing results in sterile water as compared to ZYMO. In case host contamination is not an issue to consider, QPS may be recommended, since it shows the least overall contamination in sterile water.

## Materials and methods

### Sample collection

In brief, stool samples were collected by each participant using a paper toilet-hat and a sterile collection tube with an integrated spoon. Approximately 500 mg to 1 g of stool were collected. Plaque samples were collected using 12 disposable micro applicators (Catalog No. MSF400, Microbrush International, Grafton, WI). Three interdental spaces per quadrant were brushed, and all micro applicators were placed into a single ESwab transport tube (Copan Diagnostics, Brescia, Italy), including the ESwab Amies Medium (Copan Diagnostics). Saliva samples were collected using 50-ml sterile, conic falcon tubes. Participants were asked to release uninduced saliva into the sterile falcon tube for 5 min. Conjunctiva samples were obtained using a ESwab. The lower eyelid was everted, and the conjunctiva was swabbed throughout the entire length of the lower fornix three times. Afterwards, the swab was placed in the respective transport medium and the tube was frozen at −80 °C. Sputum was induced by 7 min of inhalation with 0.9% NaCl solution. After inhalation, the participant was asked to release sputum by coughing into a sterile collection tube. Bile samples were collected during a duodenoscopy by drawing 5 ml to 10 ml into a sterile syringe.

### DNA extraction

DNA was extracted from all samples using three different, commercially available DNA extraction kits: QPS, QMK, and ZYMO. For each kit, the DNA was extracted according to the manufacturer's protocol. Briefly, 1 ml of sterile Milli-Q water was used for the negative control. The manufacturer's protocol was followed, respectively. Fecal samples were weighed, and 250 mg of stool were used for DNA extraction with QPS and QMK, and 50 mg of stool were used for ZYMO, according to the manufacturer's recommendation. For QPS and ZYMO, 1.5 ml of saliva samples were centrifuged for 5 min at 6000 *g* and the pellet was resuspended in the respective lysis buffer. For QMK, 1 ml of saliva was used directly. Interdental microbrushes and conjunctival swabs were vortexed rigorously in the eSwab Amies Medium for 3 min. The Amies Medium was then transferred to a 1.5-ml sterile Eppendorf tube and centrifuged for 5 min at 6000 *g* for further DNA extraction with QPS and ZYMO. The pellet was resuspended in the respective lysis buffer. For DNA extraction with QMK, the liquid Amies Medium was used directly. For DNA extraction with QPS and ZYMO, bile samples were vortexed rigorously and 2 ml of bile were transferred to a 2-ml sterile Eppendorf tube and centrifuged for 5 min at 6000 *g*. The supernatant was discarded, and the pellet was resuspended in the respective lysis buffer. To extract DNA from bile via QMK, 1 ml of bile was used directly. Sputum was mixed with Remel Sputasol (Oxoid L TD, Hants, England) in a 1:1 ratio. For QPS and ZYMO, 1.5 ml of sputum or sputasol was centrifuged for 5 min at 6000 *g* and the pellet was resuspended in the respective lysis buffer. For QMK, 1 ml of sample was used for DNA extraction without previous centrifuging. The mechanical lysis of bacterial cells was performed using the MP Biomedicals FastPrep-24 5G Instrument (FisherScientific

GmbH, Schwerte, Germany). For ZYMO, the velocity and duration were adjusted to 6 m/s for 45 s three times with 30 s of storage on ice in between each lysis step. For elution of DNA during the last step of each DNA extraction kit, the following elution volumes were used: 1) QPS: 40 µl; 2) ZYMO: 20 µl; 3) QMK: 50 µl. The DNA concentration was determined via NanoDrop 2000/2000c (ThermoFisher Scientific, Wilmington, DE) full-spectrum microvolume UV–Vis measurements. For each sample type and each DNA extraction method tested, we used a total of one biological replicate for sequencing. However, DNA was isolated from a total of n = 10 biological replicates for saliva, interdental plaque, and stool, a total of n = 4 for bile, a total of n = 8 for sputum samples, and a total of n = 4 for conjunctival swabs. From all samples that we extracted DNA from, we selected the most promising samples for library preparation and sequencing. We chose those samples with the highest amount of DNA, least impurities, and least fragmentations. For all samples prepared with QMK we performed an n = 2 technical replicates for library preparation and sequencing.

### Library preparation

DNA libraries were prepared using the MGIEasy Universal DNA Library Prep Set (MGI Technologies, Shenzhen, China) according to the manufacturer's recommendations. In general, 200 ng DNA was sheared into fragments using the M220 Focused-ultrasonicator (Covaris, Woburn, MA), followed by size selection using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA). For low-biomass samples, such as the conjunctival swab and the sterile water control, the entire amount of isolated DNA was used as an input for the fragmentation procedure. The fragmented DNA was used for end-repairing and A-tailing. Next, adaptors containing specific barcodes were ligated to the 3′ and 5′ ends, and the ligation products were amplified by PCR. The concentration of the PCR products was measured using Qubit 1× dsDNA HS Assay Kit (ThermoFisher Scientific, Waltham, MA). In the following, 8 different barcoded samples were pooled in equal amount and circularized to generate the single-stranded DNA library. The concentration of the library was measured using Qubit ssDNA Assay Kit (ThermoFisher Scientific, Waltham, MA). Additionally to the different biospecimen samples, a sterile DNase/RNase-free water sample was prepared using the same procedure as for all samples.

### NGS

For the short-read sequencing, all libraries were sent to BGI Group for DNA nanoball (DNB) generation and paired-end sequencing (PE100) on the DNBSEQ-G400 instrument according to manufacturer's instructions and recommendations.

### MinION library preparation and sequencing

Upon opening of the flow cell and again immediately prior to sequencing, flow cell pore count was measured using MinKNOW. Library preparation kits, flow cell, and other consumables used for the experiment are described in Table S1. DNA was quantified via Nanodrop 2000/2000c (ThermoFisher Scientific, Wilmington, DE) and the volume was determinated by using a pipette (Table S2). The library preparation was conducted according to the protocol "Native barcoding genomic DNA (with EXP-NBD 104)" provided by Oxford Nanopore Technologies (ONT), with the exception of the barcode ligation step and further the adapter ligation step for which the ligation mix was incubated for 15 min at room temperature instead of 10 min. The amount of initial DNA used for the barcoding kit was above 100 ng for the four specimen types corresponding to the DNA extraction kits shown in this study. In sum, the library consisted of 12 barcoded DNA samples. The barcoded DNA was stored at 4 °C for 3 days until adapter ligation. For barcoded libraries, volume-equal quantities of each sample were used for the final library. The amount of pooled barcoded DNA exceeded the recommended amount of 700 ng DNA by an additional 400 ng to reach a final DNA amount of about 1100 ng for adapter ligation (Table S3). For the last Agencourt Ampure XP bead clean-up step, short fragment buffer (SFB) was used. The completed library was loaded onto a R9.4 flow cell as per instructions given by ONT. Given the rapid advancement of protocols, chemicals, and the technology itself, data were generated with the most up-to-date methods and protocols available from ONT at the time of library preparation and sequencing. The Mk1B MinION device was used for data acquisition.

### Nanopore sequencing

MinION analysis was carried out at the Helmholtz Institute for Pharmaceutical Research Saarland (HIPS) at the Department Microbial Natural Products, Saarbrücken, Germany. The barcoded library, consisting of the metagenomic DNA samples, was generated in a S1 laboratory, whereas the sequencing of the samples was performed in the office. Sequencing methods performed simultaneous 1D sequencing of samples using native barcoding. The sequencing run was carried out over a time range of three days. At the time of use, the R9.4 spotON Flow Cell had a pore count exceeding the guaranteed level ( > 800 pores) by the manufacturer. Pore count was measured by the MinKNOW software with a result of 808 pores. The majority ( > 50%) of sequencing data were generated in the first 9 h of sequencing, corresponding to the time in which the first group of pores is actively sequencing. More than 99% of sequencing data were generated after 28 h of sequencing. The sequencing yield in a total number of estimated bases is displayed in Figure S2.

### Culturing of bacteria

All native samples were streaked out on four different agar plates: TSA with 5% sheep blood (TSA), MacConkey (MC), Columbia (Co), and Chocolate Blood (CB) agar plates (ThermoFisher Scientific, Wilmington, DE). All TSA, MC, and CB agar plates were incubated at 35.6 °C and 5% $CO_2$ for a minimum of 18 h and a maximum of 24 h. Co agar plates were used for the cultivation of anaerobic bacteria and therefore incubated in an anaerobic environment for a minimum of 48 h.

### MS-based identification

Bacterial colonies, obtained by culturing on different agar plates, were spotted onto the MALDI-TOF target plate, followed by overlaying with 1 μl of α-cyano-4-hydroxycinnamic acid (CHCA) matrix solution (Bruker Daltonics), composed of saturated CHCA dissolved in 50% (v/v) of acetonitrile, 2.5% (v/v) of trifluoroacetic acid, and 47.5% (v/v) of LC-MS grade water. After drying at room temperature, the plate was placed into the Microflex LT Mass Spectrometer (Bruker Daltonics) for MALDI-TOF MS. Measurements were performed using the AutoXecute algorithm in the FlexControl software (v3.4; Bruker Daltonics). For each spot, 240 laser shots in six random positions were carried out automatically to generate protein mass profiles in linear positive ion mode with a laser frequency of 60 Hz, a high voltage of 20 kV, and a pulsed ion extraction of 180 ns. Mass charge ratio range (*m*/*z*) was measured between 2 kDa and 20 kDa. Bacterial species were identified by using the MALDI BioTyper software. Identification scores above 2.0 were considered a precise identification, scores between 1.7 and 1.99 were considered as possible species identification, and all identification scores below 1.7 were considered unsuccessful identification.

### Data analysis

First, quality control was performed with MultiQC (v1.9) [31] and fastp (v0.20.1) [32]. Next, NGS data were decontaminated of host sequences using kneaddata (v0.7.4). Decontaminated data were uploaded to the Sequence Read Archive (SRA) [33]. We counted the exact number of basepairs contained in the fasta files before the individual steps to get a detailed overview on the overall information content. Once the data were fully cleaned, Mash distances were computed on all remaining read information with Mash (v2.3) [34]. Taxonomic profiling was done with Kraken (v2.1.2) [35]. Optional downsampling of reads was performed with seqtk (v1.3). The PlusPF database release from 9/19/2020 was used as Kraken2 index. As an alpha diversity measure, we used either the observed number of different taxa or the Shannon index. As the beta diversity measure, the Jaccard index was computed. For clustering analysis, species with relative species abundance below 1% in all samples were removed. Samples were then clustered using Ward's hierarchical agglomerative clustering in combination with the Euclidian distance measure. UMAP embeddings were computed on all scaffolds having a length over 3 kb. To this end, 5-mers of each scaffold were counted and assembled into a vector. Each vector was divided by its sum, scaled, and centered. The normalized counts were then passed to embedded using UMAP. Assemblies were computed with SPAdes (v3.15.2) using the --meta flag [36]. Scaffold quality assessment was made with MetaQUAST (v5.0.2) [37], enabling the splitting of scaffolds. Downstream analysis heavily relied on phyloseq (v1.36.0). Nanopore reads were basecalled with guppy (v5.0.7) [38] before undergoing taxonomic profiling.

## Ethical statement

All samples were collected at the Saarland University Medical Center, Germany, after having obtained written informed consent from all participants. The study was approved by the local ethics committee (Ärztekammer des Saarlandes) under reference 131/20.

## Data availability

Respecting the German Bundesdatenschutzgesetz, we uploaded the data after human read removal to the SRA of National Center for Biotechnology Information (NCBI). Preprocessed data can be found in SRA of NCBI (SRA: PRJNA802336), and are publicly accessible at https://www.ncbi.nlm.nih.gov/sra.

## CRediT author statement

**Jacqueline Rehner:** Methodology, Validation, Investigation, Writing - original draft. **Georges Pierre Schmartz:** Methodology, Software, Validation, Formal analysis, Data curation, Writing - original draft, Visualization. **Laura Groeger:** Methodology, Investigation, Writing - review & editing. **Jan Dastbaz:** Methodology, Investigation, Writing - review & editing. **Nicole Ludwig:** Supervision, Writing - review & editing. **Matthias Hannig:** Resources, Supervision, Writing - review & editing. **Stefan Rupf:** Resources, Supervision, Investigation, Writing - review & editing. **Berthold Seitz:** Resources, Supervision, Writing - review & editing. **Elias Flockerzi:** Supervision, Investigation, Writing - review & editing. **Tim Berger:** Investigation, Writing - review & editing. **Matthias Christian Reichert:** Supervision, Investigation, Writing - review & editing. **Marcin Krawczyk:** Resources, Supervision, Writing - review & editing. **Eckart Meese:** Resources, Supervision, Writing - review & editing. **Christian Herr:** Supervision, Investigation, Writing - review & editing. **Robert Bals:** Conceptualization, Methodology, Resources, Supervision, Project administration, Funding acquisition, Writing - review & editing. **Sören L. Becker:** Conceptualization, Methodology, Resources, Supervision, Project administration, Funding acquisition, Writing - review & editing. **Andreas Keller:** Conceptualization, Methodology, Software, Resources, Supervision, Project administration, Funding acquisition, Writing - review & editing. **Rolf Müller:** Conceptualization, Methodology, Resources, Supervision, Project administration, Funding acquisition, Writing - review & editing. All authors have read and approved the final manuscript.

## Competing interests

Georges Pierre Schmartz, Matthias Hannig, Stefan Rupf, Andreas Keller, and Rolf Müller are shareholders of MOOH GmbH.

## Acknowledgments

This work was supported by the UdS-HIPS-TANDEM Initiative.

## Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.05.006.

## ORCID

ORCID 0000-0003-1335-4898 (Jacqueline Rehner)
ORCID 0000-0002-9627-9223 (Georges Pierre Schmartz)
ORCID 0000-0003-1043-3910 (Laura Groeger)
ORCID 0000-0002-0880-9678 (Jan Dastbaz)
ORCID 0000-0003-4703-7567 (Nicole Ludwig)
ORCID 0000-0003-0669-6881 (Matthias Hannig)
ORCID 0000-0002-1551-9935 (Stefan Rupf)
ORCID 0000-0001-9701-8204 (Berthold Seitz)
ORCID 0000-0002-0423-3624 (Elias Flockerzi)
ORCID 0000-0003-2307-2237 (Tim Berger)
ORCID 0000-0002-8192-0575 (Matthias Christian Reichert)
ORCID 0000-0002-0113-0777 (Marcin Krawczyk)
ORCID 0000-0001-7569-819X (Eckart Meese)
ORCID 0000-0002-9782-1117 (Christian Herr)
ORCID 0000-0002-1472-9535 (Robert Bals)
ORCID 0000-0003-3634-8802 (Sören L. Becker)
ORCID 0000-0002-5361-0895 (Andreas Keller)
ORCID 0000-0002-1042-5665 (Rolf Müller)

## References

[1] NIHHMPA Team. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007–2016. Microbiome 2019;7:31.

[2] Eisenstein M. The hunt for a healthy microbiome. Nature 2020;577:S6–8.

[3] Drew L. Highlights from studies on the gut microbiome. Nature 2020;577:S24–5.

[4] Wang DD, Nguyen LH, Li Y, Yan Y, Ma W, Rinott E, et al. The gut microbiome modulates the protective association between a Mediterranean diet and cardiometabolic disease risk. Nat Med 2021;27:333–43.

[5] Romano S, Savva GM, Bedarf JR, Charles IG, Hildebrand F, Narbad A. Meta-analysis of the Parkinson's disease gut microbiome suggests alterations linked to intestinal inflammation. NPJ Parkinsons Dis 2021;7:27.

[6] Chalermwatanachai T, Vilchez-Vargas R, Holtappels G, Lacoere T, Jauregui R, Kerckhof FM, et al. Chronic rhinosinusitis with nasal polyps is characterized by dysbacteriosis of the nasal microbiota. Sci Rep 2018;8:7926.

[7] Koeller K, Herlemann DPR, Schuldt T, Ovari A, Guder E, Podbielski A, et al. Microbiome and culture based analysis of chronic rhinosinusitis compared to healthy sinus mucosa. Front Microbiol 2018;9:643.

[8] Mullish BH, Williams HR. *Clostridium difficile* infection and antibiotic-associated diarrhoea. Clin Med (Lond) 2018;18:237–41.

[9] Radlinski L, Rowe SE, Kartchner LB, Maile R, Cairns BA, Vitko NP, et al. *Pseudomonas aeruginosa* exoproducts determine antibiotic efficacy against *Staphylococcus aureus*. PLoS Biol 2017;15: e2003981.

[10] Zipperer A, Konnerth MC, Laux C, Berscheid A, Janek D, Weidenmaier C, et al. Human commensals producing a novel antibiotic impair pathogen colonization. Nature 2016;535:511–6.

[11] Vieco-Saiz N, Belguesmia Y, Raspoet R, Auclair E, Gancel F, Kempf I, et al. Benefits and inputs from lactic acid bacteria and their bacteriocins as alternatives to antibiotic growth promoters during food-animal production. Front Microbiol 2019;10:57.

[12] Demain AL, Fang A. The natural functions of secondary metabolites. Adv Biochem Eng Biotechnol 2000;69:1–39.

[13] Miethke M, Pieroni M, Weber T, Brönstrup M, Hammann P, Halby L, et al. Towards the sustainable discovery and development of new antibiotics. Nat Rev Chem 2021;5:726–49.

[14] Watson EJ, Giles J, Scherer BL, Blatchford P. Human faecal collection methods demonstrate a bias in microbiome composition by cell wall structure. Sci Rep 2019;9:16831.

[15] Menon RK, Gopinath D. Eliminating bias and accelerating the clinical translation of oral microbiome research in oral oncology. Oral Oncol 2018;79:84–5.

[16] Douglas CA, Ivey KL, Papanicolas LE, Best KP, Muhlhausler BS, Rogers GB. DNA extraction approaches substantially influence the assessment of the human breast milk microbiome. Sci Rep 2020;10:123.

[17] Neuberger-Castillo L, Hamot G, Marchese M, Sanchez I, Ammerlaan W, Betsou F. Method validation for extraction of DNA from human stool samples for downstream microbiome analysis. Biopreserv Biobank 2020;18:102–16.

[18] Bjerre RD, Hugerth LW, Boulund F, Seifert M, Johansen JD, Engstrand L. Effects of sampling strategy and DNA extraction on human skin microbiome investigations. Sci Rep 2019;9:17287.

[19] Mattei V, Murugesan S, Al Hashmi M, Mathew R, James N, Singh P, et al. Evaluation of methods for the extraction of microbial DNA from vaginal swabs used for microbiome studies. Front Cell Infect Microbiol 2019;9:197.

[20] Oriano M, Terranova L, Teri A, Sottotetti S, Ruggiero L, Tafuro C, et al. Comparison of different conditions for DNA extraction in sputum - a pilot study. Multidiscip Respir Med 2019;14:6.

[21] Wang JCC, Wang A, Gao J, Cao S, Samad I, Zhang D, et al. Technical brief: isolation of total DNA from postmortem human eye tissues and quality comparison between iris and retina. Mol Vis 2012;18:3049–56.

[22] Pérez-Losada M, Crandall K, Freishtat RJ. Comparison of two commercial DNA extraction kits for the analysis of nasopharyngeal bacterial communities. AIMS Microbiol 2016;2:108–19.

[23] Stinson LF, Keelan JA, Payne MS. Comparison of meconium DNA extraction methods for use in microbiome studies. Front Microbiol 2018;9:270.

[24] Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome 2018;6:42.

[25] Heravi FS, Zakrzewski M, Vickery K, Hu H. Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. J Microbiol Methods 2020;170:105856.

[26] Kwan SY, Jiao J, Joon A, Wei P, Petty LE, Below JE, et al. Gut microbiome features associated with liver fibrosis in Hispanics, a population at high risk for fatty liver disease. Hepatology 2022;75:955–67.

[27] Sinha SR, Haileselassie Y, Nguyen LP, Tropini C, Wang M, Becker LS, et al. Dysbiosis-induced secondary bile acid deficiency promotes intestinal inflammation. Cell Host Microbe 2020;27:659–70.

[28] Laczny CC, Kiefer C, Galata V, Fehlmann T, Backes C, Keller A. BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. Nucleic Acids Res 2017;45: W171–9.

[29] Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, et al. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinform 2019;20:1140–50.

[30] Schneeberger PHH, Becker SL, Pothier JF, Duffy B, N'Goran EK, Beuret C, et al. Metagenomic diagnostics for the simultaneous detection of multiple pathogens in human stool specimens from Cote d'Ivoire: a proof-of-concept study. Infect Genet Evol 2016;40:389–97.

[31] Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32:3047–8.

[32] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90.

[33] Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res 2021;49:D121–4.

[34] Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17:132.

[35] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20:257.

[36] Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res 2017;27:824–34.

[37] Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 2016;32:1088–90.

[38] McMurdie PJ, Holmes S. Phyloseq: a bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. Pac Symp Biocomput 2012;17:235–46.

## 3.2 *Effects of Resistant Starch on Symptoms, Fecal Markers, and Gut Microbiota in Parkinson's Disease — The RESISTA-PD Trial*

Genomics Proteomics Bioinformatics 20 (2022) 274–287

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

# Effects of Resistant Starch on Symptoms, Fecal Markers, and Gut Microbiota in Parkinson's Disease — The RESISTA-PD Trial

**Anouck Becker [1,#], Georges Pierre Schmartz [2,#], Laura Gröger [3,#], Nadja Grammes [2], Valentina Galata [2], Hannah Philippeit [1], Jacqueline Weiland [1], Nicole Ludwig [3], Eckart Meese [3], Sascha Tierling [4], Jörn Walter [4], Andreas Schwiertz [5], Jörg Spiegel [1], Gudrun Wagenpfeil [6], Klaus Faßbender [1], Andreas Keller [2,7,#], Marcus M. Unger [1,#,*]**

[1] *Department of Neurology, Saarland University, D-66421 Homburg, Germany*
[2] *Chair for Clinical Bioinformatics, Saarland University, D-66123 Saarbrücken, Germany*
[3] *Department of Human Genetics, Saarland University, D-66421 Homburg, Germany*
[4] *Department of Genetics/Epigenetics, Saarland University, D-66123 Saarbrücken, Germany*
[5] *Institute of Microecology, D-35745 Herborn, Germany*
[6] *Institute of Medical Biometry, Epidemiology and Medical Informatics, Saarland University, D-66421 Homburg, Germany*
[7] *Department of Neurology, Stanford University, Palo Alto, CA 94305, USA*

Received 28 October 2020; revised 31 May 2021; accepted 22 November 2021
Available online 25 November 2021

Handled by Jingyuan Fu

**KEYWORDS**

Parkinson's disease;
Short-chain fatty acid;
Microbiota;
Metagenomics;
Intestinal inflammation

**Abstract** The composition of the gut microbiota is linked to multiple diseases, including Parkinson's disease (PD). Abundance of bacteria producing **short-chain fatty acids** (SCFAs) and fecal SCFA concentrations are reduced in PD. SCFAs exert various beneficial functions in humans. In the interventional, monocentric, open-label clinical trial "Effects of Resistant Starch on Bowel Habits, Short Chain Fatty Acids and Gut **Microbiota** in **Parkinson's Disease**" (RESISTA-PD; ID: NCT02784145), we aimed at altering fecal SCFAs by an 8-week prebiotic intervention with resistant starch (RS). We enrolled 87 subjects in three study-arms: 32 PD patients received RS (PD + RS), 30 control subjects received RS, and 25 PD patients received solely dietary instructions. We performed paired-end 100 bp length metagenomic sequencing of fecal samples using the BGISEQ platform at an average of 9.9 GB. RS was well-tolerated. In the PD + RS group, fecal butyrate concentrations increased significantly, and fecal calprotectin concentrations dropped significantly after 8 weeks of RS intervention. Clinically, we observed a reduction in non-motor

[*] Corresponding author.
E-mail: marcus.unger@uks.eu (Unger MM).
[#] Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2021.08.009

symptom load in the PD + RS group. The reference-based analysis of metagenomes highlighted stable alpha-diversity and beta-diversity across the three groups, including bacteria producing SCFAs. Reference-free analysis suggested punctual, yet pronounced differences in the metagenomic signature in the PD + RS group. RESISTA-PD highlights that a prebiotic treatment with RS is safe and well-tolerated in PD. The stable alpha-diversity and beta-diversity alongside altered fecal butyrate and calprotectin concentrations call for long-term studies, also investigating whether RS is able to modify the clinical course of PD.

## Introduction

Gut microbiota composition is altered in Parkinson's disease (PD) [1–3]. An increased abundance of Enterobacteriaceae has been consistently described in the fecal samples of PD patients, whereas the abundance of *Prevotella*, *Faecalibacterium*, *Blautia*, and *Bifidobacterium* is reduced in PD [1,4–8]. This is of potential relevance since bacteria with anti-inflammatory properties (*e.g.*, synthesis of short-chain fatty acids, SCFAs) are less abundant in PD. Potentially pro-inflammatory bacteria (*e.g.*, endotoxin-containing species) are more abundant in PD. Members of the families Prevotellacae, Ruminococcacae, and Bacteroidacae are capable of fermenting resistant starch (RS), a nutritional component that arrives in the large intestine without previous degradation by human enzymes [9]. Anaerobic fermentation of RS results in SCFAs, such as butyrate [10]. Butyrate exerts essential functions in the gut: it represents the main energy source for enterocytes, enhances gut motility, and exerts immunomodulatory effects [9,10]. Animal studies have shown that butyrate interacts with colonic regulatory T cells, creating an anti-inflammatory environment [11]. Consequently, a lack of SCFA-producing bacteria and reduced colonic SCFA concentrations presumably lead to reduced gut motility as well as to a shift in the intestinal immune system toward a more pro-inflammatory environment [12]. Intestinal inflammation, as well as altered gut motility (*e.g.*, constipation), has frequently been described in PD. In addition, we have previously shown that PD patients have reduced fecal SCFA concentrations compared to matched controls [6].

With regard to techniques used to characterize the microbiome, 16S amplicon sequencing has been most frequently used in microbiome studies due to its broad availability, moderate costs, and straightforward analysis. In recent years, whole-genome sequencing (WGS) has become widely available. Compared to 16S amplicon sequencing, WGS requires more complex computational and analytical procedures but is superior in characterizing the metagenomic landscape with regard to resolution, accuracy, and functional profiling [13,14]. To characterize the metagenomic landscape, two different approaches can be used: 1) reference-free approaches characterize the metagenomic landscape based solely on sequencing data; 2) reference-based approaches rely on existing databases to compare the generated sequences against. In the present study, we computed the taxonomic profile with reference-based approaches. In addition, we also performed a comparative analysis with a hybrid approach named BusyBee [15]. BusyBee is a software combining both reference-free and reference-based approaches.

A sensitive and valid marker of intestinal inflammation is fecal calprotectin. Calprotectin is a protein in human leukocytes. In case of inflammation, leukocytes migrate into the intestinal lumen, and calprotectin can be measured in the feces as a stable marker that reflects even subclinical intestinal inflammation [16]. In accordance with the finding of prevailing pro-inflammatory bacteria in PD, elevated fecal calprotectin concentrations have been described in PD, too [17,18].

A prebiotic approach to increase SCFA concentrations is nutritional supplementation with RS. The efficacy and tolerability of a 12-week intervention with RS have already been shown in a controlled clinical trial for elderly subjects ($\geq$ 70 years old): RS was well-tolerated and, compared with placebo, elderly subjects on RS showed an altered intestinal microbiota, an increase in fecal butyrate concentrations, and a significant reduction in the use of laxatives [19].

Taken together, we set up the following hypothesis concerning a sequence of events: oral supplementation with RS enhances SCFA synthesis in the gut, probably accompanied by a shift in gut microbiota composition (due to a survival advantage for bacteria capable of fermenting RS). Consequently, the increased SCFA concentrations should lead to improved gut motility (improved constipation, respectively) and a reduction in markers of intestinal inflammation.

## Results

### The RESISTA-PD study cohort

Eighty-seven subjects participated in the trial "Effects of Resistant Starch on Bowel Habits, Short Chain Fatty Acids and Gut Microbiota in Parkinson's Disease" (RESISTA-PD). The study design and workflow illustrating subjects' allocation to study-arms, clinical visits, sample collection, and analysis are summarized in **Figure 1**. The majority of subjects (n = 76) completed the study per protocol. The median age was 64.5 years old in the PD group receiving RS (PD + RS), 66 years old in the PD group receiving dietary instruction (PD + DI), and 61.5 years old in the control group receiving RS (Co + RS). There was no significant difference regarding sex ratio between the groups. The majority of subjects were on an omnivorous diet. Additional epidemiologic and clinical data are summarized in **Table 1**, and detailed information regarding the medication of the enrolled subjects is provided in Table S1. No major side effects were reported during the 8-week intervention with RS.

### Gut microbiota composition differs between PD patients and controls at baseline

At baseline, PD patients (n = 57) and controls (n = 30) showed no significant difference with regard to alpha-diversity with neither of the two applied analytical tools (MetaPhlAn2 and mOTUs2) (Figure S1A and B; File S1). With regard to beta-diversity, we observed a significant

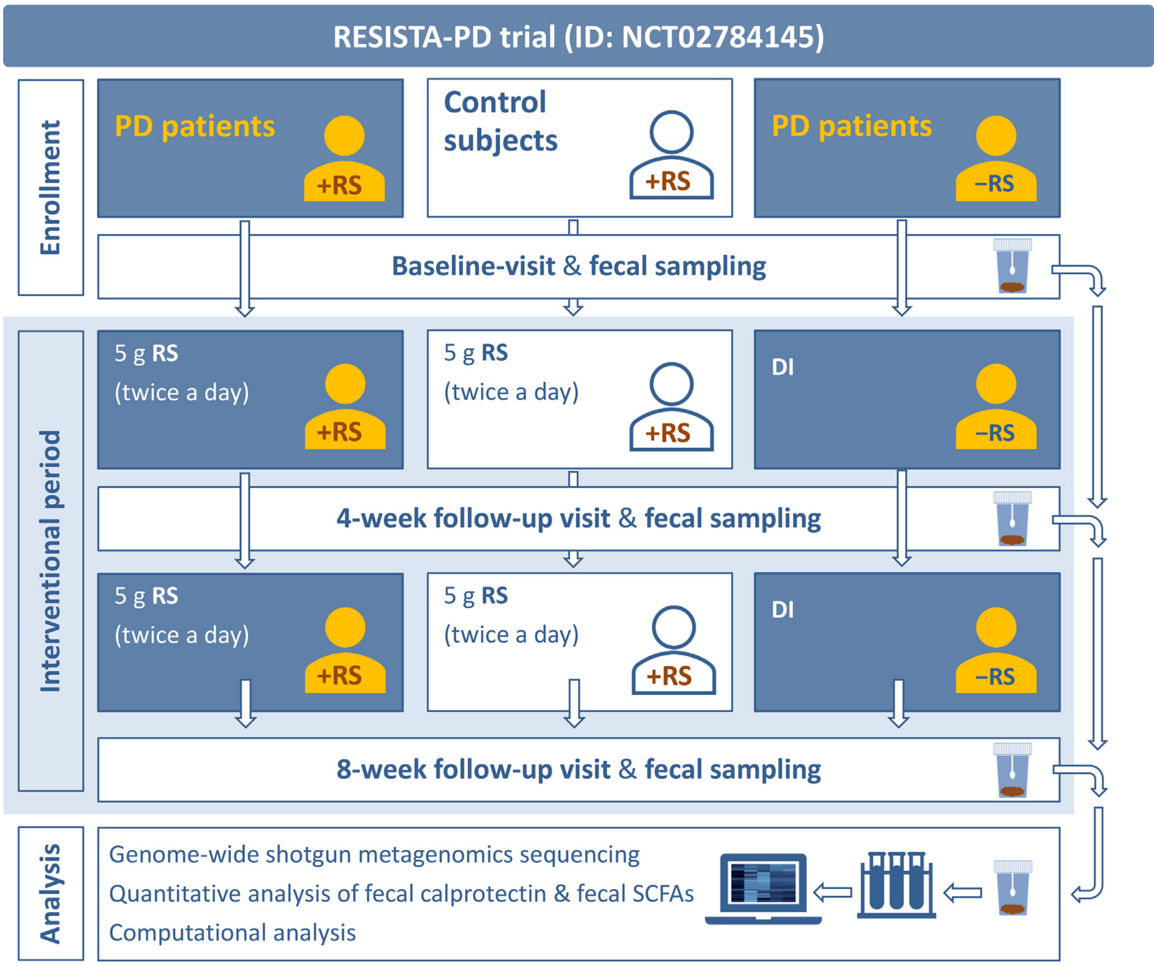

RESISTA-PD trial (ID: NCT02784145)
Enrollment
PD patients +RS
Control subjects +RS
PD patients −RS
Baseline-visit & fecal sampling
Interventional period
5 g RS (twice a day) +RS
5 g RS (twice a day) +RS
DI −RS
4-week follow-up visit & fecal sampling
5 g RS (twice a day) +RS
5 g RS (twice a day) +RS
DI −RS
8-week follow-up visit & fecal sampling
Analysis
Genome-wide shotgun metagenomics sequencing
Quantitative analysis of fecal calprotectin & fecal SCFAs
Computational analysis

**Figure 1 Study design**

Subjects were assigned to three different study-arms. One group of PD patients and a control group received 5 g RS twice a day for a total period of 8 weeks. The second group of PD patients solely received DI. Fecal samples and clinical scores were collected at baseline, after 4 weeks, and after 8 weeks for analysis. +RS in the pictograms visualizes subjects receiving RS, −RS in the pictograms visualizes subjects not receiving RS. PD, Parkinson's disease; RS, resistant starch; SCFA, short-chain fatty acid; DI, dietary instruction.

difference between PD patients and controls ($P = 0.001$) with both analytical tools applied in this study (Figure S1C and D). With regard to specific taxa, Lachnospiraceae *incertae sedis* (mOTU_v25_12240, $P = 0.017$) and *Faecalibacterium prausnitzii* (mOTU_v25_06110, $P = 0.019$) showed significantly reduced abundances after correction for multiple testing in PD patients compared to controls (Table S2). Figure 2 illustrates descriptive differences at different taxonomic levels between PD patients and controls prior to the intervention. Descriptively, taxa of the phylum Firmicutes showed higher abundances in controls (except for the class Bacilli), while most taxa of the phylum Proteobacteria, especially Enterobacteriaceae, were more abundant in PD.

### Intervention with RS alters symptom load and fecal markers in PD

We next analyzed the intervention-associated changes in subject-reported symptoms and in fecal markers. We observed a significant improvement with regard to non-motor symptoms [measured by the Non-Motor Symptoms Questionnaire (NMSQ) score, $P = 0.001$] and a significant improvement with regard to depressive symptoms [assessed by the Beck Depression Inventory (BDI), $P = 0.001$] in the PD + RS group at 8 weeks post intervention compared to the baseline (Table 2; Figure S2A). No significant changes in these parameters were identified over the 8-week intervention period for the PD + DI or Co + RS group. There was no significant change in bowel habits [assessed with the Constipation Scoring System (CSS)] between baseline and 8 weeks post intervention for any of the three investigated groups (Table 2; Figure S2A). Calprotectin concentrations dropped significantly in the PD + RS group at 8 weeks post intervention compared to the baseline ($P = 0.023$; Table 3; Figure S2B). No significant changes in fecal calprotectin concentrations were observed between baseline and 8 weeks post intervention in the Co + RS and PD + DI groups. Concerning fecal SCFAs, the concentration of the SCFA butyrate increased significantly in the PD + RS group at 8 weeks intervention compared to the baseline, for absolute fecal butyrate concentrations ($P = 0.029$) as well as

**Table 1 Epidemiological and clinical characteristics of the enrolled subjects**

| | PD + RS | Co + RS | PD + DI |
|---|---|---|---|
| Number of subjects enrolled | 32 | 30 | 25 |
| Reasons to withdraw from study by subject's request after having started intervention | Subject disliked taste of RS (n = 1)<br>General discomfort and upset stomach after starting RS (n = 3)<br>Concurrent acute disease after baseline visit (n = 1) | Acute disease of family member (n = 1) | Difficulties in handling fecal sampling kits at home (n = 1)<br>Claiming "personal problems and family issues" (n = 1) |
| Number of subjects with 4-week follow-up | 28 | 29 | 23 |
| Number of subjects with 8-week follow-up | 26 | 27 | 23 |
| Age (median, [range]) | 64.5 [42–84] | 61.5 [40–76] | 66 [47–80] |
| Sex (male / female) | 18 / 14 | 12 / 18 | 13 / 12 |
| Dietary habit | Omnivorous diet: n = 29<br>Vegetarian diet: n = 2<br>Pescetarian diet: n = 1 | Omnivorous diet: n = 27<br>Vegetarian diet: n = 2<br>Pesceatrian diet: n = 1 | Omnivorius diet: n = 25<br>Vegetarian diet: n = 0<br>Pescetarian diet: n = 0 |
| Smoker | 2 of 32 | 3 of 30 | 1 of 25 |
| Disease duration in months (median, [range]) | 111 [7–288] | Not applicable | 111 [22–265] |
| History of appendectomy | 15 of 32 | 7 of 30 | 13 of 25 |
| UPDRS I, II, III total score in on state (median, [range]) | 35 [4–74] | Not applicable | 30 [3–69] |
| MMST (median, [range]) | 29 [23–30] | 30 [28–30] | 29 [25–30] |

*Note*: PD + RS indicates PD patients receiving RS; Co + RS indicates control subjects receiving RS; PD + DI indicates PD patients receiving DI. PD, Parkinson's disease; RS, resistant starch; DI, dietary instruction; UPDRS, Unified Parkinson's Disease Rating Scale; MMST, Mini-Mental-Status-Test.

for relative fecal butyrate concentrations ($P = 0.026$), (**Table 4**; Figure S2C); however, there were no significant changes for the concentrations of other SCFAs (including acetate, propionate, valerate, isobutyrate, and isovalerate) between baseline and 8 weeks post intervention in the PD + RS group (Table 4). Moreover, no significant changes in SCFA concentrations were observed between baseline and 8 weeks post intervention in the Co + RS and PD + DI groups (Table 4).

### Reference-based analysis shows a stable gut microbiome after RS intervention

In order to investigate whether the observed changes in clinical symptoms and fecal markers are associated with an intervention-associated shift in the gut microbiome, we performed metagenomic sequencing. Quality control by FastQC indicated good data quality of metagenomic sequencing. During preprocessing, less than 1% of reads were removed for each sample. In addition to the standard quality control, we analyzed pairwise Mash distances [20] between all samples. Hereby, the Mash distance gauges similarity between sequencing libraries using the only sequence features directly derived from raw reads. Visualizing Mash distances showed that samples derived from the same individual frequently produced the lowest Mash distance, indicating correct labeling of samples and a lack of contamination (**Figure 3**). No significant intervention-associated changes with regard to either alpha-diversity or beta-diversity were detected for any of the three investigated groups (PD + RS, PD + DI, and Co + RS). No significant intervention-associated changes were detected concerning differences in distinct taxa (Table S2). Non-metric multidimensional scaling (NMDS) visualizing microbiome shifts did not reveal uniform shifts associated with the intervention (Figure S3).

### Reference-free analysis points at punctual differences in the metagenomic signature

Reference-free analysis revealed intervention-associated changes in taxonomic signatures in the PD + RS group (**Figure 4**). The majority (> 54%) of contigs forming one of the three clusters in the reference-free analysis were derived from the genus *Rhodococcus* (Figure S4). Density changes worth interpreting as clusters identified in the other cohorts (Co + RS and PD + DI) did not contain significant amounts of *Rhodococcus* sequences.

### Distinct microbial signatures are associated with fecal butyrate concentrations

In metagenomic samples, the change in the abundance of one taxon is likely to entail changes in the abundance of other taxa. We investigated our data for data compositionality using the selbal algorithm. Selbal searches for two groups of taxa whose relation (or balance) is associated with a certain response variable. The relationship is modeled as a linear or logistic regression model of the taxa on the response variable. Selbal builds multiple models containing different taxa combinations and evaluates their performances using cross-validation. In our dataset, response variables were measurements of acetate, propionate, butyrate, valerate, and calprotectin, as well as CSS

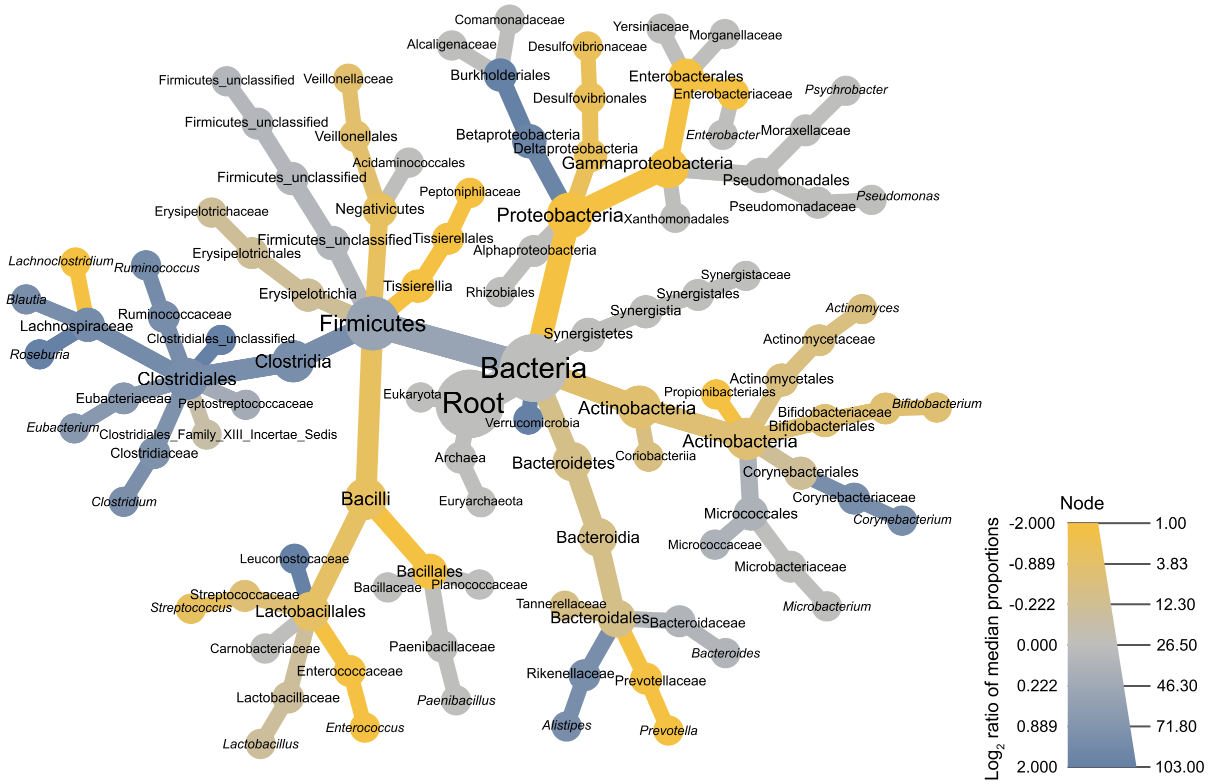


**Figure 2 Taxonomic tree illustrating differences between PD patients and controls at baseline**

This taxonomic tree illustrates the number of OTUs per taxon (visualized by the size of the radius) and the difference (visualized by color) between PD patients and controls prior to intervention (baseline). Yellow shades indicate a higher abundance in PD patients; blue shades indicate a higher abundance in controls; gray shades indicate no group-specific differences. Low abundant taxa were pruned [46]. OTU, operational taxonomic unit.

**Table 2 Scores on clinical scales at baseline and 8 weeks post intervention**

| | PD + RS | | | Co + RS | | | PD + DI | | |
|---|---|---|---|---|---|---|---|---|---|
| | Baseline | Post intervention | *P* value | Baseline | Post intervention | *P* value | Baseline | Post intervention | *P* value |
| CSS score (median [range]) | 5 [0–14] | 3.5 [0–15] | 0.257 | 1 [0–11] | 0 [0–8] | 0.125 | 2 [0–10] | 2 [0–12] | 0.674 |
| NMSQ score (median [range]) | 10.5 [3–20] | 7.5 [2–18] | 0.001 | 3 [0–9] | 3 [0–10] | 0.774 | 10 [4–19] | 10 [5–19] | 0.152 |
| BDI score (median [range]) | 6.5 [2–25] | 3 [0–12] | 0.001 | 2 [0–14] | 2 [0–20] | 0.202 | 7 [1–18] | 6 [0–13] | 0.106 |

*Note*: CSS, Constipation Scoring System; NMSQ, Non-Motor Symptoms Questionnaire; BDI, Beck Depression Inventory.

**Table 3 Fecal calprotectin concentrations at baseline and 8 weeks post intervention**

| | PD + RS | | | Co + RS | | | PD + DI | | |
|---|---|---|---|---|---|---|---|---|---|
| | Baseline | Post intervention | *P* value | Baseline | Post intervention | *P* value | Baseline | Post intervention | *P* value |
| Concentration of fecal calprotectin (median [range], μg/g feces) | 56.8 [19–327] | 20.5 [19–407] | 0.023 | 19 [19–69] | 19 [19–155] | 1.000 | 46 [19–219] | 31 [19–217] | 0.481 |

and BDI scores. Using the selbal algorithm, results for MetaPhlAn2 and mOTUs2 data (**Figure 5**) with butyrate concentrations as response variables were highly consistent. For absolute butyrate concentrations, selbal detected that higher abundances of *Fusicatenibacter saccharivorans* to *Ruthenibacterium lactatiformans* were associated with higher absolute butyrate concentrations (Figure 5A) with an association slightly below moderate (MetaPhlAn2 data, $R^2$ = 0.126). The association of *Ruthenibacterium lactatiformans* with butyrate concentrations was verified by mOTUs2 data. Here, selbal detected that higher abundances of Lachnospiraceae and *Streptococcus parasanguinis* to *Ruthenibacterium lactatifor-*

**Table 4 SCFA concentrations at baseline and 8 weeks post intervention**

| | PD + RS | | | Co + RS | | | PD + DI | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Baseline** | **Post intervention** | ***P* value** | **Baseline** | **Post intervention** | ***P* value** | **Baseline** | **Post intervention** | ***P* value** |
| Butyrate (median [range], mmol/g) | 0.24 [0.00–1.00] | 0.25 [0.01–1.92] | 0.029 | 0.46 [0.04–2.10] | 0.32 [0.07–2.45] | 0.223 | 0.29 [0.05–1.44] | 0.27 [0.06–2.38] | 0.426 |
| Butyrate (median [range], %) | 10.5 [0.9–22.2] | 11.3 [0.9–21.4] | 0.026 | 12.8 [4.5–21.1] | 11.2 [4.8–22.5] | 0.260 | 11.3 [3.9–18.8] | 10.3 [2.1–20.6] | 0.685 |
| Acetate (median [range], mmol/g) | 1.39 [0.15–4.91] | 1.80 [0.23–6.55] | 0.165 | 1.99 [0.57–6.78] | 2.04 [0.38–9.54] | 0.891 | 2.17 [0.24–36.0] | 2.04 [0.65–7.66] | 0.685 |
| Propionate (median [range], mmol/g) | 0.36 [0.03–1.74] | 0.55 [0.06–2.22] | 0.145 | 0.48 [0.08–1.50] | 0.49 [0.12–2.16] | 0.785 | 0.53 [0.11–1.66] | 0.48 [0.18–1.94] | 0.445 |
| Valerate (median [range], mmol/g) | 0.07 [0.01–0.29] | 0.07 [0.02–0.18] | 0.647 | 0.06 [0.02–0.26] | 0.07 [0.02–0.20] | 0.703 | 0.11 [0.01–0.35] | 0.07 [0.01–0.66] | 0.733 |
| Isobutyrate (median [range], mmol/g) | 0.05 [0.01–0.26] | 0.05 [0.01–0.15] | 0.829 | 0.05 [0.02–0.19] | 0.05 [0.02–0.16] | 0.715 | 0.07 [0.01–0.24] | 0.05 [0.01–0.45] | 0.501 |
| Isovalerate (median [range], mmol/g) | 0.04 [0.00–0.18] | 0.04 [0.00–0.19] | 0.461 | 0.06 [0.00–0.20] | 0.04 [0.00–0.22] | 0.584 | 0.06 [0.00–0.26] | 0.04 [0.01–0.39] | 0.537 |
| Total SCFA (median [range], mmol/g) | 2.08 [0.21–8.02] | 2.72 [0.42–10.6] | 0.142 | 3.17 [0.76–10.5] | 2.96 [0.64–14.6] | 0.973 | 3.37 [0.47–36.3] | 3.07 [0.97–13.2] | 0.745 |

*Note*: *P* values refer to the change between baseline concentrations and post-interventional concentrations. SCFA, short-chain fatty acid.

*mans* were associated with higher absolute butyrate concentrations, and the association was moderate ($R^2$ = 0.198; Figure 5B). For relative butyrate concentrations, selbal detected that higher abundances of *Dorea longicatena* (MetaPhlAn2 data) and *Blautia wexlerae* (MetaPhlAn2 and mOTUs2 data) to *Ruthenibacterium lactatiformans* (MetaPhlAn2 and mOTUs2 data) were associated with higher relative butyrate concentrations, and the association was moderate ($R^2$ = 0.238 for MetaPhlAn2 data, $R^2$ = 0.257 for mOTUs2 data; Figure 5C and D). The model itself was stable, with *Ruthenibacterium lactatiformans* and *Dorea longicatena* being included in over 95% of all models for MetaPhlAn2 data and *Blautia wexlerae* and *Ruthenibacterium lactatiformans* being included in over 96% of all models for mOTUs2 data. Other response variables did not show consistency between mOTUs2 and MetaPhlAn2 data.

### Functional profiling reveals no intervention-associated difference in metabolic pathways

In order to identify differences in the available metabolic pathways, we applied the HUMAnN2 tool to our data. The estimated pathway abundances were used for an exploratory data analysis of the samples using principal component analysis (PCA) and a differential analysis using ALDEx2. The PCA projection indicated a different tendency between the PD + RS and Co + RS groups, but no differences associated with the intervention (baseline *vs.* 8 weeks post intervention) (Figure S5). The analysis with ALDEx2 did not result in any pathway that showed a significant difference between groups nor a difference between baseline and 8 weeks post intervention (Table S3).

## Discussion

Gut microbiota composition is altered in PD [3–6] and might be a contributing factor for gastrointestinal non-motor symptoms (*e.g.*, constipation) in PD. Having recognized the relevance of the intestinal microbiome in PD, probiotics have been investigated in PD and other neurodegenerative diseases previously [21,22] and prompted us to perform the RESISTA-PD trial.

In accordance with other studies in the field [2,4,5,7,23,24], we observed a difference between PD patients and controls at baseline regarding beta-diversity. With regard to specific taxa, we detected significantly different abundances for two taxa after correction for multiple testing: abundances for Lachnospiraceae *incertae sedis* and *Faecalibacterium prausnitzii* were significantly reduced in the fecal samples of PD patients. Lachnospiraceae as well as *Faecalibacterium prausnitzii* have already been reported to be reduced in PD and have also been confirmed as altered taxa in PD in a recent meta-analysis [25]. Indeed, the lower abundance of *Faecalibacterium prausnitzii* might be one explanation for the lower fecal butyrate concentrations in PD. On a descriptive level, we also reproduced some other previously reported alterations of the gut microbiota in PD, *e.g.*, a lower abundance of Firmicutes and a higher abundance of Proteobacteria, especially Enterobacteriaceae.

For the reference-free analysis of intervention-associated changes, a metagenomic signature indicating a possible involvement of *Rhodococcus* was found in the PD + RS

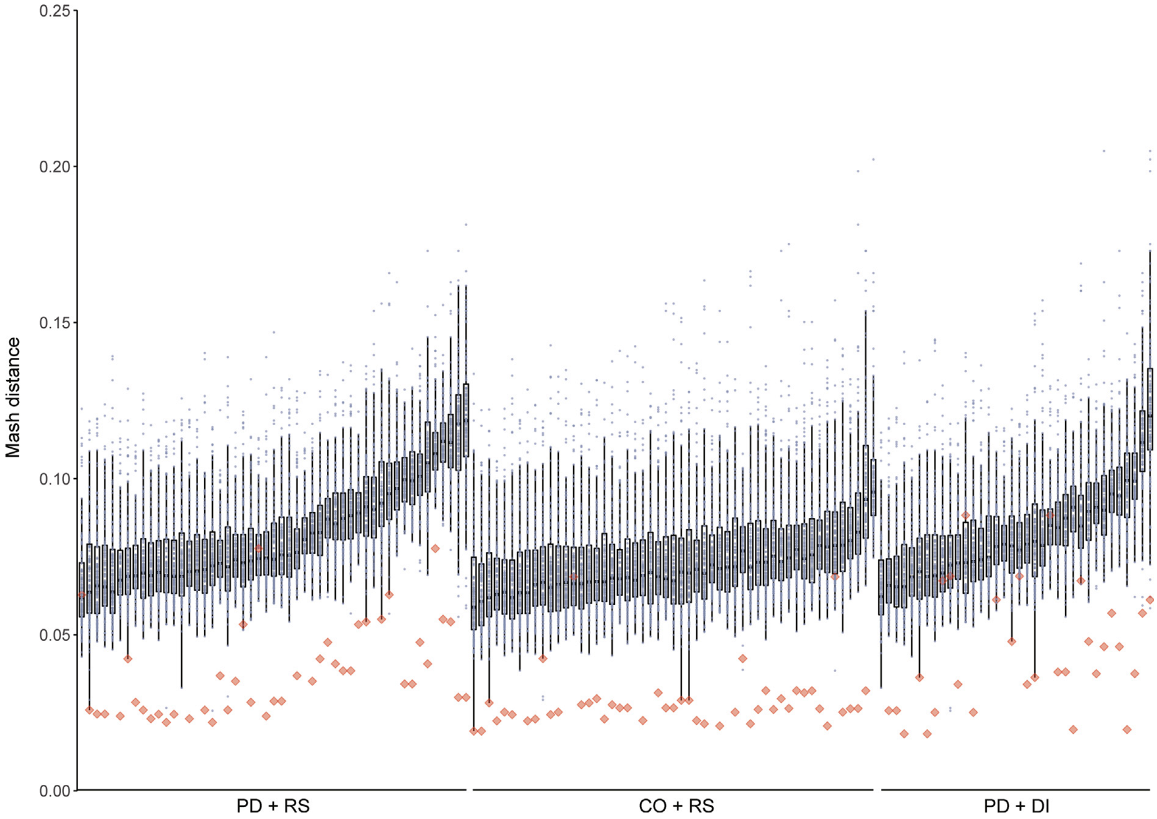


**Figure 3 High intra-individual and low inter-individual similarities of samples**
The similarity of samples visualized as Mash distance plot (grouped by study-arms). The lower the Mash distance, the higher the similarity of samples. Red diamonds represent paired samples (baseline and 8 weeks post intervention) of one subject. Dots represent samples of other subjects (unpaired).

group, despite an insignificant change in abundance during the read-based analysis. As shown in Figure 4, the blue cluster (top right) in the CO + RS group contained only sequences derived from one single sample, and the sequences were assigned to *Rhodotorula toruloides*; the two clusters with the highest density in the PD + DI group contained no sequences of the genus *Rhodococcus* (blue cluster) and less than 0.001% of sequences of the genus *Rhodococcus* (red cluster). The reference-free workflow we selected discards all quantity information after assembly. In an optimal scenario, sequences derived from identical genetic information will be collapsed into the same contig in each sample. In BusyBee, one such contig will appear as one individual point with close to no impact on the overall density distribution. The strong signal from the high-density cluster in the PD + RS group suggests the existence of multiple contigs that are dissimilar enough not to be collapsed during assembly yet qualitatively good enough to be assigned to *Rhodococcus*. Accordingly, the change in the density of the investigated cluster indicates a more complex behavior than a quantitative balance shift. Instead, an increase in genomic diversity may be postulated from this observation. The relevance of this particular finding remains unclear and requires further investigations, especially since the genus *Rhodococcus* is not a typical representative of the human gut microbiota. Given the fact that bacteria of the genus *Rhodococcus* are not typical part of the human gut microbiota and also are obligate aerobes, the identification of this genus in human fecal samples points at potential contamination. Indeed, *Rhodococcus* has been identified in human biosamples due to DNA contamination of reagents [26]. Yet, DNA contamination is mainly a problem when analyzing low microbial biomass samples (like blood or saliva). In our study that analyzed high microbial biomass samples (feces), contamination is further unlikely as contamination would have occurred solely in PD + RS samples after the intervention and not in the other two groups. One might also hypothetically consider contamination of a single batch of reagents or tubes used in our study. However, samples were analyzed in a random way and not sorted by the group prior to analysis. Hence, contamination can hardly explain this finding. As the taxonomic assignment of contigs relies on libraries, misassignment due to similar sequences of another taxon (not represented in libraries) with sequences of *Rhodococcus* should also be considered.

Our finding that a prebiotic intervention with RS significantly alters fecal butyrate concentrations and significantly reduces fecal calprotectin concentrations is also in line with a controlled clinical trial that investigated RS in mid-age and elderly subjects and reported an increase in fecal butyrate concentrations in subjects aged 70 years or older [19]. While the study by Alfa and colleagues [19] even indicated a therapeutic effect (reduction in the use of laxatives), we did not observe a significant improvement of bowel habits. This divergent obser-

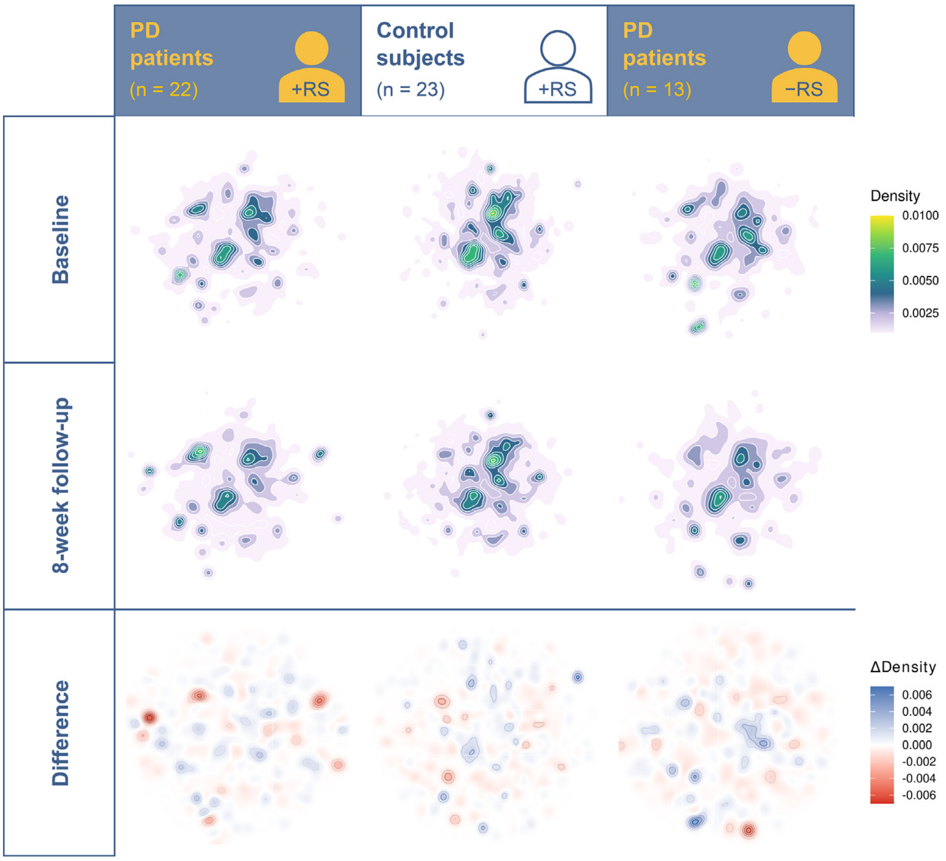


**Figure 4 Reference-free analysis points at punctual differences**
This figure shows the density distribution of the 5-mers, after dimensionality reduction with UMAP. The first row contains the baseline. The second row shows the 8-week follow-up. In the bottom row, the difference between the two previous rows is visualized; blue indicates a stronger signal at baseline, and red indicates a stronger signal at 8 weeks follow-up. UMAP, uniform manifold approximation and projection.

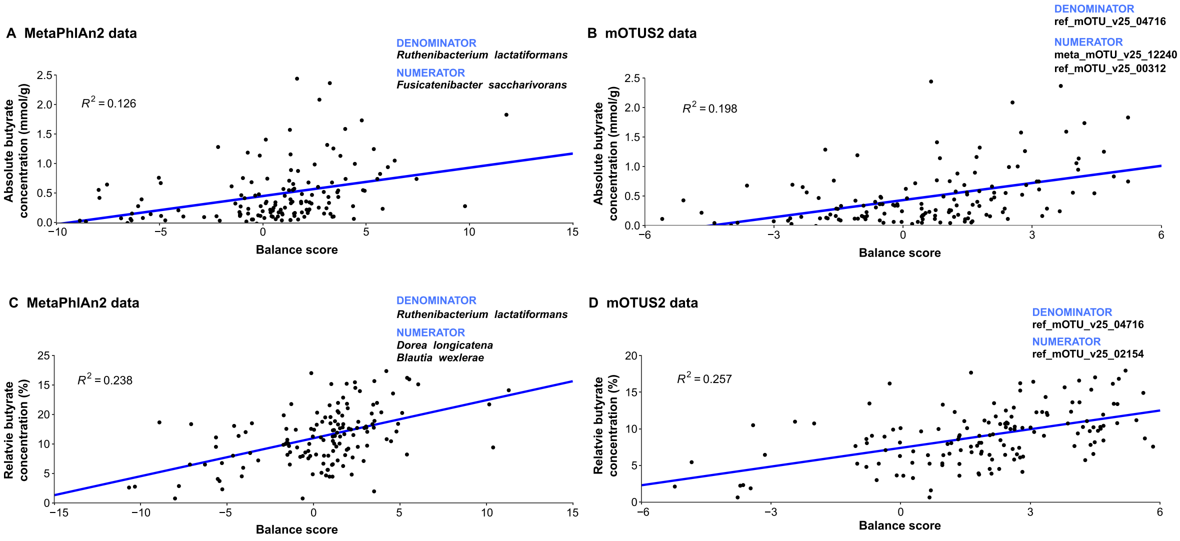


**Figure 5 Distinct microbial signatures are associated with fecal butyrate concentrations**
**A.** and **B.** Balance scores for MetaPhlAn2 data (A) and mOTUS2 data (B) with absolute butyrate concentrations as response variables. **C.** and **D.** Balance scores for MetaPhlAn2 data (C) and mOTUS2 data (D) with relative butyrate concentrations as response variables.

vation between our study and the study by Alfa and colleagues might be due to the differences in the types of RS (we used RS type 3, while Alfa and colleagues used RS type 2), the doses of RS (Alfa and colleagues administered approximately double the dosage compared to our study), and the duration of the interventional period (8 weeks in our study *versus* 12 weeks in the study by Alfa and colleagues).

Fecal butyrate concentrations and calprotectin concentrations were not altered when PD patients solely underwent nutritional counseling, including DI concerning a fiber-rich diet (PD + DI). Considering the fact that the PD + DI group underwent the same visit schedule as the PD + RS group, the effect observed with regard to clinical measures in the PD + RS group is unlikely to be completed due to unspecific effects such as attention paid to subjects during clinical visits or answering questionnaires according to social desirability.

The effect on symptoms related to depression in the PD + RS group might be explained by the observed increase in butyrate concentrations. An association between gut microbiota and depressive symptoms has been described previously [27,28]. Administration of SCFAs, including butyrate, has been shown to reduce depressive symptoms in mice [29]. Moreover, fecal SCFAs have been shown to be reduced in a cohort of female patients with depression [30]. Increasing evidence suggests a connection between depressive symptoms and fecal SCFA concentrations [31]. One explanation for the lack of a change in clinical measures in the PD + DI group might be that adherence to DI is likely to be lower compared to the more convenient approach of consuming a dietary supplement (dissolved in one glass of water) twice a day. In addition, changes in dietary habits are much more heterogeneous compared to standardized nutritional supplementation.

The fact that the Co + RS group did not show a reduction in fecal calprotectin concentrations is likely to be explained by already normal calprotectin concentrations in control subjects at baseline. The unchanged SCFA concentrations in the Co + RS group might be either explained by a ceiling effect or by a lower adherence (as controls did not expect to benefit from the intervention).

Even though the effects on fecal calprotectin and fecal butyrate were significant in the PD + RS group and also SCFAs other than butyrate showed a trend towards an increase in concentrations in the PD + RS group, our data lack a clear-cut correlate concerning specific gut microbiota. Assuming that gut microbiota composition remains stable despite the prebiotic intervention, an altered transcription might have led to the observed effects on fecal markers. The lack of a clear-cut response to the intervention with regard to gut microbiota or symptoms of constipation might also be due to various individual factors. Our study design controlled for confounding factors like age, sex, overall type of diet, comorbidities, and medication. Nevertheless, the investigated cohorts were heterogeneous (even within groups) with regard to other, more complex factors that might determine the individual response (*e.g.*, the composition of the gut microbiome prior to the intervention, adherence to the recommended RS intake, more specific dietary habits). This said, the limited sample size in this proof of concept study together with the inter-individual variability concerning potential confounding factors, is one explanation for the heterogeneous response to the intervention. Hence, larger cohorts (as well as transcriptomics and proteomics) might have been necessary to detect more subtle intervention-associated alterations in the gut microbiome (and possible changes at the transcriptomic level).

In order to identify microbial signatures associated with SCFA concentrations, we performed an *in silico* analysis (using the selbal algorithm): balance analysis of taxa and butyrate concentrations resulted in concordant results for both analytical tools (mOTUs2 and MetaPhlAn2). Moreover, we confirmed the robustness of the identified balance scores by their frequency in a cross-validation model. The microbial taxa *Blautia wexlerae*, *D. longicatena*, and *Ruthenibacterium lactatiformans* are involved in butyrate-related pathways [32]. However, all of these bacteria are not capable of directly producing butyrate from RS, but they produce lactate and succinate by fermentation which consecutively serves as substrates for other bacteria which produce butyrate [10]. Despite the fact that our *in silico* approach did not detect classical SCFA producers (like *Faecalibacterium* or *Roseburia*) as determinants for fecal butyrate concentrations, the taxa identified by selbal are indirectly involved in butyrate production (via complex interactions with other taxa) [10].

In contrast to our initial hypothesis, symptoms related to constipation (a frequent non-motor symptom in PD) were not significantly altered during the 8-week intervention. As there was at least a descriptive decline in CSS scores after RS supplementation, we suggest longer interventional periods and increased doses of RS to test such a symptomatic effect on bowel habits. Given that RS was well-tolerated in the RESISTA-PD trial, this seems to be a feasible and rational approach.

Besides the limited sample size and the relatively short interventional period, one main limitation of the RESISTA-PD trial is its open-label study design. We aimed at counteracting this shortcoming by including an additional PD control-arm (PD + DI) to control for unspecific effects (as discussed above). Adherence to the intervention was checked by patient diaries but not by more objective measures. Probably, an internal motivation to adhere to the study protocol might have been higher in the PD + RS group compared to the Co + RS group (as discussed above). The primary aim of the RESISTA-PD trial was to test the feasibility, tolerability, and efficacy of this prebiotic approach. Hence, our study protocol did not include an additional measurement of the investigated markers several weeks after withdrawal of RS. We suggest including such an assessment in future studies.

At this time, we are not able to answer the question of whether the observed anti-inflammatory effects indicated by the decline in fecal calprotectin concentrations are mediated by the increase in butyrate concentrations. Even though other studies endorse such an assumption [11], further studies are needed to clarify the exact mechanisms of this prebiotic intervention and the increase in SCFAs in detail.

A general limitation of interventions aiming at altering the gut microbiome is the question of endurance. This is why long-term studies and assessment of subjects after withdrawal of the intervention are mandatory to draw final conclusions.

## Conclusion

RS, as a dietary supplement to increase fiber intake, is safe and well-tolerated in PD. RS supplementation partially restores fecal SCFA concentrations in the PD + RS group without clear-cut changes in the gut microbiome that were attributable to the intervention. Alterations at the transcriptome level that are not captured by our approach might explain the intervention-associated significant increase in fecal markers in the PD + RS group.

In view of the good tolerability of RS, we suggest long-term studies with RS. These studies should also aim at clarifying the underlying mechanisms for the supposed anti-inflammatory effects. Based on the assumption of an RS-associated anti-inflammatory effect, these studies should also investigate whether RS supplementation is able to modify the clinical course of PD.

## Materials and methods

### Study design and registration

The interventional study RESISTA-PD is a monocentric, prospective, open-label clinical trial investigating the effects of an 8-week prebiotic intervention with the dietary supplement RS (Catalog No. P/N 03647989, SymbioIntest, SymbioPharm GmbH, Herborn, Germany) (5 g RS twice a day orally) in PD patients (PD + RS) and matched controls (Co + RS). As a third study-arm, PD patients who received solely DI (PD + DI) were enrolled in this study. DI was based on the "Food-Based Dietary Guidelines in Germany" (for further reference, see https://www.dge-medienservice.de/food-based-dietary-guidelines-in-germany.html) of the German Nutrition Society. At the baseline visit, the specified guidelines to support a health-promoting diet were explained to all subjects in the PD + DI group. These recommendations support a diet rich in whole-grain products and vegetables and moderate consumption of fat and animal products. Subjects also received a leaflet summarizing these recommendations. This leaflet included a table with practical orientation values for each food group (*e.g.*, cereal products and potatoes, vegetable and salad, and fruit). Primary outcome measures were: change (prior- *vs.* post-intervention) in (a) bowel habits, (b) fecal SCFA concentrations, and (c) gut microbiome (analyzed by whole genome-wide sequencing). Secondary outcome parameters were: differences in gut microbiome at baseline (between PD patients and controls), change (prior- *vs.* post-intervention) in clinical scales, and change in fecal calprotectin concentrations (prior- *vs.* post-intervention).

### Subjects

A total of 57 PD patients and 30 control subjects were enrolled. PD patients were assigned to two different interventional groups: PD + RS (n = 32) received 5 g RS twice per day orally over a period of 8 weeks; PD + DI (n = 25) received DI concerning high fiber intake, but no RS supplementation. Control subjects (Co + RS, n = 30) received 5 g RS twice per day orally over a period of 8 weeks. The main inclusion criteria were an age > 18 years old, diagnosis of PD (respective absence of PD or any other neurodegenerative disorder in the control group), capacity to give written informed consent. The main exclusion criteria were use of antibiotics, steroids, antimycotics or probiotic supplements (during the last 12 weeks), chronic or acute disorders of the gastrointestinal tract (other than constipation), a history of colonoscopy within the past 12 weeks, a history of gastrointestinal surgery (other than appendectomy) within the past three years.

### Clinical assessments

Subjects were assessed at baseline, 4 weeks post intervention, and 8 weeks post intervention. Baseline assessment was performed as in-person clinical visit. Assessments for 4 weeks and 8 weeks post intervention respectively, were performed as telephone visits. At baseline visit, subjects underwent rating with the Unified Parkinson's Disease Rating Scale (UPDRS) [33] and the Mini-Mental-Status-Test (MMST) [34]. Symptoms related to constipation were assessed at each of the three visits (baseline, 4 weeks post intervention, 8 weeks post intervention) with the CSS [35]. Depressive symptoms and non-motor symptoms were assessed with the BDI scores [36] and the NMSQ scores [37], respectively, at baseline and 8 weeks post intervention. In addition to collecting data on adverse events, tolerability and subjective improvement of the intervention were assessed in analogy to the seven-point Clinical Global Impression - Improvement (CGI-I) scale [38] after 4 and 8 weeks of intervention. The clinical change was rated (compared to baseline, prior to the intervention with RS respectively) as: very much improved, much improved, minimally improved, no change, minimally worse, much worse, or very much worse.

### Collection of fecal samples

At the baseline visit, all subjects received sterile containers (Calalog Nos. P/N S1000-150 and P/N H9550T, Suesse, Gudensberg, Germany) for the collection of fecal samples at home. The containers were labeled with the subject-ID and the scheduled time for collection (baseline, *i.e.*, prior to the first intake of RS; 4 weeks of intervention with RS; and 8 weeks of intervention with RS). All subjects were instructed how to collect the fecal samples at home and received a leaflet containing relevant information for sample collection. Subjects were instructed to send in two samples (collected on two consecutive days) for each time point. For metagenomic sequencing, the first baseline sample and the first 8-week sample were used. For quantitative analysis of fecal markers, the mean of the two samples was calculated for further statistical analysis. In case of missing 8-week samples, 4-week samples were analyzed (last observation carried forward, LOCF) as described below (see the "Statistical analysis of clinical data and fecal SCFA and calprotectin concentrations" section). All subjects were reminded by telephone to send in samples after 4 weeks and after 8 weeks of intervention. Stool samples were sent to the Institute of Microoecology, Herborn, Germany, and immediately frozen at −35 °C until analysis.

### Measurement of fecal SCFA and calprotectin concentrations

Quantitative analyses of fecal SCFAs and calprotectin were carried out by the Institute of Microoecology, Herborn, Germany. All persons involved in these analyses were blinded to clinical data and the diagnosis of the subjects. Fecal SCFAs were measured by gas chromatography; fecal calprotectin was measured by enzyme-linked immunosorbent assay as previously described [6,18].

### DNA isolation

DNA from fecal samples was isolated using the DNeasy PowerSoil Kit (Catalog No. P/N 47014, QIAGEN, Hilden, Germany) according to the manufacturer's instructions. To ameliorate the purity, we performed precipitation of the DNA in the presence of sodium acetate (pH = 5.5) and cold 100% ethanol at −20 °C for at least overnight. The DNA was then centrifuged, washed with 80% ethanol once, and centrifuged another time. The pellet was air-dried and resuspended in TE buffer. DNA concentration was measured using a Nanodrop 2000 spectrophotometer (Catalog No. P/N ND-2000, ThermoFisher Scientific, Waltham, MA).

### Metagenomic sequencing

DNA libraries were prepared using the MGIEasy DNA Library Prep Kit (Catalog No. P/N 940-200022-00, MGI Technologies, Shenzhen, China) according to the manufacturer's recommendations. In general, 1 μg of input DNA was sheared into fragments using the M220 Focused-ultrasonicator (Catalog No. P/N 500295, Covaris, Woburn, MA). Size selection was carried out using Agencourt AMPure XP beads (Catalog No. P/N A63882, Beckman Coulter, Krefeld, Germany). Then, 50 ng of fragmented DNA were used for end-repairing and A-tailing followed by ligation of barcode containing adaptors to the 3′- and 5′-ends. The ligation products were amplified by PCR. A total of 16 different barcoded samples were pooled in equal amounts and circularized using a specific oligo sequence, which is complementary to the sequences in the 3′- and 5′-adaptors. DNA nanoballs (DNBs) were generated by rolling circle amplification (RCA), and loaded onto a flowcell using BGIDL-50 DNB loader. Paired-end sequencing was performed according to the BGISEQ-500RS High-throughput Sequencing Set for PE100 on the BGISEQ-500RS instrument (Catalog No. P/N 940-100037-00, MGI Technologies).

### Statistical analysis of clinical data and fecal SCFA and calprotectin concentrations

In case of missing data for 8 weeks post intervention and available data for 4 weeks post intervention, we applied the LOCF method. LOCF was used for 4 subjects [PD + RS (n = 3) and PD + DI (n = 1)] to replace missing 8-week data concerning fecal markers. Concerning clinical scores, missing 8-week data of 5 subjects [PD + RS (n = 2), PD + DI (n = 1), and Co + RS (n = 2)] were replaced by 4-week data. The normal distribution of data was tested using the Shapiro-Wilk's test. Statistical significance was assumed for $P < 0.05$. The difference between groups was tested using the Mann-Whitney-U-test. Comparisons of the same group at different time points were performed with the Wilcoxon's test for paired samples and the sign test for paired samples. Pre-defined outcome measures were not adjusted for multiple testing. Spearman's correlation coefficient was used to analyze correlations between parameters.

### Sequencing data analysis

#### *Preprocessing*

FastQC (version 0.11.8) was used to validate sequence quality, and the reports were summarized using multiQC (version 1.7) [39]. Adapter contamination was controlled with the Minion tool from the Kraken package (version 16.098) [40]. None of the samples showed adapter contamination. Trimming and host contamination removal were conducted using KneadData (version 0.7.2; https://huttenhower.sph.harvard.edu/kneaddata/, accessed 30 Aug 2020).

#### *Read-based analysis*

Taxonomic composition of the samples was profiled using mOTUs2 (version 2.5.0) [41] as well as MetaPhlAn2 (2.9.19) [42]. Both methods are marker-based and were used to profile all taxonomic levels. Functional profiling was conducted using HUMAnN2 (version 2.8.1) [43]. The R-package phyloseq (version 1.28.0) [44] was used to plot the relative abundances in each sample at different taxonomic levels, ranging from kingdom to species. Alpha-diversity was computed using multiple measurements for each sample. The distributions of the alpha-diversity values were compared between patient groups for the same time point and between time points for the same patient group. Beta-diversity was calculated using the Bray-Curtis distance. Differential abundance analysis was performed by comparing the taxa abundance between groups at the same timepoint and within groups for different time points using the R-package ALDEx2 (version 1.14.1) [45]. Metacoder R-package [46] was used to visualize differences in taxa abundance between PD patients and controls. Regression-based balance analysis of the taxa was done using the R-package selbal (version 0.1.0) [47]. For analysis with the selbal algorithm, we included all samples and all time points. Mash distances were computed on the preprocessed reads using Mash (version 2.1.1) [20].

#### *Reference-free analysis*

Reference-free analysis closely resembled the BusyBee workflow [15], which is centered around *k*-mers. *De novo* assembly was performed using SPAdes (version 3.13.1) [48] for all samples with matching baseline and 8-week follow-up datasets. The obtained contigs were filtered by length, and sequences shorter than 5000 bp were discarded. Of these filtered sequences longer than 5000 bp, 5-mers and reverse complement 5-mers distributions were computed. Samples were then pooled, and a uniform manifold approximation and projection (UMAP) was computed [49]. The embedded data points were then reassigned to their respective group-time point combination. Contigs for further analysis were taken from the PD + RS group lying within the UMAP coordinates $16.8 < X < 18.2$

and 7.5 < Y < 11. The remaining contigs were analyzed with BusyBee. The reported taxonomic assignment of the filtered contigs was computed with CAT/BAT (version 5.0.3) [50].

## Ethical statement

The study was reviewed and approved by the ethics committee of the Medical Association of Saarland, Saarbruecken, Germany, and registered under the reference number 189/15. The study was registered at the clinical trials registry ClinicalTrials.gov (ID: NCT02784145). Written informed consent was obtained from all subjects prior to inclusion in the study.

## Data availability

Data obtained in this study have been deposited in the Genome Sequence Archive for Human [51] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (BioProject: PRJCA004435), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa-human/.

## CRediT author statement

**Anouck Becker:** Conceptualization, Formal analysis, Investigation, Writing - review & editing, Visualization. **Georges Pierre Schmartz:** Methodology, Software, Validation, Formal analysis, Data curation, Writing - review & editing, Visualization. **Laura Gröger:** Methodology, Investigation, Writing - review & editing. **Nadja Grammes:** Methodology, Software, Validation, Formal analysis, Data curation, Writing - review & editing, Visualization. **Valentina Galata:** Methodology, Software, Validation, Formal analysis, Data curation, Writing - review & editing, Visualization. **Hannah Philippeit:** Investigation, Formal analysis, Writing - review & editing. **Jacqueline Weiland:** Investigation, Formal analysis, Writing - review & editing. **Nicole Ludwig:** Methodology, Investigation, Writing - review & editing. **Eckart Meese:** Methodology, Resources, Writing - review & editing, Supervision. **Sascha Tierling:** Investigation, Writing - review & editing. **Jörn Walter:** Methodology, Resources, Writing - review & editing, Supervision. **Andreas Schwiertz:** Investigation, Resources, Writing - review & editing. **Jörg Spiegel:** Investigation, Writing - review & editing. **Gudrun Wagenpfeil:** Formal analysis, Writing - review & editing. **Klaus Faßbender:** Conceptualization, Resources, Writing - review & editing. **Andreas Keller:** Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data curation, Writing - review & editing, Visualization, Supervision. **Marcus M. Unger:** Conceptualization, Investigation, Writing - original draft, Supervision, Project administration, Funding acquisition. All authors have read and approved the final manuscript.

## Competing interests

AS is a consultant for SymbioPharm GmbH, Herborn, Germany. The other authors have declared that no competing interests exist.

## Acknowledgments

This study was funded by the Michael J. Fox Foundation for Parkinson's Research (Grant No. 14603) and the German Parkinson Society.

## Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.08.009.

## ORCID

ORCID 0000-0003-2465-2276 (Anouck Becker)
ORCID 0000-0002-9627-9223 (Georges Pierre Schmartz)
ORCID 0000-0003-1043-3910 (Laura Gröger)
ORCID 0000-0002-4845-2757 (Nadja Grammes)
ORCID 0000-0002-4541-427X (Valentina Galata)
ORCID 0000-0002-7998-3630 (Hannah Philippeit)
ORCID 0000-0001-5541-257X (Jacqueline Weiland)
ORCID 0000-0003-4703-7567 (Nicole Ludwig)
ORCID 0000-0001-7569-819X (Eckart Meese)
ORCID 0000-0001-9211-5291 (Sascha Tierling)
ORCID 0000-0003-0563-7417 (Jörn Walter)
ORCID 0000-0002-2876-2352 (Andreas Schwiertz)
ORCID 0000-0002-0527-0144 (Jörg Spiegel)
ORCID 0000-0002-1133-2049 (Gudrun Wagenpfeil)
ORCID 0000-0003-3596-868X (Klaus Faßbender)
ORCID 0000-0002-5361-0895 (Andreas Keller)
ORCID 0000-0003-0805-6968 (Marcus M. Unger)

## References

[1] Hasegawa S, Goto S, Tsuji H, Okuno T, Asahara T, Nomoto K, et al. Intestinal dysbiosis and lowered serum lipopolysaccharide-binding protein in Parkinson's disease. PLoS One 2015;10:e0142164.

[2] Pietrucci D, Cerroni R, Unida V, Farcomeni A, Pierantozzi M, Mercuri NB, et al. Dysbiosis of gut microbiota in a selected population of Parkinson's patients. Parkinsonism Relat Disord 2019;65:124–30.

[3] Sun MF, Shen YQ. Dysbiosis of gut microbiota and microbial metabolites in Parkinson's disease. Ageing Res Rev 2018;45:53–61.

[4] Scheperjans F, Aho V, Pereira PAB, Koskinen K, Paulin L, Pekkonen E, et al. Gut microbiota are related to Parkinson's disease and clinical phenotype. Mov Disord 2015;30:350–8.

[5] Petrov VA, Saltykova IV, Zhukova IA, Alifirova VM, Zhukova NG, Dorofeeva YB, et al. Analysis of gut microbiota in patients with Parkinson's disease. Bull Exp Biol Med 2017;162:734–7.

[6] Unger MM, Spiegel J, Dillmann KU, Grundmann D, Philippeit H, Bürmann J, et al. Short chain fatty acids and gut microbiota differ between patients with Parkinson's disease and age-matched controls. Parkinsonism Relat Disord 2016;32:66–72.

[7] Hill-Burns EM, Debelius JW, Morton JT, Wissemann WT, Lewis MR, Wallen ZD, et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. Mov Disord 2017;32:739–49.

[8] Li W, Wu X, Hu X, Wang T, Liang S, Duan Y, et al. Structural changes of gut microbiota in Parkinson's disease and its correlation with clinical features. Sci China Life Sci 2017;60:1223–33.

[9] Fu X, Liu Z, Zhu C, Mou H, Kong Q. Nondigestible carbohydrates, butyrate, and butyrate-producing bacteria. Crit Rev Food Sci Nutr 2019;59:S130–52.

[10] Louis P, Flint HJ. Formation of propionate and butyrate by the human colonic microbiota. Environ Microbiol 2017;19:29–41.

[11] Furusawa Y, Obata Y, Fukuda S, Endo TA, Nakato G, Takahashi D, et al. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. Nature 2013;504:446–50.

[12] Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature 2011;473:174–80.

[13] Brumfield KD, Huq A, Colwell RR, Olds JL, Leddy MB, Gyarmati P. Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. PLoS One 2020;15:e0228899.

[14] Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: advantages of whole genome shotgun *versus* 16S amplicon sequencing. Biochem Biophys Res Commun 2016;469:967–77.

[15] Laczny CC, Kiefer C, Galata V, Fehlmann T, Backes C, Keller A. BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. Nucleic Acids Res 2017;45: W171–9.

[16] Guardiola J, Lobatón T, Rodríguez-Alonso L, Ruiz-Cerulla A, Arajol C, Loayza C, et al. Fecal level of calprotectin identifies histologic inflammation in patients with ulcerative colitis in clinical and endoscopic remission. Clin Gastroenterol Hepatol 2014;12:1865–70.

[17] Mulak A, Koszewicz M, Panek-Jeziorna M, Koziorowska-Gawron E, Budrewicz S. Fecal calprotectin as a marker of the gut immune system activation is elevated in Parkinson's disease. Front Neurosci 2019;13:992.

[18] Schwiertz A, Spiegel J, Dillmann U, Grundmann D, Bürmann J, Faßbender K, et al. Fecal markers of intestinal inflammation and intestinal permeability are elevated in Parkinson's disease. Parkinsonism Relat Disord 2018;50:104–7.

[19] Alfa MJ, Strang D, Tappia PS, Graham M, Van Domselaar G, Forbes JD, et al. A randomized trial to determine the impact of a digestion resistant starch composition on the gut microbiome in older and mid-age adults. Clin Nutr 2018;37:797–807.

[20] Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17:132.

[21] Gazerani P. Probiotics for Parkinson's disease. Int J Mol Sci 2019;20:41212.

[22] Akbari E, Asemi Z, Daneshvar Kakhaki R, Bahmani F, Kouchaki E, Tamtaji OR, et al. Effect of probiotic supplementation on cognitive function and metabolic status in Alzheimer's disease: a randomized, double-blind and controlled trial. Front Aging Neurosci 2016;8:256.

[23] Heintz-Buschart A, Pandey U, Wicke T, Sixel-Döring F, Janzen A, Sittig-Wiegand E, et al. The nasal and gut microbiome in Parkinson's disease and idiopathic rapid eye movement sleep behavior disorder. Mov Disord 2018;33:88–98.

[24] Hopfner F, Künstner A, Müller SH, Künzel S, Zeuner KE, Margraf NG, et al. Gut microbiota in Parkinson disease in a northern German cohort. Brain Res 2017;1667:41–5.

[25] Nishiwaki H, Ito M, Ishida T, Hamaguchi T, Maeda T, Kashihara K, et al. Meta-analysis of gut dysbiosis in Parkinson's disease. Mov Disord 2020;35:1626–35.

[26] Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 2014;12:87.

[27] Pinto-Sanchez MI, Hall GB, Ghajar K, Nardelli A, Bolino C, Lau JT, et al. Probiotic bifidobacterium longum NCC3001 reduces depression scores and alters brain activity: a pilot study in patients with irritable bowel syndrome. Gastroenterology 2017;153:448–59.e8.

[28] Anderson G, Seo M, Berk M, Carvalho A, Maes M. Gut permeability and microbiota in Parkinson's disease: role of depression, tryptophan catabolites, oxidative and nitrosative stress and melatonergic pathways. Curr Pharm Des 2016;22:6142–51.

[29] van de Wouw M, Boehme M, Lyte JM, Wiley N, Strain C, O'Sullivan O, et al. Short-chain fatty acids: microbial metabolites that alleviate stress-induced brain-gut axis alterations. J Physiol 2018;596:4923–44.

[30] Skonieczna-Żydecka K, Grochans E, Maciejewska D, Szkup M, Schneider-Matyka D, Jurczak A, et al. Faecal short chain fatty acids profile is changed in Polish depressive women. Nutrients 2018;10:1939.

[31] Caspani G, Kennedy S, Foster JA, Swann J. Gut microbial metabolites in depression: understanding the biochemical mechanisms. Microb Cell 2019;6:454–81.

[32] Berni Canani R, Sangwan N, Stefka AT, Nocerino R, Paparo L, Aitoro R, et al. *Lactobacillus rhamnosus* GG-supplemented formula expands butyrate-producing bacterial strains in food allergic infants. ISME J 2016;10:742–50.

[33] Fahn S, Elton R, UPDRS. Program Members. Unified Parkinson's disease rating scale. In: Fahn S, Marsden CC, Goldstein M, Calne DB, editors. Recent developments in Parkinson's Disease (Vol. 2). New Jersey: Macmillan Healthcare Information; 1897, p.153–63.

[34] Folstein MF, Folstein SE, McHugh PR. Mini-mental state. J Psychiatr Res 1975;12:189–98.

[35] Agachan F, Chen T, Pfeifer J, Reissman P, Wexner SD. A constipation scoring system to simplify evaluation and management of constipated patients. Dis Colon Rectum 1996;39:681–5.

[36] Beck AT, Ward CH, Mendelsohn M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71.

[37] Chaudhuri KR, Martinez-Martin P, Schapira AHV, Stocchi F, Sethi K, Odin P, et al. International multicenter pilot study of the first comprehensive self-completed nonmotor symptoms questionnaire for Parkinson's disease: the NMSQuest study. Mov Disord 2006;21:916–23.

[38] Guy W. ECDEU assessment manual for psychopharmacology. Rockville: US Department of Health, Education and Welfare; 1976, p.217–20.

[39] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32:3047–8.

[40] Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. Methods 2013;63:41–9.

[41] Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun 2019;10:1014.

[42] Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012;9:811–4.

[43] Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 2018;15:962–8.

[44] McMurdie PJ, Holmes S, Watson M. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 2013;8:e61217.

[45] Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB, Parkinson J. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One 2013;8:e67019.

[46] Foster ZSL, Sharpton TJ, Grünwald NJ, Poisot T. Metacoder: an R package for visualization and manipulation of community taxonomic diversity data. PLoS Comput Biol 2017;13:e1005404.

[47] Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML, et al. Balances: a new perspective for microbiome analysis. mSystems 2018;3:e00053-18.

[48] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77.

[49] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv 2018;1802.03426.

[50] von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome Biol 2019;20:217.

[51] Chen T, Chen Xu, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578–83.

## 3.3 *The Effect of a Planetary Health Diet on the Human Gut Microbiome: A Descriptive Analysis*

nutrients

MDPI

*Article*

# The Effect of a Planetary Health Diet on the Human Gut Microbiome: A Descriptive Analysis

**Jacqueline Rehner [1], Georges P. Schmartz [2], Tabea Kramer [1], Verena Keller [3], Andreas Keller [2] and Sören L. Becker [1,*]**

[1] Institute of Medical Microbiology and Hygiene, Saarland University, 66421 Homburg, Germany; jacqueline.rehner@uks.eu (J.R.)
[2] Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany
[3] Department of Medicine II, Saarland University Medical Center, 66421 Homburg, Germany
* Correspondence: soeren.becker@uks.eu; Tel.: +49-6841-1623900; Fax: +49-6841-1423985

**Abstract:** In 2019, researchers from the EAT-*Lancet* Commission developed the 'Planetary Health (PH) diet'. Specifically, they provided recommendations pertaining to healthy diets derived from sustainable food systems. Thus far, it has not been analysed how such a diet affects the human intestinal microbiome, which is important for health and disease development. Here, we present longitudinal genome-wide metagenomic sequencing and mass spectrometry data on the gut microbiome of healthy volunteers adhering to the PH diet, as opposed to vegetarian or vegan (VV) and omnivorous (OV) diets. We obtained basic epidemiological information from 41 healthy volunteers and collected stool samples at inclusion and after 2, 4, and 12 weeks. Individuals opting to follow the PH diet received detailed instructions and recipes, whereas individuals in the control groups followed their habitual dietary pattern. Whole-genome DNA was extracted from stool specimens and subjected to shotgun metagenomic sequencing (~3 GB per patient). Conventional bacterial stool cultures were performed in parallel and bacterial species were identified with matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry. We analysed samples from 16 PH, 16 OV, and 9 VV diet patterns. The α-diversity remained relatively stable for all dietary groups. In the PH group, we observed a constant increase from 3.79% at inclusion to 4.9% after 12 weeks in relative abundance of *Bifidobacterium adolescentis*. Differential PH abundance analysis highlighted a non-significant increase in possible probiotics such as *Paraprevotella xylaniphila* and *Bacteroides clarus*. The highest abundance of these bacteria was observed in the VV group. Dietary modifications are associated with rapid alterations to the human gut microbiome, and the PH diet led to a slight increase in probiotic-associated bacteria at ≥4 weeks. Additional research is required to confirm these findings.

**Keywords:** microbiome; Planetary Health; metagenomics; diet; dietary fiber

check for updates

**Citation:** Rehner, J.; Schmartz, G.P.; Kramer, T.; Keller, V.; Keller, A.; Becker, S.L. The Effect of a Planetary Health Diet on the Human Gut Microbiome: A Descriptive Analysis. *Nutrients* **2023**, *15*, 1924. https://doi.org/10.3390/nu15081924

Academic Editor: Marloes Dekker Nitert

Received: 13 March 2023
Revised: 7 April 2023
Accepted: 13 April 2023
Published: 16 April 2023

## 1. Introduction

In 2019, the EAT-*Lancet* Commission developed the so-called 'Planetary Health diet' (PH), a diet concept framework, which could provide a healthy diet for up to 10 billion people in 2050 within the planetary boundaries from sustainably sourced food, thereby reducing the worldwide number of deaths associated with a poor diet. The main focus of this diet consists of a reduction in animal products and processed food consumption and an increase in dietary fibre uptake through plant-based products [1,2].

Dietary fibre is a non-digestible carbohydrate for humans, but a main nutrient source for bacteria, which reside in the human intestine. The human gut microbiome describes all such microorganisms and their genomic information, including bacteria, viruses, fungi, and archaea, which are located in several niches in the gastrointestinal tract [3]. Its impact on health homeostasis and risk-modulating role in developing a variety of chronic, especially inflammatory, diseases, as well as disease progression, have been evaluated in a recent study [4]. The bacterial composition within the human gut can be altered, especially in the first three years of life, but also later on during adulthood. Major microbiome-influencing factors include the mode of delivery (i.e., natural passage through the birth canal or caesarean section), early life nutrition, as well as stress and diet choices during adulthood [5,6]. Focusing on diet and its consequences for bacteria-derived metabolites produced in the gastrointestinal tract, dietary fibre has been shown to be one of the main modulating nutrients [7].

Commensal members of the gut microbiota ferment these poly- and oligosaccharides, thereby producing short-chain fatty acids (SCFAs) such as acetate, propionate, and butyrate. SCFAs have been shown to influence glucose and lipid metabolism and regulate immunity, inflammation, and blood pressure [8]. Furthermore, the presence of SCFA-producing bacteria, and thus also the presence of SCFA detected in faeces, has been correlated with a protection against allergic reactions in the respiratory tract, suggesting their important role in shaping the immune system [9]. Hence, SCFAs are key elements in health homeostasis [10–12]. Dietary fibre uptake has further been correlated with a greater gut microbiota diversity and, if compared with the Western diet, less occurrence of chronic inflammatory disease through SCFA-producing gut microbiota [13,14]. Therefore, an increase in dietary fibre intake, as suggested by the PH diet, can lead to an increase in microbial-derived SCFAs, which have a positive and protective effect on overall health.

The Mediterranean Diet, which focusses on an increase in dietary fibre uptake through plant-based foods and, similar to the PH diet, a reduction in processed foods and saturated fatty acids, moderate consumption of fish, poultry, and dairy products, as well as low consumption of red meat, was shown to positively influence the human gut microbiome and overall health status. The consumption of animal-derived foods is clearly reduced in the Mediterranean Diet when compared with the Western diet. However, the PH diet concept suggests to reduce the intake of meat and dairy even further. After following the Mediterranean Diet, an increase in microbiota diversity and microbiota-derived metabolites, in particular SCFAs, has previously been reported [15].

Another diet concept that is gaining more popularity is the plant-based diet. This diet emphasizes the consumption of plant-derived foods, such as fruits, whole grains, nuts, seeds, and vegetables, whereas animal products are minimised or strictly eliminated [16]. Similar to the Mediterranean Diet, following a plant-based diet has been shown to increase microbial diversity in the human intestine and positively affect the abundance of beneficial bacteria, such as *Prevotella* sp. [17]. Moreover, plant-based diets have been associated with reduced inflammation, lower risk of cardiovascular diseases, and improved glucose metabolism [18].

Another popular approach to maintain overall health, as well as weight management, is a low-fat diet. These diets usually focus on reducing the intake of fat to a maximum of 30% of total energy intake, while on the other hand increasing the consumption of other macronutrients, such as protein, carbohydrates, and dietary fibre [19]. Low-fat diets can be a powerful method in weight management; however, they also have been shown to decrease the diversity and abundance of several beneficial bacteria in the gut, such as *Bifidobacterium* sp. As these diet concepts vary greatly in the specific composition of the chosen foods and nutrients, positive changes within the intestinal microbiota composition have also been reported, such as an increase in beneficial *Prevotella* sp., similar to the results after following a Mediterranean Diet [20,21].

The focus of the PH diet consists of an increase in dietary fibre through the consumption of vegetables, fruits, and whole grains, and could thus lead to similar changes within the human gut microbiome as, for example, the Mediterranean Diet or plant-based diet concepts. While the PH diet concept is gaining more and more attention and support from various stakeholders, e.g., pertaining to an improved cognitive function, criticism has been raised about a relative lack of scientific evidence pertaining to its actual health effects [22–24]. To shed light on the controversial discussion about the PH diet concept, we aimed to analyse the effects of following the PH diet over the course of twelve weeks on overall biodiversity and gut microbiota composition in contrast to the most prevalent omnivorous diet (OV) and the vegan/vegetarian diet (VV). The OV Western diet followed by the participants consisted of a low intake of dietary fibre through fruits, vegetables, and wholegrains. Furthermore, individuals following this diet concept had a very high intake of highly processed foods, dairy products, meat, and refined sugars, forming the opposite of the PH diet concept. Individuals following a vegan diet are characterised by the eradication of any animal-derived products as nutrient sources; however, levels of dietary fibre intake and highly processed foods vary greatly between individuals. The abdication of meat products from an individual's diet concept is the central component of the vegetarian diet, which was included in the VV as well. Yet, similar to individuals following a vegan diet, ranges of dietary fibre uptake and highly processed foods can vary.

## 2. Materials and Methods

### 2.1. Study Design

Healthy adults aged $\geq$ 18 years were recruited to the study. Volunteers were invited to participate in the Saarland area, southwest Germany from January to April 2022. Several exclusion criteria were defined to reduce potential bias owing to the relatively small number of study participants, i.e., pregnancy, active smoking, acute and/or chronic disease conditions, and the use of antibiotics within the last 6 months prior to inclusion. We recorded a detailed medical history of each participant, including major factors that affect the microbiome, such as (i) birth condition, (ii) medication during the first three years of life, (iii) exposure to animals within the first three years of life, and (iv) breast milk or formula use. Participants were divided into three groups according to their diet: two control groups, following a VV or OV for at least one year, and the intervention group. Participants belonging to the intervention group changed from an omnivorous diet to the PH diet. Prior to the study, these participants received detailed instructions and recipes according to the guidelines developed by the EAT-*Lancet* commission (document available online at https://www.wwf.de/fileadmin/fm-wwf/Publikationen-PDF/Landwirtschaft/wwf-wochenmenue-besseresser-innen-flexitarisch.pdf, (accessed on 12 April 2023). All participants collected faecal samples in a sterile collection tube at four different time points: initiation of the study and after two, four, and twelve weeks (Figure 1). Samples were then transferred to the laboratory within 24 h and stored at −80 °C until further processing. Furthermore, we asked all participants to document whether they had an excessive alcohol intake during the course of the study, as well as the exact foods they consumed two days prior to the collection of each faecal sample in a printed food diary, in order to reduce any potential bias that might be explained by different food choices shortly before sample collection. Individuals adhering to the PH diet were asked to track any divergence from the foods recommended by the EAT-*Lancet* commission across the entire study duration.

### 2.2. Ethical Considerations

All faecal samples were collected at the Saarland University Medical Center (Homburg, Germany) after having obtained written informed consent from all participants. For this study, we obtained ethical approval from the regional ethics committee ('Ärztekammer des Saarlandes', reference no.: 116/22).

### 2.3. DNA Extraction

We extracted whole-genome DNA from all faecal samples using the ZymoBIOMICS DNA Miniprep Kit [25]. DNA was isolated and purified according to the manufacturer's protocol. Briefly, 50 mg of faecal matter was used for the mechanical lysis step of the protocol, according to the manufacturer's recommendation. The respective lysis of microbial cells was performed using the MP Biomedicals™ FastPrep-24™ 5G Instrument (FisherScientific GmbH, Schwerte, Germany). The manufacturer's protocol was adjusted in regards to the used velocity and duration of the mechanical lysis, which was increased to 6 m/s for

45 s three times with 30 s of storage on ice in between each lysis step. Finally, we eluted the DNA in 20 µL of DNase-/RNase-free water. Subsequent concentration determination of the eluted DNA was performed via NanoDrop 2000/2000c (ThermoFisher Scientific, Wilmington, NC, USA) full-spectrum microvolume UV/Vis measurements.

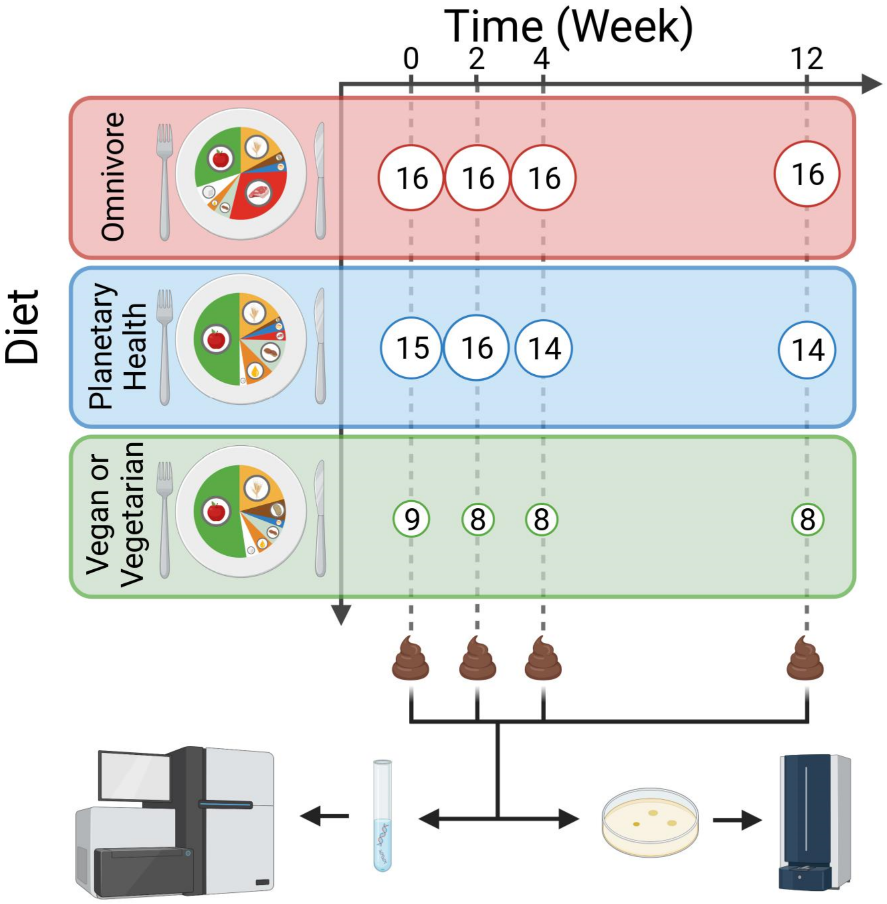


**Figure 1.** Design of the study. Participants followed three different diets over the course of twelve weeks. Stool was sampled at different time points and whole metagenome sequencing was performed. Additionally, bacteria were cultivated on different agar plates and analysed with MALDI-TOF mass spectrometry. Numbers in white circles depict the numbers of participants and respective stool samples at the different time points for each group.

### *2.4. Library Preparation and Sequencing*

Extracted whole-genome DNA was sent to Novogene Company Limited (Cambridge, UK) for library preparation and sequencing. Briefly, samples were subjected to metagenomic library preparation and further sequenced via paired-end Illumina Sequencing PE150 (HiSeq). For all samples, 3 Gb reads per sample were generated.

### *2.5. Culturing of Bacteria*

Native samples from five randomly selected participants per diet group were homogenised by vortexing after defrosting in order to achieve equal bacterial distribution within the sample without lysing the cells. Then, samples were streaked out on three different agar plates: tryptic soy agar with 5% sheep blood (TSA), MacConkey (MC), and Columbia (Co) agar plates (Becton, Dickinson and Company, Franklin Lakes, NJ, USA). We incubated all TSA and MC agar plates at 35 °C and 5% $CO_2$ for 18 h to 24 h. Anaerobic bacteria were cultivated on Co agar plates in an anaerobic environment at 35 °C for at least 48 h.

### 2.6. Mass-Spectrometry-Based Identification

After incubation of native sample material on different agar plates, grown bacterial colonies were identified on the species-level using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (MS). To this end, we picked colonies and spotted them onto the MALDI-TOF target plate and overlaid them with 1 µL of α-cyano-4-hydroxycinnamic acid (CHCA) matrix solution (Bruker Daltonics), which is composed of saturated CHCA dissolved in 50% ($v/v$) of acetonitrile, 47.5% ($v/v$) of LC-MS grade water, and 2.5% ($v/v$) of trifluoroacetic acid. The overlaid spots were then dried at room temperature and the target was subsequently placed into the Microflex LT Mass Spectrometer (Bruker Daltonics, Billerica, MA, USA) for MALDI-TOF MS analysis. We performed all measurements with the AutoXecute algorithm using FlexControl© software (version 3.4; Bruker Daltonics, Billerica, MA, USA). Each spot was automatically excited with 240 laser shots at six random positions to generate protein mass profiles in linear positive ion mode. The laser frequency was set to 60 Hz, high voltage of 20 kV, and a pulsed ion extraction of 180 ns. We measured mass charge ratio ranges ($m/z$) from 2 kDa to 20 kDa. The MALDI BioTyper software was used to identify bacterial species based on their protein mass profiles measured. In this study, we only considered identification scores $\geq$ 2.0 for analyses, which represent a precise identification on the species level, while scores between 1.7 and 1.99 were discarded as they are considered as possible species identification, and all identification scores below 1.7 were considered unsuccessful identification.

### 2.7. Data Analysis

The first step of data analysis comprised human read removal with KneadData (version (v):0.7.4, command line arguments (cla): "–trimmomatic-options='LEADING:3 TRAILING:3 MINLEN:50′ –bowtie2-options='–very-sensitive –no-discordant –reorder'") [26]. Next, we visualised the quality of the remaining reads with fastp (v:0.20.1) and MultiQC (v1.11) on default settings [27,28]. We computed a first taxonomic profile of quality-controlled reads with MetaPhlAn3 (v3.0.13, cla: "-t rel_ab_w_read_stats –unknown_estimation –add_viruses") on the ChocoPhlAn (v:mpa_v30_CHOCOPhlAn_201901) resource [29]. A second taxonomic profile was generated based on sourmash (v4.4.3, cla: "sketch dna -p k = 21, k = 31, k = 51, scaled = 1000, abund –merge") and the prepared Genome Taxonomy Database (v:GTDB R07-RS207 all genomes k51) [30,31]. Sample signatures were computed for k-mer sizes 21, 31, and 51. Distances among samples and database comparison were computed using k-mer signatures of size 31 and 51, respectively. All taxonomic profiles were then pruned and rescaled to remove viral counts.

The results of the individual samples were aggregated, and further downstream analysis was performed in R relying on the phyloseq package (v1.40.0) [31]. β-diversity was computed using the weighted UniFrac distance. Shannon diversity was used as the α-diversity measure and a two-sided unpaired Wilcoxon rank sum test was performed to test significance with a false discovery rate of 0.05. The two-dimensional embedding of sourmash sketches was performed with UMAP (v:0.2.8) [32].

Differential abundance analysis was performed with ALDEex2 (v:1.28.1) and ANCOMBC (v:1.6.2) comparing vegetarians and omnivores [33,34]. MetaPhlAn3 relative taxonomic abundances were scaled by their read count of the sample after quality control for ANCOMBC. A mean species abundance across all time points was computed for each participant, adjusting for library size if absolute counts were considered. Further, for a species to be considered for analysis, it had to be detected in over 10% of samples. Next, abundance analysis was performed, and the results were sorted by absolute effect size. We pruned the list, focusing only on the first ten percent, and intersected the sets derived from the same taxonomic profiles.

## 3. Results

### *3.1. Intestinal Microbial Diversity Stays Relatively Stable over Time*

Overall, 41 individuals from the same geographic location (Germany) were included: 16 participants following an OV, 9 following a VV, and 16 individuals who changed from an OV pattern to the PH diet at inclusion. Participating individuals were between 19 and 59 years old, with age ranges between all diet groups being non-significantly different (ANOVA $p$-value ≈ 0.84). Sex ratios differed significantly between the three diet groups (Fisher's exact test $p$-value ≈ 0.024), with more females in the VV group (8/9 individuals). General information about age, sex, and body mass index (BMI) is summarised in Table 1.

**Table 1.** General participant information. Listed below are the age ranges, BMI ranges, and sex ratio for all groups.

| | **OV** | **VV** | **PH** |
|---|---|---|---|
| **Age ranges** | 27–56 | 22–55 | 19–57 |
| **BMI ranges** | 19.8–32.8 | 19.9–40.1 | 20.0–24.4 |
| **Male** | 10 | 1 | 4 |
| **Female** | 6 | 8 | 12 |

Over the course of twelve weeks, the average α-diversity remained relatively stable for all diet groups (Figure 2A). Slight increases and decreases for individual participants were detectable between the different time points. On the one hand, investigation of the β-diversity based on dimensionality reduction of species information showed no distinct cluster formation, suggesting that, independent of the diet and time point, samples were all rather similar in their microbial composition (Figure 2B). On the other hand, reference-free diversity analysis based on sequence information alone with sourmash highlighted VV samples to be similar, whereas samples from OV and PH did not form distinct clusters, suggesting similarities between those two groups (Figure 2C) [29].

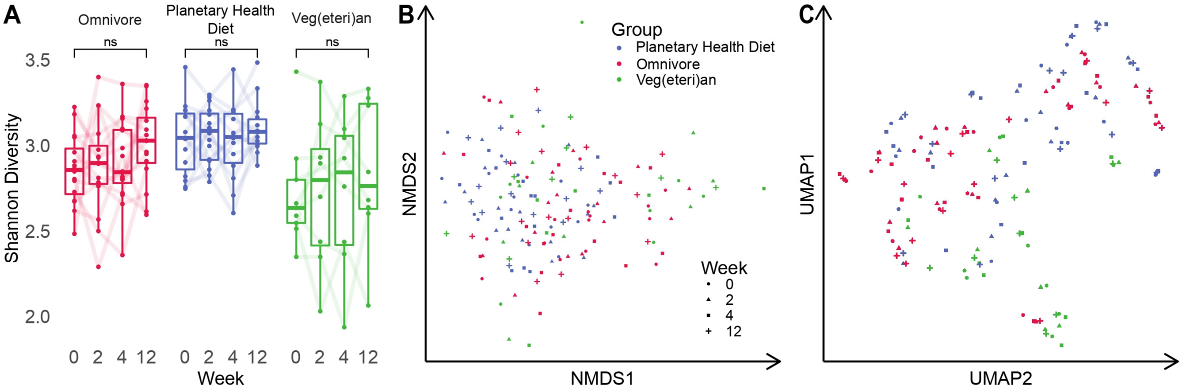


**Figure 2.** (**A**) α-diversity computed with the Shannon index for all time points and cohorts. Differences between initial and final time points were not significant for any cohort. ns = not significant (**B**) Visualised β-diversity computed with NMDS on weighted UniFrac distances among all sample pairs. (**C**) UMAP computed on sourmash distances computed among all samples.

### *3.2. Microbiota Composition Is Host-Specific and Varies between Diets*

While α-diversity describes the general number of different taxonomies present in a sample and considers the evenness of their respective abundance, taxonomic profiling enables the visualization of the exact nature of these differences. Analysis on the genus level showed variations in the microbiota composition across diets (Figures 3A and S1). In comparison with OV and PH, individuals who followed a VV diet harboured double to triple the relative amount of *Bifidobacterium* spp., *Prevotella* spp., and *Gemmiger* spp. within their intestine immediately after inclusion. *Prevotella* spp. could be detected in the OV group with a relative abundance of only 1.3%.

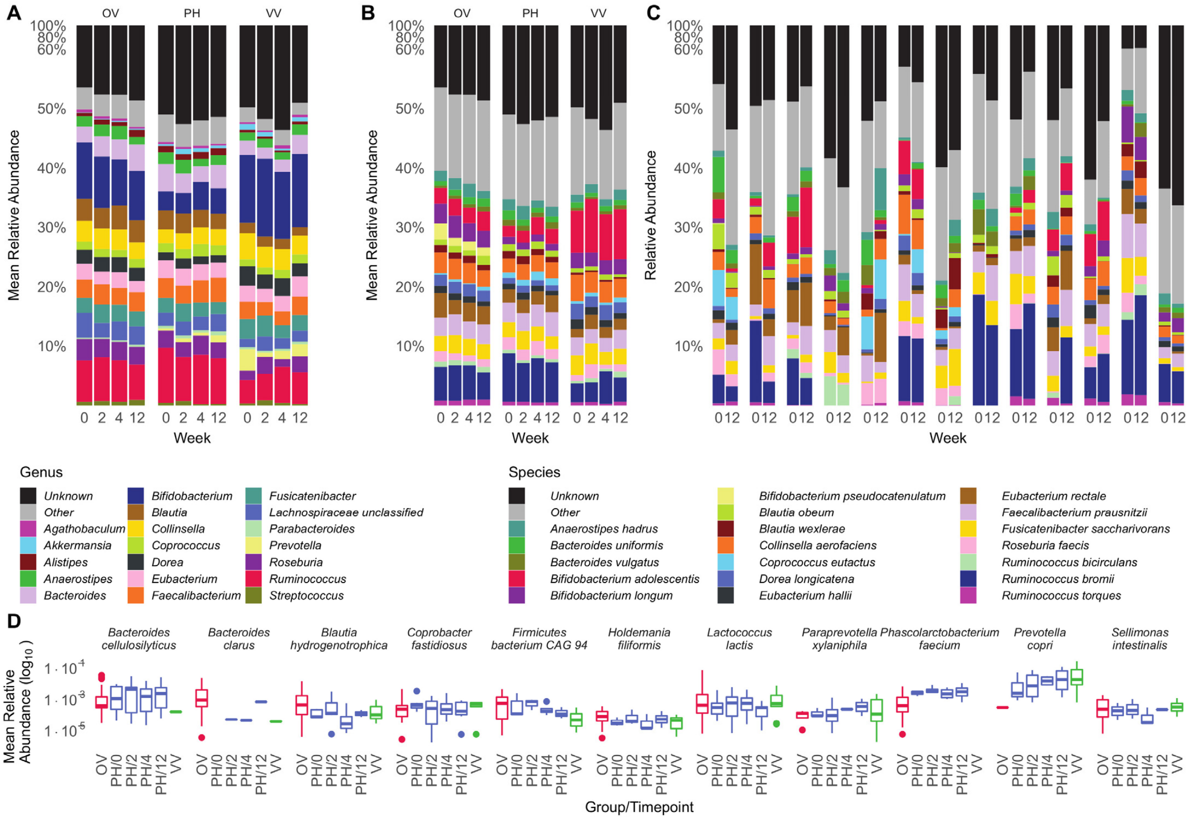


**Figure 3.** (**A**) Mean genus composition of the different dietary cohorts across different time points. Explicitly named genera were selected by looking at the highest mean relative abundances across all samples. (**B**) Identical information to panel Figure 3A, yet at species resolution. (**C**) Species composition of the PH cohort for the first and last measured time point. (**D**) Mean relative abundances of species with consistently largest effect sizes differentiating OV from VV. The results for the OV and VV cohorts were aggregated over all time points. OV, omnivore, VV, veg(etari)an, PH, Planetary Health group.

The mean relative abundance on the species level showed that the 12.1% of *Bifidobacterium* spp. in the VV consisted of 8% *Bifidobacterium adolescentis* (Figures 3B and S1). After following the PH diet for at least four weeks, we detected a two-fold increase in *Bifidobacterium adolescentis* and *Coprococcus eutactus*. These changes were not identified as significant during differential abundance analysis. We further investigated the relative abundance for each individual on the PH diet at the time of inclusion in comparison with twelve weeks after (Figures 3C and S1). Large variations in microbial composition between individuals at the time of inclusion could be observed, suggesting a partly host-specific microbiota composition.

We further analysed the differential abundance between OV and VV to highlight potentially interesting species, thereby only focusing on the top ten percent effect sizes (Figures 3D and S1). We detected a 3-fold increase in *Prevotella copri*, a 4-fold increase in *Paraprevotella xylaniphila*, and an 18-fold increase in *Bacteroides clarus*, whereas, e.g., *Firmicutes bacterium CAG 94* showed a 6-fold decrease in the PH diet over the course of the study. The differential abundance depicted in Figures 3D and S1 suggests that following the PH diet shifts parts of the microbiota composition towards a VV microbiome. However, these observed changes were not significant.

Cultivation and species identification with MALDI-TOF mass spectrometry identified 59 different bacterial species across all time points among the five randomly selected participants from each group (Figure 4). Most commonly isolated were *Escherichia coli*, *Enterococcus faecium*, *Clostridium perfringens*, and *Bifidobacterium longum*. *Enterococcus mundtii* and *Priestia megaterium* were mostly detected in the VV, whereas *Streptococcus parasanguinis*, *Streptococcus salivarius*, *Enterobacter cloacae*, and *Bacteroides uniformis* were mainly isolated from faecal samples of those participants following the PH diet. A detailed account of detected bacteria in the different groups is displayed in Supplementary Table S1. This method, however, represents only cultivable microorganisms, leaving approximately 35–65% undetected when compared with next-generation sequencing (NGS) [35].

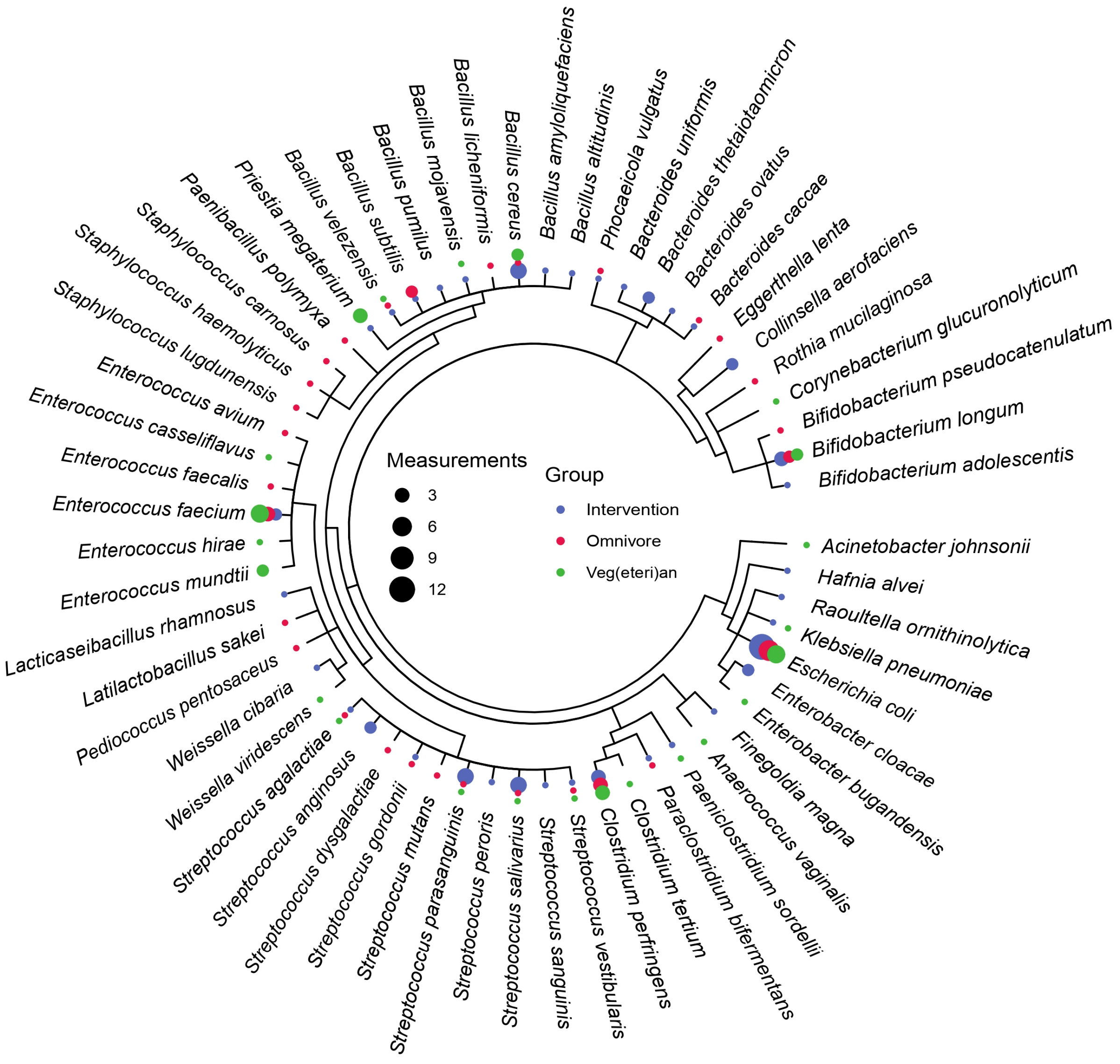


**Figure 4.** NCBI taxonomic classification of the species identified by mass spectrometry. The indicated number of measurements for the different diets represents the number of times the species has been identified in different samples at any time point.

## 4. Discussion

Our study analysed whole-genome data obtained from faecal samples after following three specific diets, i.e., OV, VV, and PH, over the course of 12 weeks to investigate the intestinal microbiota composition associated with these dietary patterns. The main difference between OV as compared with VV and PH is most likely the intake of dietary fibre. Western citizens generally ingest between 14 g (United Kingdom) and 26 g (Norway) of dietary fibre, whereas most countries recommend 25–35 g per day for adults [36]. With the PH diet suggesting 232 g of whole grains, 300 g vegetables, and 200 g fruits per day for an intake of 2500 kcal/day, participants following this diet should reach these dietary fibre recommendations [2]. A sufficient amount of fibre is directly associated with positively affecting the human intestinal microbiome, and a plant-based diet is proposed to benefit human and planetary health [15,37]. In this study, we were able to detect a trend towards an increase in *Bifidobacterium adolescentis* and *Coprococcus eutactus* (Figure 3A–C) after following the PH diet for a minimum of four weeks. An increase in *B. adolescentis* has previously been shown after supplementation with inulin, a type of dietary fibre and naturally occurring plant carbohydrate. *B. adolescentis* is capable of degrading inulin into lactate and acetate, which can be used by *Anaerostipes hadrus* and *Enterococcus rectale* to produce the SCFA butyrate [38]. In contrast to the supplementation with inulin from Baxter et al., we did not find a co-increase in *A. hadrus* when following the PH diet without tracking the exact dietary fibre composition. However, *B. adolescentis* seems to have a growth advantage after increasing inulin intake. Similarly, β-glucans have been shown to be the preferred growth substrate of *C. eutactus*, suggesting a growth advantage after increasing β-glucans consumption [39,40]. Differences in taxonomic abundances suggested that several species merit particular consideration, such as, e.g., *Prevotella copri* and *Paraprevotella xylaniphila*, for which a non-significant increase was detectable (Figure 3D and Figure S1). *P. copri* is capable of dietary fibre degradation, as they harbour vast genomic repertoires of carbohydrate active enzymes [41]. Similar to *B. adolescentis*, switching to the PH diet might favour the growth of *P. copri.* While SCFA-producing bacteria should be beneficial for the host due to their anti-inflammatory and regulatory effects, *P. copri* has also been correlated with the development of rheumatoid arthritis, although without conclusive evidence. An overgrowth of *P. copri* might also inhibit the growth of other beneficial microbiota [42]. *P. xylaniphila* can produce anti-inflammatory SCFAs, but also has the potential to synthesise pro-inflammatory metabolites, such as, for example, succinic acid. Succinic acid was previously described in close correlation with the development of hypertension, inflammatory, and metabolic diseases [43,44]. These two species, identified by differential abundance, might harbour beneficial potential, but need to be studied more extensively to analyse their exact effect on health homeostasis and their function within the complex gut microbiome. However, computing the differential abundance is a powerful tool to identify both pathogenic species and beneficial bacteria. To the best of our knowledge, no genomic or phenotypic analyses have been performed to identify the biochemical properties of *Firmicutes bacterium CAG 94*, making this species an interesting target for further research.

Several limitations of our study are offered for consideration. First, we performed a monocentric analysis with a limited number of individuals. Second, participants of this study received recipes and detailed instructions on what to consume, but we did not implement exact meal plans. For future studies, we recommend standardised meal plans to avoid any potential participant compliance issues. Third, we did not perform culture-based bacteriological analysis in all study participants. Fourth, the VV contained mostly biologically female participants, thereby creating significant differences in sex ratios between the groups. Fifth, the study groups were relatively small, and robust statistical analyses of individual groups at different time points would require a larger study population in future studies.

In conclusion, this work provides the first metagenomics-sequencing-based appraisal of the PH diet. While no significant changes were observed within the overall intestinal microbial composition of individuals opting to follow the PH diet, we identified several potentially interesting bacterial species. Indeed, when focusing on differentially abundant species between OV and VV, non-significant trends of the PH cohort towards VV were noted. Specific bacterial species are capable of producing anti-inflammatory metabolites and might be an interesting target for novel probiotics, beneficial bacteria that can be taken supplementary to a healthy diet [45]. Hence, we encourage further microbiota-targeted research pertaining to the PH diet, ideally through multi-country longitudinal and larger-scaled studies.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/nu15081924/s1. Table S1: List of bacterial species identified by culturing and MALDI-TOF. Listed are all bacterial species that were cultured from native stool samples and identified via MALDI-TOF. The table is divided in the three dietary groups.; Figure S1: Identical information as displayed in Figure 3 using sourmash for taxonomic profiling of metagenomic reads. Note, the relative amount of unknown taxonomies was removed and information was rescaled. Further, the selected species from Figure 3D were adopted and not recomputed as to highlight abundance differences among workflows.

**Author Contributions:** S.L.B., T.K. and J.R. had the idea for this study, together with G.P.S., and had full access to all the data and take responsibility for the integrity of the data and the accuracy of data analysis. T.K. and J.R. collected samples and extracted whole-genome DNA and performed culturing and MALDI-TOF analysis. Computational data analysis was performed by G.P.S. and A.K. J.R., G.P.S., V.K. and S.L.B. drafted the manuscript. All authors critically reviewed the paper for important intellectual content and agreed to submit the final version for publication. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki, and approved by the regional ethics committee) of the 'Ärztekammer des Saarlandes' (reference no.: 116/22).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Deidentified participant data and study documents (clinical questionnaire, food documentation, informed consent form, and sequencing data) will be made available to other researchers upon reasonable request directed to the corresponding author via e-mail.

**Acknowledgments:** The authors would like to thank all study participants and express their gratitude to all laboratory members involved in this study for their excellent work. This work was further supported by Saarland University and the UdS-HIPS TANDEM initiative.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Willett, W.; Rockström, J.; Loken, B.; Springmann, M.; Lang, T.; Vermeulen, S.; Garnett, T.; Tilman, D.; DeClerck, F.; Wood, A.; et al. Food in the anthropocene: The EAT-*Lancet* commission on healthy diets from sustainable food systems. *Lancet* **2019**, *393*, 447–492. [CrossRef] [PubMed]
2. Willett, W.; Rockström, J.; Loken, B.; Springmann, M.; Lang, T.; Vermeulen, S.; Garnett, T.; Tilman, D.; DeClerck, F.; Wood, A.; et al. Summary Report of the EAT-*Lancet* Commission. Available online: https://eatforum.org/eat-lancet-commission/eat-lancet-commission-summary-report/ (accessed on 7 November 2022).
3. Marchesi, J.R.; Ravel, J. The vocabulary of microbiome research: A proposal. *Microbiome* **2015**, *30*, 31. [CrossRef] [PubMed]
4. Singh, R.K.; Chang, H.W.; Yan, D.I.; Lee, K.M.; Ucmak, D.; Wong, K.; Abrouk, M.; Farahnik, B.; Nakamura, M.; Zhu, T.H.; et al. Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med.* **2017**, *15*, 73. [CrossRef]
5. Kumbhare, S.V.; Patangia, D.V.; Patil, R.H.; Souche, Y.S.; Patil, N.P. Factors influencing the gut microbiome in children: From infancy to childhood. *J. Biosci.* **2019**, *44*, 49. [CrossRef]
6. Dong, T.S.; Gupta, A. Influence of early life, diet, and the environment on the microbiome. *CGH* **2019**, *17*, 231–242. [CrossRef]
7. Tanes, C.; Bittinger, K.; Gao, Y.; Friedman, E.S.; Nessel, L.; Paladhi, U.R.; Chau, L.; Panfen, E.; Fischbach, M.A.; Braun, J.; et al. Role of dietary fiber in the recovery of the human gut microbiome and its metabolome. *Cell Host Microbe* **2021**, *29*, 394–407.e5. [CrossRef]
8. He, J.; Zhang, P.; Shen, L.; Niu, L.; Tan, Y.; Chen, L.; Zhao, Y.; Bai, L.; Hao, X.; Li, X.; et al. Short-chain fatty acids and their association with signalling pathways in inflammation, glucose and lipid metabolism. *Int. J. Mol. Sci.* **2020**, *21*, 6356. [CrossRef]

9. Arrieta, M.C.; Stiemsma, L.T.; Dimitriu, P.A.; Thorson, L.; Russell, S.; Yurist-Doutsch, S.; Kuzeljevic, B.; Gold, M.J.; Britton, H.M.; Lefebvre, D.L.; et al. Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* **2015**, *7*, 307ra152. [CrossRef] [PubMed]
10. Salvin, J. Fiber and prebiotics: Mechanisms and health benefits. *Nutrients* **2013**, *5*, 1417–1435. [CrossRef]
11. Maslowski, K.M.; Mackay, C.R. Diet, gut microbiota and immune responses. *Nat. Immunol.* **2011**, *12*, 5–9. [CrossRef]
12. Becker, A.; Schmartz, G.P.; Gröger, L.; Grammes, N.; Galata, V.; Philippeit, H.; Weiland, J.; Ludwig, N.; Meese, E.; Tierling, S.; et al. Effects of resistant starch on symptoms, fecal markers, and gut microbiota in Parkinson's Disease–The RESISTA-PD trial. *GPB* **2022**, *20*, 274–287. [CrossRef] [PubMed]
13. Yatsunenko, T.; Rey, F.E.; Manary, M.J.; Trehan, I.; Dominguez-Bello, M.G.; Contreras, M.; Magris, M.; Hidalgo, G.; Baldassano, R.N.; Anokhin, A.P.; et al. Human gut microbiome viewed across age and geography. *Nature* **2021**, *486*, 222–227. [CrossRef] [PubMed]
14. Vijay, A.; Valdes, A.M. Role of the gut microbiome in chronic disease: A narrative review. *Eur. J. Clin. Nutr.* **2022**, *76*, 489–501. [CrossRef] [PubMed]
15. Nagpal, R.; Shively, C.A.; Register, T.C.; Craft, S.; Yadav, H. Gut microbiome-mediterranean diet intercations in improving host health. *F1000research* **2019**, *8*, 699. [CrossRef]
16. Clem, J.; Barthel, B. A look at plant-based diets. *Mo. Med.* **2021**, *118*, 233–238.
17. Jeffery, I.B.; O'Toole, P.W. Diet-microbiota interactions and their implications for healthy living. *Nutrients* **2013**, *5*, 234–252. [CrossRef]
18. Watzl, B. Anti-inflammatory effects of plant-based foods and of their constituents. *Int. J. Vitam. Nutr. Res.* **2008**, *78*, 293–298. [CrossRef]
19. David, L.A.; Maurice, C.F.; Carmody, R.N.; Gootenberg, D.B.; Button, J.E.; Wolfe, B.E.; Ling, A.V.; Devlin, A.S.; Varma, Y.; Fischbach, M.A.; et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **2014**, *505*, 559–563. [CrossRef] [PubMed]
20. Hjorth, M.F.; Blædel, T.; Bendtsen, L.Q.; Lorenzen, J.K.; Holm, J.B.; Kiilerich, P.; Roager, H.M.; Kristiansen, K.; Larsen, L.H.; Astrup, A. Prevotella-to-Bacteroides ratio predicts body weight and fat loss success on 24-week diets varying in macronutrient composition and dietary fiber: Results from a post-hoc analysis. *Int. J. Obes.* **2018**, *43*, 149–157. [CrossRef]
21. Salonen, A.; Lahti, L.; Salojärvi, J.; Holtrop, G.; Korpela, K.; Duncan, S.H.; Date, P.; Farquharson, F.; Johnstone, A.M.; Lobley, G.E.; et al. Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. *ISME J.* **2014**, *8*, 2218–2230. [CrossRef]
22. Zagmutt, F.J.; Pouzou, J.G.; Costard, S. The EAT-*Lancet* commission's dietary composition may not prevent noncommunicable disease mortality. *J. Nurt.* **2020**, *150*, 985–988. [CrossRef] [PubMed]
23. Dalile, B.; Kim, C.; Challinor, A.; Geurts, L.; Gibney, E.R.; Galdos, M.V.; La Fata, G.; Layé, S.; Mathers, J.C.; Vauzour, D.; et al. The EAT-*Lancet* reference diet and cognitive function across the life course. *Lancet Planet Health* **2022**, *6*, e749–e759. [CrossRef]
24. Cacau, L.T.; De Carli, E.; de Carvalho, A.M.; Lotufo, P.A.; Moreno, L.A.; Bensenor, I.M.; Marchioni, D.M. Development and validation of an index based on EAT-*Lancet* recommendations: The Planetary Health Diet Index. *Nutrients* **2021**, *13*, 1698. [CrossRef] [PubMed]
25. Rehner, J.; Schmartz, G.P.; Groeger, L.; Dastbaz, J.; Ludwig, N.; Hannig, M.; Rupf, S.; Seitz, B.; Flockerzi, E.; Berger, T.; et al. Systematic cross-biospecimen evaluation of DNA extraction kits for long- and short-read multi-metagenomic sequencing studies. *GPB* **2022**, *20*, 405–417. [CrossRef]
26. Beghini, F.; McIver, L.J.; Blanco-Míguez, A.; Dubois, L.; Asnicar, F.; Maharjan, S.; Mailyan, A.; Manghi, P.; Scholz, M.; Thomas, A.M.; et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **2021**, *10*, e65088. [CrossRef] [PubMed]
27. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef] [PubMed]
28. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef]
29. Brown, C.T.; Irber, L. sourmash: A library MinHash sketching of DNA. *J. Open Source Softw.* **2016**, *1*, 27. [CrossRef]
30. Parks, D.H.; Chuvochina, M.; Rinke, C.; Mussig, A.J.; Chaumeil, P.A.; Hugenholtz, P. GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank nonrmalized and complete genome-based taxonomy. *Nucl. Acids Res.* **2022**, *50*, 785–794. [CrossRef]
31. McMurdie, P.J.; Holmes, S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **2013**, *8*, e61217. [CrossRef]
32. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
33. Gloor, G. Alex2: Anova-like differential expression tool for compositional data. *ALDEX Man. Modul.* **2015**, *20*, 1–11.
34. Lin, H.; Das Peddada, S. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **2020**, *11*, 3514. [CrossRef]
35. Lagkouvardos, I.; Overmann, J.; Clavel, T. Cultured microbes represent a substantial fraction of the human and mouse gut microbiota. *Gut Microbes* **2017**, *8*, 493–503. [CrossRef]

36. Stephen, A.M.; Champ, M.M.J.; Cloran, S.J.; Fleith, M.; Van Lieshout, L.; Mejborn, H.; Burley, V.J. Dietary fibre in Europe: Current state of knowledge on definitions, sources, recommendations, intakes and relationships to health. *Nutr. Res. Rev.* **2017**, *30*, 149–190. [CrossRef]
37. Craig, M.F. A planet based diet benefits personal and planetary health. *BMJ* **2022**, *379*, o2651. [CrossRef] [PubMed]
38. Baxter, N.T.; Schmidt, A.W.; Venkataraman, A.; Kim, K.S.; Waldron, C.; Schmidt, T.M. Dynamics of human gut microbiota and short-chain fatty acids in response to dietary interventions with three fermentable fibers. *mBio* **2019**, *10*, e02566-18. [CrossRef] [PubMed]
39. Louis, P.; Solvang, M.; Duncan, S.H.; Walker, A.W.; Mukhopadhya, I. Dietary fibre complexity and its influence on functional groups of the human gut microbiota. *Proc. Nutr. Soc.* **2021**, *80*, 386–397. [CrossRef]
40. Alessi, A.M.; Gray, V.; Farquharson, F.M.; Flores-López, A.; Shaw, S.; Stead, D.; Wegmann, U.; Shearman, C.; Gasson, M.; Collie-Duguid, E.S.R.; et al. Β-Glucan is a major growth substrate for human gut bacteria related to *Coprococcus eutactus*. *Environ. Microbiol.* **2020**, *22*, 2150–2164. [CrossRef]
41. Yeoh, Y.K.; Sun, Y.; Ip, L.Y.T.; Wang, L.; Chan, F.K.; Miao, Y.; Ng, S.C. *Prevotella* species in the human gut is primarily comprised of *Prevotella copri*, *Prevotella stercorea* and related lineages. *Sci. Rep.* **2022**, *12*, 9055. [CrossRef]
42. Scher, J.U.; Sczesnak, A.; Longman, R.S.; Segata, N.; Ubeda, C.; Bielski, C.; Rostron, T.; Cerundolo, V.; Pamer, E.G.; Abramson, S.B.; et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *eLife* **2013**, *2*, e01202. [CrossRef] [PubMed]
43. Gutiérrez-Calabrés, E.; Ortega-Hernández, A.; Modrego, J.; Gómez-Gordo, R.; Caro-Vadillo, A.; Rodríguez-Bobada, C.; González, P.; Gómez-Garre, D. Gut microbiota profile identifies transition from compensated cardiac hypertrophy to heart failure in hypertensive rats. *Hypertension* **2020**, *76*, 1545–1554. [CrossRef] [PubMed]
44. Morotomi, M.; Nagai, F.; Sakon, H.; Tanaka, R. *Paraprevotella clara* gen. Nov.; spo. Nov. and *Paraprevotella xylaniphila* sp. Nov.; members of the family 'Prevotellaceae' isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **2009**, *59*, 1895–1900. [CrossRef] [PubMed]
45. Grenda, T.; Grenda, A.; Domaradzki, P.; Krawczyk, P.; Kwiatek, K. Probiotic potential of *Clostridium* spp.—Advantages and doubts. *Curr. Issues Mol. Biol.* **2022**, *44*, 3118–3130. [CrossRef] [PubMed]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

## 3.4 *Occurrence, resistance patterns, and management of carbapenemase-producing bacteria in war-wounded refugees from Ukraine*

International Journal of Infectious Diseases 132 (2023) 89–92

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Case Report

# Occurrence, resistance patterns, and management of carbapenemase-producing bacteria in war-wounded refugees from Ukraine

Fabian K. Berger[1,#], Georges P. Schmartz[2,#], Tobias Fritz[3], Nils Veith[3], Farah Alhussein[1], Sophie Roth[1], Sophie Schneitler[1], Thomas Gilcher[4], Barbara C. Gärtner[1], Vakhtang Pirpilashvili[3], Tim Pohlemann[3], Andreas Keller[2], Jacqueline Rehner[1,#], Sören L. Becker[1,#,*]

[1] *Institute of Medical Microbiology and Hygiene, Saarland University, Homburg/Saar, Germany*
[2] *Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany*
[3] *Department of Trauma, Hand and Reconstructive Surgery, Saarland University Medical Center, Homburg/Saar, Germany*
[4] *Hospital Pharmacy, Saarland University, Homburg/Saar, Germany*

ARTICLE INFO

*Article history:*
Received 19 February 2023
Accepted 11 April 2023

*Keywords:*
Antimicrobial resistance
Carbapenemase
Cefiderocol
Ceftazidime/avibactam
*Klebsiella pneumoniae*
*Pseudomonas aeruginosa*

ABSTRACT

We analyzed consecutive clinical cases of infections due to carbapenemase-producing gram-negative bacteria detected in war-wounded patients from Ukraine who were treated at one university medical center in southwest Germany between June and December 2022. The isolates of multiresistant gram-negative bacteria were subjected to a thorough microbiological characterization and whole genome sequencing (WGS). We identified five war-wounded Ukrainian patients who developed infections with New Delhi metallo-$\beta$-lactamase 1-positive *Klebsiella pneumoniae*. Two isolates also carried OXA-48 carbapenemases. The bacteria were resistant to novel antibiotics, such as ceftazidime/avibactam and cefiderocol. The used treatment strategies included combinations of ceftazidime/avibactam + aztreonam, colistin, or tigecycline. WGS suggested transmission during primary care in Ukraine. We conclude that there is an urgent need for thorough surveillance of multiresistant pathogens in patients from war zones.

## Introduction

Antimicrobial resistance (AMR) was associated with an estimated 4.95 million deaths worldwide in 2019 alone, and carbapenem-resistant gram-negative bacteria were among the major contributors to this enormous disease burden [1]. All-age death rates were the highest in sub-Saharan Africa, followed by South Asia and Eastern Europe. The war in Ukraine has led to significant migration movements, with ≥7.8 million refugees across Europe until December 2022 [2]. We report AMR patterns in war-wounded patients from Ukraine.

## Case descriptions

### *Patient P1*

A male patient aged 34 years experienced a femoral shaft fracture after an explosion. Swabs from an inserted external fixator grew *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*. Both pathogens were highly resistant to antibiotics, including carbapenems, ceftazidime/avibactam, and cefiderocol. A New Delhi metallo-$\beta$-lactamase (NDM-1) was detected in the *K. pneumoniae* strain. Both pathogens remained susceptible to colistin, whereas tigecycline showed a comparatively low minimal inhibitory concentration (MIC) of 1.5 mg/L (Table 1). The patient developed an infection and was successfully treated with colistin plus tigecycline.

[*] Corresponding author: Tel: +49 6841 16 23901.
*E-mail address:* soeren.becker@uks.eu (S.L. Becker).
[#] These authors contributed equally.

https://doi.org/10.1016/j.ijid.2023.04.394
1201-9712/© 2023 The Author(s). Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

**Table 1**
Resistance testing results and carbapenemases detected in gram-negative bacteria isolated from war-wounded patients from Ukraine in a university medical center in southwest Germany, June to November 2022.

| Patient | Bacterium | Carbapenemase | Resistance to several antimicrobials (minimal inhibitory concentration, expressed as mg/L) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MER | IMI | ERT | TIG | AMI | GEN | TOB | COL | CEF/AVI | CEF/TAZ | CEFI | | |
| | | | | | | | | | | | | | Disk diffusion | Microdilution | Epsilometry |
| P1.1 | *Klebsiella pneumoniae* | NDM-1 | R (>32) | R (>32) | R (>32) | IE (=1.5) | R (>256) | R (>1024) | R (>256) | S (0.25) | R (>8) | R (>256) | R (9 mm) | R (64) | R (3) |
| P1.2 | *Pseudomonas aeruginosa* | None | R (>32) | R | – | – | R (>256) | – | R (>256) | S (=0.5) | R (>256) | R (>256) | R (20 mm) | – | – |
| P2.1 | *Klebsiella pneumoniae* | NDM-1 | R (>32) | I (=6) | R (>32) | IE (=1.5) | R (>32) | S (=1) | R (=24) | S (=0.25) | R (>8) | R (>256) | ATU (18 mm) | S (2) | S (0.36) |
| P2.2 | *Pseudomonas aeruginosa* | None | R (>32) | I (=4) | – | – | R (=48) | – | R (>256) | S (=1) | R (>256) | R (>256) | S (25 mm) | – | – |
| P3.1 | *Klebsiella pneumoniae* | NDM-1 | R (>32) | R (>32) | – | IE (=0.75) | S (=8) | R (=64) | R (=64) | S (=2.0) | R (>8) | R (>256) | R (10 mm) | R (4) | R (12) |
| P4.1 | *Providencia stuartii* | NDM-1 | S (=2) | I (=6) | – | IE (=3) | R (>256) | R (>1024) | R (>256) | R (>16) | R (>8) | R (>256) | S (26 mm) | S (1) | S (0.047) |
| P4.2 | *Klebsiella pneumoniae* | NDM-1 & OXA-48 | R (>32) | R (>32) | – | IE (=0.5) | R (>256) | R (>1024) | R (>256) | S (=0.25) | R (>8) | R (>256) | R (14 mm) | R (8) | S (1.5) |
| P5.1 | *Klebsiella pneumoniae* | NDM-1 & OXA-48 | R (>32) | R (>32) | – | IE (=0.75) | R (>256) | R (>1024) | R (>256) | S (=0.25) | R (>8) | R (>256) | R (9 mm) | R (4) | S (1.5) |
| P5.2 | *Providencia stuartii* | NDM-1 | S (=2) | I (=6) | – | IE (=3) | – | – | R (>256) | R (>16) | – | R (>256) | – | – | – |

AMI, amikacin; ATU, Area of technical uncertainty according to the European Committee on Antimicrobial Susceptibility Testing (EUCAST) definition; CEF/AVI, ceftazidime/avibactam; CEF/TAZ, ceftolozane/tazobactam; CEFI, cefiderocol; COL, colistin; ERT, ertapenem; GEN, gentamicin; IMI, imipenem; MER, meropenem; TOB, tobramycin; TIG, tigecyclin.

All carbapenem-resistant isolates reported here were also uniformly resistant to penicillins, cephalosporins (first to fourth generation), and fluoroquinolones.

Interpretation of testing results: S, Sensitive (standard exposure); I, Sensitive (increased exposure); R, resistant; IE, insufficient evidence (no established clinical breakpoint); –, not done.

Resistance testing was performed on a MicroScan WalkAway system. Carbapenem resistance was confirmed by epsilometry. Resistance testing for colistin was performed using broth microdilution, whereas epsilometry was applied to ceftazidime/avibactam and ceftolozane/tazobactam. For cefiderocol testing, we performed broth microdilution, epsilometry (except for *P. aeruginosa*), and agar disk diffusion.

### *Patient P2*

A male patient aged 43 years had complex fractures of the tibia and humerus, which were caused by gunshots and blast injuries. The patient developed fever, and a highly resistant NDM-1-positive *K. pneumoniae* was detected in the blood cultures and wound swabs. Furthermore, a multiresistant *P. aeruginosa* strain and *Enterococcus faecalis* were recovered. The patient was successfully treated with combined colistin and high-dose imipenem (MIC of the isolate: 6 mg/L). The infection improved, but the patient required further wound debridement and vacuum-assisted closure therapy.

### *Patient P3*

A female patient aged 58 years had severe wound infection after having undergone unilateral below-knee amputation. Wound smears revealed an NDM-1-positive *K. pneumoniae* strain. The antimicrobial treatment comprised colistin and tigecycline (MIC 0.75 mg/L). The patient's course improved and she was finally discharged to a rehabilitation center.

### *Patient P4*

A female patient aged 64 years was admitted for severe blast injuries of the chest and above-elbow amputation with wound infection. *K. pneumoniae* grew in wound smears, which were positive for OXA-48 and NDM-1. Initial therapy included tigecycline (MIC 0.75 mg/L) and colistin. Colistin was discontinued due to acute kidney failure. Further smears grew NDM-positive *Providencia stuartii* strains. After surgical and antimicrobial treatment, the patient's condition improved and she was discharged.

### *Patient P5*

A male patient aged 56 years was admitted for thoracic blast injuries and extensive soft tissue damage of one leg after gunshot injury. Combined femoral and tibial fractures showed signs of infection (osteitis) and maggot infestation. The patient developed a systemic infection with an OXA-48 and NDM-1-positive *K. pneumoniae* strain, which was isolated from the blood cultures and wounds. Initial therapy consisted of tigecycline and colistin. Furthermore, an NDM-positive *P. stuartii* grew in the wound swabs. Because the patient's condition did not improve, treatment was switched to ceftazidime/avibactam in combination with aztreonam. Later, the clinical course was complicated by candidemia and the patient underwent amputation of the infected leg due to persistent infection.

## Microbiological characterization and whole genome sequencing (WGS) analysis

Details on the microbiological methods [3], sequencing techniques, and analysis models can be found in Supplement 1. Seven gene multilocus sequence typing with mlst (v2.22.1) suggested at least three different strains by predicting three different sequences types (STs) for *K. pneumoniae* isolates: ST395 (P3.1), ST147 (P2.1), and ST23 (P1.1, P4.2, and P5.1) (Supplementary Figure 1A). The three STs shared only one house-keeping gene, whereas ST147 and ST23 shared two house-keeping genes. Altogether, 25.484 single-nucleotide variants (SNVs) were shared by all three ST23 isolates, whereas 359 SNVs were unique to P1.1, 253 SNVs were unique to P4.2, and 231 SNVs were only detected for P5.1. Hence, the shared ST23, in combination with an increased SNV agreement, suggest a common epidemiological background and clonality of the *K. pneumoniae* isolates of patients 1, 4, and 5.

Resistance detection with ABRicate (v1.0.1) revealed a wide range of different resistance genes (Supplementary Table 1). Regarding carbapenem resistance, the NDM-1 coding gene *blaNDM-1* was found in all analyzed isolates. The gene coding for OXA-48 was detected in P4.2 and P5. The *blaNDM-1* plasmids of P2.1 and P3.1 were identical, as well as the plasmids isolated from P4.2 and P5, for which even chromosomal comparison showed high similarities, thus suggesting at least partial transmission between patients. Details on the plasmid genes are displayed in Supplementary Figure 1B.

Regarding the resistance against ceftazidime/avibactam and cefiderocol, respectively, neither $bla_{KPC\text{-}2}$ nor $bla_{NDM\text{-}35}$ could be detected.

## Discussion

The ongoing war in Ukraine has a profound negative impact on the country's health care system, including the fight against infectious diseases. Due to the high number of Ukrainian migrants having fled the country, specific challenges have also arisen for public health and appropriate surveillance measures in other countries [4,5]. Indeed, a French practice guideline has put forth a host of recommendations to health care providers who care for migrants from Ukraine, which prioritize communicable diseases, vaccination catch-up, and psychological sequelae, etc. [6]. After anecdotal reports of "how war is spreading drug resistant superbugs across Ukraine and beyond", [7] recent genomic surveillance data highlighted the considerable challenges arising from carbapenemase-producing gram-negative bacteria [8]. Here, we report on a series of Ukrainian patients with contaminated wounds who were found to be colonized and/or infected with NDM-1- and NDM-1/OXA-48-positive *Enterobacterales* and nonfermentative bacteria. WGS findings and the rapid detection of pathogens in swabs taken on admission to our hospital suggest a previously established colonization with these bacteria.

The epidemiology of carbapenemases found in hospitalized patients varies considerably across Europe, with the highest rates being reported from southern and southeastern Europe. In Germany, OXA-48 was the most commonly detected carbapenemase in 2021, followed in descending order by VIM-1, KPC-2, and NDM-1 [9]. Of note, bacterial strains carrying more than one carbapenemase were a rarity (<5%). Several international guidance documents have been published on the treatment of infections caused by carbapenem-resistant gram-negative bacilli, which recommend the preferred use of ceftazidime/avibactam, if susceptible *in vitro*. For metallo-$\beta$-lactamases, such as NDM-1, the use of cefiderocol monotherapy or the combination of ceftazidime/avibactam plus aztreonam is conditionally recommended [10,11]. Although these recommendations are supported by compelling susceptibility data, none of the carbapenemase-producing strains in our investigation was susceptible to ceftazidime/avibactam, and cefiderocol was resistant in four of the six tested isolates. Of note, cefiderocol susceptibility is notoriously difficult to test [12], and we also observed some discrepancy depending on the testing method. The combination of ceftazidime/avibactam and aztreonam restores activity against NDM-1-producing *K. pneumoniae* and other *Enterobacterales*, but its routine use outside clinical studies is currently hampered by the unavailability of intravenous aztreonam in some European countries and the absence of a licensed fixed dose combination of these compounds. Furthermore, this combination seems much less promising in NDM-1-producing *P. aeruginosa* isolates [13].

Our WGS data suggest a high relatedness of the different carbapenemase-producing strains, which point to one or multiple common origins, *e.g.*, in the field hospital, where the patients had received emergency medical care before being transferred abroad. Our investigation is limited by the absence of environmental samples from the field hospital so that the exact transmission pathways cannot be reconstructed.

## Conclusion

There is an urgent need for a thorough surveillance of multiresistant gram-negative bacteria in patients from Ukraine with war-related wounds in Europe and elsewhere. These pathogens should be subjected to a broad antimicrobial susceptibility testing because previously unknown rates of resistance to 'last-line' and novel antibiotics are to be expected.

## Declaration of competing interest

Fabian K. Berger has received consultant fees from MSD and Pfizer (pertaining both to *Clostridioides difficile*). B.C. Gärtner has received honoraria from Pfizer, outside the submitted work. S.L. Becker has received speaker fees and advisory board participation fees from Pfizer (pertaining to ceftazidime/avibactam) and Shionogi (pertaining to cefiderocol). All other authors have no competing interests to declare.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Ethical statement

Written informed consent was obtained from the patients reported in this manuscript.

## Author contributions

Patient treatment: TF, NV, VP, TP. Microbiological diagnostics: FKB, FA, SR, SS, BCG, SLB. WGS and data analysis: GPS, JR, AK. Drafting the manuscript: FKB, GPS, JR, SLB. All authors have read and approved the final version of the manuscript.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijid.2023.04.394.

## References

[1] Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 2022;**399**:629–55. doi:10.1016/S0140-6736(21)02724-0.

[2] World Health Organization. *WHO Response to the Ukraine crisis: annual report, 2022*. Geneva: World Health Organization; 2022.

[3] Roth S, Berger FK, Link A, et al. Application and clinical impact of the RESIST-4 O.K.N.V. rapid diagnostic test for carbapenemase detection in blood cultures and clinical samples. *Eur J Clin Microbiol Infect Dis* 2021;**40**:423–8. doi:10.1007/s10096-020-04021-4.

[4] Beauté J, Kramarz P. Public health surveillance in countries hosting displaced people from Ukraine. *Euro Surveill* 2022;**27**:2200430. doi:10.2807/1560-7917.ES.2022.27.22.2200430.

[5] Rzymski P, Falfushynska H, Fal A. Vaccination of Ukrainian refugees: need for urgent action. *Clin Infect Dis* 2022;**75**:1103–8. doi:10.1093/cid/ciac276.

[6] Vignier N, Halley des Fontaines V, Billette de Villemeur A, et al. Public health issues and health rendezvous for migrants from conflict zones in Ukraine: a French practice guideline. *Infect Dis Now* 2022;**52**:193–201. doi:10.1016/j.idnow.2022.04.006.

[7] Melwani M. How war is spreading drug resistant superbugs across Ukraine and beyond. *BMJ* 2022;**379**:o2731. doi:10.1136/bmj.o2731.

[8] Sandfort M, Hans JB, Fischer MA, et al. Increase in NDM-1 and NDM-1/OXA-48-producing *Klebsiella pneumoniae* in Germany associated with the war in Ukraine, 2022. *Euro Surveill* 2022;**27**:2200926. doi:10.2807/1560-7917.ES.2022.27.50.2200926.

[9] Pfennigwerth N, Schauer J. Bericht des Nationalen Referenzzentrums für Gram-negative Krankenhauserreger – Zeitraum 1. Januar 2021 bis 31. Dezember 2021. *Epid Bull* 2022;**19**:3–9 [in German].

[10] Paul M, Carrara E, Retamar P, et al. European Society of Clinical Microbiology and Infectious Diseases (ESCMID) guidelines for the treatment of infections caused by multidrug-resistant Gram-negative bacilli (endorsed by European Society of Intensive Care Medicine). *Clin Microbiol Infect* 2022;**28**:521–47. doi:10.1016/j.cmi.2021.11.025.

[11] Tamma PD, Aitken SL, Bonomo RA, Mathers AJ, van Duin D, Clancy CJ. Infectious Diseases Society of America 2022 Guidance on the Treatment of Extended-Spectrum beta-lactamase Producing Enterobacterales (ESBL-E), Carbapenem-Resistant Enterobacterales (CRE), and *Pseudomonas aeruginosa* with Difficult-to-Treat Resistance (DTR-P. aeruginosa). *Clin Infect Dis* 2021;**72**:e169–83. doi:10.1093/cid/ciac268.

[12] Simner PJ, Patel R. Cefiderocol antimicrobial susceptibility testing considerations: the Achilles' heel of the Trojan horse? *J Clin Microbiol* 2020;**59** e00951–20. doi:10.1128/JCM.00951-20.

[13] Mauri C, Maraolo AE, Di Bella S, Luzzaro F, Principe L. The revival of aztreonam in combination with avibactam against metallo-beta-lactamase-producing gram-negatives: a systematic review of *in vitro* studies and clinical cases. *Antibiotics (Basel)* 2021;**10**:1012. doi:10.3390/antibiotics10081012.

## 3.5 *BusyBee Web: towards comprehensive and differential composition-based metagenomic binning*

*W132–W137 Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 30 April 2022*
*https://doi.org/10.1093/nar/gkac298*

# BusyBee Web: towards comprehensive and differential composition-based metagenomic binning

**Georges P. Schmartz[1], Pascal Hirsch[1,2], Jérémy Amand[1,2], Jan Dastbaz[3,4], Tobias Fehlmann[1], Fabian Kern[1,2], Rolf Müller[3,4] and Andreas Keller[1,2,*]**

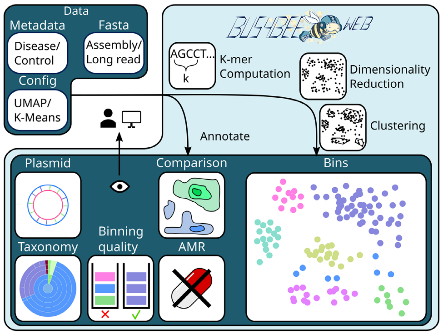
[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Clinical Bioinformatics (CLIB), Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research, 66123 Saarbrücken, Germany, [3]Microbial Natural Products (MINS), Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research, 66123 Saarbrücken, Germany and [4]Deutsches Zentrum für Infektionsforschung (DZIF), Standort Hannover-Braunschweig, 38124 Braunschweig, Germany

Received March 08, 2022; Revised April 07, 2022; Editorial Decision April 12, 2022; Accepted April 14, 2022

## ABSTRACT

**Despite recent methodology and reference database improvements for taxonomic profiling tools, metagenomic assembly and genomic binning remain important pillars of metagenomic analysis workflows. In case reference information is lacking, genomic binning is considered to be a state-of-the-art method in mixed culture metagenomic data analysis. In this light, our previously published tool BusyBee Web implements a composition-based binning method efficient enough to function as a rapid online utility. Handling assembled contigs and long nanopore generated reads alike, the webserver provides a wide range of supplementary annotations and visualizations. Half a decade after the initial publication, we revisited existing functionality, added comprehensive visualizations, and increased the number of data analysis customization options for further experimentation. The webserver now allows for visualization-supported differential analysis of samples, which is computationally expensive and typically only performed in coverage-based binning methods. Further, users may now optionally check their uploaded samples for plasmid sequences using PLSDB as a reference database. Lastly, a new application programming interface with a supporting python package was implemented, to allow power users fully automated access to the resource and integration into existing workflows. The webserver is freely available under: https://www.ccb.uni-saarland.de/busybee.**

## GRAPHICAL ABSTRACT



## INTRODUCTION

State-of-the-art metagenomics data analysis predominantly depends on reference databases. Reads are compared against well-characterized sequences and in case of sufficient sequence similarity, a read may be assigned to a taxonomy, an associated operational taxonomic unit count is incremented, or a genomic function is deduced (1–4). However, metagenomic studies operating at the boundary of what is known to humankind, e.g. investigating extreme maritime or volcanic environments, will inevitably come to the point where reference data is incomplete or of insufficient quality (5–7). While the overall possibilities for analysis are limited, a lack of reference information does not necessarily prevent any analysis. Instead, metagenomic short-read assembly or long-read metagenomic sequencing is frequently performed to allow for further hypothesizing, analysis, and discovery. However, due to high species diversity, sequencing errors, and other conflicts during assembly, metagenomic assemblies frequently yield multiple thousands of contigs of variable lengths and qualities (8,9).

[*]To whom correspondence should be addressed. Tel: +49 681 30268611; Fax: +49 681 30268610; Email: andreas.keller@ccb.uni-saarland.de

Since short-read metagenomic read assembly and long-read metagenome sequencing output a mix of sequences of all the present species, structured analysis of results remains difficult. Therefore, longer sequences are usually grouped using binning methods to separate sequences into taxonomic units. Two features are frequently used to achieve informed separation into groups. Coverage-based binning uses coverage profiles of sequences, computed across multiple samples, to cluster into bins. Composition-based binning utilizes the conservation of sequence features like tetranucleotide profiles and derives bins from the input sequence (10). Many of the state-of-the-art binning methods such as MaxBin2 are hybrid methods using both kinds of features (11–13). However, coverage profiles provide limited information if only one individual sample is analyzed, and they may even be not applicable depending on the selected sequencing method. Accordingly, new methods that do not require coverage profiles are further developed (14,15).

In 2018, we proposed BusyBee Web as a reference-free composition-based binning tool efficient enough to function as a webserver (16,17). The underlying pipeline trains a classifier on a subset of the input data which is then used to assign sequences into bins. The used features are normalized k-mer profiles of length four or five. The tool optionally provides various functional and taxonomic annotations with Prokka and Kraken respectively allowing for taxonomic binning (18,19). Five years after initial publication, the community used BusyBee Web to analyze >2500 individual samples and perform >4500 runs. Here, we present a major update to the binning resource.

## MATERIALS AND METHODS

Developing an update to an existing resource allowed us to revisit some of the already available functionality and cover a broad list of minor improvements. Accordingly, the taxonomic annotation was updated to support Kraken 2 with a newer database and marker genes for bin quality assessment were extended to include the Archaea genes from the anvi'o project (20). Further, a sunburst plot was added and several new expert settings for clustering and embedding methods were implemented. Namely, we included t-SNE (21), Fit-SNE (22), UMAP (arXiv:1802.03426), PHATE (23) and TriMap (arXiv:1910.00204) as embedding and DBSCAN (24), HDBSCAN (25), *k*-means and spectral clustering (26) as new clustering methods. From the list of new features, we want to highlight three major changes with higher visibility to newer users.

### Plasmids annotations

Due to the random sampling involved in shotgun sequencing experiments, metagenomic data often includes plasmid fragments that may also end up in assemblies, potentially impacting downstream analysis. BusyBee Web now optionally compares input sequences to the most recent version of PLSDB using mash screen (27,28). In case plasmid signatures are found, the most relevant information about the plasmids is displayed. From here, users can take a deeper look into the findings by continuing their analysis on PLSDB.

### Comparative metagenomics

Group comparison is a frequently requested analysis that is often neglected in composition-based methods. In BusyBee Web, we compute a differential density between two user-defined classes, by first applying a Gaussian 2D kernel to the embedded sequences for both classes separately. Bandwidth and grid size used in the computation can be modified by the user, within given boundaries. Next, the difference between both densities is visualized. This usually results in a picture where various areas are dominated by different classes. While this method does not directly provide statistics on coverage differences, it remains indicative of different phenomena. On the one hand, if long reads are directly embedded, higher density regions should represent a higher relative number of sequences with a similar k-mer spectrum in the sample. On the other hand, if assembled contigs were provided, interpretation becomes more complex. First, the number of embedded sequences is expected to increase simply due to technological errors, resulting in higher density regions for higher sequence counts similar to the long-read interpretation. Second, increased phylogenetic diversity is captured since identical sequences should ideally be collapsed already during assembly. The difference in density can be retrieved for each cluster allowing the user to further analyze potentially interesting patterns and areas.

### Application programming interface

To allow programmatic access to BusyBee Web, we implemented an application programming interface (API). The API complies with the Open API 3.0.2 standard (29). Users can start jobs, check their status, and download individual results over the API. Additionally, a python package is supported and distributed via conda, which allows for easy integration into R scripts using reticulate. The package is available on: https://github.com/CCB-SB/busybee_api.

### Case studies

In order to benchmark BusyBee Web on a mock community in the first case study, we downloaded the *ERR3152364* dataset from the sequence read archive and converted the fastq files into fasta files while also adapting the header names. Due to the high sequencing depth of the experiment, the sample had to be pruned to comply with the constraints imposed by the webserver. Thus, we shuffled the fasta file randomly and selected the first 200 Mb of data, corresponding exactly to the upload limit and which accounts for <2% of the initial file. The resulting file contained a total of 50 679 reads. After data generation, we started analysis with default parameters changing only the embedding to UMAP. For comparison, various embeddings with different dimension reduction methods were computed (Supplementary Figure S1).

The second case study discussing differences between sequencing technologies was conducted with newly generated data. Both datasets were derived from the same 1mL of bile sample of a healthy human individual and DNA was extracted with the same QiAamp DNA Microbiome Kit allowing for comparison between technologies. Next-generation sequencing DNA libraries were prepared using the MGIEasy Universal DNA Library Prep set following the recommendations of the manufacturer. The DNBSEQ-G400 was used as short-read sequencing platform. Oxford nanopore sequencing was prepared with the SQK-LSK109 Ligation Sequencing kit before sequencing on an FLO-MIN106D flow cell in a MinION Mk1B. Basecalling was performed with Guppy v5.0.7. For both datasets, human-read contamination was removed by first running kneaddata v0.7.4, followed by sra-human-scrubber v1.0.2021_05_05 (1,30). After removal of human reads, the ONT fastq was converted to fasta and read names were shortened to generic header names. For the short-read sequencing data, reads were assembled to scaffolds with metaSPAdes v3.15.2 and scaffolds were retained (8). Before analysis with BusyBee Web, both datasets, short- and long-read, were combined and a mapping to the original fasta entries was generated. Next, data was passed to BusyBee Web with default settings, but selecting UMAP as embedding algorithm.

## RESULTS

With the increasing popularity of whole shotgun metagenome and long-read sequencing competing with amplicon sequencing, dedicated analysis of plasmids from metagenomics data is becoming increasingly tempting to the metagenomics community (31). However, shared sequences between chromosomes and plasmids, variable sizes, and a wide range of other factors render plasmid assembly from short reads an algorithmic challenging task often entailing high misassembly rates (31,32). Similarly, the prediction of both plasmid reads, and plasmid sequences remains an intensively debated field of research, also affecting long-read sequencing technology (33–38). Attributed to these difficulties, plasmid sequences frequently appear in binning inputs where they may be difficult to interpret. With the newly added plasmid annotation, BusyBee Web explicitly notifies the user about the presence of already known putative plasmid signatures. Further, the newly adopted differential density-based visualization allows for visual interpretation of similarity between aggregated samples. Since cohort and interventional studies comparing healthy against diseased patients, elderly against young, or different treatment conditions are increasingly performed in biomedical research, the field also faces an upsurge of comparative metagenomic studies. However, many of the conclusions drawn from cohort studies are either based on differential taxonomic counts or the functional aspect of sequences. In both cases, the comparison relies on reference information. One method to alleviate this constraint is to assess differential coverage profiles of binned sequences. However, similar to coverage-based binning, coverage profiles are required for this approach, which may not be available. Moreover, minor differences in binning outcomes may largely impact conclusions weakening the stability of this approach. The embedding followed by subsequent kernel application that we implemented alleviates these drawbacks and the volatility of results is bound to the characteristics of the selected dimensionality reduction method. Lastly, with the added application programming interface (API) BusyBee Web can easily be integrated into new and existing data analysis pipelines. In combination with workflow managing tools such as Nextflow or Snakemake, the API increases experiment throughput and reproducibility of results (39,40).

In order to highlight the improved functionality of BusyBee Web, we analyzed two datasets of varying ground truth information. While at the core BusyBee uses a reference-free algorithm for binning, here we make use of reference-based taxonomic annotations that were added after binning. Combining these annotations with the knowledge of well-characterized microbial environments allows us to better gauge binning quality.

### Mock community benchmark

To assess the binning quality of BusyBee Web on a well-characterized example, we used a dataset by Nicholls et al. as ground truth (41). This nanopore sequencing data represents a mock community composed of exactly ten known species. The output of BusyBee Web consists of 27 bins (Figure 1A and B). However, 14 of these bins each contained <1% of sequences and may be discarded from further analysis. Of the remaining 13 bins, five bins, namely 1, 5, 8, 22 and 24, were mostly composed of unclassified sequences. We postulate that these bins are mostly made up of *Cryptococcus neoformans* and *Saccharomyces cerevisiae* which are not included in the selected Kraken 2 database. The taxonomic composition of bin 9, which is the smallest remaining bin composing only 515 sequences, is highly fractioned indicating a low binning quality. While not exempt from cross-contamination, all the remaining bins (2, 11, 13, 16, 17, 21 and 27) can clearly be attributed to the distinct species from the mock community, indicating that despite only using a fraction of the input, BusyBee Web is able to successfully recover the contained major species.

### Sequencing technology comparison

To highlight the new analysis functionality added in this update, we compared the suitability of long reads with short-read assembled scaffolds. With the 19,262 input sequences passing the default length filter a total of 19 bins were predicted of which five (5, 7, 8, 15 and 19) contained <1% of sequences. A total of 340 sequence similarities to potentially relevant plasmids were identified where the majority was reported in *Enterococcus faecium*. Looking at the new differential density plot from Figure 1C we observe six clusters (1, 2, 4, 6, 16 and 17) that are specific to the short-read sequencing experiment. The taxonomic profile of cluster 1 has a high relative number of unclassified sequences, pointing towards potentially unreliable assemblies (Figure 1D). Nevertheless, we note that within this bin a few long reads were found at a relative proportion of ∼15.5%. Four of the remaining five clusters (2, 4, 16 and 17) have low contaminations. These four clusters presumably consist mostly of Lachnospiraceae, *Enterobacteriaceae*, *Actinomycetaceae* and *Micrococcaceae* respectively. Potentially, due to biological random sampling or decreased sequencing depth, these genomic signatures mostly escaped the nanopore sequencing.

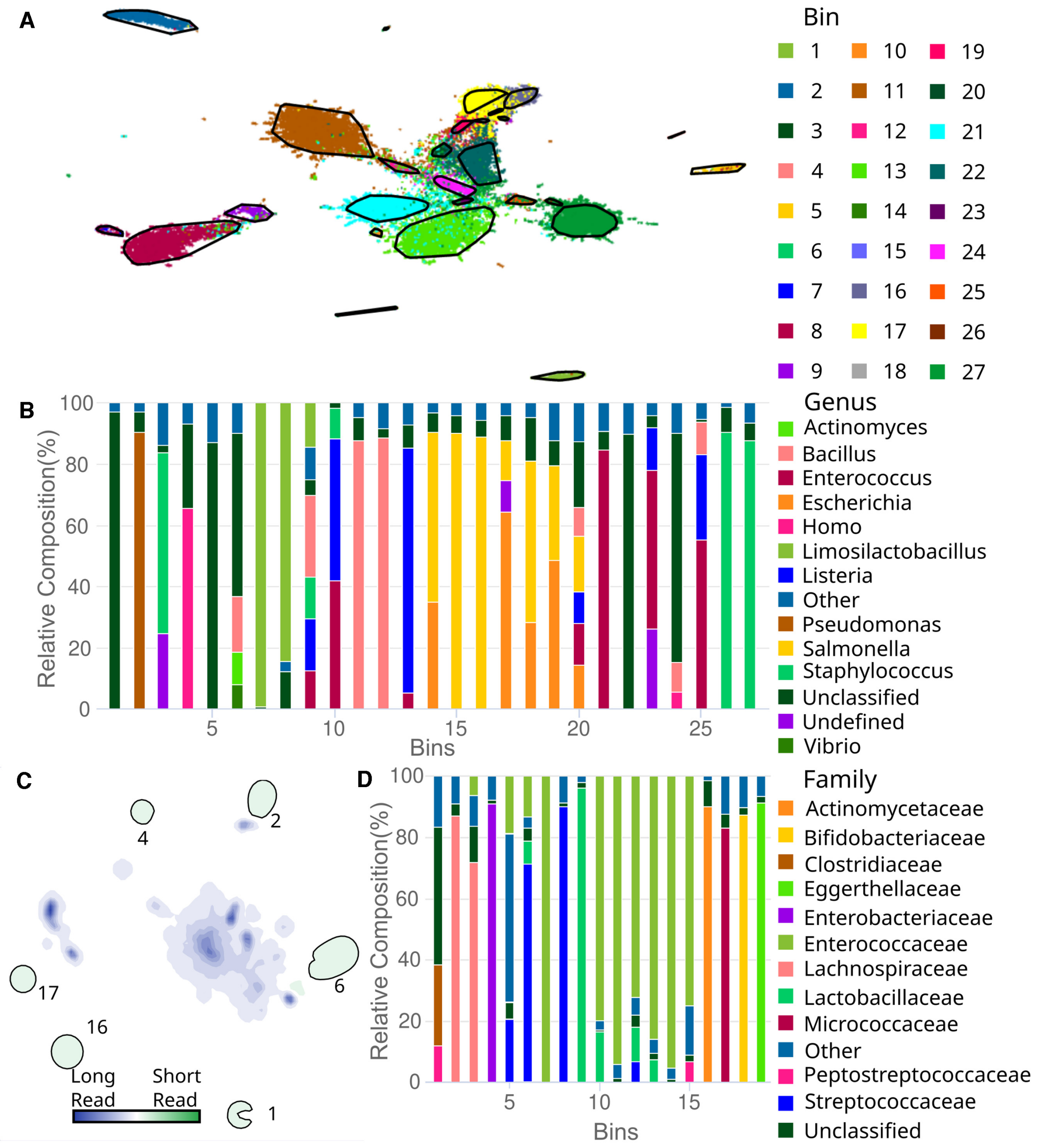


**Figure 1.** (**A**) Embedding of the mock community dataset, using UMAP with default settings. (**B**) Taxonomic profile at genus level of the different bins computed on a mock community composed of ten different species. (**C**) Differential density embedding of a bile sample sequenced with Oxford Nanopore MinION (Long Read) and DNBSEQ-400 (Short Read) respectively. (**D**) Taxonomic annotation of bins computed on the comparison dataset.

## CONCLUSION

With the new update, we substantially extended the capabilities of BusyBee Web as a versatile composition-based binning tool. On the one hand, with the newly added clustering methods, embedding algorithms, and API, we increased the data analysis possibilities for expert users. On the other hand, we hope to widen our user base by providing new visualizations and annotations. While we always strive for maximal flexibility, the ease of use of BusyBee Web as an installation-free webservice comes at a cost. For example, the data upload is limited to 200Mb per sample which can quickly be reached if multiple samples are being analyzed. Moreover, some of the presented clustering and embedding options will not be able to handle the theoretical maximal number of contigs that fit into a 200Mb file, due to time and memory constraints. Therefore, BusyBee Web provides an option for compressing information before embedding computation, alleviating some of these limitations. Nonetheless, visualization of the embedding in the local browser for many data points may become slow or irresponsive on less powerful hardware. Here, we recommend to prefer API usage instead. Moreover, with sufficient coverage information available, state-of-the-art coverage-based and hybrid metagenomic binning tools are expected to outperform composition-based tools on short-read sequencing data in larger projects.

Potential future development efforts may further focus on the identification of mobile genetic elements. However, with large disagreements already observed across plasmid classification tools, potential counter-strategies, e.g. automated removal of putative sequences from user input, are likely unstable and thus currently not advisable. Further, by extending the BusyBee Web server to allow for a selection of different embedding and clustering methods, it will be easier in the future to integrate newer algorithms into the generalized framework.

## DATA AVAILABILITY

BusyBee Web is freely available at: https://www.ccb.uni-saarland.de/busybee.

## ACCESSION NUMBERS

Respecting the German federal privacy law, we uploaded the short- and long-read data after human read removal to the Sequence Read Archive. Preprocessed data can be found in NCBI SRA using the accession numbers SRX14022915 and SRX14435297.

The mock community dataset was made available by Nicholls et al. in the Sequence Read Archive under the accession: ERR3152364.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

This work was supported by the UdS-HIPS TANDEM Initiative. We thank Valentina Galata for fruitful discussions on potential case studies.

G.P.S.: Software, Formal Analysis, Writing - Original Draft, Visualization; P.H.: Software, Writing - Review & Editing; J.A.: Methodology, Software, Writing - Review & Editing; J.D.: Investigation, Data Curation; F.K.: Writing - Review & Editing, Supervision; T.F.: Software, Supervision; R.M.: Conceptualization, Resources, Funding acquisition; A.K.: Conceptualization, Resources, Supervision, Project administration, Funding acquisition.

## FUNDING

Funding for open access charge: Saarland University.

*Conflict of interest statement.* G.P.S., R.M., and A. K. are shareholders of MOOH GmbH.

## REFERENCES

1. Beghini,F., McIver,L.J., Blanco-Miguez,A., Dubois,L., Asnicar,F., Maharjan,S., Mailyan,A., Manghi,P., Scholz,M., Thomas,A.M. *et al.* (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife*, **10**, e65088.
2. Wood,D.E., Lu,J. and Langmead,B. (2019) Improved metagenomic analysis with kraken 2. *Genome Biol.*, **20**, 257.
3. Bharti,R. and Grimm,D.G. (2021) Current challenges and best-practice protocols for microbiome analysis. *Brief. Bioinform*, **22**, 178–193.
4. Milanese,A., Mende,D.R., Paoli,L., Salazar,G., Ruscheweyh,H.J., Cuenca,M., Hingamp,P., Alves,R., Costea,P.I., Coelho,L.P. *et al.* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.*, **10**, 1014.
5. Spieck,E., Spohn,M., Wendt,K., Bock,E., Shively,J., Frank,J., Indenbirken,D., Alawi,M., Lucker,S. and Hupeden,J. (2020) Extremophilic nitrite-oxidizing chloroflexi from yellowstone hot springs. *ISME J.*, **14**, 364–379.
6. Wibowo,M.C., Yang,Z., Borry,M., Hubner,A., Huang,K.D., Tierney,B.T., Zimmerman,S., Barajas-Olmos,F., Contreras-Cubas,C., Garcia-Ortiz,H. *et al.* (2021) Reconstruction of ancient microbial genomes from the human gut. *Nature*, **594**, 234–239.
7. Almeida,A., Nayfach,S., Boland,M., Strozzi,F., Beracochea,M., Shi,Z.J., Pollard,K.S., Sakharova,E., Parks,D.H., Hugenholtz,P. *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, **39**, 105–114.
8. Nurk,S., Meleshko,D., Korobeynikov,A. and Pevzner,P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, **27**, 824–834.
9. Li,D., Liu,C.M., Luo,R., Sadakane,K. and Lam,T.W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, **31**, 1674–1676.
10. Teeling,H., Meyerdierks,A., Bauer,M., Amann,R. and Glockner,F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, **6**, 938–947.
11. Kang,D.D., Li,F., Kirton,E., Thomas,A., Egan,R., An,H. and Wang,Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, **7**, e7359.
12. Mallawaarachchi,V., Wickramarachchi,A. and Lin,Y. (2020) GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics*, **36**, 3307–3313.
13. Wu,Y.W., Simmons,B.A. and Singer,S.W. (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, **32**, 605–607.
14. Wickramarachchi,A. and Lin,Y. (2021) *21st International Workshop on Algorithms in Bioinformatics (WABI 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
15. Wickramarachchi,A., Mallawaarachchi,V., Rajan,V. and Lin,Y. (2020) MetaBCC-LR: metagenomics binning by coverage and composition for long reads. *Bioinformatics*, **36**, i3–i11.
16. Laczny,C.C., Kiefer,C., Galata,V., Fehlmann,T., Backes,C. and Keller,A. (2017) BusyBee web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res.*, **45**, W171–W179.
17. Benson,G. (2017) Editorial: the 15th annual nucleic acids research web server issue 2017. *Nucleic Acids Res.*, **45**, W1–W5.
18. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
19. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
20. Eren,A.M., Esen,O.C., Quince,C., Vineis,J.H., Morrison,H.G., Sogin,M.L. and Delmont,T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, **3**, e1319.
21. Van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
22. Linderman,G.C., Rachh,M., Hoskins,J.G., Steinerberger,S. and Kluger,Y. (2019) Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods*, **16**, 243–245.
23. Moon,K.R., van Dijk,D., Wang,Z., Gigante,S., Burkhardt,D.B., Chen,W.S., Yim,K., Elzen,A.V.D., Hirn,M.J., Coifman,R.R. *et al.*

(2019) Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.*, **37**, 1482–1492.

24. Xu,X., Ester,M., Kriegel,H.-P. and Sander,J. (1998), *Proceedings 14th International Conference on Data Engineering*. IEEE, pp. 324–331.
25. Campello,R.J., Moulavi,D. and Sander,J. (2013), *Pacific-Asia conference on knowledge discovery and data mining*. Springer, pp. 160–172.
26. Von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.
27. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol.*, **17**, 132.
28. Schmartz,G.P., Hartung,A., Hirsch,P., Kern,F., Fehlmann,T., Muller,R. and Keller,A. (2022) PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res.*, **50**, D273–D278.
29. Tarkowska,A., Carvalho-Silva,D., Cook,C.E., Turner,E., Finn,R.D. and Yates,A.D. (2018) Eleven quick tips to build a usable REST API for life sciences. *PLoS Comput. Biol.*, **14**, e1006542.
30. Katz,K.S., Shutov,O., Lapoint,R., Kimelman,M., Brister,J.R. and O'Sullivan,C. (2021) STAT: a fast, scalable, minhash-based k-mer tool to assess sequence read archive next-generation sequence submissions. *Genome Biol.*, **22**, 270.
31. Antipov,D., Raiko,M., Lapidus,A. and Pevzner,P.A. (2019) Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.*, **29**, 961–968.
32. Pellow,D., Zorea,A., Probst,M., Furman,O., Segal,A., Mizrahi,I. and Shamir,R. (2021) SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome*, **9**, 144.
33. Krawczyk,P.S., Lipinski,L. and Dziembowski,A. (2018) PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.*, **46**, e35.
34. Laczny,C.C., Galata,V., Plum,A., Posch,A.E. and Keller,A. (2019) Assessing the heterogeneity of in silico plasmid predictions based on whole-genome-sequenced clinical isolates. *Brief Bioinform*, **20**, 857–865.
35. Pellow,D., Mizrahi,I. and Shamir,R. (2020) PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.*, **16**, e1007781.
36. Pradier,L., Tissot,T., Fiston-Lavier,A.S. and Bedhomme,S. (2021) PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC Bioinf.*, **22**, 349.
37. Wickramarachchi,A. and Lin,Y. (2022) GraphPlas: refined classification of plasmid sequences using assembly graphs. *IEEE/ACM Trans. Comput. Biol. Bioinform*, **19**, 57–67.
38. Zhou,F. and Xu,Y. (2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, **26**, 2051–2052.
39. Di Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
40. Molder,F., Jablonski,K.P., Letcher,B., Hall,M.B., Tomkins-Tinch,C.H., Sochat,V., Forster,J., Lee,S., Twardziok,S.O., Kanitz,A. *et al.* (2021) Sustainable data analysis with snakemake. *F1000Res.*, **10**, 33.
41. Odahara,M., Nakamura,K., Sekine,Y. and Oshima,T. (2021) Ultra-deep sequencing reveals dramatic alteration of organellar genomes in physcomitrella patens due to biased asymmetric recombination. *Commun. Biol.*, **4**, 633.

## 3.6 PLSDB: advancing a comprehensive database of bacterial plasmids

*Published online 25 November 2021*

*Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D273–D278***
*https://doi.org/10.1093/nar/gkab1111*

# PLSDB: advancing a comprehensive database of bacterial plasmids

**Georges P. Schmartz[1], Anna Hartung[1], Pascal Hirsch[1,2], Fabian Kern[1], Tobias Fehlmann[1], Rolf Müller[2,3] and Andreas Keller[1,2,*]**

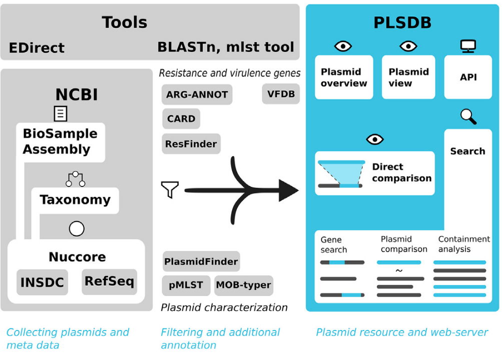
[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Department of Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), Campus E8 1, 66123 Saarbrücken, Germany and [3]Department of Pharmacy, Saarland University, 66123 Saarbrücken, Germany

Received September 15, 2021; Revised October 20, 2021; Editorial Decision October 21, 2021; Accepted October 22, 2021

## ABSTRACT

**Plasmids are known to contain genes encoding for virulence factors and antibiotic resistance mechanisms. Their relevance in metagenomic data processing is steadily growing. However, with the increasing popularity and scale of metagenomics experiments, the number of reported plasmids is rapidly growing as well, amassing a considerable number of false positives due to undetected misassembles. Here, our previously published database PLSDB provides a reliable resource for researchers to quickly compare their sequences against selected and annotated previous findings. Within two years, the size of this resource has more than doubled from the initial 13,789 to now 34,513 entries over the course of eight regular data updates. For this update, we aggregated community feedback for major changes to the database featuring new analysis functionality as well as performance, quality, and accessibility improvements. New filtering steps, annotations, and preprocessing of existing records improve the quality of the provided data. Additionally, new features implemented in the web-server ease user interaction and allow for a deeper understanding of custom uploaded sequences, by visualizing similarity information. Lastly, an application programming interface was implemented along with a python library, to allow remote database queries in automated workflows. The latest release of PLSDB is freely accessible under https://www.ccb.uni-saarland.de/plsdb.**

## GRAPHICAL ABSTRACT



## INTRODUCTION

Plasmids are extrachromosomal DNA sequences that are short in comparison to chromosomes and frequently found in circular form within prokaryotes. They can harbor a wide range of genes such as antibiotic resistance and virulence factors (1,2). Due to the appearance of such clinically relevant phenotypes, the analysis of plasmid sequences is widely acknowledged and often performed in the context of microbiome sequencing studies (3,4). On the one hand, associative connections between clinical conditions and plasmids may allow untangling specific disease and treatment patterns. On the other hand, plasmid research furthermore plays a significant role on a population level (5). Due to several mechanisms, e.g., horizontal gene transfer via conjugation, antibiotic resistance may spread calling for a readjustment of focus in pharmaceutical research on new innovative antibiotics (6). However, to allow monitoring global distributions of plasmids within populations, a general-purpose database is required, providing easy access to previously reported plasmids. Here, PLSDB (7) supports researchers with an easy-to-use web interface since 2018.

[*]To whom correspondence should be addressed. Tel: +49 681 30268611; Fax: +49 681 30268610; Email: andreas.keller@ccb.uni-saarland.de

The original PLSDB was created to complement NCBI's plasmid collection on RefSeq, which is partially incomplete, inconsistent, lacking in functionality, and contains several chromosomal sequences. PLSDB gathers data from NCBI & INSDC based on the query formulated by Orlek A et al. (8) and adds further filtering and annotation steps. The filtering hereby focuses on deduplication, Mash distances (9), and identification of putative chromosomal sequences using 53rps genes from PubMLST (10). The additional annotations consist of resistance and virulence factors from ARG-ANNOT (11), CARD (12), ResFinder (13) and VFDB (14). Apart from the dataset, PLSDB also provides a web interface to present the data in a simple but powerful manner. A core function of PLSDB is to allow users to upload their own sequences and compare them to the database contents, thereby selecting from established search methods such as Mash (9) or blastn (15).

With the rising popularity of whole metagenome shotgun sequencing slowly superseding 16S rRNA sequencing, more plasmids are getting discovered. Furthermore, dedicated algorithms for plasmid extraction from short read sequencing are gaining attention allowing for more efficient automated analysis of sequencing data (16–18). Similarly, new databases emerge trying to manage and overcome the resulting flood of plasmid data. A new plasmid collection by Brooks et al., for example, tries to bundle NCBI plasmid information in a collection (19). Another recent database, mMGE (20) has the advantage of unifying phage and plasmid information in a single catalog. However, the database creation workflow is solely focused on the human microbiome. The COMPASS database (21), is of comparable scope to PLSDB and focuses extensively on replicon typing. Due to the scope, size, functionality, and quality of its content, PLSDB is widely used in the scientific community as a central resource for reference data on natural occurring plasmids. By focusing on this domain, the resource finds extensive usage in environmental studies (22). Further, antibiotic resistance analyses with a diverse scope profit frequently from the resistance annotations found in PLSDB (23,24). Based on aggregated feedback of this primary expert userbase, we conducted a major update. A sizable portion of the update focuses on improving future maintenance, aggregation, and quality of the data. Further, we saw this as an opportunity to implement various new utilities into the online resource (Graphical Abstract).

## MATERIALS AND METHODS

The updated version of PLSDB provides easier access to sequence files, additional visualizations, further data export options and other improvements. We want to highlight two major changes for user interaction and two changes impacting the contents of the database.

### Plasmid data collection

PLSDB is prohibitively large to manually curate and is steadily growing. Based on frequent-user feedback, two additional filtering rules were added to satisfy new community demands. The first rule is a simple threshold cutoff constraining the minimal size of a sequence to be considered a plasmid. This is necessary since the data retrieval rules set by Orlek et al. require sequences to be complete, yet smaller sequences not surpassing this threshold were observed to indicate incorrect labeling. The second rule was implemented to address incomplete assemblies and is a result of the computation of the Mash distance. The underlying designation of the Jaccard index as in Mash computation does not aim to test inclusion properties. Even in the case of perfectly covered subsequences, the set similarity threshold may not be reached leading to a retainment of both sequences. In order to address this issue, another filtering step was set in place specially focused to capture these hierarchical relationships. To this end, a blastn search querying for exact matches between plasmid pairs of the same Biosample or the same Nucleotide database description was implemented. Hereby, we considered that plasmid matches may split into two perfect matches, due to the linearization of circular plasmids in file formats. In addition to the new filtering rules, new annotations were integrated into the data collection pipeline. First, disease information from the BioSample database is now supported. Second, MOB-typer (18) has been added to annotate mobility families and mating pair formation classes.

### Annotation preprocessing

While the NCBI BioSample database offers highly relevant additional information for data analysis, it is widely accepted that the quality of the annotated meta data is lacking in several ways (25). This is because the database does not constrain meta data input upon submission. While this simplifies and encourages data upload, it complicates data analysis for users (26). Data preprocessing fixing typographical errors, annotations running under an incorrect header, incorrectly typed values, etc. is a time-consuming procedure. Yet, skipping it may negatively impact downstream analysis (25). Accordingly, a first step users often had to take when trying to make further use of the meta data of PLSDB, was to clean it up. In this update, we integrated a part of this process into our data collection pipeline, to shift some of that workload away from our users. With this goal in mind, additional processed annotations were added, while also leaving the original meta data intact. The first column we support in the new workflow indicates the host of the biosample. Here, we try to link the entry to a valid NCBI taxonomy entry. If an entry has already been resolved, we recycle the mapping. If this is not the case, we first split the text at any common separating characters, and then query the individual components in the NCBI taxonomy browser using the taxize (27) R package. If an assignment was uniquely mapped to a taxonomy, we consider it to be processed. For each biosample, we start from the host description and proceed with the isolation source column in case no result is found. The second meta data type we process indicates a potential disease of the host. As reference terminology, we use the Disease Ontology (28). Here, we start again by removing various separating characters, numerals, and stop words in the ontology terms. Afterward, we compute the case insensitive complete Levenshtein distance ratio between all terms and the query using the fuzzywuzzy python package (https://github.com/seatgeek/fuzzywuzzy) and keep the best match. Once the result surpasses a given threshold, we keep the result, for which we found a threshold of 80 to work well. In case the threshold is not met, we compute the token set ratio instead and retain the best match surpassing a threshold of 95.

### Sequence comparison

PLSDB allows a user to compare their sequences against the database choosing among a variety of search strategies. While it is useful to filter and sort detected plasmids, the displayed similarity scores may be perceived as abstract and unintuitive. In case a deeper understanding of the results had been desired, downloading of sequences and a manual investigation was previously necessary. As to improve user experience, PLSDB now allows users to visualize similarities directly and interactively in the web interface as a bipartite graph. To this end, a blastn search is run and visualization is made using Kablammo (29). Kablammo allows for filtering of blastn results to display only the most relevant similarities by adjusting blastn cutoff values. The same view can also be used to compare two selected plasmids included in the database. Further, a tblastn search is used if a user uploads protein sequences.

### Application programming interface

In the age of massively parallel sequencing automation and reproducibility is key in bioinformatics. Accordingly, PLSDB now provides an application programming interface (API). The main functionalities are focused on data retrieval and automated sequence search. A user may download information for a specific plasmid, filter the entire database for relevant subsets, or get sequence information. Users can upload their sequences through the API and remotely search for sequence similarities in PLSDB.

The API complies with OpenAPI guidelines and is further extended by an open-source python wrapper for straightforward integration into custom workflows (30). Similarly, a wrapper based on reticulate allows portability to R applications. For data analyses exceeding the throughput of the API, the open data policy of PLSDB offers the user to freely download the entire database as well as any matching sequence information.

## RESULTS

### Content summary

Since the release of PLSDB, the number of contained sequences has been continuously growing by ∼250% from the initial 13 789 to now 34 513 entries (Figure 1A). The current update consists of 350 Mb in sequence information. Focusing only on those sequences where geolocation information is available, substantial portions of the data come from China (21%), USA (21%) and the UK (8%). On the South American and African continents, Mexico and Egypt provide most of entries with 1% and less than 1%, respectively. Further, the sequences are not only sampled unevenly from a geolocational perspective, but also on a taxonomic level (Figure 1B). The phylum with most entries in PLSDB is *Proteobacteria* at 70%. Largely this is due to *Escherichia coli* species, making up ∼29% of all *Proteobacteria*, being the most predominantly represented species in the database, likely due to its model organism status. The second leading phylum, *Firmicutes,* has only a proportion of 30% compared to *Proteobacteria* with the most prevalent species being *Enterococcus faecium*. The largest relative growth compared to the first version is observed in *Actinobacteria.* Due to a low presence in the first version, relative growth is over 300-fold with a little more than one thousand entries in the current version. The least represented phyla in the current version are *Synergistetes* with one and *Chlorobi*, *Deferribacteres*, *Gemmatimonadetes*, and *Nitrospirae* with two plasmids, respectively. We observe on the gene level that the number of annotated sequences where genes involved in antimicrobial resistance are observed is growing faster than the overall number of plasmids (Figure 1C). The underlying cause for this observation may not necessarily be due to a spreading of antibiotic resistance genes. A confounding factor influencing this numerical growth might be attributed to a stronger focus of researchers on clinically relevant plasmids. In contrast to antibiotic resistances, virulence annotations decrease in relative frequency. Finally, manual analysis of the new annotation preprocessing feature shows the quality of processed annotations to be robust (Figure 1D). For the host disease meta data, available information is sparse with only 4368 entries. With our preprocessing pipeline, we were able to link 1795 entries to a valid Disease Ontology term. Considering host information, a total of 13 050 entries originally contained annotations. Taken together, we provide 12 877 processed terms, where 602 annotations derived information, despite an initially empty meta data field. Nevertheless, we note that there remain many plasmids where the automated annotation was not reliable enough to link either an ontology or taxonomy term. Here, no processed annotation is given, leaving users the choice to modify, transcribe, or drop original annotations. To gauge the overall quality of the database we assessed all sequences with various external tools used for differentiating sequences into plasmid- or chromosome-derived sequences. PlasClass (31), PlasFlow (32) and PlasForest (33) labeled 91%, 82% and 100% of sequences as plasmids respectively, indicating low contamination from e.g., chromosomal information. We do note that for PlasClass we used a threshold of 0.5 and that the tool was trained with an older version of PLSDB, likely biasing these results. To explore the completeness of the database, we compared contents to alternative databases with CD-HIT-EST-2D v4.8.1 (34). We found that 94% and 92% of sequences from the database of Brooks et al. and COMPASS, respectively were replicated in PLSDB at a sequence identity of 100%.

### Case example analysis

PLSDB is frequently used in a wide range of antimicrobial resistance-focused analyses as reference material (35). When observing antimicrobial resistance in a clinical scenario metagenomic sequencing and assembly of whole samples may not be desired due to cost or time constraints. Instead, it may be more interesting to narrow down the resistance to a few potential candidate plasmids responsible for the resistance, then to either validate it experimentally with PCR or adjust treatment (36). For demonstration pur-

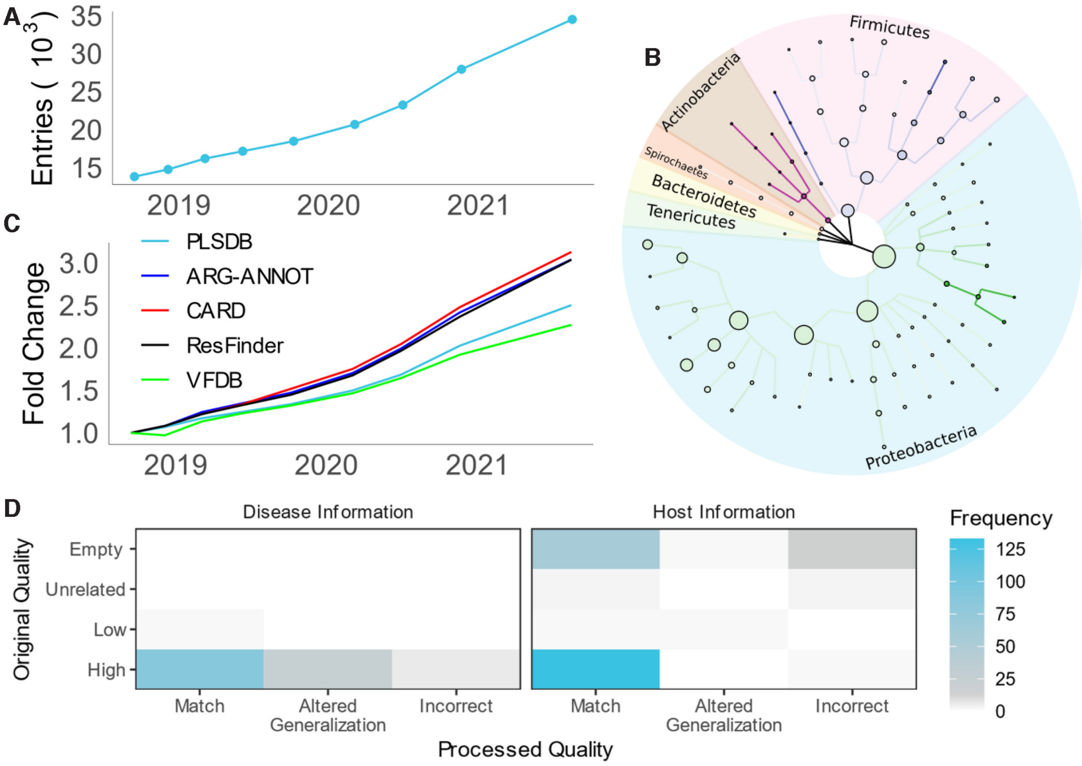


**Figure 1.** New Data of PLSDB: (**A**) Growth of the PLSDB data collection over time. (**B**) Taxonomic tree capturing the main composition of PLSDB in terms of quantity across several taxonomic ranks. Node size indicates frequency in the current database. Color fade represents relative growth compared to the first release of PLSDB. (**C**) Yearly growth of annotation data per source collection. Fold change is always computed with respect to the first release. (**D**) Manual validation of automatic preprocessing results. For each information type, annotations are compared before and after preprocessing. To generate the heatmap, all unique lowercase representatives of descriptions were extracted from the current version of PLSDB. Entries were then manually evaluated. For the preprocessed description, the comparison was drawn to the respective ontologies.

poses, we investigate a resistance qualification in *Staphylococcus aureus,* which is infamous for e.g., methicillin resistance (37). For the analysis, we search in PLSDB for available data by filtering for known plasmids in *Staphylococcus aureus*. Further, we may narrow down our findings by using the geolocational information available in PLSDB. Therefore, we filter for results originating from Germany. The number of potential plasmid candidates is now reasonably small to investigate individual plasmids. Since longer plasmids have better odds for containing an interesting gene, we sort by length and check plasmids iteratively for resistances using the plasmid overview of PLSDB. Using the newly added minimum spanning tree visualization, we may navigate among similar plasmids searching for a viable candidate, once a first interesting plasmid is identified. With few candidates selected, further investigation can be done by direct comparison. Following this general analysis gist for demonstration, we quickly identified two pivotal plasmids, NZ_CP022909.1 and NZ_CP022907.1, harboring several β-lactam resistance genes.

## CONCLUSION AND FUTURE DIRECTION

With this update, we aim to prepare PLSDB from a widely accepted data resource to a recognized reference database in the field of naturally occurring plasmids. The added features address both power and casual users alike and with them, we hope to invite more researchers into the analysis of plasmids and therein included clinically relevant resistances. Further, with the new processed meta data and content improvements to PLSDB, existing users will find important quality-of-life changes. At last, changes in the data gathering pipeline will allow us to provide regular content updates to the database for an extended amount of time.

Upcoming development efforts on the web-server will be invested into speeding up existing functionalities that may be affected by the rapid data growth such, as the search by sequence. Considering the database, future work is centered around the improvement of annotations and quality assurance of regularly scheduled data releases. Qualitatively, the goal is to further improve annotation processing by extending the natural language processing techniques with manual curation of the most frequent correction issues. Quantitatively, we aim to add more specific information if desired by the community while still balancing the user experience for new researchers entering the field. To facilitate and further automate the rolling out of future releases, we advocate for automated outlier detection in our data generation pipeline. We observed that more advanced methods were unable to provide a concise decision on plasmid classification. Yet, rule-based methods would already be able to signal suspicious annotations and potentially chromosome-contaminated sequences.

We will continuously adapt our web-server to the needs of the research community and provide accurate plasmid information in future data updates. We therefore encourage users to remain vocal on data quality and feature requests.

## DATA AVAILABILITY

The PLSDB web-server is freely accessible at: https://www.ccb.uni-saarland.de/plsdb. The entire data collection of PLSDB can be found on the website. The dedicated python package for API access is available on GitHub https://github.com/CCB-SB/plsdbapi. Finally, the data collection pipeline can be found on GitHub https://github.com/VGalata/plsdb where we are also welcoming any user feedback.

## ACKNOWLEDGEMENTS

We thank Nadja Grammes for her maintenance work on PLSDB. Further, we are grateful for the high-quality documentation by her and Valentina Galata on the architecture of PLSDB.

## FUNDING

Funding for open access charge: Saarland University and Helmholtz-Institute for Pharmaceutical Research Saarland.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Botelho,J. and Schulenburg,H. (2021) The role of integrative and conjugative elements in antibiotic resistance evolution. *Trends Microbiol.*, **29**, 8–18.
2. Yang,X., Dong,N., Chan,E.W., Zhang,R. and Chen,S. (2021) Carbapenem resistance-encoding and virulence-encoding conjugative plasmids in *Klebsiella pneumoniae*. *Trends Microbiol.*, **29**, 65–83.
3. Gancz,A., Kondratyeva,K., Cohen-Eli,D. and Navon-Venezia,S. (2021) Genomics and virulence of *Klebsiella pneumoniae* Kpnu95 ST1412 harboring a novel incf plasmid encoding Blactx-M-15 and Qnrs1 causing community urinary tract infection. *Microorganisms*, **9**, 1022.
4. Peter,S., Bosio,M., Gross,C., Bezdan,D., Gutierrez,J., Oberhettinger,P., Liese,J., Vogel,W., Dorfel,D., Berger,L. *et al.* (2020) Tracking of antibiotic resistance transfer and rapid plasmid evolution in a hospital setting by nanopore sequencing. *mSphere*, **5**, e00525-20.
5. Lerminiaux,N.A. and Cameron,A.D.S. (2019) Horizontal transfer of antibiotic resistance genes in clinical environments. *Can. J. Microbiol.*, **65**, 34–44.
6. Miethke,M., Pieroni,M., Weber,T., Bronstrup,M., Hammann,P., Halby,L., Arimondo,P.B., Glaser,P., Aigle,B., Bode,H.B. *et al.* (2021) Towards the sustainable discovery and development of new antibiotics. *Nat. Rev. Chem.*, **5**, 726–749.
7. Galata,V., Fehlmann,T., Backes,C. and Keller,A. (2019) PLSDB: a resource of complete bacterial plasmids. *Nucleic. Acids. Res.*, **47**, D195–D202.
8. Orlek,A., Phan,H., Sheppard,A.E., Doumith,M., Ellington,M., Peto,T., Crook,D., Walker,A.S., Woodford,N., Anjum,M.F. *et al.* (2017) Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid*, **91**, 42–52.
9. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
10. Jolley,K.A., Bray,J.E. and Maiden,M.C.J. (2018) Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.*, **3**, 124.
11. Gupta,S.K., Padmanabhan,B.R., Diene,S.M., Lopez-Rojas,R., Kempf,M., Landraud,L. and Rolain,J.M. (2014) ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.*, **58**, 212–220.
12. Alcock,B.P., Raphenya,A.R., Lau,T.T.Y., Tsang,K.K., Bouchard,M., Edalatmand,A., Huynh,W., Nguyen,A.V., Cheng,A.A., Liu,S. *et al.* (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **48**, D517–D525.
13. Bortolaia,V., Kaas,R.S., Ruppe,E., Roberts,M.C., Schwarz,S., Cattoir,V., Philippon,A., Allesoe,R.L., Rebelo,A.R., Florensa,A.F. *et al.* (2020) ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.*, **75**, 3491–3500.
14. Keppler,D., Hagmann,W. and Denzlinger,C. (1987) Leukotrienes as mediators in endotoxin shock and tissue trauma. *Prog. Clin. Biol. Res.*, **236A**, 301–309.
15. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
16. Antipov,D., Hartwick,N., Shen,M., Raiko,M., Lapidus,A. and Pevzner,P.A. (2016) plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*, **32**, 3380–3387.
17. Pellow,D., Zorea,A., Probst,M., Furman,O., Segal,A., Mizrahi,I. and Shamir,R. (2021) SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome*, **9**, 144.
18. Robertson,J. and Nash,J.H.E.(2018) MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom*, **4**, e000206.
19. Brooks,L., Kaze,M. and Sistrom,M. (2019) A curated, comprehensive database of plasmid sequences. *Microbiol Resour Announc*, **8**, e01325-18.
20. Lai,S., Jia,L., Subramanian,B., Pan,S., Zhang,J., Dong,Y., Chen,W.H. and Zhao,X.M. (2021) mMGE: a database for human metagenomic extrachromosomal mobile genetic elements. *Nucleic Acids Res.*, **49**, D783–D791.
21. Douarre,P.E., Mallet,L., Radomski,N., Felten,A. and Mistou,M.Y. (2020) Analysis of COMPASS, a new comprehensive plasmid database revealed prevalence of multireplicon and extensive diversity of incf plasmids. *Front. Microbiol*, **11**, 483.
22. Bleichenbacher,S., Stevens,M.J.A., Zurfluh,K., Perreten,V., Endimiani,A., Stephan,R. and Nuesch-Inderbinen,M. (2020) Environmental dissemination of carbapenemase-producing Enterobacteriaceae in rivers in Switzerland. *Environ. Pollut.*, **265**, 115081.
23. Kieffer,N., Royer,G., Decousser,J.W., Bourrel,A.S., Palmieri,M., Ortiz De La Rosa,J.M., Jacquier,H., Denamur,E., Nordmann,P. and Poirel,L. (2019) mcr-9, an inducible gene encoding an acquired phosphoethanolamine transferase in escherichia coli, and its origin. *Antimicrob. Agents Chemother.*, **63**, e00965-19.
24. Taxt,A.M., Avershina,E., Frye,S.A., Naseer,U. and Ahmad,R. (2020) Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing. *Sci. Rep.*, **10**, 7622.
25. Goncalves,R.S. and Musen,M.A. (2019) The variable quality of metadata about biological samples used in biomedical experiments. *Sci. Data*, **6**, 190021.
26. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
27. Chamberlain,S.A. and Szocs,E. (2013) taxize: taxonomic search and retrieval in R. *F1000Res*, **2**, 191.
28. Schriml,L.M., Mitraka,E., Munro,J., Tauber,B., Schor,M., Nickle,L., Felix,V., Jeng,L., Bearer,C., Lichenstein,R. *et al.* (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.

29. Wintersinger,J.A. and Wasmuth,J.D. (2015) Kablammo: an interactive, web-based BLAST results visualizer. *Bioinformatics*, **31**, 1305–1306.
30. Tarkowska,A., Carvalho-Silva,D., Cook,C.E., Turner,E., Finn,R.D. and Yates,A.D. (2018) Eleven quick tips to build a usable REST API for life sciences. *PLoS Comput. Biol.*, **14**, e1006542.
31. Pellow,D., Mizrahi,I. and Shamir,R. (2020) PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.*, **16**, e1007781.
32. Krawczyk,P.S., Lipinski,L. and Dziembowski,A. (2018) PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.*, **46**, e35.
33. Pradier,L., Tissot,T., Fiston-Lavier,A.S. and Bedhomme,S. (2021) PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC Bioinformatics*, **22**, 349.
34. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
35. Naorem,R.S., Urban,P., Goswami,G. and Fekete,C. (2020) Characterization of methicillin-resistant *Staphylococcus aureus* through genomics approach. *3 Biotech*, **10**, 401.
36. Takayama,Y., Tanaka,T., Oikawa,K., Fukano,N., Goto,M. and Takahashi,T. (2018) Prevalence of blaZ gene and performance of phenotypic tests to detect penicillinase in *Staphylococcus aureus* isolates from japan. *Ann. Lab. Med*, **38**, 155–159.
37. Lee,A.S., de Lencastre,H., Garau,J., Kluytmans,J., Malhotra-Kumar,S., Peschel,A. and Harbarth,S. (2018) Methicillin-resistant *Staphylococcus aureus*. *Nat. Rev. Dis. Primers*, **4**, 18033.

## 3.7 *Auritidibacter ignavus, an Emerging Pathogen Associated with Chronic Ear Infections*

SYNOPSIS

# *Auritidibacter ignavus*, an Emerging Pathogen Associated with Chronic Ear Infections

Sophie Roth, Maximilian Linxweiler, Jacqueline Rehner,
Georges-Pierre Schmartz, Sören L. Becker, Jan Philipp Kühn

JOINTLY ACCREDITED PROVIDER™
WITH COMMENDATION
INTERPROFESSIONAL CONTINUING EDUCATION
Medscape CME Activity

In support of improving patient care, this activity has been planned and implemented by Medscape, LLC and Emerging Infectious Diseases. Medscape, LLC is jointly accredited with commendation by the Accreditation Council for Continuing Medical Education (ACCME), the Accreditation Council for Pharmacy Education (ACPE), and the American Nurses Credentialing Center (ANCC), to provide continuing education for the healthcare team.

Medscape, LLC designates this Journal-based CME activity for a maximum of 1.00 ***AMA PRA Category 1 Credit(s)*™**. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Successful completion of this CME activity, which includes participation in the evaluation component, enables the participant to earn up to 1.0 MOC points in the American Board of Internal Medicine's (ABIM) Maintenance of Certification (MOC) program. Participants will earn MOC points equivalent to the amount of CME credits claimed for the activity. It is the CME activity provider's responsibility to submit participant completion information to ACCME for the purpose of granting ABIM MOC credit.

All other clinicians completing this activity will be issued a certificate of participation. To participate in this journal CME activity: (1) review the learning objectives and author disclosures; (2) study the education content; (3) take the post-test with a 75% minimum passing score and complete the evaluation at http://www.medscape.org/journal/eid; and (4) view/print certificate. For CME questions, see page 211.

NOTE: It is Medscape's policy to avoid the use of Brand names in accredited activities. However, in an effort to be as clear as possible, the use of brand names should not be viewed as a promotion of any brand or as an endorsement by Medscape of specific products.

**Release date: December 20, 2023; Expiration date: December 20, 2024**

**Learning Objectives**

Upon completion of this activity, participants will be able to:

- Analyze the microbiology of *Auritidibacter ignavus*
- Assess risk factors for *A. ignavus* otitis
- Distinguish antimicrobial resistance patterns of *A. ignavus*
- Identify physical findings associated with *A. ignavus* otitis

**CME Editor**

**Thomas J. Gryczan, MS,** Technical Writer/Editor, Emerging Infectious Diseases. *Disclosure: Thomas J. Gryczan, MS, has no relevant financial relationships.*

**CME Author**

**Charles P. Vega, MD,** Health Sciences Clinical Professor of Family Medicine, University of California, Irvine School of Medicine, Irvine, California. *Disclosure: Charles P. Vega, MD, has the following relevant financial relationships: served as an advisor or consultant for Boehringer Ingelheim; GlaxoSmithKline; Johnson & Johnson Services, Inc.*

**Authors**

***Sophie Roth, MD; Maximilian Linxweiler, MD; Jacqueline Rehner, MSc; Georges-Pierre Schmartz, MSc; Sören L. Becker, MD, PhD; Jan Philipp Kühn, MD.***

Author affiliation: Saarland University Institute of Medical Microbiology and Hygiene, Homburg/Saar, Germany

DOI: https://doi.org/10.3201/eid3001.230385

We describe detection of the previously rarely reported gram-positive bacterium *Auritidibacter ignavus* in 3 cases of chronic ear infections in Germany. In all 3 cases, the patients had refractory otorrhea. Although their additional symptoms varied, all patients had an ear canal stenosis and *A. ignavus* detected in microbiologic swab specimens. A correct identification of *A. ignavus* in the clinical microbiology laboratory is hampered by the inability to identify it by using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Also, the bacterium might easily be overlooked because of its morphologic similarity to bacterial species of the resident skin flora. We conclude that a high index of suspicion is warranted to identify *A. ignavus* and that it should be particularly considered in patients with chronic external otitis who do not respond clinically to quinolone ear drop therapy.

*Auritidibacter ignavus* is an aerobic gram-positive, rod-shaped bacterium that was described by Yassin et al. in 2011 after isolation from an ear swab specimen (*1*). Thus far, all published cases with microbiological detection of *A. ignavus* were associated with ear infection that clinically manifested as otitis externa with otorrhea, which indicates a specific role of this pathogen in inflammatory diseases of the outer ear (*1*–*3*). However, only a limited number of cases have been published, and scant data are hampering valid conclusions on the clinical relevance and therapeutic implications of this pathogen. In addition, there are discrepant results with regard to susceptibility testing (*1*,*2*).

We describe 3 cases of patients with otorrhea caused by *A. ignavus* detected during March 2021 and October 2022 at the Saarland University Institute of Medical Microbiology and Hygiene (Homburg/Saar, Germany); the total number of ear swab specimens analyzed for diagnostic purposes in the institute's microbiology laboratory during 2021 and 2022 was 922. We provide an in-depth description of the clinical isolates, including their antimicrobial drug susceptibility patterns and strain comparison by whole-genome sequencing. Furthermore, we review the available literature pertaining to *A. ignavus*.

## Case Reports

Written informed consent was obtained from the 3 patients to publish this case report. Patient 1 was a 50-year-old man who sought care for a chronic right-sided otorrhea caused by treatment-resistant external otitis, which had caused symptoms for several months. An outpatient topical treatment with ciprofloxacin ear drops for several weeks did not result in clinical improvement. At initial examination, the patient described persistent itching and otalgia on the affected ear. Clinical examination showed an extensive stenosis of the external auditory canal caused by multiple exostoses that narrowed the lumen by >50%. The ear canal appeared swollen and red by ear microscopy (Figure 1, panel A). The eardrum was covered with black fungal spores. Microbiological wound swab specimens showed *A. ignavus* and the dematiaceous fungus *Exophiala dermatitidis*. Thus, an alternating topical therapy with povidone-iodine drops and ethanol drops was initiated. Four weeks later, the patient reported major clinical improvement and absence of any symptoms. The examination showed a dry ear canal without any abnormal findings.

Patient 2 was a 72-year-old woman who sought care for slowly progressing conductive hearing loss of the right ear and occasional otorrhea. She denied any pain, dizziness, or tinnitus. Although an otologic examination of the left ear showed unremarkable findings, the right side showed a fibrotic, moist auditory canal with stenosis, which was suggestive of a postinflammatory acquired atresia of the external auditory canal (Figure 1, panel B). Audiometry showed an air bone gap of up to 20 dB on the right side with bilateral sensorineural normacusis. To exclude middle and inner ear affection or malformations, computed tomography was performed and showed a partial obstruction of the right external auditory canal by fibrous tissue without any additional pathologic findings. To widen the external auditory canal and to help with outer ear drainage, we performed a meatoplasty. Because the otorrhea did not subside postoperatively, we obtained a microbiological swab specimen, which grew *A. ignavus.* A topical therapy with ethanol drops and nourishing oil drops led to a long-lasting improvement of symptoms without recurring otorrhea.

Patient 3 was a 76-year-old man who had lichen planus and sought care for recurrent otorrhea of both ears for >2 months. He reported no otalgia, vertigo, or tinnitus. A symmetric presbycusis had remained unchanged for years and was treated with conventional hearing aids. On examination, both auditory canals were moist and constricted, clinically manifesting as inflammatory meatal fibrosis, a common finding in patients who have lichen planus. Result of a computed tomography scan showed a bilateral circumferential bony overgrowth of the osseous external auditory canal. A microbiological swab specimen led to the identification of *A. ignavus* in both ears. Thus, a topical therapy with ethanol drops and a tincture of isopropyl alcohol, glycerin, acetic acid, and peppermint

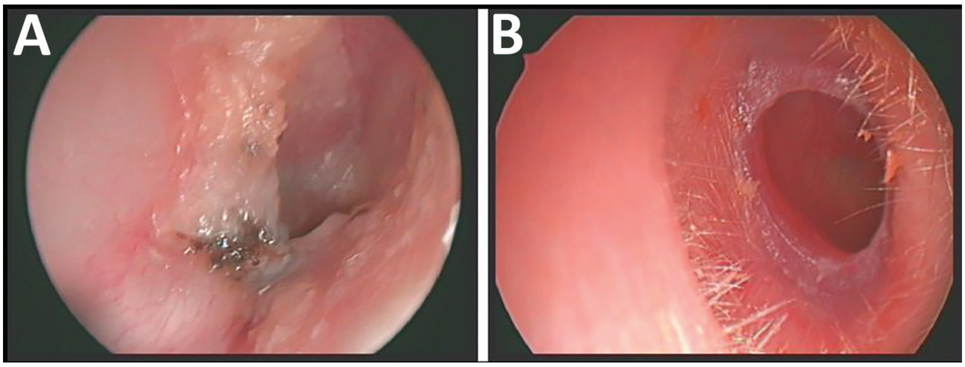


**Figure 1.** Right ears of 2 patients with chronic ear infections who were infected with *Auritidibacter ignavus*, Germany. A) Patient 1. Auditory canal was swollen and red and contained fungal spores. B) Patient 2. Fibrotic stenosis in the cartilaginous part of the ear canal, which was suggestive of a postinflammatory acquired atresia of the external auditory canal.

oil was initiated. At follow-up after 3 weeks, both auditory canals were dry and without signs of acute infection but with an unchanged fibrotic stenosis.

### Microbiological Characteristics of *A. ignavus*

In all 3 cases, microbiological ear swab specimens were subjected to standard microbiological culture methods (i.e., incubation on tryptic soy blood agar and chocolate agar for ≥48 hours). After 1 day of incubation, small white-gray colonies appeared (Figure 2), which changed to a gray-yellow appearance with a slimy surface over the course of few days. On Gram staining, gram-positive rods were observed, with a partially coccoid morphology.

No distinct identification was achieved by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (Bruker Daltonics). Thus, we performed a 16S broad-range PCR and subsequent Sanger sequencing. Analysis using a BLAST search (https://www.ncbi.nlm.nih.gov/BLAST) based on the National Center for Biotechnology Information genome database showed a sequence homology ≥99% for *A. ignavus* in all 3 cases.

We performed antimicrobial susceptibility testing using epsilometry on Mueller-Hinton agar with 5% sheep blood. In the absence of specific species-related clinical breakpoints for *A. ignavus*, we assessed the MICs by using the non–species-related breakpoints put forth by the European Committee on Antimicrobial Susceptibility Testing (https://www.eucast.org). We consistently noted high MICs for ciprofloxacin, which are likely to be associated with clinical failure of this drug. In contrast, all isolates were susceptible to β-lactam antimicrobial drugs and vancomycin (Table).

We extracted whole-genome DNA from isolates of *A. ignavus* by using the ZymoBIOMICS DNA Miniprep Kit (Zymo Research Corp.). We performed subsequent whole-genome sequencing by using Illumina PE150 (HiSeq), conducted by Novogene UK Ltd.. We performed quality control of sequencing output by using Fastp version 0.23.2 and MultiQC version 1.13a. We aligned reads against the reference genome of *A. ignavus* (CP031746.1 *Auritidibacter* sp. NML130574) by using Bowtie2 version 2.4. Variant calling using Freebayes version 1.3.2, filtering using Vcftools version 0.1.16 with a set threshold of 20, and comparison with Vcftools suggested that all 3 isolates were unrelated and had only 5,246 single-nucleotide polymorphisms in common (Figure 3).

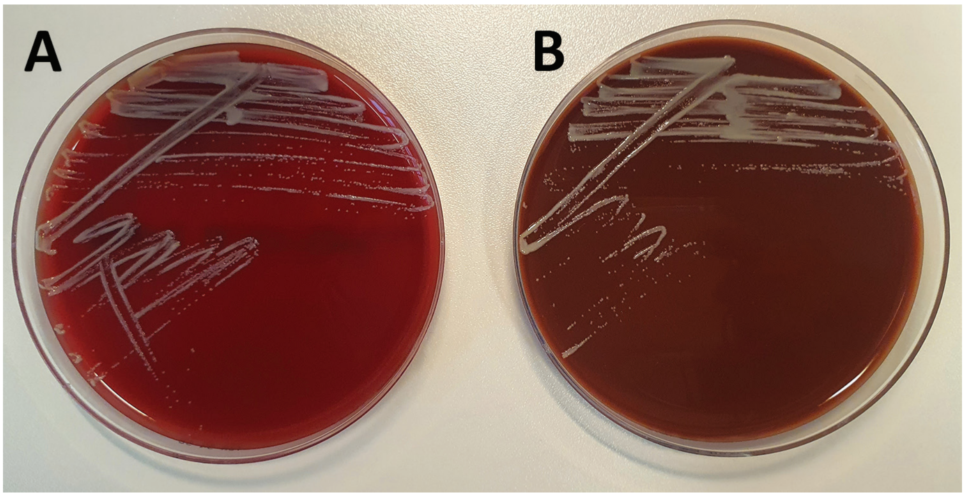


**Figure 2.** Small white-gray colonies of *Auritidibacter ignavus* in a sample from a chronic ear infection patient, Germany. Colonies are shown after 2 days of incubation at 37°C on tryptic soy blood agar (A) and chocolate agar (B).

**Table.** Antimicrobial drug susceptibility patterns for 10 drugs of 3 *Auriditibacter ignavus* isolates from patients with chronic ear infections, Germany*

| | MIC, mg/L | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Isolate | PEN | CRX | AMS | MEM | VAN | LIN | CLI | DOX | SXT | CIP |
| 1 | 0.38 | 0.5 | 0.25 | 1.5 | 0.064 | 0.5 | 32 | 0.5 | 0.094 | 32 |
| 2 | 0.19 | 0.094 | 0.125 | 0.38 | 0.125 | 0.75 | 2 | 0.064 | 0.008 | 16 |
| 3 | 0.125 | 0.125 | 0.25 | 0.5 | 0.064 | 0.38 | 1.5 | 0.125 | 0.19 | 12 |

*Testing was performed by using epsilometry on Mueller-Hinton-Agar with 5% sheep blood. AMS, ampicillin/sulbactam; CIP, ciprofloxacin; CLI, clindamycin; CRX, cefuroxime; DOX, doxycycline; LIN, linezolid; MEM, meropenem; PEN, penicillin, SXT, trimethoprim/sulfamethoxazole; VAN, vancomycin.

## Discussion

*Auritidibacter* spp. infections have rarely been reported in the literature. A systematic PubMed/MEDLINE search using the search term "*Auritidibacter*" yielded only 3 results. In 2011, Yassin et al. (*1*) provided a detailed account of this bacterium with a microbiological, biochemical, and phylogenic characterization. The phenotypic culture morphology pattern described in their work matched our own observations. Eight years later, Seth-Smith et al. (*3*) published a complete genome assembly of an isolate from Switzerland and compared it with 4 global genomes, which showed a high diversity within the species. That finding is consistent with our findings of only 24.4%–29.1% single-nucleotide polymorphism identity between the 3 different isolates from the 3 case-patients (Ai_01, 29.1%; Ai_02, 24.4%; Ai_03, 26.5%). More recently, Bernard et al. (*2*) investigated 4 isolates of the genus *Auriditibacter* by microbiological and biochemical detection methods, as well as whole-genome sequencing, to assess their relatedness to the species *A. ignavus*.

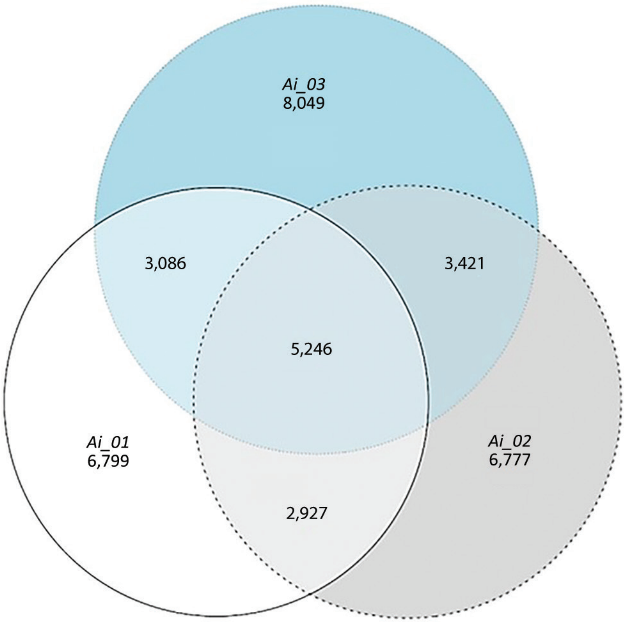


**Figure 3.** Venn diagram showing overlapping single-nucleotide polymorphism information among *Auritidibacter ignavus* isolates (Ai_01, Ai_02, and Ai_03) from 3 chronic ear infection patients, Germany.

All of those studies reported only little clinical data of the included patients. We present a report that includes details on the patients' clinical courses, including the clinical treatment response. Whereas no clear associations of *A. ignavus* infections with predisposing factors was found, outer ear canal stenosis was observed in all 3 patients. This anatomic feature seems to favor the colonization and probably also the infection with this pathogen. However, limited data make it difficult to explicitly establish a causal link between both conditions. Thus, additional studies or case series of a larger number of patients, including a control group of patients with ear canal stenosis and no clinical symptoms suggestive of acute inflammation, would be necessary to distinguish between colonization and infection.

According to Yassin et al. (*1*), *A. ignavus* is usually susceptible to β-lactam antimicrobial drugs, whereas Bernard et al. (*2*) reported resistance to cefepime. Such discrepancies might partially be explained by different antimicrobial testing methods, which underscores the need for coordinating testing recommendations for rare bacteria such as *A. ignavus*. Particular attention should be paid to our observation of ciprofloxacin resistance in all isolates, a finding that is consistent with the report by Bernard et al. (*2*).

Ciprofloxacin ear drops are commonly prescribed in clinical practice. Although MICs enable only limited conclusions on the clinical effectiveness of local antimicrobial drug therapy, we suggest that patients with therapeutic failure after empiric topical treatment with ciprofloxacin ear drops should be assessed for *A. ignavus* by using microbiological tests. The clinical suspicion should be reported to the microbiology laboratory because there is a serious risk of overlooking *A. ignavus* caused by its morphologic similarity to bacterial species belonging to the residential skin flora.

No specific request for an in-depth analysis was made by the treating clinicians in the cases we describe. Thus, increased awareness among the clinical microbiologists was caused by the repeated receipt of ear swab specimens from the patients with the clinical information otorrhea in context with the bacterial growth of presumed physiologic flora in large quantities, which led to a low threshold to submit bacterial colonies to additional testing for species identification. Finally, the absence of *A. ignavus* in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry databases poses an additional threat to correct identification in the clinical microbiology laboratory, as has been reported for other pathogens (*4*).

In conclusion, *A. ignavus* is a novel, potentially underrecognized pathogen that seems to be associated with a distinct clinical pattern in patients with ear infections. A high level of disease awareness and accurate microbiological diagnostics are required for correct identification. In patients who have a clinical course of chronic external otitis and who do not respond to empirical treatment with quinolone ear drops, *Auritidibacter* infection should be considered and further investigated.

## About the Author

Dr. Roth is a physician at the Institute of Medical Microbiology and Hygiene, Saarland University, Homburg/Saar, Germany. Her primary research interest is improved and faster diagnostic of bacterial infections.

## References

1. Microbiology Society. *Auritidibacter ignavus* gen. nov., sp. nov., of the family Micrococcaceae isolated from an ear swab of a man with otitis externa, transfer of the members of the family Yaniellaceae Li et al. 2008 to the family Micrococcaceae and emended description of the suborder Micrococcineae [cited 2023 Feb 24]. https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.019786-0#tab2
2. Bernard KA, Pacheco AL, Burdz T, Wiebe D, Beniac DR, Hiebert SL, et al. Emendation of the genus *Auritidibacter* Yassin et al. 2011 and *Auritidibacter ignavus* Yassin et al. 2011 based on features observed from Canadian and Swiss clinical isolates and whole-genome sequencing analysis. Int J Syst Evol Microbiol. 2020;70:83–8. https://doi.org/10.1099/ijsem.0.003719
3. Seth-Smith HM, Goldenberger D, Bernard KA, Bernier AM, Egli A. Complete genome assembly of an *Auritidibacter ignavus* isolate obtained from an ear infection in Switzerland and a comparison to global isolates. Microbiol Resour Announc. 2019;8:e00291–19. https://doi.org/10.1128/MRA.00291-19
4. Chen XF, Hou X, Xiao M, Zhang L, Cheng JW, Zhou ML, et al. Matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) analysis for the identification of pathogenic microorganisms: a review. Microorganisms. 2021;9:1536. https://doi.org/10.3390/microorganisms9071536

Address for correspondence: Sophie Roth, Institute of Medical Microbiology and Hygiene, Saarland University, Kirrberger Strasse, Bldg 43, Homburg 66421, Germany; email: mikrobiologie@uks.eu

## 3.8 Between Cages and Wild: Unraveling the Impact of Captivity on Animal Microbiomes and Antimicrobial Resistance

# Between Cages and Wild: Unraveling the Impact of Captivity on Animal Microbiomes and Antimicrobial Resistance

Georges P. Schmartz[1, #], Jacqueline Rehner[2, #], Miriam J. Schuff[2], Sören L. Becker[2], Marcin Krawczyk[3], Azat Tagirdzhanov[1, 4], Alexey Gurevich[4, 5], Richard Francke[6], Rolf Müller[4], Verena Keller[1], Andreas Keller[1, 4]

1 Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany
2 Institute of Medical Microbiology and Hygiene, 66421 Saarland University, Homburg, Germany
3 Department of Medicine II, 66421 Saarland University, Homburg, Germany
4 Helmholtz Institute for Pharmaceutical Research Saarland, Helmholtz Center for Infection Research, 66123 Saarbrücken, Germany
5 Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany
6 Zoo Saarbücken, 66121 Saarbrücken, Germany
# Authors contributed equally

## ABSTRACT

Understanding human, animal, and environmental microbiota is essential for advancing global health and combating antimicrobial resistance (AMR). We investigated the oral and gut microbiota of 48 animal species in captivity, comparing them to those of wildlife animals. Specifically, we characterized the microbiota composition, metabolic pathways, AMR genes, and biosynthetic gene clusters (BGCs). We described 585 novel species-level genome bins (SGBs), predicted 484 complete BGCs, and observed diet-dependent metabolic pathway variations. Furthermore, in comparison to wildlife-derived microbiomes, we noticed examples of converging microbiota. Importantly, our study identified AMR genes against common veterinary antibiotics and resistance to vancomycin, a critical antibiotic in human medicine. The study contributes to a better

understanding of the complexity of the animal microbiome, highlights its BGC diversity relevant to the discovery of novel antimicrobial compounds, and underlines the importance of the ‘One Health’ approach due to the potential for zoonotic transmission of pathogenic bacteria and AMR.

## INTRODUCTION

The Human Microbiome Project, launched by the United States National Institutes of Health in 2007, aimed to describe the microbiota compositions at various body sites in its initial phase (1). This led to the first characterizations of the human microbiome in healthy individuals and paved the way for phase two, which focused on analyzing the role of the human microbiome in several disease states in detail (2). Since the project's launch, researchers worldwide have focused on understanding which microorganisms are present in and on the human body, their contributions to disease development, progression, and exacerbation, and their potential to protect us, especially against chronic-inflammatory diseases (3-6). The dynamic nature of the microbiome, which can be altered by various factors such as diet, changes in the environment, and frequent exposure to other microbiota, including those of pets, emphasizes the importance of analyzing environmental microorganisms and animal microbiota, along with the human microbiome, as part of the World Health Organization's 'One Health' approach (7-10).

This approach acknowledges the interconnectedness of human, animal, and environmental health and highlights the need to analyze the three components to understand and address antimicrobial resistances (AMR) along with other factors (11). As AMR continues to increase globally, it is crucial to investigate the role of microorganisms in humans, animals, and the environment to contribute to overall homeostatic ecosystems, control diseases, and secure global health (12). Moreover, the genomic and functional characterization of microorganisms from all three components of the 'One Health' approach contributes significantly to the identification of novel antimicrobial compounds to tackle the AMR crisis and, moreover, identify novel species before they disappear from the earth's diversity again (13-16). Biosynthetic gene clusters (BGCs) encode the biosynthetic repertoire of microbes resulting in the context-dependent production of natural products such as antibiotics. They are thus a potential target for the discovery of antimicrobial compounds as the secondary metabolites they encode aid in inter-species competition between microbes (17, 18). BGC-derived metabolites have the potential to offer specificity against selected species, in contrast to commonly used broad-spectrum antibiotics (19, 20). Additionally, their existence for multiple millennia minimizes the risk of spontaneous resistance development, making them an attractive option for human and veterinarian medicine (21). While many studies have focused on identifying novel BGCs derived from human microbiota and soil bacteria that still represent a major source of new active compounds, few have investigated the occurrence of BGCs in animal microbiota (22-25).

In this study, we investigate the oral and intestinal microbiota residing within various animal classes in captivity, a setting closely intertwined with the presence of zookeepers and visitors. Our objective was to unravel the intricacies of these microbial communities, illuminating their composition and establishing connections with the animals' diets. Beyond mere identification, we sought to comprehend the functional contributions of microorganisms to health and disease, exploring metabolic pathways, AMR genes, and BGCs. However, our investigation extended beyond the confines of captive environments, as we embarked on a comparative analysis between the microbiomes of captive zoo animals and those of their wild counterparts (***Fig. 1a***). By scrutinizing the microbiota of these captivated creatures, our study not only aims to enrich our understanding of the microbiome's complexity but also holds the promise of unearthing novel antimicrobial compounds sourced from animal microbiota.

# RESULTS

## Deep sequencing and quality control results in 64 metagenomes from 45 species

First, we assessed the quality of the metagenomics sequencing results in light of the diversity of species and sample types included and characterized the robustness of our data. We collected a total of 55 stool and 16 saliva samples, representing an extensive range of 48 and 15 distinct zoo animal species (mammals, birds, and reptiles), respectively (***Fig. 1b***). Subsequently, after sequencing and quality control, we obtained a final dataset comprising 52 stool and 14 saliva samples, reflecting 45 and 13 species (***Fig. 1c***). Our quality control measures, including host DNA decontamination, yielded minimal read losses during the process, with an average loss of only 6.6% and a standard deviation (SD) of ±13.2%. We retained an average of 5.3 gigabases of sequencing data (SD: 1.7 GB), ensuring a reliable dataset for further analysis.

To account for the species for which a reference assembly was not available on RefSeq, we employed substitute assemblies that were taxonomically close. Notably, this substitution did not significantly impact the relative number of filtered reads (two-sided Wilcoxon p-value of 0.36, Supplementary Table 1), supporting our methodology. Utilizing reference-free ordination analysis, we performed an in-depth examination of the cleaned reads, unveiling distinct patterns of sample clustering primarily based on biospecimen (PERMANOVA p-value < 0.001, ***Extended Data Figure 1***). This finding underscores the significance of differentiating between stool and saliva samples and highlights the influence of the animal's specific microbiota on each biospecimen.

**De-novo analysis reveals 585 novel genomes and enhances taxonomic assignment**

We encountered an expected – yet significant – challenge when performing taxonomic profiling based on the Genome Taxonomy Database (GTDB) (21). The assignment rate using this database was low, with an average of less than 17% (SD: ±16.6%) matches. This scarcity of read assignments prompted us to adopt a de-novo analysis workflow. Applying this de-novo analysis workflow proved to be instrumental in overcoming some limitations of the taxonomic profiling from existing databases and uncovered the hidden microbial diversity within our dataset. Through this approach, we successfully recovered a total of 786 dereplicated species-level genome bins (SGBs) exceeding the criteria of at least medium MIMAG quality (namely, less than 10% contamination and a minimum of 50% completeness) (22). Among these SGBs, 585 genomes (74%) had no representatives in the GTDB with ANI (Average Nucleotide Identity) less than 95%

(***Fig. 2a, Extended Data Figure 2, Supplemental Table 2***). Specifically, when examining the stool samples, we found that out of the 616 dereplicated SGBs, 446 had no representatives (72%). In the case of saliva samples, the ratio increased to 139 out of 170 (82%). Saliva samples, accounting for 21% of the overall samples, contributed 22% and 23% of all the dereplicated SGBs and novel dereplicated SGBs, respectively, suggesting the importance of the oral microbiome in uncovering microbial diversity to be on par with the gut microbiome. Analyzing all the recovered SGBs, we observed an average scaffold length of 13 kb (SD: ±2.5kb). Additionally, we conducted searches for tRNA sequences as well as 5S, 16S, and 23S rRNA sequences within the SGBs. In total 11801 tRNAs and 205 rRNAs were detected in the SGBs averaging at 15 tRNAs and 0.3 rRNAs per SGB. Whereas these functional gene statistics are indicative of the overall quality of the assemblies, they also highlight the challenges of reliably assembling ribosomal RNA genes.

Importantly, the integration of our SGBs into the GTDB prior to taxonomic profiling yielded a substantial improvement in the read assignment rate (paired two-sided Wilcoxon p-value $< 1.7 \times 10^{-12}$, ***Extended Data Figure 3***). Nevertheless, for 17 samples the assignment rate remained below the low threshold of 20%. This highlights the significance of including the novel microbial genomes discovered in this study to enhance the accuracy and comprehensiveness of taxonomic assignments. This analysis is also necessary to assess compositional and functional differences between microbiomes and to uncover the distribution of BGCs.

## Culture-based taxonomic assignment yields differences between herbivores and carnivores

In our metagenomic data, the measured alpha-diversity, a sign of the microbial complexity of a sample, appears stable for biological replicates. In contrast, the alpha-diversity fluctuates significantly across species (***Fig. 3***). Astonishingly, we observed a negative Spearman correlation of –0.38 between assignment rates and diversity. Moreover, the reference-based ordination analysis does not yield clear clusters, reflecting neither zoological classification nor diet compositions. Nevertheless, specific zoological proximities are reflected in the clustering hierarchy such as similar patterns between sheep and goat, or between zebra and horse. But in sum, the overall assessment is that the reference-based ordination analysis remains inconclusive with respect to identifying sub-groups of animals. One likely reason for this result is the high variability of assignment rates and missing SGBs. Because differences in the gut microbiota between herbivores and carnivores are known from the literature, we asked whether a more targeted approach involving culturing of bacteria highlights such differences (26, 27).

Culturing of 11 saliva and 49 stool samples on TSA, Chocolate blood, Columbia, and MacConkey agar followed by subsequent MALDI-TOF analysis enabled the identification of overall 79 different bacterial species (***Extended Data Figure 4, Supplemental Table 3***). While we identified a total of 29 species in saliva (37%), only 8 of them (28%) were also detected in stool samples, where 6 of these 8 were of the genus *Staphylococcus*. In total, 32 species (40%) were only detected in the 38 samples of herbivore animals (including species such as *Enterococcus mundtii, Bacteroides ovatus,* and *Bacillus pumilus)*. In contrast, 8 species (10.1%) were observed only in the 7 samples from carnivore animals (including *Citrobacter braakii, Plesiomonas shigelloides, and Staphylococcus simulans).* Moreover, 17 bacterial species (21.3%) were uniquely detected in 15 samples of omnivore animals (including *Neisseria zoodegmatis* and *Staphylococcus hominis* depicting the highest frequency across samples). Across all samples, 7 species (8.8%) are present in all three diet forms, including prevalent intestinal microbiota such as *Enterococcus faecalis, Escherichia coli,* and *Enterococcus faecium*, as well as *Clostridium perfringens* and *Bacillus cereus.* Before adjustment for multiple hypothesis testing, 11 species were significantly unevenly distributed within the cohorts ($\chi^2$ test p-value<0.05). After the Benjamini-Hochberg adjustment, no p-value remained significant. Performing the same test over all stool samples and cohorts did not display significant differences between diets ($\chi^2$ test p-value=0.51).

As for culturing, only a selection of media was used, bias is introduced by excluding the growth of certain bacteria, that cannot grow on the selected media. However, all samples were treated the same, which makes these results at least comparable. It is worth mentioning, that only 30-60 % of microorganisms are cultivatable under laboratory conditions, making the metagenomic analysis a more powerful and more precise tool to investigate the microbiome. Nevertheless, the considerably different repertoire of microbiota suggests unique functional characteristics that might be connected to the dietary origin. We thus performed a functional in-silico gene analysis of the respective microbiota.

The statistically significant results from this functional gene analysis highlight elevated creatinine degradation I pathway in herbivore animals (***Extended Data Figure 5***). Contrastingly, the superpathway of tetrahydrofolate biosynthesis and salvage is more prevalent in microbiota from carnivore animals. Enriched in both, carnivore and omnivore animals, are bacteria carrying the genomic information for flavin-dependent thymidylate synthase (thyX), which is required to synthesize pyrimidine deoxyribonucleotides de novo. Most notably, this gene and the encoded protein are present in human and animal pathogens, such as *Helicobacter pylori*, *Borrelia burgdorferi*, and *Chlamydia trachomatis* (28-30).

**Differences in 484 complete biosynthetic gene clusters depending on the diet**

After the initial general functional gene analysis of the different animal microbiota, we looked into the specific metabolite landscapes of individual members of the microbiomes. We performed genome mining of the previously defined SGBs and identified 1,588 potential BGCs. Of those, 1,104 remained partial and 484 were identified as full BGC clusters of various categories (**Fig. 2a, Fig. 2b, Supplemental Table 4**). Further analysis with BiG-SCAPE categorized BGCs into 1,482 families, out of which 1,407 families were singletons containing only one BGC (31). A total of 5 families compromising 6 BGCs are linked to annotated gene clusters from the MIBiG 3.1 database (32). But interestingly, BiG-SCAPE did not form any clans of the families. Together with a high number of singleton families, this suggests a high diversity of BGCs in the collected dataset.

With 604 (38%) BGCs, *Clostridia* was the class where we predicted most BGCs. However, this is mostly due to *Clostridia* making up about 36% of our recovered dereplicated SGBs. If we look at the average number of BGCs per SGB and exclude singletons we observe that on average most BGCs were predicted for the class of *Planctomycetia*. Averaged over 15 genomes, we observed 3.73 BGCs per SGB. With only 4 BGCs in 33 SGBs, *Saccharimonadia* had the lowest non-singleton ratio of BGCs to SGBs. Concerning disparities between the oral and gut microbiome, we observed a total of 450 BGCs (28%) in the 170 saliva-derived SGBs averaging 2.65 BGCs per SGB, which compares to 1,138 BGCs in 616 SGBs at a ratio of 1.85 in the stool samples. Focusing only on the stool-derived BGCs we observed an average of 2.01, 1.65, and 1.47 BGCs per SGB for herbivores, omnivores, and carnivores, respectively. These differences were confirmed to be significant (Kruskal-Wallis p-value < 0.0053). Specifically, the average number of NRPS was 2.87 and 3.46 times higher in herbivore SGBs compared to carnivores and omnivores respectively.

Through a comparative analysis of predicted BGCs and known annotated BGCs from the MIBiG database, we observed 37 BGCs (2%) within our SGBs that shared a similarity of over 50% with known entries. Among these annotations, various compounds may be of relevance to the host organism (***Extended Data Figure 6)***. We detected virulence factors, such as the toxin tolaasin I, within an SGB derived from tapir saliva. Furthermore, we uncovered various annotations associated with health benefits, including the bacteriocin salivaricin CRL 1328, present in an SGB derived from a mandrill stool sample (33). We encountered two further compounds with noteworthy properties: α-galactosylceramide, an immunostimulating compound found in an SGB derived from horse stool, and rhizomide, identified in an SGB derived from tapir saliva, exhibiting anti-tumor and antimicrobial properties in vitro (34).

Having captured differences in the repertoire of bacteria from animals with different diets in the gut and oral cavity along with unique functional characteristics and novel BGCs raises the question of whether captivity has an influence on the microbiota or whether wildlife animals reveal similar patterns.

**Animals in captivity present different antimicrobial resistance gene patterns**

Comparing microbiome differences between captive and wildlife animals and addressing the complexities in the sample extraction process, we conducted a comparative analysis with data from Youngblut et al. (35), the – as of now – most complete study of animal gut microbiota. Their dataset consisted of 289 samples from 180 different host species, including humans. The large differences between both studies in the selection of animal species, call for a balanced and stratified analysis approach. Therefore, we implemented a matching scheme that carefully selects a subset of samples with close zoological similarity from both studies (***Supplementary Table 1***). We excluded the oral microbiomes of the zoo animals from this analysis because no oral microbiota from wildlife animals were present.

It is important to acknowledge the easier collection process in a controlled environment as a zoo in comparison to a wildlife setting, likely leading to differences in the sample quality. To quantify these differences and ensure methodological consistency, we thus applied our analysis workflow to the selected metagenomes from Youngblut et al. (35). We observed a significant decrease in read quantity after decontamination compared to the present data, which is explained by the above-mentioned challenges in wildlife sampling (***Extended Data Fig. 7a***). This reduction also influences the assembly quality, which was lower in the wildlife samples, finally leading to overall shorter fragments (***Extended Data Fig. 7b***). Consequently, fewer SGBs were recovered in the wildlife samples compared to the zoo dataset (***Extended Data Fig. 7c***). While the samples from animals kept in captivity retained an average of 9.8 SGBs per sample, the wildlife dataset yielded just 1.9 SGBs. Similar differences also apply to the number and abundance distribution of BGCs. Here, SGB derived from wildlife animals present on average 13 fewer BGCs per SGB (**Extended Data Fig. 7d**). We were only able to recover partial BGCs in the wildlife samples compared to 50 complete BGCs in the matching zoo samples. Further, only one BGC was annotated to have a similarity >10% to any known MIBiG BGC. It has a 28% similarity to a carotenoid cluster derived from an *Algoriphagus* species. Again, the latter results might seem counterintuitive, and we might expect more BGCs in wildlife, yet the results are likely biased by the challenges of wildlife sampling. Most importantly, the quality of the wildlife samples is still sufficient to enable reference-free comparison.

As one first aspect, we asked whether the microbiomes between zoo and wildlife animals present a conserved proximity-dependent on the relatedness of host animal species. For the selected samples, we thus performed reference-free FracMinHash comparisons (***Fig. 4***a). On average, we computed a large dissimilarity between any compared pairs. In detail, the average dissimilarity amounts to 0.98 (SD: 0.018), which is close to the maximal dissimilarity value of 1. Importantly, the dissimilarity distributions within the wildlife and zoo animals do not differ significantly (two-sided Wilcoxon p-value > 0.37). Nevertheless, zoo animals display several strong similarities between gut microbiota. These include mostly inter-replicate comparisons of zebra, camel, and giraffes, yielding an overall significantly lower dissimilarity index as compared to the other zoo animals (two-sided Wilcoxon p-value $< 9.44 \times 10^{-7}$). Of note, no replicates for the wildlife animals are available, explaining the missing similarities within those samples. Interestingly, several of the zoo animal species including the yak, giraffe, camel, and goat displayed increased similarities in gut microbiota. The same applies to two kangaroo species that also show similarities in the gut microbiota. Of note, such similarities are not present in the wildlife animals and may suggest an influence e.g. of the nutrition in this controlled environment. Further, the results clearly argue for combining the advantages of studies in wildlife animals (being closer to nature) and controlled environments (facilitating higher sample quality).

One immediate question in comparing wildlife to captivity set-ups concerns the presence of AMR. AMR gene analysis of zoo animals revealed potential resistances against antibiotics that are commonly used in veterinary medicine, such as tetracyclines, macrolides, and lincosamides (***Fig. 4b***) (36). However, we also observed resistance genes against vancomycin, which is a last-resort antibiotic against infections with Gram-positive bacteria in human medicine (37). Specifically, we documented the well-known resistance clusters *vanD* and *vanG* (38, 39). However, we also detected the *vanO* operon, which has not been identified from animal- or human-derived samples yet (40). As outlined in the One Health concept, such resistant bacteria could be transferred from zoo animals to zookeepers, increasing the global spreading of such organisms. When we compared our matching stool zoo samples to the wildlife samples, we observed a significantly smaller number of antimicrobial compound classes that are targeted by at least one resistance gene in the wildlife samples (two-sided Wilcoxon p-value < 0.001). Overall, we only observed a total of four resistance genes in all analyzed wildlife samples. This suggests that wild animals overall suffer from less AMR. Nevertheless, we want to highlight that this result is again to be interpreted in the light of the inferior assembly quality of the wildlife samples which impact the quality of AMR gene detection.

## DISCUSSION

Our findings, in line with the study by Youngblut et al. (35), indicate that the microbial dark matter within animal microbiomes remains inadequately characterized in existing data repositories. Despite our extensive efforts and the generation of several novel SGBs, we encountered 17 samples with a low estimated assignment rate below 20%. This deficiency significantly impacts state-of-the-art reference-based analysis, as evident in our own investigation.

The microbial richness we detect, despite the accompanying challenges, presents an intriguing opportunity for the discovery of BGCs associated with antimicrobial natural compounds within these samples. In this context, it is worth emphasizing the advantages of combining different study setups. While our focus lies on samples from a highly controlled environment, specifically a zoo, complementary studies like that of Youngblut et al. (35) provide valuable insights into wildlife microbiomes, which are closer to the natural microbiota. By integrating findings from diverse settings, we can gain a more comprehensive understanding of the animal microbiome and potentially uncover novel microbial resources with therapeutic potential.

Specifically, the zoo animals present higher numbers of SGBs and BGCs per SGBs but also higher proximity of gut microbiota as compared to the wildlife animals. It is important to acknowledge that the number of BGCs within SGBs can vary, depending on the specific species discovered. However, the improved assembly statistics highlight the advantages of easier sample collection in captivity compared to wild animals, at the cost of BGCs that might only be present in wildlife animals.

When comparing studies, one limitation we encountered was the need to perform inter-species comparisons, which involved species from different continents with potentially diverse diets. This aspect adds complexity to the analysis, as the microbiomes of zoo animals, despite sharing similar diets such as local seasonal vegetables, still exhibit considerable differences. The convergence of microbiome composition across zoo animals appears to be limited, yet measurable.

Furthermore, the presence of AMR genes in animal microbiomes is of considerable importance from the One Health perspective. In addition to detecting AMR genes against commonly used antibiotics in veterinary medicine, we also identified resistance genes against vancomycin in certain animals, including prosimians. Considering their close contact with zookeepers, there is a potential risk of transferring vancomycin-resistant bacteria to humans. As transmission of multi-resistant bacteria has been observed in clinical settings, these findings emphasize the need for comprehensive surveillance and management of AMRs in zoo settings to mitigate potential health risks and maintain a safe environment for both animals and humans (41, 42).

## METHODS

### Study design

For docile animals such as horses, dwarf goats, and tapirs, buccal swabs were easily taken from the oral cavity to collect saliva samples. Concurrently, fresh fecal samples were collected from the enclosures or stables and immediately transported to the veterinary station. Using a spoon from a stool sample tube, feces from the inner portion of the excreta were transferred into sample tubes. Subsequently, all samples were promptly frozen at -20°C in the freezer compartment of a refrigerator. Typically, samples were frozen within 30 minutes of collection.

For non-docile animals, such as primates and large or small carnivores, the same sample collection methods were employed during necessary anesthesia, which occurred for veterinary examinations, treatment, transport, or gender determination. For small animals, fecal samples were collected rectally as swabs, following the same protocol described above, and stored frozen until further analysis.

### DNA extraction

We extracted whole-genome DNA from all fecal and salivary swabs using the Qiagen QiAamp Microbiome Kit (Qiagen, Hilden, Germany).[8] The DNA extraction procedure was conducted according to the manufacturer's protocol. Briefly, all swabs containing native samples were vortexed in 1 ml PBS for 2 minutes. The PBS containing the microbes from each sample was then used for DNA extraction according to the manufacturer's recommendation. We used the MP Biomedicals™ FastPrep-24™ 5G Instrument (FisherScientific GmbH, Schwerte, Germany) for mechanical lysis of bacterial cells. The velocity and duration were adjusted to the "hard-to-lyse" protocol, meaning 6.5 m/s for 45 s two times and 5 minutes storage on ice in between each lysis step. DNA was eluted in 50 µl elution buffer. The DNA concentration after elution was determined via NanoDrop 2000/2000c (ThermoFisher Scientific, Wilmington, DE) full-spectrum microvolume UV-Vis measurements (43).

**Library preparation and sequencing**

Extracted whole-genome DNA was sent to Novogene Company Limited (Cambridge, UK) for library preparation and sequencing. Briefly, samples were subjected to metagenomic library preparation and further sequenced via paired-end Illumina Sequencing PE150 (HiSeq). For all samples, 5 Gb reads per sample were generated.

## Culturing of bacteria

Native fecal samples were streaked out using the swab they were taken with, on three different agar plates: TSA with 5 % sheep blood (TSA), MacConkey (MC), and Columbia (Co) agar plates (Becton, Dickinson and Company, Heidelberg, Germany). Oral samples were streaked out on TSA, Co and Chocolate blood (CB) agar plates (Becton, Dickinson and Company, Heidelberg, Germany). All TSA, CB, and MC agar plates were incubated at 35.6 °C and 5 % CO2 for a minimum of 18 h and a maximum of 24 h. Co agar plates were used for the cultivation of anaerobic bacteria and therefore incubated in an anaerobic environment for a minimum of 48 h at 35.6 °C (43).

## Mass spectrometry-based identification

Bacterial colonies obtained by culturing native fecal and oral samples on different agar plates were subjected to species identification using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry. To this end, colonies were spotted on the MALDI-TOF target plate, dried, and then overlayed with 1 µl of α-cyano-4-hydroxycinnamic acid (CHCA) matrix solution (Bruker Daltonics, Bremen, Germany). The matrix solution is composed of saturated CHCA dissolved in 50 % (v/v) acetonitrile, 47.5 % (v/v) LC-MS grade water, and 2.5 % (v/v) trifluoroacetic acid. After drying the matrix solution at room temperature, each spot was overlayed with 70 % formic acid to pre-disrupt the cells. Followed by drying at room temperature, the plate was placed into the Microflex LT Mass Spectrometer (Bruker Daltonics) for MALDI-TOF MS. All measurements were performed with the AutoXecute algorithm in the FlexControl© software version 3.4 (Bruker Daltonics). Each spot was excited with 240 laser shots in six random positions. Measurements were carried out automatically to generate protein mass profiles in linear positive ion mode using a laser frequency of 60 Hz, high voltage of 20 kV, and pulsed ion extraction of 180 ns. Mass charge ratio ranges (m/z) were measured between 2 kDa and 20 kDa. We identified bacterial species using the software MALDI BioTyper. Identification scores above 2.0 were considered a precise identification on the species level, scores between 1.7 and 1.99 were considered as possible species identification and precise genus identification, and all identification scores below 1.7 were considered unsuccessful identification. In this study, we only considered scores ≥ 2 for analyses (26).

### Next-generation sequencing preprocessing

The first step of data analysis was host read removal with KneadData (version (v):0.10.0; command line arguments (cla): "--trimmomatic-options='LEADING:3 TRAILING:3 MINLEN:50' --bowtie2-options='--very-sensitive --no-discordant -reorder'") using the respective genomes as specified in Supplementary Table 1

(44). The selected, publicly available, host genomes were downloaded with the ncbi-datasets-cli (v13.35.0). For several animal species, no exact sequenced genome of sufficient quality was available and instead, a taxonomically close substitute was selected. Bowtie2 (v2.4.5; -s) databases were prepared for each reference (45). After decontamination, we performed sequence overrepresentation analysis and quality assurance with fastp (v:0.23.2; cla: --overrepresentation_analysis) and visualized results with MultiQC (v1.13a) (46, 47). The two-sided Wilcoxon rank sum test was performed on the relative loss attributed to host DNA removal. To reduce bias, replicates were averaged. Saliva and stool samples were not averaged.

**Metagenome assembly**

We assembled each sample with SPAades (v3.15.4; cla: --meta) and monitored assembly quality with QUAST (v5.0.2; cla: -s ) (48, 49). Next, we aligned each host decontaminated sample against each set of assembled scaffolds with BWA-MEM2 (v2.2.1) and generated abundance profiles for each combination (50). We extracted coverage information to bin scaffolds with MetaBAT2 (v2.15; cla:-l --seed 420 --unbinned ), MaxBin2 (v2.2.7), and DAS Tool(v1.1.5; --search_engine diamond)(51-53)￼. MAGs across all samples were aggregated and dereplicated with dRep (v:3.4.0; cla: -comp 50 -con 10 --checkM_method lineage_wf --S_algorithm fastANI --

S_ani 0.95 -nc 0.5). At last, we used GTDB-Tk (v:2.1.1; cla: classify_wf), tRNAscanSE (v:2.0.11;--brief -Q), and barrnap(v:0.9; cla: -q) to taxonomically classify MAGs and annotate them with tRNA and rRNA information based on their classified kingdom(54, 55)￼.

### Reference-based compositional analysis

FracMinHash profiles were computed for all samples with sourmash (v:4.4.3; cla: -k51) (56). After FracMinHash profile generation, samples were compared with sourmash compare. Dissimilarities were computed by subtracting the resulting similarities from one. Samples were embedded with UMAP (v:0.2.8) (57). Further, for each SGB, FracMinHash profiles were computed as well, and an index was generated. The PERMANOVA analysis treated samples and replicates as independent (58). Taxonmic profiling was performed with sourmash (cla: -k51) our previously generated indices, GTDB (v:GTDB R07-RS207 all genomes k51), and host decontaminated reads. Shannon index was used as the alpha-diversity measure and computed with phyloseq (v:1.40.0) (59, 60). Relative abundances were averaged if replicates were available. Clustering was performed with average hierarchical clustering on Bray-Curtis distances computed with the vegan package on mean relative abundances (v: 2.6.2) (61). Tanglegram was optimized for visual clarity with “step2side” algorithm of the R dendextend package (v: 1.16.0) (62).

Differential abundance analysis was made with ANCOMBC (v:1.6.2) comparing herbivores and the union of omnivores and carnivores (63).

### Functional analysis

In order to incorporate our own SGBs into the functional profiling step, we updated an existing GTDB207-based database with Struo2 (v2.3.0) (64). After database generation, functional profiling was performed with HUMAnN 3 (v3.6; cla: --bypass-nucleotide-index) (44). We also used ANCOMBC for exploration of differences in function. The default setting of Holm-Bonferroni p-value adjustment was employed. Genes were predicted with prodigal (-p meta) and passed to antiSMASH (v6.1.1; cla -cb-knownclusters --cb-subclusters --asf) for BGC detection (65, 66). A BGC was classified as partial if it is shorter than 5 kbp or located on a contig edge and as full otherwise. Clustering of all BGCs was performed with BiG-SCAPE (v1.1.5; --mibig) using Pfam (v35.0). BiG-SCAPE failed to process two BGCs and removed them from further analysis (32, 67).

### Wildlife comparison

Samples specified in Supplementary Table 1 were downloaded from the European Nucleotide Archive and processed identically to our dataset, from host DNA removal to BGC prediction (68). We subsetted our data to only the paired samples specified in the aforementioned table. Pairings were manually selected based on taxonomic similarity. Paired comparison to our data was done based on FracMinHash dissimilarities.

## Data availability

Raw unfiltered sequencing reads as well as dereplicated SGBs were uploaded to the Sequence Read Archive under the accession: PRJNA983076.

## Funding

This work was further supported financially by Saarland University and the UdS-HIPS TANDEM initiative. The compute infrastructure for this project was funded by the DFG [469073465].

## Contributors

AK, VK, and RF had the idea for this study. GPS, JR, SLB, and AK developed the methodology. RF collected samples and managed the animals. JR and MS performed whole-genome DNA extraction, culturing, and MALDI-TOF analysis. BGC interpretation was performed by GPS, AG, and RM. SLB, VK, RM, MK, and AK provided financial resources and supervised the project. JR supervised experimental investigations and project coordination. GPS, AT, and AK performed computational analysis and data processing. GPS and JR drafted the manuscript.

All authors have read and critically reviewed the manuscript, in particular intellectual content, and agreed to the submission of the manuscript.

**Declaration of interest**

G.P.S., R.M., and A.K. are co-founders of MooH GmbH, a company developing metagenomic-based oral health tests.

**Figure legends**

**Figure 1, Study setup and data quality: a)** The sampling strategy of the study focuses on the comparison of saliva and stool samples of different zoo animals. Extension with the dataset by Youngblut et al. (35). further allows a comparison to wildlife-derived samples. **b)** Map of the Zoo Saarbrücken with the position of each individual animal species. Co-located animals are encircled in blue. Silhouette-species mappings are elaborated in Figure 1c. Silhouettes were taken from PhyloPic (phylopic.org) **c)** Species included in the study after quality control and introduction of their silhouettes for a large portion of the remaining plots in this study. **d)** Statistics on host-derived read decontamination of the metagenomic samples. For datapoints in green, a species-level genome was available to perform read decontamination. Violet datapoints used a taxonomic close substitute genome instead. The p-value indicates the significance of the two-sided Wilcoxon rank sum test on the relative read loss attributed to host contamination.

**Figure 2, Species-level genome bins: a)** Phylogenetic tree of species-level genome bins as classified by the GTDB-Tk. The colored background of clades indicates class ranks. The innermost ring named *Novel* indicates if the GTDB-Tk found a species-level assignment. The second, third, and fourth rings discuss bin quality by displaying detected rRNAs, tRNAs, and scaffold length distribution respectively. The two outer rings indicate the BGCs that were detected in the respective bins. BGCs are classified by type and by completion. A more richly annotated version of this visualization is available in Extended data Figure 2 **b)** Number of SGBs and BGCs recovered from each sample.

**Figure 3, Reference-based analysis:** Summary statistics on quality, diversity, composition, and compositional similarity of microbiomes. Starting from the left, the taxonomic classification of host animals is displayed. **S**ilhouettes represent the host species and their color represent the different specimen. If multiple replicates were available, multiple pie charts are displayed, where each pie chart indicates the overall quality of the reference-based analysis. Further, diet classification is provided for each species which is consistently used throughout the manuscript. Three diets are being distinguished: herbivore, carnivore, and omnivore. Alpha-diversity of each sample is indicated using the Shannon index, to visualize microbiome complexity. On the rightmost side, hierarchical clustering based on Bray-Curtis distances is displayed. The optimized tanglegramm displays the accordance between taxonomic class and membership based on predicted microbial composition. The edges are colored by the taxonomic class of the host.

**Figure 4, Potential consequences of captivity: a)** FracMinHash dissimilarity between samples within our dataset and the dataset of Youngblut et al. (35). The cross-comparison matches sample pairs as elaborated in Supplementary Table 1. For reference, zoo replicates and their dissimilarity are visualized alongside. **b)** Presence of antimicrobial resistance genes for each of the zoo and wildlife samples classified by antimicrobial compound.

**Extended data figure legends**

**Extended data Figure 1, Ordination analysis:** Two-dimensional embedding of the dataset generated with UMAP computed on FracMinHash dissimilarities. Silhouettes represent the different animals as depicted in Figure 1c. Colors represent the different specimen and diet combinations.

**Extended data Figure 2, SGBs:** Complete version of Figure 2a. Visualization of dereplicated SGBs with color-encoded BGC and class information.

**Extended data Figure 3, Assignment rate:** Relative amount of reads after quality control assigned during taxonomic profiling using GTDB and GTDB extended by our SGBs respectively. The indicated p-value is the statistical significance of a paired two-sided Wilcoxon rank sum test.

**Extended data Figure 4, Mass spectrometry:** Bacterial species identified by performing mass spectrometry after culturing in different mediums. The bacterial taxonomic classes are provided as a tree structure. Counts indicate in how many different animals the associated bacterial species was detected.

**Extended data Figure 5, Differential pathway abundance:** Statistically significant results of the differential pathway abundance analysis after the Benjamini-Hochberg p-value adjustment. During analysis omnivores and carnivores were agglomerated and compared against herbivores.

**Extended data Figure 6, Known BGCs:** Selection of detected BGCs with a >75% similarity to MIBiG annotated clusters. Comparisons are minor adaptations of the figures directly reported by antiSMASH.

**Extended data Figure 7, Study QC Comparison:** Detailed comparison of the matched samples in our dataset and the dataset of Youngblut et al. (35). Two-sided Wilcoxon rank sum tests were performed to estimate statistical significance. **a)** Number of reads after host DNA removal and quality control. **b)** Contig length distribution after metagenomic assembly. **c)** Number of SGBs generated with the two dataset subsets. **d)** Number of BGCs predicted from each initial input sample.

## Supplementary data

**Supplementary data Table 1, Metadata:** Table aggregating metadata of the different samples including animal species, reference genome used for decontamination, suitability of the reference, specimen, relative loss during host decontamination, and the assigned diet label for data analysis.

**Supplementary data Table 2, SGB data:** Information on each dereplicated SGB that matched at least medium MIMAG quality. Displayed information includes rRNA, tRNA, and scaffold counts as well as completeness, contamination, and overall genome size for quality information. Further two classification schemes are provided. First, GTDB lineage as provided by GTDB-Tk is given. Second, a best-matching NCBI taxonomy classification is provided.

**Supplementary data Table 3, Mass spectrometry results:** Aggregated results of the mass spectrometry data. Providing an overview of presence by diet and specimen type for each bacterial species on the first sheet. The second sheet lists each unique bacterial species – zoo sample combination that was detected.

**Supplementary data Table 4, BGC summary:** Overview data of the observed partial and full BGCs including positional, type, and location SGB information.

## References

1. Rup L. 2012. The human microbiome project. Indian J Microbiol 52:315.
2. Integrative HMPRNC. 2014. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe 16:276-89.
3. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, Zhou J, Ni S, Liu L, Pons N, Batto JM, Kennedy SP, Leonard P, Yuan C, Ding W, Chen Y, Hu X, Zheng B, Qian G, Xu W, Ehrlich SD, Zheng S, Li L. 2014. Alterations of the human gut microbiome in liver cirrhosis. Nature 513:59-64.
4. Song H, Yoo Y, Hwang J, Na YC, Kim HS. 2016. Faecalibacterium prausnitzii subspecies-level dysbiosis in the human gut microbiome underlying atopic dermatitis. J Allergy Clin Immunol 137:852-60.
5. Matsha TE, Prince Y, Davids S, Chikte U, Erasmus RT, Kengne AP, Davison GM. 2020. Oral Microbiome Signatures in Diabetes Mellitus and Periodontal Disease. J Dent Res 99:658-665.
6. Becker A, Schmartz GP, Groger L, Grammes N, Galata V, Philippeit H, Weiland J, Ludwig N, Meese E, Tierling S, Walter J, Schwiertz A, Spiegel J, Wagenpfeil G, Fassbender K, Keller A, Unger MM. 2022. Effects of Resistant Starch on Symptoms, Fecal Markers, and Gut Microbiota in Parkinson's Disease - The RESISTA-PD Trial. Genomics Proteomics Bioinformatics 20:274-287.
7. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559-63.
8. Gacesa R, Kurilshikov A, Vich Vila A, Sinha T, Klaassen MAY, Bolte LA, Andreu-Sanchez S, Chen L, Collij V, Hu S, Dekens JAM, Lenters VC, Bjork JR, Swarte JC, Swertz MA, Jansen BH, Gelderloos-Arends J, Jankipersadsing S, Hofker M, Vermeulen RCH, Sanna S, Harmsen HJM, Wijmenga C, Fu J, Zhernakova A, Weersma RK. 2022. Environmental factors shaping the gut microbiome in a Dutch population. Nature 604:732-739.
9. Tun HM, Konya T, Takaro TK, Brook JR, Chari R, Field CJ, Guttman DS, Becker AB, Mandhane PJ, Turvey SE, Subbarao P, Sears MR, Scott JA, Kozyrskyj AL, Investigators CS. 2017. Exposure to household furry pets influences the gut microbiota of infant at 3-4 months following various birth scenarios. Microbiome 5:40.
10. Kuthyar S, Reese AT. 2021. Variation in Microbial Exposure at the Human-Animal Interface and the Implications for Microbiome-Mediated Health Outcome. mSystems 6:e0056721.
11. Aggarwal D, Ramachandran A. 2020. One Health Approach to Address Zoonotic Diseases. Indian J Community Med 45:S6-S8.
12. Aljeldah MM. 2022. Antimicrobial Resistance and Its Spread Is a Global Threat. Antibiotics (Basel) 11.
13. Berglund F, Osterlund T, Boulund F, Marathe NP, Larsson DGJ, Kristiansson E. 2019. Identification and reconstruction of novel antibiotic resistance genes from metagenomes. Microbiome 7:52.

14. Garcia-Gutierrez E, Mayer MJ, Cotter PD, Narbad A. 2019. Gut microbiota as a source of novel antimicrobials. Gut Microbes 10:1-21.

15. Peixoto RS, Voolstra CR, Sweet M, Duarte CM, Carvalho S, Villela H, Lunshof JE, Gram L, Woodhams DC, Walter J, Roik A, Hentschel U, Thurber RV, Daisley B, Ushijima B, Daffonchio D, Costa R, Keller-Costa T, Bowman JS, Rosado AS, Reid G, Mason CE, Walke JB, Thomas T, Berg G. 2022. Harnessing the microbiome to prevent global biodiversity loss. Nat Microbiol 7:1726-1735.

16. Miethke M, Pieroni M, Weber T, Bronstrup M, Hammann P, Halby L, Arimondo PB, Glaser P, Aigle B, Bode HB, Moreira R, Li Y, Luzhetskyy A, Medema MH, Pernodet JL, Stadler M, Tormo JR, Genilloud O, Truman AW, Weissman KJ, Takano E, Sabatini S, Stegmann E, Brotz-Oesterhelt H, Wohlleben W, Seemann M, Empting M, Hirsch AKH, Loretz B, Lehr CM, Titz A, Herrmann J, Jaeger T, Alt S, Hesterkamp T, Winterhalter M, Schiefer A, Pfarr K, Hoerauf A, Graz H, Graz M, Lindvall M, Ramurthy S, Karlen A, van Dongen M, Petkovic H, Keller A, Peyrane F, Donadio S, Fraisse L, et al. 2021. Towards the sustainable discovery and development of new antibiotics. Nat Rev Chem 5:726-749.

17. Behsaz B, Bode E, Gurevich A, Shi YN, Grundmann F, Acharya D, Caraballo-Rodriguez AM, Bouslimani A, Panitchpakdi M, Linck A, Guan C, Oh J, Dorrestein PC, Bode HB, Pevzner PA, Mohimani H. 2021. Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery. Nat Commun 12:3225.

18. Galvan AE, Paul NP, Chen J, Yoshinaga-Sakurai K, Utturkar SM, Rosen BP, Yoshinaga M. 2021. Identification of the Biosynthetic Gene Cluster for the Organoarsenical Antibiotic Arsinothricin. Microbiol Spectr 9:e0050221.

19. Bitschar K, Sauer B, Focken J, Dehmer H, Moos S, Konnerth M, Schilling NA, Grond S, Kalbacher H, Kurschus FC, Gotz F, Krismer B, Peschel A, Schittek B. 2019. Lugdunin amplifies innate immune responses in the skin in synergy with host- and microbiota-derived factors. Nat Commun 10:2730.

20. Martinet L, Naome A, Deflandre B, Maciejewska M, Tellatin D, Tenconi E, Smargiasso N, de Pauw E, van Wezel GP, Rigali S. 2019. A Single Biosynthetic Gene Cluster Is Responsible for the Production of Bagremycin Antibiotics and Ferroverdin Iron Chelators. mBio 10.

21. Alegado RA, King N. 2014. Bacterial influences on animal origins. Cold Spring Harb Perspect Biol 6:a016162.

22. Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J, Mitreva M, Clardy J, Linington RG, Fischbach MA. 2014. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell 158:1402-1414.

23. Aleti G, Baker JL, Tang X, Alvarez R, Dinis M, Tran NC, Melnik AV, Zhong C, Ernst M, Dorrestein PC, Edlund A. 2019. Identification of the Bacterial Biosynthetic Gene Clusters of the Oral Microbiome Illuminates the Unexplored Social Language of Bacteria during Health and Disease. mBio 10.

24. Tang L. 2019. Exploring the chemical space of the human microbiome. Nat Methods 16:1201.

25. Charlop-Powers Z, Pregitzer CC, Lemetre C, Ternei MA, Maniko J, Hover BM, Calle PY, McGuire KL, Garbarino J, Forgione HM, Charlop-Powers S, Brady SF. 2016. Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. Proc Natl Acad Sci U S A 113:14811-14816.

26. Rehner J, Schmartz GP, Kramer T, Keller V, Keller A, Becker SL. 2023. The Effect of a Planetary Health Diet on the Human Gut Microbiome: A Descriptive Analysis. Nutrients 15.

27. Nishida AH, Ochman H. 2018. Rates of gut microbiome divergence in mammals. Mol Ecol 27:1884-1897.

28. Zhong J, Skouloubris S, Dai Q, Myllykallio H, Barbour AG. 2006. Function and evolution of plasmid-borne genes for pyrimidine biosynthesis in Borrelia spp. J Bacteriol 188:909-18.

29. Sodolescu A, Dian C, Terradot L, Bouzhir-Sima L, Lestini R, Myllykallio H, Skouloubris S, Liebl U. 2018. Structural and functional insight into serine hydroxymethyltransferase from Helicobacter pylori. PLoS One 13:e0208850.

30. Escartin F, Skouloubris S, Liebl U, Myllykallio H. 2008. Flavin-dependent thymidylate synthase X limits chromosomal DNA replication. Proc Natl Acad Sci U S A 105:9948-52.

31. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glockner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Genome Standards C, Lapidus A, Meyer F, Yilmaz P, Parks DH, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725-731.

32. Navarro-Munoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 2020. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16:60-68.

33. Vera Pingitore E, Hebert EM, Nader-Macias ME, Sesma F. 2009. Characterization of salivaricin CRL 1328, a two-peptide bacteriocin produced by Lactobacillus salivarius CRL 1328 isolated from the human vagina. Res Microbiol 160:401-8.

34. Wang X, Zhou H, Chen H, Jing X, Zheng W, Li R, Sun T, Liu J, Fu J, Huo L, Li YZ, Shen Y, Ding X, Muller R, Bian X, Zhang Y. 2018. Discovery of recombinases enables genome mining of cryptic biosynthetic gene clusters in Burkholderiales species. Proc Natl Acad Sci U S A 115:E4255-E4263.

35. Youngblut ND, de la Cuesta-Zuluaga J, Reischer GH, Dauser S, Schuster N, Walzer C, Stalder G, Farnleitner AH, Ley RE. 2020. Large-Scale Metagenome Assembly Reveals

Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity. mSystems 5.

36. De Briyne N, Atkinson J, Pokludova L, Borriello SP. 2014. Antibiotics used most commonly to treat animals in Europe. Vet Rec 175:325.

37. Wilhelm MP. 1991. Vancomycin. Mayo Clin Proc 66:1165-70.

38. Depardieu F, Kolbert M, Pruul H, Bell J, Courvalin P. 2004. VanD-type vancomycin-resistant Enterococcus faecium and Enterococcus faecalis. Antimicrob Agents Chemother 48:3892-904.

39. McKessar SJ, Berry AM, Bell JM, Turnidge JD, Paton JC. 2000. Genetic characterization of vanG, a novel vancomycin resistance locus of Enterococcus faecalis. Antimicrob Agents Chemother 44:3224-8.

40. Gudeta DD, Moodley A, Bortolaia V, Guardabassi L. 2014. vanO, a new glycopeptide resistance operon in environmental Rhodococcus equi isolates. Antimicrob Agents Chemother 58:1768-70.

41. Thorpe HA, Booton R, Kallonen T, Gibbon MJ, Couto N, Passet V, Lopez-Fernandez S, Rodrigues C, Matthews L, Mitchell S, Reeve R, David S, Merla C, Corbella M, Ferrari C, Comandatore F, Marone P, Brisse S, Sassera D, Corander J, Feil EJ. 2022. A large-scale genomic snapshot of Klebsiella spp. isolates in Northern Italy reveals limited transmission between clinical and non-clinical settings. Nat Microbiol 7:2054-2067.

42. van Schaik W. 2022. Baas Becking meets One Health. Nat Microbiol 7:482-483.

43. Rehner J, Schmartz GP, Groeger L, Dastbaz J, Ludwig N, Hannig M, Rupf S, Seitz B, Flockerzi E, Berger T, Reichert MC, Krawczyk M, Meese E, Herr C, Bals R, Becker SL, Keller A, Muller R, Consortium I. 2022. Systematic Cross-biospecimen Evaluation of DNA Extraction Kits for Long- and Short-read Multi-metagenomic Sequencing Studies. Genomics Proteomics Bioinformatics 20:405-417.

44. Beghini F, McIver LJ, Blanco-Miguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife 10.

45. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-9.

46. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884-i890.

47. Ewels P, Magnusson M, Lundin S, Kaller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047-8.

48. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824-834.

49. Mikheenko A, Saveliev V, Gurevich A. 2016. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32:1088-90.

50. Vasimuddin M, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems, p 314-324. *In* (ed),

51. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359.

52. Wu YW, Singer SW. 2021. Recovering Individual Genomes from Metagenomes Using MaxBin 2.0. Curr Protoc 1:e128.

53. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 3:836-843.

54. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2022. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics 38:5315-5316.

55. Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res 49:9077-9096.

56. Irber L, Brooks PT, Reiter T, Pierce-Ward NT, Hera MR, Koslicki D, Brown CT. 2022. Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. bioRxiv doi:10.1101/2022.01.11.475838:2022.01.11.475838.

57. McInnes L, Healy J, Melville J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426.

58. Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32-46.

59. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. 2021. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research 50:D785-D794.

60. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217.

61. Dixon P. 2003. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science 14:927-930.

62. Galili T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 31:3718-20.

63. Lin H, Peddada SD. 2020. Analysis of compositions of microbiomes with bias correction. Nat Commun 11:3514.

64. Youngblut ND, Ley RE. 2021. Struo2: efficient metagenome profiling database construction for ever-expanding microbial genome datasets. PeerJ 9:e12198.

65. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

66. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T. 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res 49:W29-W35.

67. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res 49:D412-D419.

68. Burgin J, Ahamed A, Cummins C, Devraj R, Gueye K, Gupta D, Gupta V, Haseeb M, Ihsan M, Ivanov E, Jayathilaka S, Balavenkataraman Kadhirvelu V, Kumar M, Lathi A, Leinonen R, Mansurova M, McKinnon J, O'Cathail C, Pauperio J, Pesant S, Rahman N, Rinck G, Selvakumar S, Suman S, Vijayaraja S, Waheed Z, Woollard P, Yuan D, Zyoud A, Burdett T, Cochrane G. 2023. The European Nucleotide Archive in 2022. Nucleic Acids Res 51:D121-D125.

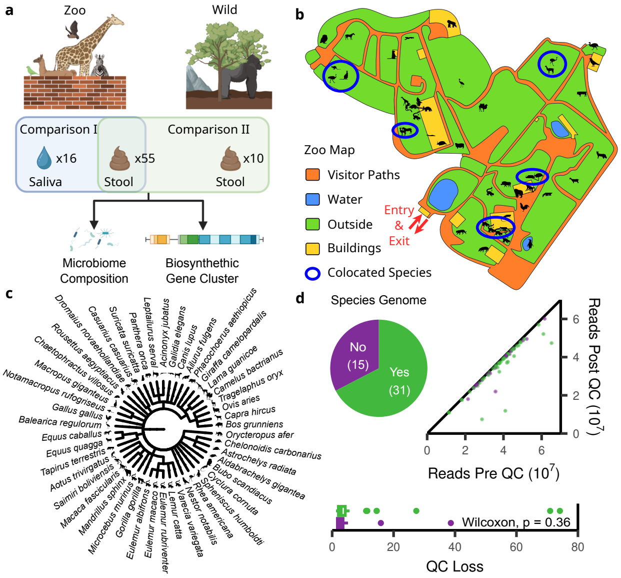


**Figure 1, Study setup and data quality: a)** The sampling strategy of the study focuses on the comparison of saliva and stool samples of different zoo animals. Extension with the dataset by Youngblut et al. (35). further allows a comparison to wildlife-derived samples. **b)** Map of the Zoo Saarbrücken with the position of each individual animal species. Co-located animals are encircled in blue. Silhouette-species mappings are elaborated in Figure 1c. Silhouettes were taken from PhyloPic (phylopic.org) **c)** Species included in the study after quality control and introduction of their silhouettes for a large portion of the remaining plots in this study. **d)** Statistics on host-derived read decontamination of the metagenomic samples. For datapoints in green, a species-level genome was available to perform read decontamination. Violet datapoints used a taxonomic close substitute genome instead. The p-value indicates the significance of the two-sided Wilcoxon rank sum test on the relative read loss attributed to host contamination.

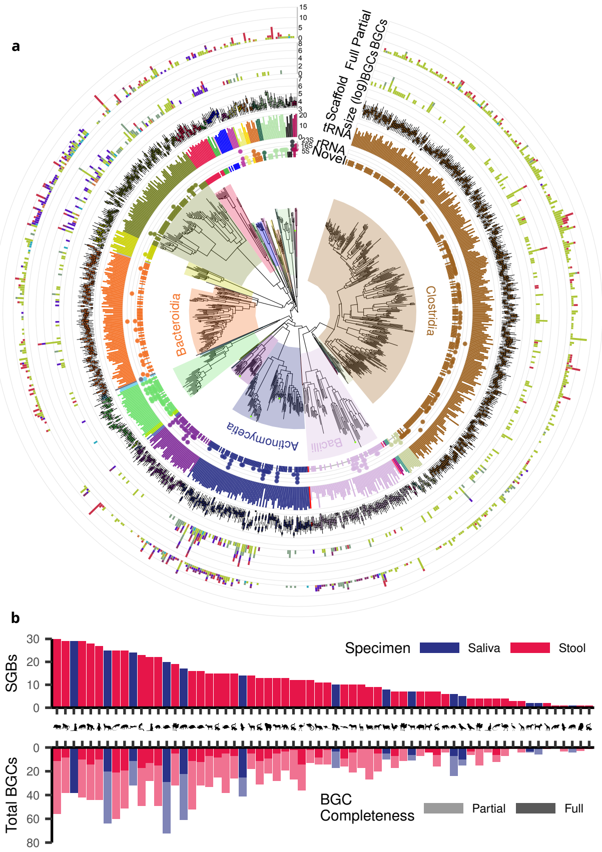


**Figure 2, Species-level genome bins: a)** Phylogenetic tree of species-level genome bins as classified by the GTDB-Tk. The colored background of clades indicates class ranks. The innermost ring named Novel indicates if the GTDB-Tk found a species-level assignment. The second, third, and fourth rings discuss bin quality by displaying detected rRNAs, tRNAs, and scaffold length distribution respectively. The two outer rings indicate the BGCs that were detected in the respective bins. BGCs are classified by type and by completion. A more richly annotated version of this visualization is available in Extended data Figure 2 **b)** Number of SGBs and BGCs recovered from each sample.

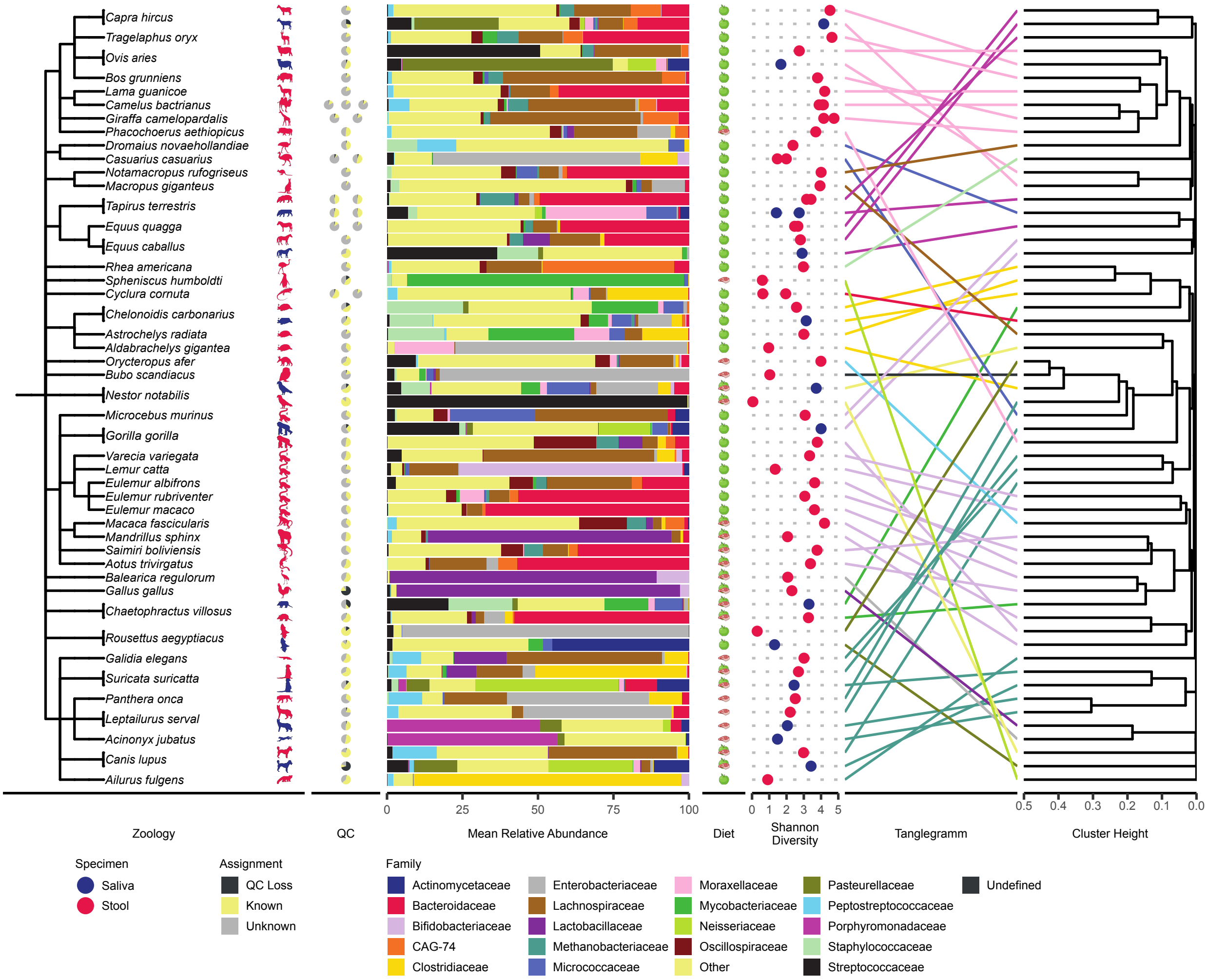


**Figure 3, Reference-based analysis:** Summary statistics on quality, diversity, composition, and compositional similarity of microbiomes. Starting from the left, the taxonomic classification of host animals is displayed. Silhouettes represent the host species and their color represent the different specimen. If multiple replicates were available, multiple pie charts are displayed, where each pie chart indicates the overall quality of the reference-based analysis. Further, diet classification is provided for each species which is consistently used throughout the manuscript. Three diets are being distinguished: herbivore, carnivore, and omnivore. Alpha-diversity of each sample is indicated using the Shannon index, to visualize microbiome complexity. On the rightmost side, hierarchical clustering based on Bray-Curtis distances is displayed. The optimized tanglegramm displays the accordance between taxonomic class and membership based on predicted microbial composition. The edges are colored by the taxonomic class of the host.

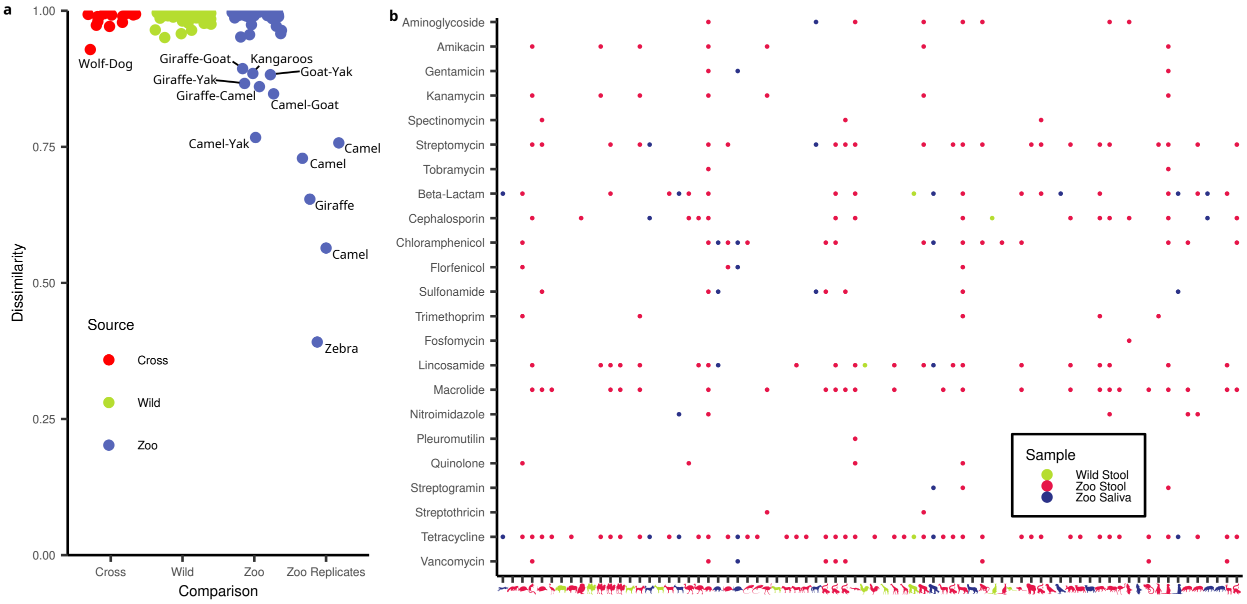


**Figure 4, Potential consequences of captivity: a)** FracMinHash dissimilarity between samples within our dataset and the dataset of Youngblut et al. (35). The cross-comparison matches sample pairs as elaborated in Supplementary Table 1. For reference, zoo replicates and their dissimilarity are visualized alongside. **b)** Presence of antimicrobial resistance genes for each of the zoo and wildlife samples classified by antimicrobial compound.

## 3.9 *Decoding the diagnostic and therapeutic potential of microbiota using pan-body pan-disease microbiomics*

# Decoding the diagnostic and therapeutic potential of microbiota using pan-body pan-disease microbiomics

Georges P. Schmartz[1,#], Jacqueline Rehner[2,#], Madline Gund[3], Stefan Rupf[3,4], Matthias Hannig[3], Tim Berger[5], Elias Flockerzi[5], Berthold Seitz[5], Sara Fleser[6], Sabina Schmitt-Grohé[6], Sandra Kalefack[6], Michael Zemlin[6], Michael Kunz[7], Felix Götzinger[7], Caroline Gevaerd[8], Thomas Vogt[8], Jörg Reichrath[8], Leidy Alejandra Gonzalez Molano[1], Lisa Diehl[1], Anne Hecksteden[9], Tim Meyer[9], Christian Herr[10], Alexey Gurevich[11], Daniel Krug[11], Julian Hegemann[11,12], Kenan Bozhueyuek[11], Olga Kalinina[11], Anouck Becker[13], Marcus Unger[13], Nicole Ludwig[1], Martina Seibert[10], Marie-Louise Stein[10], Nikolas Loka Hanna[10], Marie-Christin Martin[10], Felix Mahfoud[7], Verena Keller[14], Marcin Krawczyk[14], the IMAGINE consortium, Sören L. Becker[2,#], Rolf Müller[11,15,#], Robert Bals[10, 15,#], Andreas Keller[1,11,15,#]

1 Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany
2 Institute of Medical Microbiology and Hygiene, Saarland University, Homburg/Saar, Germany
3 Clinic of Operative Dentistry, Periodontology and Preventive Dentistry, Saarland University, 66421 Homburg, Germany
4 Synoptic Dentistry, Saarland University, 66421 Homburg, Germany
5 Department of Ophthalmology, Saarland University Medical Center, 66421 Homburg, Germany
6 Clinic of Pediatry and Neonatalogy, Saarland University, 66421 Homburg, Germany
7 Department of Internal Medicine III, Cardiology, Angiology, Intensive Care Medicine, Saarland University Hospital, 66421 Homburg, Germany
8 Clinic for Dermatology, Venereology, and Allergology, 66421 Homburg, Germany
9 Institute for Sport and Preventive Medicine, Saarland University, 66123 Saarbrücken, Germany
10 Department of Internal Medicine V - Pulmonology, Allergology, Intensive Care Medicine, Saarland University,
11 Helmholtz Institute for Pharmaceutical Research Saarbrücken, 66123 Saarbrücken, Germany
12 Saarland University, Department of Pharmacy, Campus E8 1, 66123 Saarbrücken, Germany
13 Department for Neurologie, Saarland University Medical Center, 66421 Homburg, Germany
14 Department of Medicine II, Saarland University Medical Center, 66421 Homburg, Germany
15 PharmaScienceHub, 66123 Saarbrücken, Germany
66421 Homburg, Germany.
# Authors contributed equally

## ABSTRACT

The human microbiome emerges as a promising reservoir for diagnostic markers and therapeutics. Given that a human has many microbiomes at various body sites and that diseases do not occur in isolation, a comprehensive analysis strategy is desirable. This strategy should encompass diverse specimen types and various diseases to unveil their intricate potential. To ensure robust data quality and comparability across specimen types and diseases, we employed standardized protocols to generate sequencing data from 1,931 prospectively collected specimens, including from saliva, plaque, skin, throat, eye, and stool, with an average sequencing depth of 5.3 gigabases. Collected from 515 patients, these samples yield an average of 3.7 metagenomes per patient. Our results suggest significant microbial variations across diseases and specimen types, including unexpected anatomical sites. Expanding beyond known species, we identify 729 new species-level genome bins (SGBs) of which 314 are significantly associated with disease. Of note, the existence of microbial resistance genes in one specimen of a patient was indicative of the same resistance genes in other samples of the same patient. Annotated and new SGBs collectively harbor 28,315 potential Biosynthetic Gene Clusters (BGCs), with 1,050 significant correlations to diseases. Our combinatorial approach identifies new SGBs and BGCs, emphasizing the value of pan-body pan-disease microbiomics as a source for diagnostic and therapeutic strategies.

## INTRODUCTION

Recent advancements in the field of microbiota research have spotlighted the complex interplay between non-communicable diseases and the human microbiome, offering novel avenues for understanding disease pathogenesis, identifying biomarkers, and developing new therapies[1-7]. Respective studies demonstrate site-specific changes in microbiota linked to disease development and progression. Large-scale metagenomic initiatives, such as the Human Microbiome Project, International Human Microbiome Consortium[8], the American Gut Project, and other hallmark studies[9-11], have compiled thousands of microbiota samples leading to results arguing all in the very same direction. Data on the impact of an organ-confined disease process in the microbiota of remote body sites remain, however, sparse. Understanding the dynamic relationship between health and disease, especially in the context of frequent co-morbidities, necessitates a holistic exploration of the human microbiome. One example is the discovery of the gut-brain axis, a refined communication between the intestinal microbiota, intestinal host cells, and the central nervous system via the vagal nerve[12]. Overall, the microbiome can be modified at sites remote from the primary disease process[13]. Thus, understanding the composition of metagenomes in health and disease in a systemic manner might be key to improving patient care.

Especially considering the growing health challenges spurred by demographic shifts and social influences, the imperative for enhanced treatment strategies becomes increasingly evident. For example, chronic inflammatory diseases encompass a diverse array of conditions such as periodontitis, chronic obstructive pulmonary disease (COPD), cystic fibrosis, cardiovascular diseases, heart failure, and ulcerative colitis, among others [14-19]. Respective ailments collectively pose a significant burden on individuals and healthcare systems alike. This calls for an in-depth exploration of underlying molecular mechanisms[20,21]. These and other non-communicable diseases are associated with chronic local and systemic inflammation and often occur as multimorbidities. Processes driving the development of multimorbidity are largely unknown but include a state of systemic hyperinflammation, metabolic changes, and senescence. While the underlying mechanisms leading to these (co)morbidities are not yet fully explored, emerging evidence suggests that chronic inflammatory diseases are intrinsically linked to perturbations in the composition and function of the human microbiota[22]. Imbalances in microbial communities residing at various body sites, including the oral cavity, respiratory tract, gastrointestinal tract, and skin, have been implicated in the initiation and perpetuation of inflammatory cascades[23-26]. This has led to the paradigm-shifting realization that the microbiota, once considered a bystander, plays a pivotal role in disease progression and resolution.

The pharmaceutical landscape has long drawn inspiration from nature, with a significant - but reducing - proportion (35% percent) of drugs on the market derived from natural products (NPs) and their producers. In 2021, 50 drugs were approved by the FDA[27]. Of those, 14 approvals represent biologics and 36 small molecules. Only four of the drugs are based on NPs, indicating the need to improve strategies to identify new NPs. These NPs are typically encoded by so-called biosynthetic gene Clusters (BGCs). Considering specific microbial interactions and their underlying chemical processes promises to identify new NPs that remain hidden in classical culture-based studies. In a recent investigation (Salazar et al.), a novel dimension of microbial interactions was unveiled through the examination of the interplay between *Staphylococcus epidermidis* and *S. aureus* isolated from the human nasal region. Although most isolates are classified as commensals, the heightened prevalence of specific strains, notably methicillin-resistant *S. aureus* (MRSA), constitutes a substantial risk factor for severe and frequently fatal infections and finally lead to epifadin as a new antimicrobial compound class.

Highly accurate, fast, and inexpensive next-generation sequencing paired with computational tools like antiSMASH[28,29] are now used at scale in genome mining approaches to identify novel BGCs from microbial communities. As a consequence, important resources such as the BiG-FAM[30] database are increasing in size and complexity, now hosting 1.2 Million BGCs, and being potentially the dominating source for BGCs. However, the annotation of BGCs originating from humans, especially in the context of BGCs present at different body sites, is largely absent, presenting a gap in evidence-based prioritization strategies aimed at identifying BGCs with the highest therapeutic potential. As an initial effort to address this gap, we developed ABC-HuMi[31], a database featuring BGCs identified across various human body sites. Nevertheless, to provide a systematic disease context for diverse specimen types and diseases, data e.g. from meta-analyses might not be sufficient. Here, metagenomes from deeply phenotyped cohorts must be analyzed in a standardized manner.

The aim of this study is to fill current gaps by comprehensively characterizing the pan-body alterations of the microbiome in single-organ disease and multimorbidity. The inclusion of different diseases is crucial in several aspects. The analysis of multiple disorders allows to understand the specificity of different abundances of species and BGCs for single diseases[32]. At the same time, it tremendously improves the definition of a "healthy" or "normal" microbiome. Admittedly, the combination of multiple diseases affecting multiple organs and collecting multiple specimen types per patient bears significant challenges in including patients, measuring the sample with the least possible bias, and of course computational complexity and interpretation. Yet, a respective data set is the basis for identifying new diagnostic and therapeutic strategies as the basis for AI tools with increasing importance in NP development[33].

## RESULTS

### Standardized workflow considering multi-morbid patients

Our study results base on a clinical, experimental, and computational part (**Fig. 1A**). Between 2021 and 2023 we collected a total of 3,483 samples from 657 individuals spanning a broad spectrum of diseases (chronic inflammatory diseases of the lung, heart, eyes, intestine, skin, and oral cavity). Nine different departments at the Saarland University Hospital (dentistry, dermatology, cardiology, gastroenterology, pulmonology, ophthalmology, pediatry, periodontology, and sports medicine) recruited patients. The standardized sample collection and medical assessment across disciplines (e.g., each patient, independently of the enrolling clinics, medical assessment, and oral examination) was a key criterion within the clinical workflow. The enrolling sites transferred the specimens to the Department of Microbiology for biobanking and metagenomic sequencing and stored medical data in a database after manual curation. A respective approach specifically requires standardized protocols for metagenomic sequencing that we previously developed[34]. The computational analysis team obtained both, the raw sequencing data as well as the clinical information after blinded measurement of the metagenomes. We analyzed the microbial composition of multiple body sites, including the oral cavity (saliva, interdental plaque), the skin, the throat swabs, the gastro-intestinal tract (stool), and the eye (conjunctiva swabs), largely matching the affected organs (**Fig. 1B**). The standardized sequencing results and broad and well-curated medical annotations are the basis for the computational analyses, linking known and not yet annotated bacterial species and their repertoire of BGCs to human pathologies, finally representing a unique resource for new NPs.

**Metagenomic sequencing and clinical annotation yield 1,931 high-quality samples**

The perfect world scenario encompasses a complete dataset where for each patient each sample is collected, and each sample results in a high-quality metagenome. The real-world situation presents a more complex scenario. Not each patient consented to collecting all specimen types and not all collected specimens yielded high-quality metagenomes. To avoid bias in analyzing the profiles calls for stringent quality control. From 3,483 samples we obtained 1,931 high-quality metagenomes following quality assessment. These stem from 515 individuals, marking an average of 3.7 microbiomes for each proband (**Supplemental Table 1**). We excluded samples for various reasons, most importantly an insufficient amount of DNA or lacking DNA quality (c.f. Methods, **Fig. 1C**). The sample removal process was unevenly distributed with significantly lower losses of stool, saliva, plaque, and throat specimens and conversely significantly increased loss of skin and eye specimens. For each metagenome surviving the quality control assessment we obtained independent of the specimen type an average of 5.3 gigabases (standard deviation +/- 2.3 gigabases) of metagenomic information after removing ambient human DNA (**Fig. 1D**). Together, the sequencing efforts generated 10.2 terabases of non-human sequencing information. We used this curated and annotated dataset of 1,931 metagenomes and clinical data through the remaining analyses. This dataset excels in that over 90% of all samples belong to a patient with at least three different metagenomes available (**Supplemental Figure 1A**).

As a first analysis, we considered the observed comorbidity pattern of the enrolled individuals (**Fig. 1E**). The majority of 437 individuals were patients, i.e. probands diagnosed with at least one disease. We included further 46 participants as controls, encompassing individuals without known disease affection and an additional 36 competitive athletes. The five most frequent disease entities in the cohort were hypertension (32.4%), diabetes (14.2%), obesity (21.7%), periodontitis (18.8%), and caries (16.7%). We assessed comorbidities, with many patients suffering from two (23.7%) or three (17.4%) concurrent diseases, respectively. The most frequently observed association was diabetes and cardiac failure. Of note, covering a broad spectrum of diseases intrinsically and intentionally leads to a broad age spectrum (**Supplemental Fig. 1B**). Similarly, we observed an uneven gender distribution leaning towards more males than females, with a ratio of 58.1% to 41.9%. Together, these factors led to a slight yet statistically significant gender disparity persisting across different age groups ($\chi 2$ test p-value=0.02). To reach sufficient statistical power we propose a three-tiered ontology framework for organizing our cohort analysis (**Fig. 1F**). By categorizing samples into hierarchical tiers, we ensured to enhance power while acknowledging the potential for increased heterogeneity and confounding factors within the data (**Supplemental Table 2**). We defined the associated cohort of a specimen as the second-level ontology category encompassing all diseases logically linked to the disease itself; for instance, all oral diseases are associated with the plaque specimen. This second ontology level gave the best overall balance between specificity and still having sufficient samples in the respective cohorts. We thus used this level throughout the manuscript and explicitly mention when another ontology level is the basis of a result

### Compositional analysis identifies a complex pattern of microbiome-disease associations

To get insights into the general distribution of metagenomes we computed an embedding based on the MinHash distances (**Fig. 2A**). This embedding highlights a separation of metagenomes with respect to the specimen types, with samples from the oral cavity clustering together and being most different from the stool samples. Samples with lower bacterial and DNA abundance from the skin and the eye clustered between these two groups, however splitting up in two separate clusters. Those two clusters have no common confounding factors such as sequencing batches, age, gender, and others. The only difference is within the duplication rate of reads (Two-tailed unpaired Wilcoxon p-value < $10^{-41}$, **Supplemental Figure 2**). To account for this effect in downstream analyses, we adjusted our models for the duplication rate within samples. Taking the Shannon entropy as a measure for the diversity of the sequenced microbiota, we recognized a broader spread of the specimen types with lower DNA yield (**Fig. 2B**). Within the oral cavity samples throat-derived microbiomes are characterized by a decreased Shannon entropy as compared to saliva and plaque samples. In this regard, the throat samples followed a similar distribution as the stool samples. This observation posed the question of which bacterial species are driving the distances and whether those species are correlated to disease patterns. We thus computed the relative microbiome compositions at the genus level for each specimen type within each disease (**Fig. 2C**). The main driver for differences in microbial composition was the specimen types, prompting us to group the patterns within each type for the different diseases. Across the specimen types, our results suggested varying signatures for the most abundant bacterial genera with the competitive athletes frequently being the most deviating group. Of note, the patterns of the athletes only partially matched the standard control cohort. Especially for the skin deviating microbiome compositions are present. Another obvious shift is a differential stool microbiome composition present in patients suffering from digestive tract disorders. To find significant changes in microbial compositions between disease cohorts and species, we carried out a differential abundance analysis at the species level. Splitting higher and lower abundant species in the different diseases confirms the complex patterns (**Fig. 2D**). For metabolic disorders and heart diseases the largest number of significant species were recorded. Both are marked by a trend towards increased presence of species scattered between the different specimen types. Among the specimen types, the oral cavity had overall the highest shares. Digestion disorders are characterized by a lower abundance of species in the stool samples at an increased frequency in the oral cavity samples. A similar pattern of higher abundance of species in the oral cavity is present in eye diseases. But these are also showing decreased species specifically in the samples taken from the eye. To get a broader overview, we grouped the patterns per specimen type and per disease (**Fig. 2E, Supplemental Table 3**). After adjusting p-values with respect to the number of species present in each specimen type, statistically significant differentially abundant hits emerged in nearly all cohort comparisons. Of note, the effects extended to specimen types not directly linked to the diseases; for example, 63 significant species were identified in the saliva in the case of neural disorders. Overall, for each disease group over 200 significant species are present and saliva samples yielding the largest number of significant hits across the diseases. Remarkable was the small number of significant results between controls and competitive athletes following adjustment for multiple testing. The comorbidity patterns might impact the results, making us split the dataset into those patients with more than one diagnosed disease (**Fig. 2F**) and those with exactly one diagnosed disease (**Fig. 2G**). In the latter case, a reduced number of significant hits remained. Because this might be partially due to the smaller number of patients with exactly one disease and the different number of samples per specimen type might impact the results, we further analyzed the effect sizes as a robust measure in addition to the p-values. The results provided evidence for the validity of the reported patterns in general, however, also demonstrated substantial effects (absolute Cohens D > 0.5) in cases where the regression-based hypothesis test did not yield significance (**Supplemental Figure 3**).

One motivation for including multiple diseases in our study was to define a healthy microbiome. To this end we analyzed the highest hierarchy level in our disease ontology: all patients versus all controls **(Fig. 2H, Supplemental Table 4)**. Our results suggested that *Staphylococcus epidermidis* and *Corynebacterium pseudogenitalium* are significantly more frequent in skin swabs from patients suffering from coronary artery disease. Moreover, we detected a significant increase of *Capnocytophaga gingivalis* in saliva samples from aniridia patients, as well as *Porphyromonas endodontalis* in interdental plaque from obesity patients, and *Lachnoanaerobaculum saburreum* in interdental plaque from aniridia patients. Furthermore, *Streptococcus australis. Lachnospiraceae bacterium oral taxon 096, Streptococcus sp. A12, Streptococcus sp. HMSC067H01,* and *Aggregatibacter aphrophilus* appeared in decreased abundance in saliva in most analyzed diseases. These commensal bacteria are the five species, which across all analyzed diseases displayed a significant decrease. Only two species demonstrated differential abundance across the disease comparisons. Specifically, *Bacteroides cellulosilyticus* exhibited a significant decrease in stool samples from patients with digestive ailments, while *Streptococcus vestibularis* displayed a significant decrease in saliva samples from aniridia patients and an increase in saliva samples from Parkinson's disease patients. Overall, the abundance analysis of known and annotated bacteria yielded significant disease annotations. However, the body still seems to harbor microbes that are not characterized and annotated in databases[35]. This motivated the following analysis of assembled metagenomic data, specifically in the context of existing antimicrobial resistance (AMR) genes.

## AMR analyses suggest pan-microbiome resistance within patients

From the 10.2 terabases of sequencing information, we generated 450 million scaffolds by computing metagenomic assemblies for each sample separately. Following the nature of short-read metagenomic sequencing, metagenomics assemblies yield short contigs. Nonetheless, 19 million fragments exceeded 1 kilobase and 300,000 fragments surpassed the 50 kb mark (**Fig. 3A, Supplemental Table 5**). Intriguingly, long contigs were present in samples from all specimen types, including the skin and the eye. One of the main reasons for metagenomic assemblies is to search for the presence of antimicrobial resistance genes. We thus performed resistance gene profiling based on the assembled fragments (**Fig. 3B**), pinpointing a consistent prevalence of the *mef*(A) resistance gene. This gene was present in 1,484 of 1,931 samples (77%), spanning a set of 493 of 515 distinct individuals (96%).

Overall, our data suggested that when a resistant gene is identified in one specimen of a patient, there is an increased likelihood of detecting the same resistance gene in other specimens from the same patient. (**Fig. 3C**). We report 120 of such significantly associated biospecimen and gene combinations after p-value adjustment (Fisher exact test p-value < 0.05; **Supplemental Table 6**). Particularly noteworthy is the significant prevalence of resistance genes observed within skin samples, which maintain continual contact with the external environment. In this context, our results suggested that the detection of a resistance gene within either the arm or forehead microbiome corresponds to a probability of 16.1% for the arm microbiome and 13.1% for the forehead microbiome to encounter the same resistance gene within the patient's stool. Furthermore, we screened for emerging resistance genes in Gram-negative bacteria against carbapenems and colistin, which display a global health threat. In our data, we detected several *bla-Oxa* genes, encoding various β-Lactamases in *Acinetobacter sp., Klebsiella sp., Pseudomonas sp.,* and more (**Fig. 3B**). Of note, we did not detect the most prevalent *bla-Oxa-48* across the study cohort. In conjunctiva swabs, we however observed the New Delhi metallo β-lactamase-1 (NDM-1) in *Citrobacter* sp. Moreover, the plasmid-mediated resistance to colistin, *mcr-1,* was found in one conjunctiva swab and one stool sample. When correlating the AMR genes to the different cohorts, we found surprisingly few statistically significant hits, suggesting that the presence of resistance genes is similar in cases and controls independent of the specimen type.

### 314 species genome bins associated with diseases

In light of the observed disease associations of known bacteria and the limited association patterns of AMRs in health and disease states, it is reasonable to ask for disease annotations of the not yet annotated bacteria. Therefore, we generated species genome bins (SGBs) and probed their potential links to diseases. Among 4,380 dereplicated SGBs, 729 (16.6%) lacked species-level assignments. Utilizing the available coverage data, we assessed SGB enrichment within cohort-specimen combinations, revealing 10,170 statistically significant combinations with an absolute log fold change exceeding two (**Supplemental Table 7**). Among these combinations, 1,364 involved novel species. Finally, 314 SGB were associated with diseases (absolute log fold change >2 & p-value <0.05). The pattern of disease associations in known and unannotated species, and the limited number of significant AMR genes between patients and diseases suggest other factors that might impact physiological or pathophysiological conditions in the host caused by bacteria. Here, the potential of microbes as natural producers that carry BGCs with broad functional scope must be recognized.

### Coverage-guided genome mining highlights 814 disease-related core biosynthetic genes

Among all metagenomic assembled scaffolds surpassing the 50 kb length threshold, we predicted a total of 28,315 BGCs. To identify pertinent candidates for further examination, we tailored the BigMAP[36] workflow to harmonize with our data strategy (**Fig. 4A**). Employing coverage profiles for each sample-predicted core biosynthetic gene pair, we acquired an informed assessment of whether a BGC exhibited enrichment or depletion within specific disease cohorts. Notably, our observations reveal that numerous BGCs demonstrate remarkable specificity to the originating samples. Nonetheless, upon focusing exclusively on matching specimen-disease cohort pairings, we unveiled a total of 1,050 statistically significant differentially altered coverages following p-value adjustment for the plethora of tested genes (**Fig. 4B**). Further exploring these findings and specifically examining individual BGCs, we searched for potential pathogenic or protective attributes (**Fig. 4C**), including the coverage information (i.e. whether higher or lower coverage within the control cohort was observed). It is crucial to note, however, that this methodology offers insights into the genomic presence but does not encompass the transcriptional activity of BGC genes or the overall concentration of potentially bioactive compounds. Nevertheless, the presented data strongly advocate for prioritizing future investigations into the functionality of the identified BGCs concerning their association with diseases.

**Impact of confounders on the microbiota showcased by the diet of individuals**

As the last aspect of our study, we emphasize the importance of considering confounding factors. Inherent to study designs, such as the one applied in this study, it is a broad spectrum of confounders that might impact the results. Correlations between individuals' early-life breastfeeding experience, gender, and educational attainment in relation to the microbial communities across various body sites exists[37]. Especially the sex has a large impact[38] but also factors such as ethnicity and geography[39-41]. The regional proximity and a largely shared ethnology of individuals in our study account for these factors but other obvious and non-obvious confounders remain, potentially impacting our results. One of those is the diet, that impacts microbial compositions[42]. Because the diet was one of the variables included in our questionnaire, we performed a specific analysis of the diet, testing the dietary information related to the disease context. To this end we also added data from a longitudinal investigation of the planetary health diet on the gut microbiome[43]. In the vegetarian stool cohort, we identified eleven significantly diminished microbial species, including those previously linked with alternative diets. Notably, species like Bifidobacterium animals, *Alistipes inops*, and *Phascolarctobacterium faecium*, known for producing short-chain fatty acids through dietary fiber fermentation, were more abundant in omnivorous participants. In contrast, *Dialister sp. CAG 357*, associated with inflammation, exhibited higher levels in omnivores. With respect to the SGBs, only one hit with respect to the diet remained, derived from plaque: *Saccharimonas sp013333645*. The absence of statistically significant differences in our coverage analysis might be due to the limited number of vegetarians/vegans. Nonetheless, we asked for shared signatures concerning disease correlations. Accounting for limitations in using p-values, we again evaluated effect sizes for each BGC and correlated them with disease effect sizes. In this analysis, we identified negative Spearman coefficients such as -0.35 for coronary artery disease in the forehead skin microbiome, -0.30 for heart diseases in the eye, and -0.28 in the plaque of diabetes patients (**Fig. 4D, Supplemental Table 8**). In sum, our results provide evidence that confounding factors do have an influence on the metagenomic patterns, but despite these relevant factors, disease signals seem to remain.

## DISCUSSION

In our extensive metagenomic sequencing investigation, we analyzed 1,931 samples following rigorous quality control. We maintained a standardized data generation protocol across multiple biospecimen samples obtained from the same individuals. Having the different microbiomes measured from the same patient offers a fairer comparison of differentially abundant microbes between different sample types and disease entities. We deliberately included and categorized a diverse spectrum of dominantly chronic inflammatory diseases, as well as globally widespread diseases. While the standardized sampling strategy and the inclusion of multiple disease cohorts represent a core strength of our study, we acknowledge the challenges that remain. One of those is the impact of obvious and less obvious confounders. The broad spectrum of diseases with different ages of onset leads to a broad age distribution. With an additional gender distribution leaning towards more males than females, with a ratio of 58.1% to 41.9% we have a second confounding factor. Others include concomitant medications, ethnicity, geographic location, and the diet. Of note, respective confounding factors are correlated to each other (e.g. the nutrition is linked to the geographic origin), making the local sampling characteristics an advantage of our study. Further, standardized SoPs add to the stability of the results. To investigate the impact of one confounder in detail, we compared the dietary association with microbiomes in the context of disease associations with the microbiomes. These results suggest that this confounder has an impact on the metagenomes, but that the disease trajectories remained despite this influence.

One major aim of our study was the identification of diagnostic patterns. Indeed, our results suggest a complex pattern of disease-to-microbiome associations depending on the specimen types. We reached the highest diagnostic power from gut and oral cavity samples. Here, the low abundance of DNA in the eye or the skin and smaller cohort sizes might lead to lower overall diagnostic values. Still, several interesting hits remained in those specimen types, especially in the case of acne inversa and the skin. It is important to highlight that all associations discovered in this study need in-depth considerations and validation. Examples of associations include the increased presence of *otrichia sp. oral taxon 225* species in the saliva of patients suffering from Parkinson´s disease*.* Parkinson´s disease remains challenging to diagnose, for which additional testing for biomarkers in easily accessible body fluids, such as saliva, would provide great potential for improved diagnostic procedures[44]. However, the question of what comes first - the microbes or the disease, remains to be solved in functional studies.

Beyond diagnostic associations of microbiota to diseases, one aim of our study was the examination of antimicrobial resistances because we speculate that microbial dark matter carries resistance gene information that needs to be monitored. The most prevalent resistance gene identified across all specimens was *mef(A),* encoding a resistance against macrolide antibiotics. This macrolide efflux gene was first described in 1996 and has emerged rapidly in *Streptococcus* sp. worldwide[45-48]. Therefore, it is not surprising that we observed such a high prevalence in our study cohort. Furthermore, we identified emerging resistance genes against carbapenem and colistin, both used to treat infections with Gram-negative bacteria. These resistances display a global health threat as treatment options are limited. The most prevalent carbapenem resistance genes are related to Oxacillin-hydrolysing (OXA) carbapenemases and New Delhi metallo beta-lactamases (NDM)[49,50]. Colistin had been abandoned for the treatment of Gram-negative infections for many decades but has been reintroduced as a last-resort antibiotic in the last decade. First described in 2011, the plasmid-mediated colistin resistance gene *mcr-1* displays another challenging global health threat, as it spreads rapidly and decreases the options for last-resort antibiotics in case of multi-resistant Gram-negative infections [51,52]. In our study cohort, including 515 patients from southwest Germany, we identified two patients colonized by NDM-1 positive *Citrobacter freundii*, two patients colonized by *mcr-1* carrying *Gammaproteobacteria*, and a variety of OXA mediated resistances against carbapenems. Interesting are *blaOXA-50* carrying *Pseudomonas aeruginosa*, *blaOXA-270* carrying *Acinetobacter pittii,* and *blaOXA-58* carrying *Acinetobacter baumannii*. The carbapenem hydrolyzing activity of blaOXA-50 and blaOXA-270 has neither been confirmed nor denied. The carbapenemase blaOXA-58, however, was first described in 1995 and has spread globally ever since, posing one of the major carbapenem resistance genes in Acinetobacter baumannii[53]. We did not observe any bacterial *blaOXA-48*, which displays a now emerging resistance against carbapenems[54]. Our study setup also allows us to compare resistance genes across different specimens of the same patient. Here, resistance genes on the skin were indicative of carrying the same resistance genes in the gut.

Another important aim was to explore the disease association of biosynthetic gene clusters BGCs. Such BGCs encode for molecular machineries, building natural products that are screened as a source of therapies. Our study setup was thought to enable a prioritization of BGCs with respect to therapeutic potential. By categorizing the data into distinct cohorts at various disease ontology levels, we identified BGCs that exhibited differential abundance and coverage. Beyond potential pathogenic species markers, we uncovered benign BGCs that displayed heightened coverage in healthy control groups. These BGCs warrant further exploration *in vitro,* offering promising avenues for medical discoveries, including the potential development of antibiotic compounds. As a next step, we plan to thoroughly investigate these promising BGCs for their potential beneficial properties.

## METHODS

**Clinical sampling:** Clinical samples were obtained from study participants after having obtained written informed consent at Saarland University Medical Center in Homburg, Germany. Approval for the study was granted by the ethics committee of the local medical association (Ärztekammer des Saarlandes) with the identification number 131/20. Per patient, an extensive medical examination was conducted by medical staff and diseases of interest for this study were identified. If such a disease was present, an in-depth medical history was obtained, including major factors that might influence microbial compositions in and on the human body, such as medication, lifestyle choices pertaining to diet, activity, smoking and for example alcohol uptake, as well as co-morbidities. After, clinical samples were obtained: saliva, interdental plaque, conjunctiva swab, throat swab, stool, and skin swabs of the forehead and arm region, as well as in case of Acne inversa – affected skin areas. Concisely, fecal samples were procured from participants through utilization of a paper toilet-hat and a sterile collection tube complete with an integrated spoon, yielding an approximate range of 500 mg to 1 g of stool. Plaque samples were gathered through the use of twelve disposable micro applicators (Catalog No. MSF400, Microbrush International, Grafton, WI). Each quadrant involved brushing three interdental spaces, and subsequent transfer of all micro applicators to an ESwab transport tube (Copan Diagnostics, Brescia, Italy) along with ESwab Amies Medium (Copan Diagnostics). Saliva samples were obtained using 50-ml sterile, conic falcon tubes. Participants were instructed to deposit unstimulated saliva into the sterile falcon tube for a duration of 5 minutes. Conjunctiva specimens were acquired utilizing an ESwab. Eversion of the lower eyelid facilitated swabbing along the complete extent of the lower fornix thrice. Throat swabs were taken cautiously by medical staff, avoiding contact to saliva, tonsils, gums, and teeth, by streaking the throat-area 3-5 times. The forehead and arm were swabbed using an ESwab and wetting the nylon-flocked swab with the provided Amies Medium prior to contact with the skin. The areas were swabbed roughly to ensure removing bacterial mass not just from the surfaces, but also from hair follicles for example. Subsequent placement of the swabs into the designated transport medium was followed by freezing the tube at -80 °C.

**DNA extraction:** DNA was extracted from all native samples using the Qiagen QiAamp Microbiome Kit (Qiagen, Hilden, Germany). The DNA was extracted according to the manufacturer’s protocol. Briefly, swabs were vortexed in 1.5 ml Amies Medium for 2 minutes. The Amies medium containing the microbial mass from each sample was then used for DNA extraction according to the manufacturer’s recommendation. For fecal samples, 250 mg of stool was used and mixed with 500 µl of buffer AHL. Micro applicators used to collect interdental plaque were mixed with 1 ml 1x PBS (pH = 7.4) and vortexed rigorously for 2 minutes. 1 ml PBS containing bacterial cells was then used for DNA extraction. Saliva samples were vortexted briefly to allow homogenization of the sample. Then, 1 ml of saliva was used for DNA extraction according to the manufacturer`s recommendations. Utilizing the MP Biomedicals™ FastPrep-24™ 5G Instrument (FisherScientific GmbH, Schwerte, Germany), mechanical disruption of bacterial cells was carried out. The operational parameters were set to a velocity of 6.5 m/s for 45 seconds, executed twice, with intervals of 5 minutes of ice storage separating each lysis cycle. After the lysis procedure, DNA was extracted into 50 µl elution buffer. To determine the DNA concentration, comprehensive microvolume UV-Vis measurements were performed using the NanoDrop 2000/2000c (ThermoFisher Scientific, Wilmington, DE)[34].

**Library Preparation, sequencing, and quality control:** Extracted DNA from all native samples was sent to Novogene Company Limited (Cambridge, UK) for metagenomic library preparation and subsequent paired-end (PE150) Illumina sequencing (HiSeq). From the study we excluded sparsely collected biospecimens (n=47), substandard (n=1304), and anomalous samples (n=201).

**Next-generation sequencing data preprocessing:** The first step of data analysis was host read removal with KneadData [55] (version (v): 0.7.4; command line arguments (cla): “--trimmomatic-options=’LEADING:3 TRAILING:3 MINLEN:50’ --bowtie2-options=’--very-sensitive --no-discordant --reorder’”). Due to the high contamination load among skin and eye samples, we additionally ran the human sra-human-scrubber [56] (v: 1.0.2021_05_05) after KneadData. Paired-end reads were only kept if none of the read pairs mapped to the human reference. After decontamination, we performed sequence overrepresentation analysis and quality assurance with fastp [57] (v: 0.20.1; cla: "--overrepresentation_analysis") and visualized results with MultiQC [58] (v: 1.11).

**Reference-free analysis and outlier removal:** Mash [59] (v: 2.3; cla: "sketch -S 23 -k 31 -s 5000 -r -m 2") was used to compute and compare MinHash distances. A two-dimensional embedding was generated in R with the UMAP package [60] (v: 0.2.8). After noticing the separation of the low-input biospecimen into two clusters, we split the low-input samples by their clustering behavior during outlier removal. During this outlier removal, we performed for each biospecimen a Grubb's test on the mean of all pairwise MinHash distances and removed the most significant outlier. This procedure was repeated iteratively until no more significant outliers were left. In the case of low-input biospecimen, this algorithm was performed for each subcluster instead.

**Reference-based compositional analysis:** MetaPhlAn3 [55] (v: 3.0.13; cla: "-t rel_ab_w_read_stats --unknown_estimation –add_viruses") on the mpa_v30_CHOCOPhlAn_201901 database was used to profile quality controlled samples. Relative counts were rescaled to absolute counts based on the number of reads and virus counts were removed. Shannon diversity was used as an alpha-diversity measure. Reference-based beta-diversity was assessed with non-metric multidimensional scaling on Bray-Curtis distances. Differential abundance analysis was performed with ANCOM-BC [61] (v: 1.6.2). P-values were adjusted via Benjamini-Hochberg adjustment. Note, that we only tested specimen-cohort combinations with more than ten samples in each cohort. While the athletic and sports cohorts were tested against healthy controls, all other diseases were tested against the union of healthy control and sports cohorts, i.e. the healthy cohort. Samples that were part of the control and disease cohort during testing, such as athletes with diseases, were removed. Based on the differential abundance analysis results, we searched for interesting pathogens and commensal bacteria to further investigate. To this end, we defined the pathogenicity score as the number of various diseases where a pathogen is predicted to be significantly more abundant in the diseased cohort in at least one biospecimen. Similarly, we define the commensal score as the number of disease cohorts where a bacterial species is significantly reduced in the diseased cohort for at least one biospecimen. In visualizations of abundances, absolute counts were center log ratio normalized.

**Metagenomic Assembly:** We assembled each sample with SPAdes [62] (v: 3.15.4; cla: "--meta") and monitored assembly quality with QUAST [63] (v: 5.0.2; cla: "-s"). Scaffolds were binned with MetaBAT2 [64] (v: 2.15; cla: "--seed 420"). MAGs across all samples were aggregated and dereplicated with dRep [65] (v: 3.4.2; cla: "-comp 50 -con 5 --checkM_method lineage_wf --S_algorithm fastANI --S_ani 0.95 -nc 0.5"). GTDB-TK [66] (v: 2.3.0; cla: "classify_wf") in combination with GTDB database release 214 [67] was used to annotate SGBs with taxonomic information. On each dereplicated SGB, we ran PathoFact [68] (commit v: 55d8240). To capture resistances that did not end up in any final bins, we ran AMRFinderPlus [69] (v: 3.11.4) and complemented the information with Kraken 2 [70] (v:2.1.2; cla: "--use-mpa-style"; database version: k2-pluspf_20220908) taxonomic classifications for each contig. Differential coverage analysis on SGBs was performed by first, aggregating all bins creating one reference file, and then aligning all samples against the newly created reference file with Bowtie2 [71] (v: 2.3.4.3; cla: "-a"). Afterward, coverage information was extracted from each alignment with SAMtools idxstats [72] (v: "1.16.1"). Differential coverage analysis was organized identically to the differential abundance analysis, i.e. cohort-specimen multiplicity needed to be larger than ten and the control cohorts altered if sports or athletes were considered for testing. However, here, a Wilcoxon rank sum test was executed, performing Benjamini-Hochberg adjustment to adjust for the total number of tested dereplicated SGBs.

**Genome mining:** After assembly, all contigs with a length > 50,000 bp were mined for BGCs with antiSMASH [73] (v: 6.1.1; cla: "--genefinding-tool prodigal --cb-knownclusters --cb-subclusters –asf"). Next, all core biosynthetic genes that were annotated by antismash were extracted and aggregated into one file. Reads of all samples were then aligned against this reference and coverage information was extracted for each contig-sample combination, following the procedure described for the SGB analysis. Similarly, differential coverage analysis was repeated identically to the SGB analysis. P-value adjustment was performed using Benjamini-Hochberg adjustment, adjusting for the number of tested core biosynthetic genes.

**Dietary alterations comparison:** All day zero samples from the dataset of Rehner et al. [43] were taken and processed according to the new data. The dataset was then extended by our dataset, however, only using our healthy controls. The dataset was from then on analyzed independently. If comparisons were performed, e.g. in statistical tests, the vegan/vegetarian cohort was always compared to the omnivore cohort.

**Data Availability:** All sequencing data are freely available from the sequence read archive (SRA). Please note that only data after removing ambient human DNA can be made available.

**Acknowledgements:** Our appreciation extends to all individuals associated with Saarland University, Saarland University Medical Center, and Helmholtz Institute for Pharmaceutical Research Saarbrücken, whose contributions significantly enriched this study, albeit not reflected in authorship. Furthermore, we extend our gratitude to the participants of this study, whose consent allowed us to acquire and scrutinize clinical samples, thereby unravelling insights into their intricate microbial composition and functions.

**Funding:** This work was further supported financially by the Saarland University and the UdS-HIPS TANDEM initiative. The compute infrastructure for this project was funded by the DFG [469073465].

**Contributors:** RM, SLB, AK, RB supervised the research and obtained research funding. SLB, AK, GPS and JR designed the experiments. JR performed the experiments, and together with GPS analysed the data. GPS designed the computational analysis. LAGM added to the interpretation of results. MH, BS, SK, FM, VK, MK, TM, TB, EF, BS, SF, SSG, SK, MZ, FG, JR, TV, CH, MS, MLS, MU, AB, NLH, MCM, and RB supervised clinical sample collection and medical assessment of participants and contributed to the medical and experimental workflow. MG, SR, TB, EF, SF, SSG, MK, FG, VK and AH collected clinical samples and performed medical assessment of participants health. JR, GPS and AK wrote the paper. All authors reviewed and edited the paper.

**Declaration of competing interest:** G.P.S., R.M., and A.K. are co-founders of MooH GmbH, a company developing metagenomic based oral health tests. FM is supported by Deutsche Gesellschaft für Kardiologie (DGK), Deutsche

Forschungsgemeinschaft (SFB TRR219, Project-ID 322900939), and Deutsche Herzstiftung. His institution (Saarland University) has received scientific support from Ablative Solutions, Medtronic and ReCor Medical. He has received speaker honoraria/consulting fees from Ablative Solutions, Amgen, Astra-Zeneca, Bayer, Boehringer Ingelheim, Inari, Medtronic, Merck, ReCor Medical, Servier, and Terumo.

## Figure Legends

**Figure 1, Study set up, metagenomics data and clinical information: A)** Schematic Workflow describing the sample (upper arrow) and data flow (lower arrow) between clinicians, microbiology, and data science. The clinical data were kept separated from the measurement of microbiomes and only combined after measurement in the computational analysis. **B)** Clinical sampling was focused on seven biospecimens (left blue part). We included patients from a wide range of clinical diseases that allows us analyzing the diagnostic potential of different specimen types across diseases. **C)** Sankey plot for the number of samples included in the study at different intervals of the data generation process in relation to our quality control strategy. Specimen types are ordered vertically at each step in the pipeline by frequency of the respective specimen. **D)** Number of reads for each sample colored by specimen. The horizontal line represents the 5 gigabase threshold at a paired-end read length of 150bp. **E)** Pruned upset plot displaying the most frequent co-occurrence of diseases within the dataset. The combinations are ordered with decreasing frequency, marking the combination of Hypertension and obesity as most common comorbidity in our study. **F)** Ontology used throughout the study grouping diseases by biological systems and separating healthy control from diseased patients. Areas are proportional to the number of patients falling into each category. Patients may be represented multiple times if multiple diseases are diagnosed.

**Figure 2, Compositional analysis, and link of microbiota to diseases: A)** Two-dimensional UMAP embedding of pairwise computed mash distances, colored by biospecimen of the sample **B)** Alpha-diversity of all samples, colored by specimen. As a measure of species richness, we selected the Shannon diversity. **C)** Relative genus abundance for each cohort of the second ontology level, divided by biospecimen. **D)** Sorted log-fold changes of differentially abundant species matching the visualized results of the previous panel. Each panel is split vertically separating positive and negative log-fold changes. **E-G)** Number of differentially abundant species after p-value adjustment revealed during analysis across all cohorts and specimen combinations. Numbers in the circles represent the number of specimens included in the respective analysis. **H)** Center-log ratio normalized abundance counts of selected species-cohort-specimen combinations. The visualized diseased cohort is indicated by the text above each panel, whereas the selected biospecimen is indicated by the color of the writing. The first row of panels displays potential pathogen candidates with the highest statistical significance and a pathogen score of one. The second row of panels displays saliva samples of commensal bacteria candidates with a commensal score larger than eighteen.

**Figure 3, Assembly and resistance gene analysis: A)** Distribution of the number of scaffolds in each sample at various length limits, colored by specimen as box-whisker plot. **B)** Sequence of pie charts indicating the presence of emerging antimicrobial resistance genes. Panels are subdivided by genus that was assigned to the contig where resistance genes have been detected. Pie charts scale with the number of measurements in different samples and are colored by the relative frequency of the sample's biospecimen. **C)** Network visualization of counts of shared resistance genes among different biospecimen samples derived from the same patient. Note, any resistance gene annotated by AMRFinderPlus was used for this plot. **D)** Dereplicated SGBs defined from our data. Visualized information includes biospecimen of the initial sample where the SGB was derived from, selected resistance information taken from Pathofact, and effect size of differential coverage analysis for selected cohorts. Note, the visualized differential coverage focuses only on the biospecimen of the initial sample where the SGB has been defined from that is also visualized in the central ring.

**Figure 4, Evidence-supported genome mining and disease association: A)** Schematic representation of our proposed BGC prioritization strategy representing an adapted version of the BiGMAP workflow. Metagenomic assembly is performed for each sample, followed by BGC prediction. Next, all samples are aligned against all core biosynthetic genes of predicted BGCs. Coverage information is extracted, and downstream analysis is performed. **B)** Volcano plot of the differential BGC coverage analysis results. In this visualization, only matching biospecimen – initial BGC contig combinations are visualized, constituting only a fraction of all results. **C)** Randomly drawn BGC examples of each region of the previous volcano plot. The visualized coverage includes all biospecimens. **D)** Comparison of the highest correlating effect sizes, comparing differential BGC coverage results between alternative diets and diseases. The effect size of the vegetarian-omnivore comparison is visualised on the y-axis. On the x-axis, the cohort named above the panel is compared against the healthy cohort. For the fourth panel, the minimum effect size across all cohort comparisons is taken for each BGC and compared against the diet comparison.

## References

1 Potrykus, M., Czaja-Stolc, S., Stankiewicz, M., Kaska, L. & Malgorzewicz, S. Intestinal Microbiota as a Contributor to Chronic Inflammation and Its Potential Modifications. *Nutrients* **13**, doi:10.3390/nu13113839 (2021).

2 Kahrstrom, C. T., Pariente, N. & Weiss, U. Intestinal microbiota in health and disease. *Nature* **535**, 47, doi:10.1038/535047a (2016).

3 Becker, A. *et al.* Effects of Resistant Starch on Symptoms, Fecal Markers, and Gut Microbiota in Parkinson's Disease - The RESISTA-PD Trial. *Genomics Proteomics Bioinformatics* **20**, 274-287, doi:10.1016/j.gpb.2021.08.009 (2022).

4 Puschhof, J. & Elinav, E. Human microbiome research: Growing pains and future promises. *PLoS Biol* **21**, e3002053, doi:10.1371/journal.pbio.3002053 (2023).

5 Katsanos, A. H. *et al.* in *Biomarkers for Endometriosis: State of the Art* (ed Thomas D'Hooghe) 41-75 (Springer International Publishing, 2017).

6 Hajjo, R., Sabbah, D. A. & Al Bawab, A. Q. Unlocking the Potential of the Human Microbiome for Identifying Disease Diagnostic Biomarkers. *Diagnostics (Basel)* **12**, doi:10.3390/diagnostics12071742 (2022).

7 Li, M. *et al.* Performance of Gut Microbiome as an Independent Diagnostic Tool for 20 Diseases: Cross-Cohort Validation of Machine-Learning Classifiers. *Gut Microbes* **15**, 2205386, doi:10.1080/19490976.2023.2205386 (2023).

8 Integrative, H. M. P. R. N. C. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16**, 276-289, doi:10.1016/j.chom.2014.08.014 (2014).

9 Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174-180, doi:10.1038/nature09944 (2011).

10 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).

11 Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-546, doi:10.1038/nature12506 (2013).

12 Fan, Y. *et al.* The gut microbiota contributes to the pathogenesis of anorexia nervosa in humans and mice. *Nature Microbiology* **8**, 787-802, doi:10.1038/s41564-023-01355-5 (2023).

13 Worby, C. J. *et al.* Longitudinal multi-omics analyses link gut microbiome dysbiosis with recurrent urinary tract infections in women. *Nature Microbiology* **7**, 630-639, doi:10.1038/s41564-022-01107-x (2022).

14 Wang, R., Li, Z., Liu, S. & Zhang, D. Global, regional and national burden of inflammatory bowel disease in 204 countries and territories from 1990 to 2019: a systematic analysis based on the Global Burden of Disease Study 2019. *BMJ Open* **13**, e065186, doi:10.1136/bmjopen-2022-065186 (2023).

15 Redondo, M., Keyt, H., Dhar, R. & Chalmers, J. D. Global impact of bronchiectasis and cystic fibrosis. *Breathe (Sheff)* **12**, 222-235, doi:10.1183/20734735.007516 (2016).

16 Bhattacharya, S., Heidler, P. & Varshney, S. Incorporating neglected non-communicable diseases into the national health program-A review. *Front Public Health* **10**, 1093170, doi:10.3389/fpubh.2022.1093170 (2022).

17 Pakdin, M., Zarei, L., Bagheri Lankarani, K. & Ghahramani, S. The cost of illness analysis of inflammatory bowel disease. *BMC Gastroenterol* **23**, 21, doi:10.1186/s12876-023-02648-z (2023).

18 Zannetos, S., Zachariadou, T., Zachariades, A., Georgiou, A. & Talias, M. A. The economic burden of adult asthma in Cyprus; a prevalence-based cost of illness study. *BMC Public Health* **17**, 262, doi:10.1186/s12889-017-4184-0 (2017).

19 Cortaredona, S. & Ventelou, B. The extra cost of comorbidity: multiple illnesses and the economic burden of non-communicable diseases. *BMC Med* **15**, 216, doi:10.1186/s12916-017-0978-2 (2017).

20 Ramesh, S. & Kosalram, K. The burden of non-communicable diseases: A scoping review focus on the context of India. *J Educ Health Promot* **12**, 41, doi:10.4103/jehp.jehp_1113_22 (2023).

21 Bloom, D. E. *et al.* The global economic burden of noncommunicable diseases. (Program on the Global Demography of Aging, 2012).

22 Campbell, C. *et al.* Crosstalk between Gut Microbiota and Host Immunity: Impact on Inflammation and Immunotherapy. *Biomedicines* **11**, doi:10.3390/biomedicines11020294 (2023).

23 Giordano-Kelhoffer, B. *et al.* Oral Microbiota, Its Equilibrium and Implications in the Pathophysiology of Human Diseases: A Systematic Review. *Biomedicines* **10**, doi:10.3390/biomedicines10081803 (2022).

24 Bjerre, R. D. *et al.* Skin dysbiosis in the microbiome in atopic dermatitis is site-specific and involves bacteria, fungus and virus. *BMC Microbiol* **21**, 256, doi:10.1186/s12866-021-02302-2 (2021).

25 Santana, P. T., Rosas, S. L. B., Ribeiro, B. E., Marinho, Y. & de Souza, H. S. P. Dysbiosis in Inflammatory Bowel Disease: Pathogenic Role and Potential Therapeutic Targets. *Int J Mol Sci* **23**, doi:10.3390/ijms23073464 (2022).

26 Yang, D., Xing, Y., Song, X. & Qian, Y. The impact of lung microbiota dysbiosis on inflammation. *Immunology* **159**, 156-166, doi:10.1111/imm.13139 (2020).

27 de la Torre, B. G. & Albericio, F. The Pharmaceutical Industry in 2021. An Analysis of FDA Drug Approvals from the Perspective of Molecules. *Molecules* **27**, doi:10.3390/molecules27031075 (2022).

28 Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res* **51**, W46-W50, doi:10.1093/nar/gkad344 (2023).

29 Blin, K., Shaw, S., Medema, M. H. & Weber, T. The antiSMASH database version 4: additional genomes and BGCs, new sequence-based searches and more. *Nucleic Acids Res*, doi:10.1093/nar/gkad984 (2023).

30 Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res* **49**, D490-D497, doi:10.1093/nar/gkaa812 (2021).

31 Hirsch, P. *et al.* ABC-HuMi: the Atlas of Biosynthetic Gene Clusters in the Human Microbiome. *Nucleic Acids Res*, doi:10.1093/nar/gkad1086 (2023).

32 Keller, A. *et al.* Toward the blood-borne miRNome of human diseases. *Nature Methods* **8**, 841-843, doi:10.1038/nmeth.1682 (2011).

33 Mullowney, M. W. *et al.* Artificial intelligence for natural product drug discovery. *Nat Rev Drug Discov* **22**, 895-916, doi:10.1038/s41573-023-00774-7 (2023).

34 Rehner, J. *et al.* Systematic Cross-biospecimen Evaluation of DNA Extraction Kits for Long- and Short-read Multi-metagenomic Sequencing Studies. *Genomics Proteomics Bioinformatics* **20**, 405-417, doi:10.1016/j.gpb.2022.05.006 (2022).

35 Kowarsky, M. *et al.* Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proceedings of the National Academy of Sciences* **114**, 9623-9628, doi:10.1073/pnas.1707009114 (2017).

36 Pascal Andreu, V. *et al.* BiG-MAP: an Automated Pipeline To Profile Metabolic Gene Cluster Abundance and Expression in Microbiomes. *mSystems* **6**, e0093721, doi:10.1128/mSystems.00937-21 (2021).

37 Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the human body. *Nature* **509**, 357-360, doi:10.1038/nature13178 (2014).

38 Kim, Y. S., Unno, T., Kim, B. Y. & Park, M. S. Sex Differences in Gut Microbiota. *World J Mens Health* **38**, 48-60, doi:10.5534/wjmh.190009 (2020).

39 Deschasaux, M. *et al.* Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nature Medicine* **24**, 1526-1531, doi:10.1038/s41591-018-0160-1 (2018).

40 He, Y. *et al.* Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature Medicine* **24**, 1532-1535, doi:10.1038/s41591-018-0164-x (2018).

41 Gaulke, C. A. & Sharpton, T. J. The influence of ethnicity and geography on human gut microbiome composition. *Nature Medicine* **24**, 1495-1496, doi:10.1038/s41591-018-0210-8 (2018).

42 Singh, R. K. *et al.* Influence of diet on the gut microbiome and implications for human health. *J Transl Med* **15**, 73, doi:10.1186/s12967-017-1175-y (2017).

43 Rehner, J. *et al.* The Effect of a Planetary Health Diet on the Human Gut Microbiome: A Descriptive Analysis. *Nutrients* **15**, doi:10.3390/nu15081924 (2023).

44 Tolosa, E., Garrido, A., Scholz, S. W. & Poewe, W. Challenges in the diagnosis of Parkinson's disease. *Lancet Neurol* **20**, 385-397, doi:10.1016/S1474-4422(21)00030-2 (2021).

45 Clancy, J. *et al.* Molecular cloning and functional analysis of a novel macrolide-resistance determinant, mefA, from Streptococcus pyogenes. *Mol Microbiol* **22**, 867-879, doi:10.1046/j.1365-2958.1996.01521.x (1996).

46 Daly, M. M., Doktor, S., Flamm, R. & Shortridge, D. Characterization and prevalence of MefA, MefE, and the associated msr(D) gene in Streptococcus pneumoniae clinical isolates. *J Clin Microbiol* **42**, 3570-3574, doi:10.1128/JCM.42.8.3570-3574.2004 (2004).

47 Ardanuy, C. *et al.* Distribution of subclasses mefA and mefE of the mefA gene among clinical isolates of macrolide-resistant (M-phenotype) Streptococcus pneumoniae, viridans group streptococci, and Streptococcus pyogenes. *Antimicrob Agents Chemother* **49**, 827-829, doi:10.1128/AAC.49.2.827-829.2005 (2005).

48 Harimaya, A. *et al.* High prevalence of erythromycin resistance and macrolide-resistance genes, mefA and ermB, in Streptococcus pneumoniae isolates from the upper respiratory tracts of children in the Sapporo district, Japan. *J Infect Chemother* **13**, 219-223, doi:10.1007/s10156-007-0528-5 (2007).

49 Codjoe, F. S. & Donkor, E. S. Carbapenem Resistance: A Review. *Med Sci (Basel)* **6**, doi:10.3390/medsci6010001 (2017).

50 Khan, A. U., Maryam, L. & Zarrilli, R. Structure, Genetics and Worldwide Spread of New Delhi Metallo-beta-lactamase (NDM): a threat to public health. *BMC Microbiol* **17**, 101, doi:10.1186/s12866-017-1012-8 (2017).

51 Wang, R. *et al.* The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat Commun* **9**, 1179, doi:10.1038/s41467-018-03205-z (2018).

52 Gregoire, N., Aranzana-Climent, V., Magreault, S., Marchand, S. & Couet, W. Clinical Pharmacokinetics and Pharmacodynamics of Colistin. *Clin Pharmacokinet* **56**, 1441-1460, doi:10.1007/s40262-017-0561-1 (2017).

53 Coelho, J., Woodford, N., Afzal-Shah, M. & Livermore, D. Occurrence of OXA-58-like carbapenemases in Acinetobacter spp. collected over 10 years in three continents. *Antimicrob Agents Chemother* **50**, 756-758, doi:10.1128/AAC.50.2.756-758.2006 (2006).

54 Fursova, N. K. *et al.* The spread of bla OXA-48 and bla OXA-244 carbapenemase genes among Klebsiella pneumoniae, Proteus mirabilis and Enterobacter spp. isolated in Moscow, Russia. *Ann Clin Microbiol Antimicrob* **14**, 46, doi:10.1186/s12941-015-0108-y (2015).

55 Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, doi:10.7554/eLife.65088 (2021).

56 Katz, K. S. *et al.* STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biol* **22**, 270, doi:10.1186/s13059-021-02490-0 (2021).

57 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

58 Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048, doi:10.1093/bioinformatics/btw354 (2016).

59 Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132, doi:10.1186/s13059-016-0997-x (2016).

60 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

61 Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat Commun* **11**, 3514, doi:10.1038/s41467-020-17041-7 (2020).

62 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824-834, doi:10.1101/gr.213959.116 (2017).

63 Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088-1090, doi:10.1093/bioinformatics/btv697 (2016).

64 Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359, doi:10.7717/peerj.7359 (2019).

65 Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* **3**, 836-843, doi:10.1038/s41564-018-0171-1 (2018).

66 Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315-5316, doi:10.1093/bioinformatics/btac672 (2022).

67 Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* **50**, D785-D794, doi:10.1093/nar/gkab776 (2021).

68 de Nies, L. *et al.* PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* **9**, 49, doi:10.1186/s40168-020-00993-9 (2021).

69 Feldgarden, M. *et al.* AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci Rep* **11**, 12728, doi:10.1038/s41598-021-91456-0 (2021).

70 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257, doi:10.1186/s13059-019-1891-0 (2019).

71 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

72 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, doi:10.1093/gigascience/giab008 (2021).

73 Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* **49**, W29-W35, doi:10.1093/nar/gkab335 (2021).

844

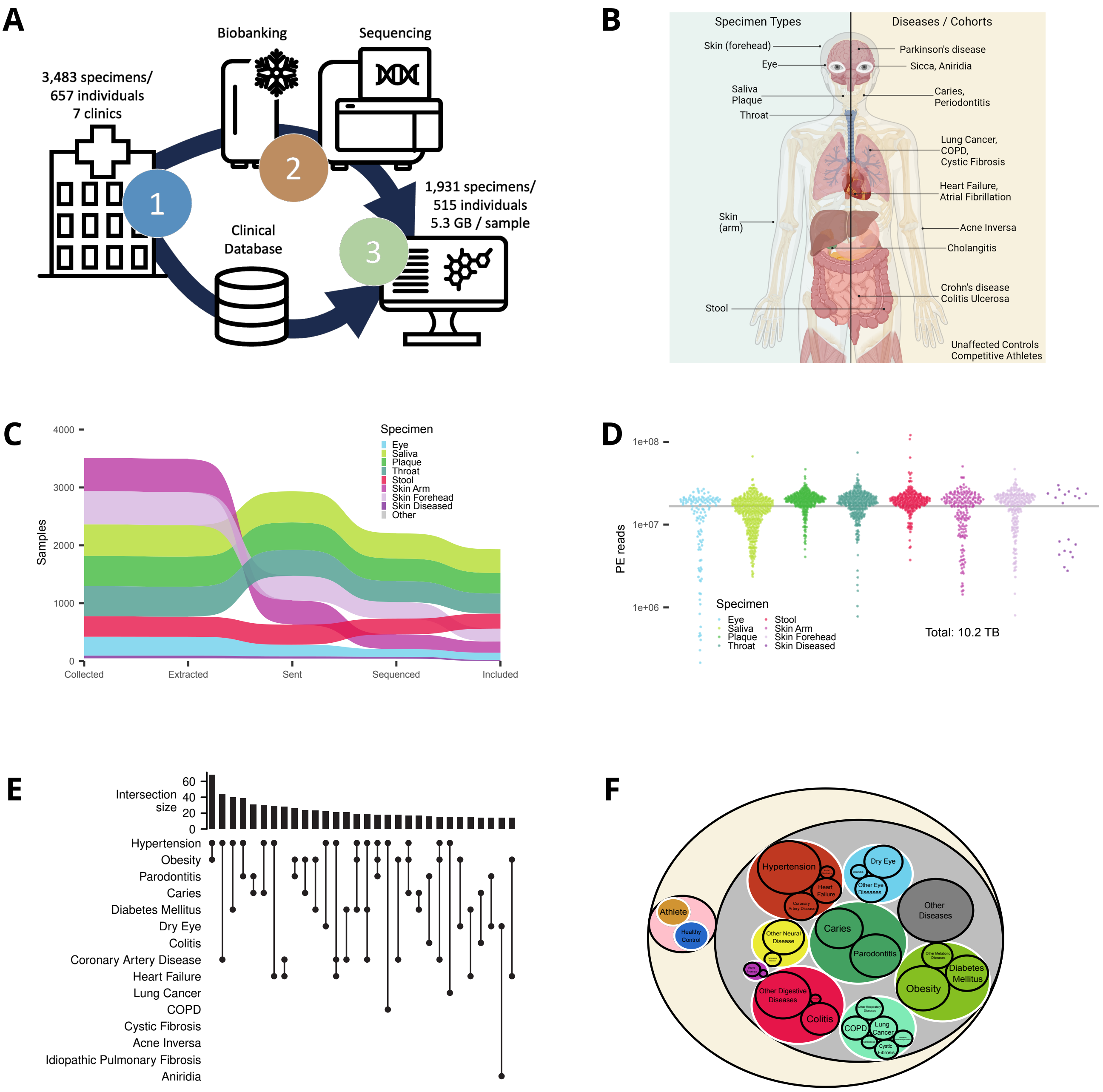


**Figure 1, Study set up, metagenomics data and clinical information: A)** Schematic Workflow describing the sample (upper arrow) and data flow (lower arrow) between clinicians, microbiology, and data science. The clinical data were kept separated from the measurement of microbiomes and only combined after measurement in the computational analysis. **B)** Clinical sampling was focused on seven biospecimens (left blue part). We included patients from a wide range of clinical diseases that allows us analyzing the diagnostic potential of different specimen types across diseases. **C)** Sankey plot for the number of samples included in the study at different intervals of the data generation process in relation to our quality control strategy. Specimen types are ordered vertically at each step in the pipeline by frequency of the respective specimen. **D)** Number of reads for each sample colored by specimen. The horizontal line represents the 5 gigabase threshold at a paired-end read length of 150bp. **E)** Pruned upset plot displaying the most frequent co-occurrence of diseases within the dataset. The combinations are ordered with decreasing frequency, marking the combination of Hypertension and obesity as most common comorbidity in our study. **F)** Ontology used throughout the study grouping diseases by biological systems and separating healthy control from diseased patients. Areas are proportional to the number of patients falling into each category. Patients may be represented multiple times if multiple diseases are diagnosed.

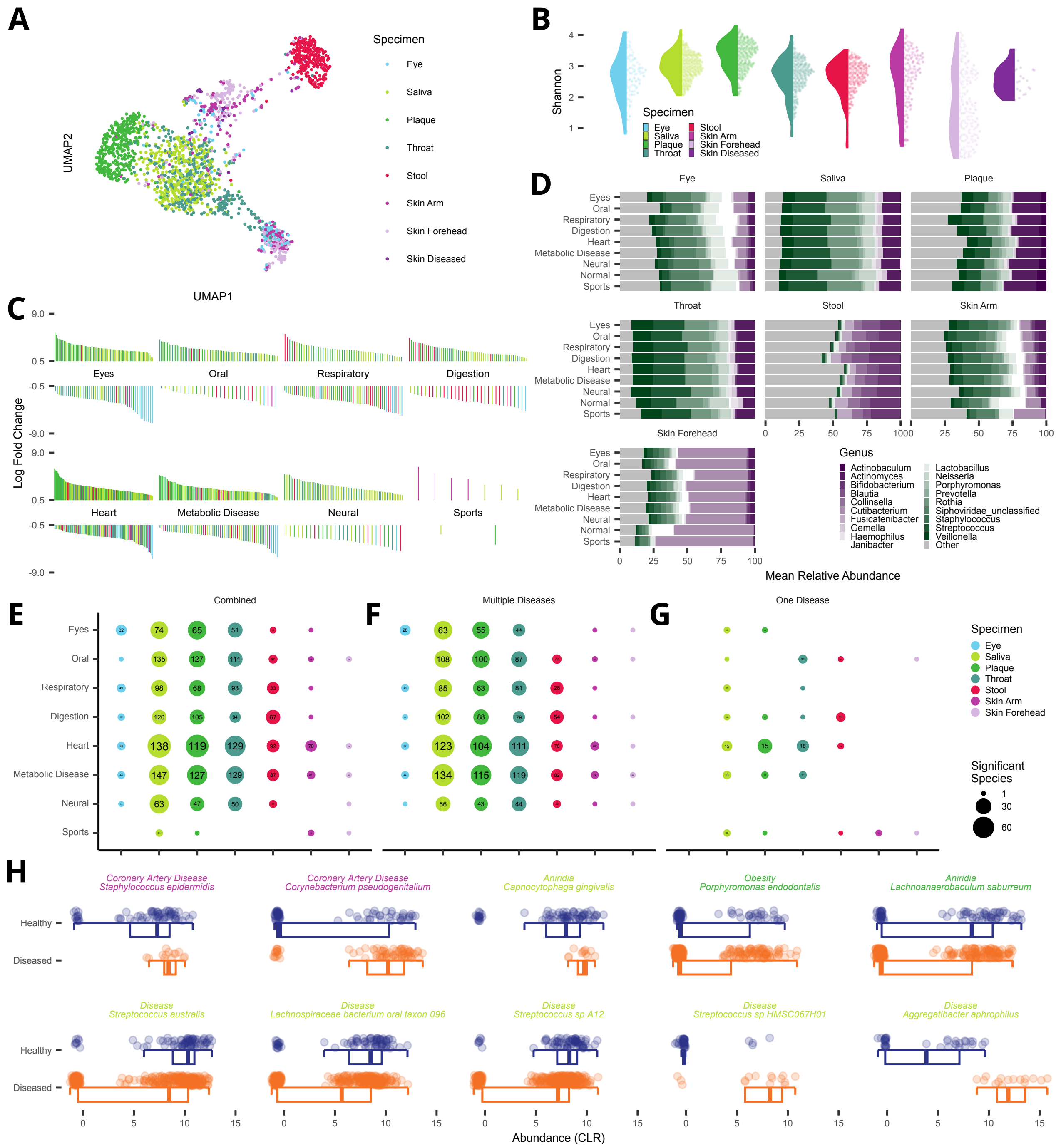


**Figure 2, Compositional analysis, and link of microbiota to diseases: A)** Two-dimensional UMAP embedding of pairwise computed mash distances, colored by biospecimen of the sample **B)** Alpha-diversity of all samples, colored by specimen. As a measure of species richness, we selected the Shannon diversity. **C)** Relative genus abundance for each cohort of the second ontology level, divided by biospecimen. **D)** Sorted log-fold changes of differentially abundant species matching the visualized results of the previous panel. Each panel is split vertically separating positive and negative log-fold changes. **E-G)** Number of differentially abundant species after p-value adjustment revealed during analysis across all cohorts and specimen combinations. Numbers in the circles represent the number of specimens included in the respective analysis. **H)** Center-log ratio normalized abundance counts of selected species-cohort-specimen combinations. The visualized diseased cohort is indicated by the text above each panel, whereas the selected biospecimen is indicated by the color of the writing. The first row of panels displays potential pathogen candidates with the highest statistical significance and a pathogen score of one. The second row of panels displays saliva samples of commensal bacteria candidates with a commensal score larger than eighteen.

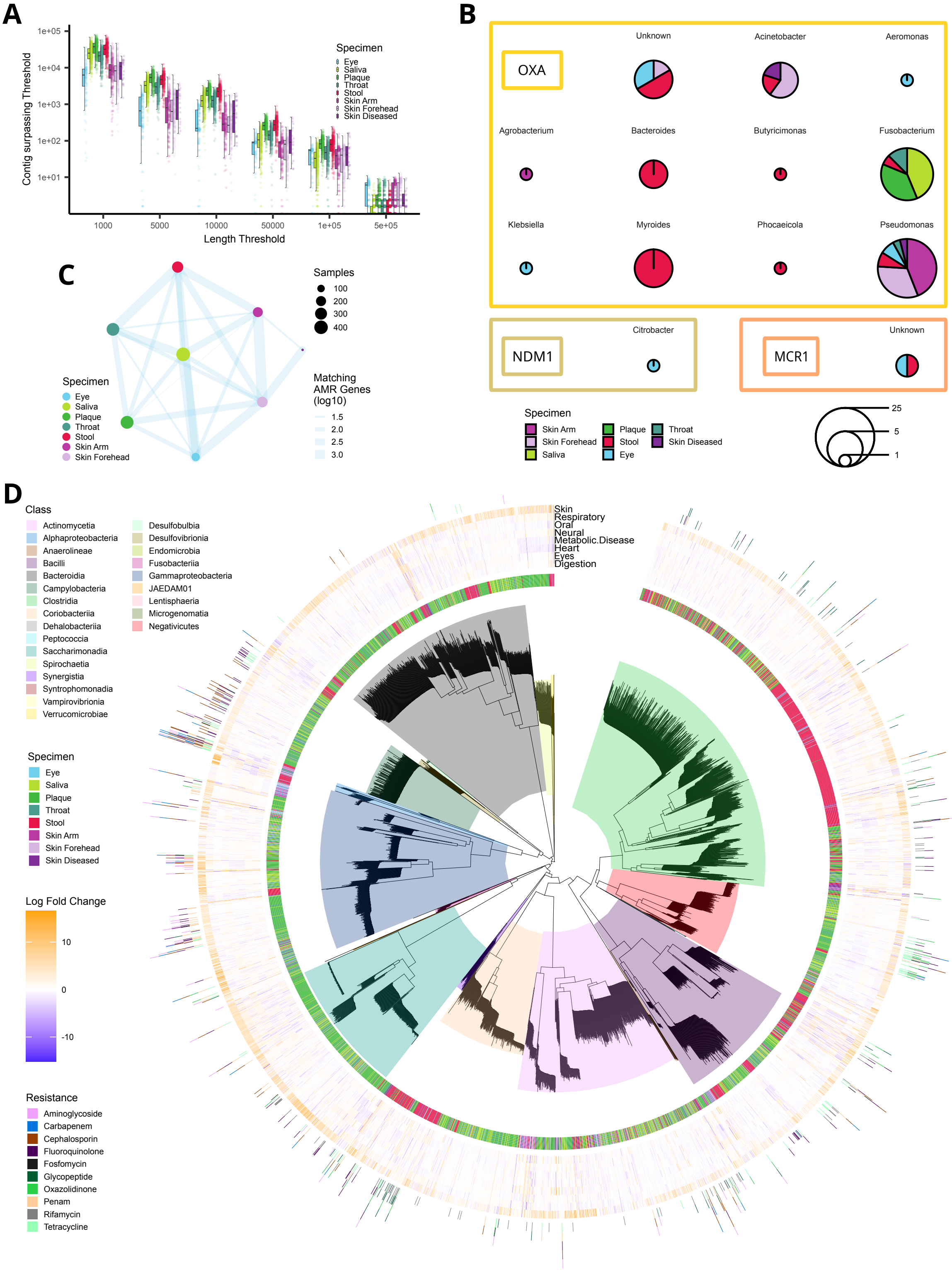


**Figure 3, Assembly and resistance gene analysis: A)** Distribution of the number of scaffolds in each sample at various length limits, colored by specimen as box-whisker plot. **B)** Sequence of pie charts indicating the presence of emerging antimicrobial resistance genes. Panels are subdivided by genus that was assigned to the contig where resistance genes have been detected. Pie charts scale with the number of measurements in different samples and are colored by the relative frequency of the sample's biospecimen. **C)** Network visualization of counts of shared resistance genes among different biospecimen samples derived from the same patient. Note, any resistance gene annotated by AMRFinderPlus was used for this plot. **D)** Dereplicated SGBs defined from our data. Visualized information includes biospecimen of the initial sample where the SGB was derived from, selected resistance information taken from Pathofact, and effect size of differential coverage analysis for selected cohorts. Note, the visualized differential coverage focuses only on the biospecimen of the initial sample where the SGB has been defined from that is also visualized in the central ring.

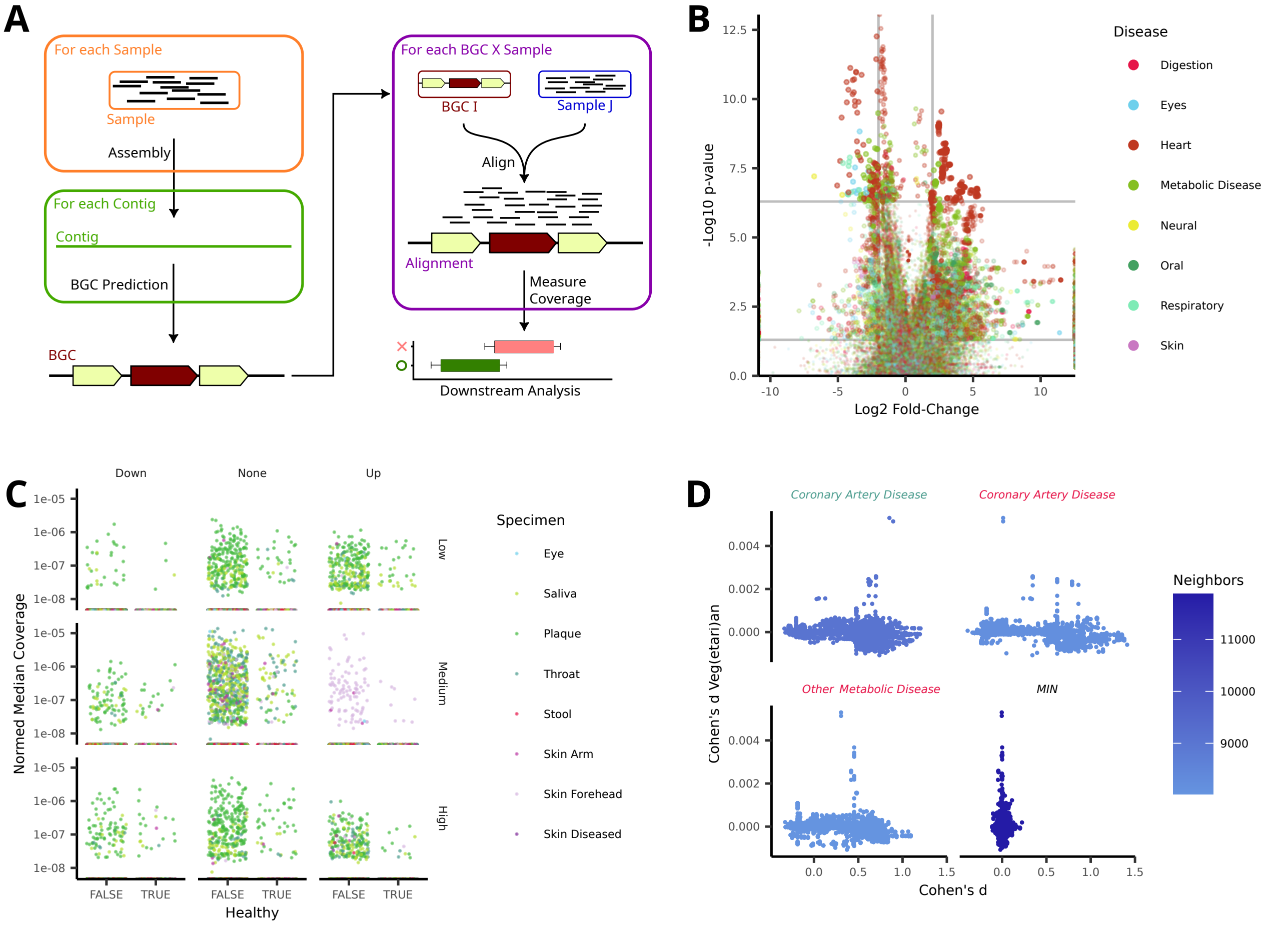


**Figure 4, Evidence-supported genome mining and disease association: A)** Schematic representation of our proposed BGC prioritization strategy representing an adapted version of the BiGMAP workflow. Metagenomic assembly is performed for each sample, followed by BGC prediction. Next, all samples are aligned against all core biosynthetic genes of predicted BGCs. Coverage information is extracted, and downstream analysis is performed. **B)** Volcano plot of the differential BGC coverage analysis results. In this visualization, only matching biospecimen – initial BGC contig combinations are visualized, constituting only a fraction of all results. **C)** Randomly drawn BGC examples of each region of the previous volcano plot. The visualized coverage includes all biospecimens. **D)** Comparison of the highest correlating effect sizes, comparing differential BGC coverage results between alternative diets and diseases. The effect size of the vegetarian-omnivore comparison is visualised on the y-axis. On the x-axis, the cohort named above the panel is compared against the healthy cohort. For the fourth panel, the minimum effect size across all cohort comparisons is taken for each BGC and compared against the diet comparison.

## 3.10 Mibianto: An Online Microbiome Analysis Tool

# Mibianto: ultra-efficient online microbiome analysis through k-mer based metagenomics

Pascal Hirsch[1], Georges P. Schmartz[1], Alejandra Leidy Gonzales [1], Annika Engel [1], Jens Zentgraf [2], Sven Rahmann [2], Matthias Hannig [3], Rolf Müller [4,5,6], Fabian Kern[1,4], Andreas Keller[1,4,6,*]

1 Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany

2 Algorithmic bioinformatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany

3 Clinic of Operative Dentistry, Periodontology and Preventive Dentistry, Saarland University Hospital, Saarland University, Kirrberger Str. 100, Building 73, 66421, Homburg, Saar, Germany

4 Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research, 66123 Saarbrücken, Germany

5 Deutsches Zentrum für Infektionsforschung (DZIF), Standort Hannover-Braunschweig, 38124 Braunschweig, Germany

6 PharmaScienceHub, 66123 Saarbrücken, Germany

* To whom correspondence should be addressed. Tel: +49 681 30268611; Fax: +49 681 30268610; Email: andreas.keller@ccb.uni-saarland.de

## ABSTRACT

K-mer-based approaches in bioinformatics were long established to expedite the analysis of large sequence data in genomic studies of higher organisms but are now widely applied in annotating metagenomic data. Here, we make use of k-mer counting techniques for efficient and effective compositional analysis of microbiota in whole metagenome sequencing data. The quantification of microbiome species composition, a fundamental step in metagenomic studies, often poses challenges due to its time-consuming nature and computational complexity. The Mibianto web server addresses these challenges by enabling direct operations on read files, thus eliminating the need for preprocessing or comprehensive data exchange. It handles diverse sequencing platforms, including short single-end, paired-end, and long read technologies. Our sketch-based workflow significantly reduces the data volume transferred from the user to the server (0.41% of the original FASTQ file size) to subsequently perform taxonomic profiling, enhancing both efficiency and data privacy. Mibianto offers functionalities beyond k-mer quantification; it supports advanced community composition analyses, including diversity, ordination, and differential abundance analysis. Our tool aids in the standardization of computational workflows, thus supporting the reproducibility of scientific sequencing studies. Its adaptability to small- and large-scale experimental designs, coupled with its user-friendly interface, makes it an invaluable tool for both clinical and research-oriented metagenomic studies. Mibianto is freely available without the need for a login: https://www.ccb.uni-saarland.de/mibianto.

## INTRODUCTION

Technological advancements are rapidly transforming the research practices of molecular microbiology and ecology. Although the initial application of high-throughput sequencing focused on bacterial isolates only, further refinement of DNA preparation protocols has enabled the profiling of entire micro-ecosystems via metagenomics (1). It appears, if the challenges arising from whole metagenome sequencing data analysis can be overcome, unprecedented valuable insights into diverse research areas such as bioremediation, natural compound discovery, and human health can be gained (2-5). In practice however, whole metagenome sequencing experiments generate enormous quantities of sequencing data, rendering sufficient expertise in computational analysis indispensable (6). To support researchers in their data evaluation workflow, a wide variety of online data processing tools have emerged. MG-RAST (7), MGnify (8), Galaxy (9), and similar tools accept uploads of whole datasets and process them on their servers through custom data analysis pipelines. While this method of data management is intuitive and allows for in-depth analysis, it requires a fast internet connection on the user side, a solid infrastructure on the service provider side, and a considerable amount of processing time. For studies involving sensitive clinical data, users must consider all legal aspects before trusting any third parties. Furthermore, with continuously decreasing sequencing costs, most cohort studies steadily increase their data output, rendering large data exchange pipelines a major inconvenience. One way to circumvent this issue is to request users to preprocess data on their end before uploading it to a service for further downstream analysis. For instance, our tool BusyBee Web provides comprehensive binning functionality at the expense that users need to upload a complete assembly (10). In a similar manner, the tool MicrobiomeAnalyst, performs extensive downstream analysis on taxonomic counts (11), for which users are required to generate and upload a count matrix from the raw sequencing data. Even though many out-of-the-box solutions exist to solve both mentioned user-imposed challenges, lack of sufficient computational expertise prevents many researchers from accessing them.

Here, we propose Mibianto, an online whole metagenome sequencing data analysis web server centered around compositional analysis. It is light on user connection requirements, trivial to get started, and capable of quickly processing small to medium-sized studies. The tool leverages recent progress made in MinHash-based data analysis to compute and transmit a compressed representation of the data to the remote server where it performs taxonomic profiling (12,13). Once the job submission is finished, it provides a wide range of state-of-the-art analysis options, visualizations, and recommendations for further interactive analysis.

## MATERIALS & METHODS

Mibianto is composed of three main components. First, the initial submission interface, where users are prompted to select metagenomic reads and metadata. Second, our server-side data analysis pipeline, which handles most of the computationally intensive tasks. Lastly, the result interface offers interactive data exploration, sharing and exporting capabilities.

### Taxonomic profiling

Mibianto prevents transfering sequencing reads directly. Instead, a selection of specifically calculated hashes is sent to the server. A hash function takes one read at a time and generates a numerical value based on the actual sequence. When working with sets of subsequences that have the same length as the original sequence, a MinHash is the smallest value obtained from hashing each subsequence (14). FracMinHashes are a subset of MinHashes collected from multiple sequences and below a specified threshold value. We calculate FracMinHashes from user provided FASTQ files with a WebAssembly version of sourmash (15). To be able to perform this computation on the client side, sourmash was compiled from Rust to WebAssembly. We achieved this conversion with the rust package manager cargo (v:1.65.0), rustc (v:1.65.0), wasm-pack (v:0.9.1), and sourmash (v:4.6.1). Via JavaScript the user input files get decompressed and streamed to the WebAssembly package after which it is transmitted to the server. For reference, the computation of a 150 bp paired-end sample containing 8 Gbp takes around 15 minutes on a standard consumer laptop. Users may select to save their metadata and hashes on the web server for later reusage. Once the data upload is completed, taxonomic profiling starts, and the user may enter a waiting queue and receive a unique job identifier. To ease software maintenance, we implemented the taxonomic profiling pipeline in snakemake (v:7.18.2) (16). Data processing closely follows the sourmash documentation from GitHub at a fixed k-mer size of 51. Sketches of each sample are compared against the Genome Taxonomy Database (GTDB) (v:rs207) to collect feature abundances (17). Next, taxonomic counts of all samples are aggregated, taxonomic annotations, sample data, and a phylogenetic tree are attached, and a phyloseq (v: 1.42.0) object is saved (18). Upon successful completion of the server-side processing pipeline, the user is forwarded from the queue to the results page, where a wide range of further analyses can be performed, and two download options are available. A phyloseq object and an Excel file with taxonomic counts can be downloaded, aiming to serve users with and without programming skills, respectively.

### Compositional analysis

Users that prefer additional support can refer to our results page for state-of-the-art compositional analyses supported with rich interactive visualization and customization options. We also implemented various data normalization options, namely compositional, z-score, log10, log10p, hellinger, centered log-ratio and additive log-ratio. Moreover, users may filter their data by removing individual samples or operational taxonomic units (OTU) based on abundance criteria. Alpha diversity can be visualized and checked for significant differences among cohort or sample groups. Further, ordination analysis is supported by a range of dimensionality reduction methods and dissimilarity measures. To this end, we integrated major parts of microViz, which allows for an interactive selection of samples in the embedding and displays their relative composition (19). We also provide a table view where estimated abundances are numerically displayed for each taxonomy. Individual OTUs can be selected, and their normalized abundances are displayed across samples. Most relevant for clinical applications, we keep a manually curated list of potentially pathogenic species automatically highlighted in the table. In case higher taxonomic ranks are of interest, we highlight an OTU as a potential pathogen when it contains at least one potentially pathogenic species. Finally, we implemented differential abundance analysis with ANCOMBC (v:2.0.1) (20).

### Proxies for Quality Control

Quality control (QC) is a crucial step of every sequencing analysis workflow. Performing QC on the user side would aggravate the computational burden on their end. However, since Mibianto only transfers hashed information to the server, QC on the server side becomes a challenge. We addressed this issue by computing several proxies for QC and forwarding users to further online analysis with e.g. BusyBee Web in case of apparent data anomalies. First, the estimate of the overall assignment rate from sourmash is displayed to the user on the results page indicating how robustly the community of each sample was quantified. Low assignment rates indicate either that the submitted community is not adequately represented in the database or that quality issues exist. Second, we compare user hashes against a precomputed QC dataset by computing distances with the built-in comparison from sourmash. This QC dataset was built by downloading metagenomics samples along with annotations from SRA. Adapter and host DNA contaminations were estimated with trimgalore (v: 0.6.10) (21) and bowtie2 (v: 2.5.1) (22) alignment against the human genome, respectively. Spatial proximity in co-embeddings of QC samples with increased contamination might be indicative of similar issues in user data. For clinical cohort studies, we annotate potential outliers using the local outlier factor (LOF) from DescTools (v: 0.99.47) (23). We note that the local outlier factor highlights abnormal clustering behavior of individual data points. Yet, depending on the experimental setup, this may be expected.

### Case studies

To explore the potential and limitations of our new online tool, we provide an example analysis of two different datasets, comprising one classical cohort study, seven biospecimens, four DNA extraction protocols, and two different sequencing technologies. The datasets Rehner et al. (24) and Becker et al. (25) were fully processed with our implemented snakemake pipeline to ensure that results can be replicated. Original metadata classes were curated. Pipeline outputs were integrated into the web server and explored using the results page. In the cohort study, the 4-week timepoint was removed.

## RESULTS

The impact of the human microbiome on host health through various exogenous molecules such as peptides and metabolites are well established (26,27). While extensive research is improving our understanding of the causal mechanisms linking microbiota and the immune system, disease associations with microbiome composition can provide valuable insight for clinicians (28,29). However, quantifying microbiome composition from metagenomic sequencing reads remains time-consuming and computationally challenging. Recent developments in the field of sketch-based taxonomic profiling have resulted in an efficient solution for exchanging large quantities of metagenomic sequencing data between client and server. By leveraging sourmash as a taxonomic profiling backbone, we have developed Mibianto, an online solution for convenient microbiome composition analysis. We provide a wide range of state-of-the-art downstream analyses with many customization options that are partially based on the microViz package. The functionality includes assessment of different taxonomic ranks, data filtering, diversity analysis, pathogen highlighting, and more. We are aware of the workflow's QC limitations and have therefore provided several indicators to detect potentially contaminated samples.

### Mibianto handles metagenomes from saliva, skin, plaque, stool and eye samples.

We aimed to rigorously evaluate the performance of Mibianto across diverse experimental setups. A critical factor influencing metagenomic outcomes is the type of sample being analyzed. Metagenomes derived from saliva, gut, dental plaque, skin, or the eye exhibit significant variations in microbial composition, reflecting the unique microbiota habitats of these biological sources. Moreover, the choice of DNA extraction protocols can greatly impact the variability of results. Different methodologies may preferentially extract certain microbial groups, leading to variation in observed community structures. Therefore, comprehensive testing across these varying conditions is crucial to ensure tool robustness and reliability in accurately capturing and reflecting the intricate diversity and dynamics of microbial communities. In the first dataset by Rehner et al. we thus compare three different DNA extraction kits on seven different biospecimens compiling a total of 30 data points (24). DNA was extracted from bile, stool, saliva, plaque, sputum, conjunctiva, and water control samples with Qiagen DNeasy PowerSoil Pro (QPS), Qiagen QiAamp DNA Microbiome Kit (QMK), and ZymoBIOMICS DNA Miniprep Kit (ZYMO). For each biospecimen, all samples were derived from the same biological sample. Additionally, to the short read MGI sequencing datasets, matched Oxford nanopore sequencing is available for saliva and bile samples. Full details on experimental design may be found in the corresponding manuscript.

As QC proxies, Mibianto displayed the assignment rates of the almost sterile samples, water and conjunctiva, which are low with only one sample surpassing 25% (**Figure 1A**). Further, in Saliva and bile samples, nanopore reads have lower assignment rates compared to their short-read counterpart. The MinHash-embedding without Mibianto's internal data indicated QPS sputum as an outlier (**Figure 1B**). We note that the samples we investigate here were already human-read decontaminated by Rehner et al. beforehand, accordingly the samples did not cluster with our selection of highly contaminated samples. Partially reconstructing the original analysis from the manuscript with Mibianto on a species level, the highest number of observed species is found in the oral cavity in QMK and ZYMO, which corroborates the findings in the results of the original manuscript (**Figure 1C**). Water contamination is highest with ZYMO. Principal coordinate analysis on Bray-Curtis distances of short reads clusters samples by biospecimen (**Figure 1D**). Bile and stool cluster closely together. Following the decision in the original manuscript, we do not perform differential abundance analysis due to the high number of confounding variables and missing replicates.

### Mibianto identifies significantly de-regulated gut microbiome species in Parkinson's disease.

After testing its robust performance across various species types and DNA extraction methods, we next evaluate Mibianto's efficacy in executing case-control metagenomic studies. It is designed to facilitate the dissection of microbial variations between control and case groups, offering valuable insights into microbial dynamics. In that, we aim positioning Mibianto as a powerful tool for medical and life-scientific researchers with an interest in understanding differences within microbiomes across diverse research settings. The second dataset is a next-generation sequencing dataset of an interventional cohort study on Parkinson's disease by Becker et al.￼￼. The dataset consists of 140 stool samples and three cohorts, namely Parksinon's Disease (PD), Parkinson's Disease with a resistant starch intervention (PD+RS), and the control cohort. The PD cohort received dietary instructions, whereas the control and PD+RS cohorts received resistant starch as a nutritional supplement. Measurements were taken at different timepoints. Full details may be found in their manuscript.

Data exploration with Mibianto on the species level indicates a high assignment rate with a few outliers. We observed statistically significant differences in alpha diversity before adjustment comparing PD+RS against PD (Wilcoxon Mann-Whitney p-value ≈ 0.0306) and the control (p-value ≈ 0.0002), yet no significant difference was observed between PD and control (p-value ≈ 0.3185) (**Figure 2A**). Ordination analysis clusters control samples visibly closer together (**Figure 2B**). Differential abundance analysis with ANCOMBC and Benjamini-Hochberg p-value adjustment highlighted 33 and 17 OTUs as significantly differentially abundant among control and PD+RS, and control and PD, respectively. No significant differences were detected when comparing PD with PD+RS. We want to highlight that the significance of all p-values mentioned in this section is inflated since the samples are not statistically independent due to the aggregation across timepoints. In both contrasts, the most significantly differentially abundant OTU was Faecalibacterium prausnitzii_C (**Figure 2C**). F. prausnitzii is described to have anti-inflammatory effects, produce butyrate, and known to be depleted in Parkinson's disease patients (30,31).

### Mibianto compresses metagenomic data sets by a factor of X

An important aspect of Mibianto is to facilitate the online analysis of larger studies by reducing the transferred data set at the client site. From the previous studies we estimate how many bytes are transferred from the user site to the server site of Mibianto, after the k-mer spectra are generated on the local computer. XXX

### Enabling specific analyses using BusyBee Web

While Mibianto excels at managing large-scale studies through efficient data compression, it is recognized that certain in-depth analyses fall outside the scope. To address this, we have seamlessly integrated Mibianto with our previously developed BusyBee Web platform. BusyBee Web is tailored for a distinct purpose: conducting extensive in-depth analyses of a smaller number of metagenomic samples. Unlike Mibianto, which operates on compressed data, BusyBee Web requires a full upload of metagenomic datasets. This complementary approach allows Mibianto to identify and propose a subset of samples that warrant more detailed examination. Leveraging both platforms in tandem enables researchers to navigate from broader metagenomic surveys to focused, in-depth analyses with ease and precision. XXX

### Outlook: ultra-high processing performance for eukaryotic reads

We demonstrate that Mibianto can handle small- to large-scale metagenomic studies using small footprint k-mer spectra and complement its application scope in combination with BusyBee web. Towards a more complete set of tools, we evaluated the possibility to extend the concept of Mibianto – namely to use k-mer spectra with a reduced size for performing web-based analyses at client side – to human nucleic acid data sets. As one of the most frequent use-cases we considered gene expression profiling. We selected XX.

## CONCLUSION

Mibianto is a web server that specializes in the compositional data analysis of metagenomic sequencing data. It distinguishes itself from existing online solutions by input flexibility, ease of use, and minimal connection requirements. However, incorporating functional analysis, de-novo assembly, genome mining, or any analysis requiring access to larger pieces of DNA sequence is currently not possible due to the design of our client-server data exchange model. Additionally, further research is required in the field of MinHash-based taxonomic profiling to refine results for nanopore sequencing reads (32). Nonetheless, Mibianto already provides a wide range of features and functionalities that enable rapid insights into microbial communities with extensive result visualization and the ability to customize to individual workflows. Based on two previous publications, we demonstrated how our tool was able to confirm central findings in metagenomic experiments without the need for any bioinformatics expertise. We are confident that Mibianto will serve as a valuable tool for researchers in metagenomics and related fields and we invite users to suggest additional desired features or ideas on our GitHub project page (https://github.com/CCB-SB/mibianto).

## AVAILABILITY

Mibianto is freely available without any login requirement at:
https://www.ccb.uni-saarland.de/mibianto

## ACCESSION NUMBERS

Data for the cohort study was made available by Becker et al. On NGDC GSA under the accession: HRA000635

Data for the protocol study was made available by Rehner et al. on NCBI SRA under the accession: PRJNA802336

## FUNDING

Funding for open access charge: Saarland University and Helmholtz-Institute for Pharmaceutical Research Saarland.

This compute infrastructure for this project was funded by the DFG [469073465].

## CONFLICT OF INTEREST

G.P.S., R.M., and A.K. are co-founders of MooH GmbH, a company developing metagenomic based oral health tests.

## ACKNOWLEDGEMENT

We thank Jacqueline Rehner for insightful discussion on the internally used pathogen list.

## AUTHOR CONTRIBUTIONS

P.H., G.P.S. designed and developed the Mibianto web server under the supervision of F.K., R.M., A.K. E.A-P. and A.E. curated the data. G.P.S. drafted the manuscript. P.H., E.A-P., A.E., F.K., R.M., A.K contributed to manuscript writing.

## REFERENCES

1. Stein, J.L., Marsh, T.L., Wu, K.Y., Shizuya, H. and DeLong, E.F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol*, **178**, 591-599.
2. Hauptfeld, E., Pelkmans, J., Huisman, T.T., Anocic, A., Snoek, B.L., von Meijenfeldt, F.A.B., Gerritse, J., van Leeuwen, J., Leurink, G., van Lit, A. *et al.* (2022) A metagenomic portrait of the microbial community responsible for two decades of bioremediation of poly-contaminated groundwater. *Water Res*, **221**, 118767.

3. Huo, L., Hug, J.J., Fu, C., Bian, X., Zhang, Y. and Müller, R. (2019) Heterologous expression of bacterial natural product biosynthetic pathways. *Nat Prod Rep*, **36**, 1412-1436.
4. Ko, K.K.K., Chng, K.R. and Nagarajan, N. (2022) Metagenomics-enabled microbial surveillance. *Nat Microbiol*, **7**, 486-496.
5. (2019) The Integrative Human Microbiome Project. *Nature*, **569**, 641-648.
6. Elworth, R.A.L., Wang, Q., Kota, P.K., Barberan, C.J., Coleman, B., Balaji, A., Gupta, G., Baraniuk, R.G., Shrivastava, A. and Treangen, T.J. (2020) To Petabytes and beyond: recent advances in probabilistic and signal processing algorithms and their application to metagenomics. *Nucleic Acids Res*, **48**, 5217-5234.
7. Meyer, F., Bagchi, S., Chaterji, S., Gerlach, W., Grama, A., Harrison, T., Paczian, T., Trimble, W.L. and Wilke, A. (2019) MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Brief Bioinform*, **20**, 1151-1159.
8. Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M.L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L.J. *et al.* (2023) MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res*, **51**, D753-d759.
9. (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res*, **50**, W345-351.
10. Schmartz, G.P., Hirsch, P., Amand, J., Dastbaz, J., Fehlmann, T., Kern, F., Müller, R. and Keller, A. (2022) BusyBee Web: towards comprehensive and differential composition-based metagenomic binning. *Nucleic Acids Res*, **50**, W132-137.
11. Dhariwal, A., Chong, J., Habib, S., King, I.L., Agellon, L.B. and Xia, J. (2017) MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res*, **45**, W180-w188.
12. Irber, L. and Brown, C.T. (2020) Lightweight compositional analysis of metagenomes with sourmash gather. *Manubot. Available at: https://dib-lab. github. io/2020-paper-sourmash-gather/(Accessed: 16 December 2020)*.
13. Irber, L., Brooks, P.T., Reiter, T., Pierce-Ward, N.T., Hera, M.R., Koslicki, D. and Brown, C.T. (2022) Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. *bioRxiv*, 2022.2001.2011.475838.
14. Broder, A.Z. (1997), *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, pp. 21-29.
15. Pierce, N.T., Irber, L., Reiter, T., Brooks, P. and Brown, C.T. (2019) Large-scale sequence comparisons with sourmash. *F1000Res*, **8**, 1006.
16. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O. and Kanitz, A. (2021) Sustainable data analysis with Snakemake. *F1000Research*, **10**.
17. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.A. and Hugenholtz, P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res*, **50**, D785-d794.
18. McMurdie, P.J. and Holmes, S. (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, **8**, e61217.
19. Barnett, D.J., Arts, I.C. and Penders, J. (2021) microViz: an R package for microbiome data visualization and statistics. *Journal of Open Source Software*, **6**, 3201.
20. Lin, H. and Peddada, S.D. (2020) Analysis of compositions of microbiomes with bias correction. *Nat Commun*, **11**, 3514.
21. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, **17**, 10-12.
22. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**, 357-359.

23. Breunig, M.M., Kriegel, H.-P., Ng, R.T. and Sander, J. (2000), *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93-104.
24. Rehner, J., Schmartz, G.P., Groeger, L., Dastbaz, J., Ludwig, N., Hannig, M., Rupf, S., Seitz, B., Flockerzi, E., Berger, T. *et al.* (2022) Systematic Cross-biospecimen Evaluation of DNA Extraction Kits for Long- and Short-read Multi-metagenomic Sequencing Studies. *Genomics Proteomics Bioinformatics*, **20**, 405-417.
25. Becker, A., Schmartz, G.P., Gröger, L., Grammes, N., Galata, V., Philippeit, H., Weiland, J., Ludwig, N., Meese, E., Tierling, S. *et al.* (2022) Effects of Resistant Starch on Symptoms, Fecal Markers, and Gut Microbiota in Parkinson's Disease - The RESISTA-PD Trial. *Genomics Proteomics Bioinformatics*, **20**, 274-287.
26. Aho, V.T., Ostaszewski, M., Martin-Gallausiaux, C., Laczny, C.C., Schneider, J.G. and Wilmes, P. (2022) SnapShot: The Expobiome Map. *Cell Host & Microbe*, **30**, 1340-1340. e1341.
27. Wilmes, P., Martin-Gallausiaux, C., Ostaszewski, M., Aho, V.T., Novikova, P.V., Laczny, C.C. and Schneider, J.G. (2022) The gut microbiome molecular complex in human health and disease. *Cell Host & Microbe*, **30**, 1201-1206.
28. Ni, J., Wu, G.D., Albenberg, L. and Tomov, V.T. (2017) Gut microbiota and IBD: causation or correlation? *Nature reviews Gastroenterology & hepatology*, **14**, 573-584.
29. Beam, J.E., Wagner, N.J., Shook, J.C., Bahnson, E.S., Fowler Jr, V.G., Rowe, S.E. and Conlon, B.P. (2021) Macrophage-produced peroxynitrite induces antibiotic tolerance and supersedes intrinsic mechanisms of persister formation. *Infection and Immunity*, **89**, e00286-00221.
30. Zhang, M., Zhou, L., Wang, Y., Dorfman, R.G., Tang, D., Xu, L., Pan, Y., Zhou, Q., Li, Y., Yin, Y. *et al.* (2019) Faecalibacterium prausnitzii produces butyrate to decrease c-Myc-related metabolism and Th17 differentiation by inhibiting histone deacetylase 3. *Int Immunol*, **31**, 499-514.
31. Wallen, Z.D., Demirkan, A., Twa, G., Cohen, G., Dean, M.N., Standaert, D.G., Sampson, T.R. and Payami, H. (2022) Metagenomics of Parkinson’s disease implicates the gut microbiome in multiple disease mechanisms. *Nature Communications*, **13**, 6958.
32. Portik, D.M., Brown, C.T. and Pierce-Ward, N.T. (2022) Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics*, **23**, 541.

## FIGURE LEGENDS

**Graphical Abstract:** Mibianto is a web server that accepts metagenomic short- and long-reads as input, performs taxonomic profiling, and reports on compositional analysis. We provide numerous downstream options for interactive exploration and support cohort studies.

**Figure 1:** Mibianto results of the protocol comparison after minor adjustments to the visualizations downloaded from the server. **A)** Assignment rate for all samples. Samples with the prefix SRR18 were sequenced with nanopore sequencing. **B)** Quality control proxy computed on FracMinHash-based dissimilarities without co-embedding of our precomputed samples. **C**) Number of observed species in each sample, split by DNA extraction kit. **D)** Short-read sequencing samples were embedded with principal coordinate analysis on Bray-Curtis distances computed on the species level.

**Figure 2:** Mibianto results of the cohort study after minor adjustments to the visualizations downloaded from the server. **A)** Shannon diversity computed on species level grouped by cohort, split by timepoint. **B)** Ordination analysis using non-metric multidimensional scaling on Bray-Curtis distances. **C)** Abundance of F. *prausnitzii_C* in the different cohort groups after center log-ratio normalization.

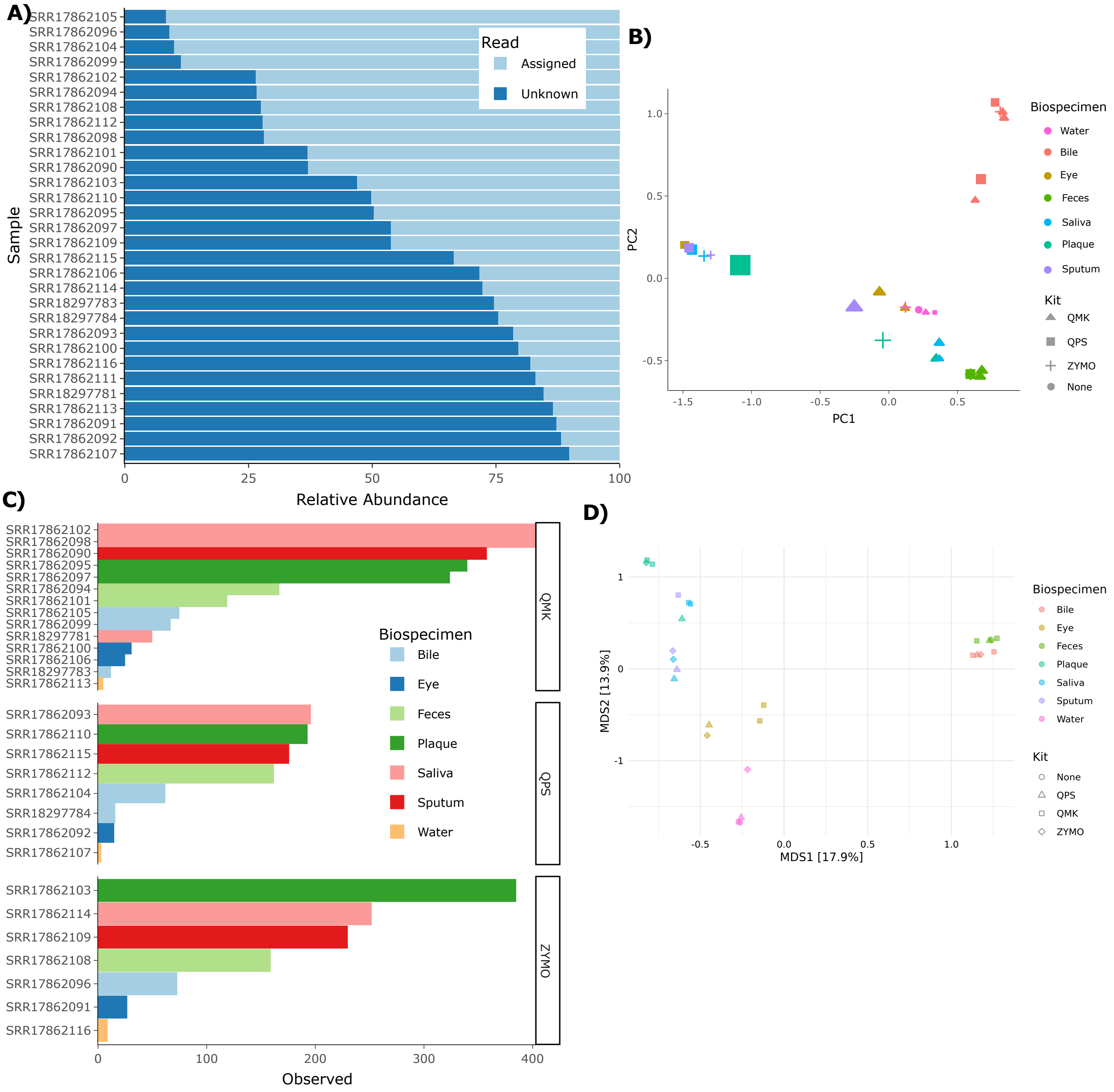


**Figure 1:** Mibianto results of the protocol comparison after minor adjustments to the visualizations downloaded from the server. **A)** Assignment rate for all samples. Samples with the prefix SRR18 were sequenced with nanopore sequencing. **B)** Quality control proxy computed on FracMinHash-based dissimilarities without co-embedding of our precomputed samples. **C)** Number of observed species in each sample, split by DNA extraction kit. **D)** Short-read sequencing samples were embedded with principal coordinate analysis on Bray-Curtis distances computed on the species level.

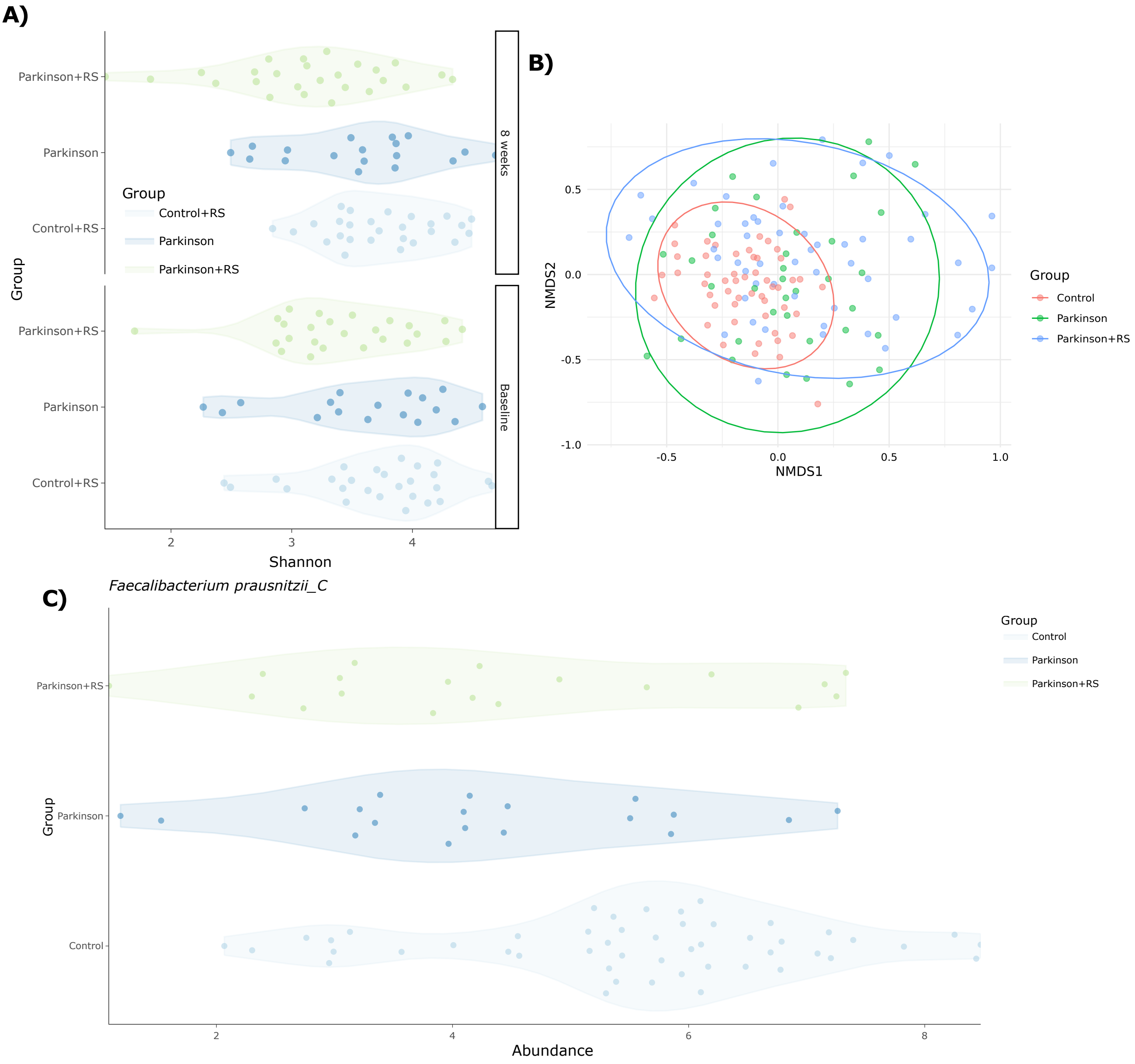


**Figure 2:** Mibianto results of the cohort after minor adjustments to the visualizations downloaded from the server. **A)** Shannon diversity computed on species level grouped by cohort, split by timepoint. **B)** Ordination analysis using non-metric multidimensional scaling on Bray-Curtis distances. **C)** Abundance of *F. prausnitzii_C* in the different cohorts after center log ratio normalization.

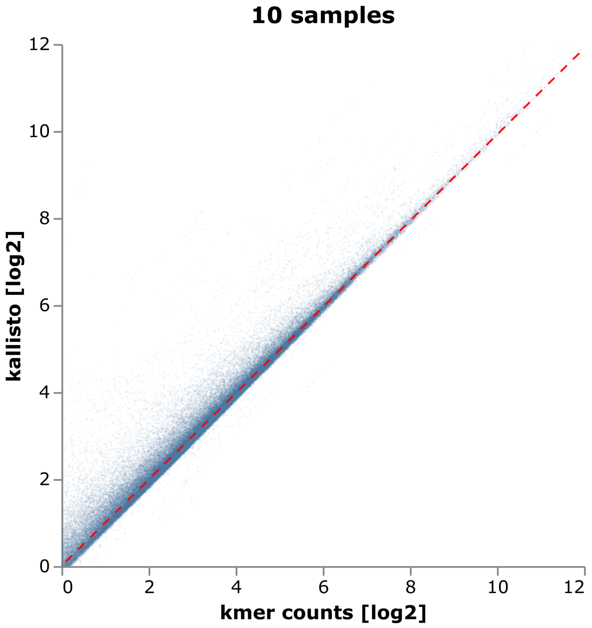


**Figure 3:** Scatter plot of normalized base-2 logarithms of k-mer counts vs. kallisto FPKM values; one point per expressed gene per sample from 10 different samples (103.415 points overall), with a trend line (dashed red) obtained by robust regression. Partial transparency was used to visualize regions of low vs. high point density. Overall, a strong correlation is visible (Pearson correlation coefficient 0.983).

# 4
# Discussion

The role of bioinformatics in the field of microbiology as well as clinical microbiology seems to be ever increasing also beyond the scope of research as a diagnostic tool in clinics. While adequate bioinformatic analysis provides immense gains in insights and a better understanding of the causal mechanisms of underlying infections, it comes at the cost of high complexity, workload, and resource requirements. In the works included in this thesis, we aimed to improve accessibility to bioinformatic evaluations and resources with our implemented and updated web services and databases. Further, we supported ongoing efforts in a wide range of applied research projects in the context of antimicrobial resistance, emerging pathogens, and dietary interventions. Lastly, we aided in the earliest steps of natural compound research by searching for disease-associated BGCs across many biospecimens, diseases, and host species.

The applied interventional clinical research metagenomics projects presented within this work, yielded several differentially abundant species at baseline condition. As we mentioned in subsection 1.4.4 the stability of these results may be second-guessed. Not only did we select ANCOM-BC and ALDEX2 for our differential analysis that have been consistently been recommended across benchmarks, but there have also been several studies confirming our findings e.g. for the Parkinson's disease study underlining the validity of these results [524; 570; 633–636]. Furthermore, as can be seen in in both of our studies, the dietary interventions did not yield any significantly differentially abundant species highlighting the specificity of the analysis after all.

The herein presented various clinical research projects were not exclusively centered around metagenomic studies. Instead within the bacterial isolates of the Ukrainian war-wounded refugees, we were able to conjecture about the relatedness of several infections. Even more crucial, we detected likely causal drivers for the resistance phenotype highlighting the importance of plasmid research. Concerning the *Auritidibacter ignavius* isolates, since the initial submission of our manuscript Gatermann et al. published eight more strains isolated from ear infections, further cementing its position as a pathogen implicated in ear infections [637].

Busybee Web and Mibianto are both services we provide to the scientific community as web servers. Initially designed for assembled contigs of short-reads, Busybee Web remained relevant with the introduction of long-read sequencing. Here, we implemented a reference-free approach to identify differences among cohorts based on an unsupervised embedding method. Further analysis may then assess the exact contigs responsible for the difference in signal. Moreover, aware of the challenge of large amounts of data for Busybee Web, we developed Mibianto. Unlike other online methods requiring a count matrix as input, Mibianto is able to efficiently compress NGS data directly and transmit hashed information to the web server where additional downstream analysis is possible. Open challenges we face with Mibianto are mostly focused on quality control and potential preprocessing. A major challenge hereby is that these steps would have to be performed on the user side. Similar to the compression, k-mer-based approaches may be used for quality control, however, methods such as the BBTools are notorious for having a major memory footprint to store the reference genome k-mers [638]. Thus, more research is required on this end. Nevertheless, we believe that the concept of using FracMinHashes for efficient data transmission of metagenomic data bears a lot of unexplored potential.

Plasmids from PLSDB have served as ground truth in sequence classifier training and as reference material for large-scaled studies on mobile DNA [639; 640]. With the immense potential of metagenomics sequencing several voices in the scientific community advocate to integrate metagenomic assembled plasmids into plasmid databases and also successfully published these results [641]. However as we previously demonstrated, these plasmids assembled from metagenomic experiments are frequently misassembled and incorrectly classified [642]. Due to recent advancements, in nanopore sequencing e.g. with the transition to the new R10.4 flowcell, the long-read sequencing technology tremendously improved accuracy upon previous developments [643]. Currently, this improvement is coming at higher costs. Yet expenses are likely going to decrease, following the general trend of other sequencing methods over the years. With reliable long-read sequencing at scale, a wide range of challenges in metagenomics data analysis would see improvements, including misassemblies. Likely plasmid sequencing and assembly from metagenomic experiments will profit from these developments, potentially justifying their integration into PLSDB in future iterations.

The two metagnome studies discussing BGCs yield their own merit displaying a wide range of compositional microbiome associations with the inclusion of many confounding factors. Here, they mostly serve as associative biomarkers and are not to be interpreted in a causal context. To fully leverage this data, additional research efforts are already ongoing in close coordination between medical experts and bioinformaticians. However, to see the main value of the two studies, interpretation in a larger research context is necessary. As discussed in subsection 1.2.9, natural products deriving from BGCs are known to encode toxins and are discussed as a potential alley to find new drug candidates. In this study, we predicted over 28,000 BGCs. A manual scan of these predictions implicates elevated costs, labor, and time. In order to find health or disease-associated BGCs, our adapted coverage-based prioritization method aims to highlight the most promising BGCs for further assessment. As of the time of writing, the first BGCs have been selected for further investigation in close coordination with biotechnologist, experts in different natural product classes. Furthermore, considerable value lies in the biobanking of the original samples that can be leveraged later in the laboratory pipeline.

### 4.0.1 Future directions

Challenges and opportunities lie ahead for the field of bioinformatics with the seemingly ever-increasing scales of sequencing data, refinement of existing as well as development of new protocols, and plentiful innovations in the realm of machine learning. Improvements of multimodal, single cell, and spatial omics approaches are continuously being reported also for prokaryotes which is most relevant also for microbiome research [644]. Steadily increasing in scale while reducing the costs, these methods allow us to observe the microbiome from a new perspective. Sequencing of individual cells could decrease the number of misassemblies in metagenomic experiments and provide better estimates for the true community composition on strain level. Similarly, spatial information as well as other omics types, bring insight into the host-microbiome interactions at tissue interfaces. This improved understanding can hopefully be leveraged to reveal more causal mechanisms of community-born diseases and help in the design of targeted probiotic treatment plans to avoid dysbiosis. While the many innovations from the field of molecular biology deeply open up the realm of new possibilities, new developments in data science and computational sciences are also not to be underestimated. Landmark studies such as AlphaFold as well as progresses in graph neural networks, propelled drug discovery research into a blooming field full of potential [464; 645; 646]. Lastly, quantum computing promises massively parallel computing as never seen before and may serve as a valuable resource for drug discovery in the more distant future [647].

### 4.0.2 Conclusions

As a final conclusion, the herein presented work contributed short, intermediate, and long-term benefits to the combat against bacterial pathogens by directly supporting clinical microbiologists in their applied research, highlighting potential antibiotic leads to biotechnologists, and enabling future researchers with bioinformatic tools.

# *Bibliography*

[1] Jacqueline Rehner, Georges Pierre Schmartz, Laura Groeger, Jan Dastbaz, Nicole Ludwig, et al. Systematic cross-biospecimen evaluation of dna extraction kits for long-and short-read multi-metagenomic sequencing studies. *Genomics, Proteomics and Bioinformatics*, 20(2):405–417, 2022.

[2] Anouck Becker, Georges Pierre Schmartz, Laura Gröger, Nadja Grammes, Valentina Galata, et al. Effects of resistant starch on symptoms, fecal markers, and gut microbiota in parkinson's disease—the resista-pd trial. *Genomics, Proteomics & Bioinformatics*, 20(2):274–287, 2022.

[3] Jacqueline Rehner, Georges P Schmartz, Tabea Kramer, Verena Keller, Andreas Keller, and Sören L Becker. The effect of a planetary health diet on the human gut microbiome: A descriptive analysis. *Nutrients*, 15(8):1924, 2023.

[4] Fabian K Berger, Georges P Schmartz, Tobias Fritz, Nils Veith, Farah Alhussein, et al. Occurrence, resistance patterns, and management of carbapenemase-producing bacteria in war-wounded refugees from ukraine. *International Journal of Infectious Diseases*, 132:89–92, 2023.

[5] Sophie Roth, Maximilian Linxweiler, Jacqueline Rehner, Georges-Pierre Schmartz, Sören L Becker, and Jan Philipp Kühn. Auritidibacter ignavus, an emerging pathogen associated with chronic ear infections. *Emerging Infectious Diseases*, 30(1):8, 2024.

[6] Georges P Schmartz, Pascal Hirsch, Jérémy Amand, Jan Dastbaz, Tobias Fehlmann, Fabian Kern, Rolf Müller, and Andreas Keller. Busybee web: towards comprehensive and differential composition-based metagenomic binning. *Nucleic Acids Research*, 50(W1):W132–W137, 2022.

[7] Georges P Schmartz, Anna Hartung, Pascal Hirsch, Fabian Kern, Tobias Fehlmann, Rolf Müller, and Andreas Keller. Plsdb: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Research*, 50(D1):D273–D278, 2022.

[8] Pascal Hirsch, Alejandra Leidy Gonzales, Ernesto Aparicio-Puerta, Annika Engel, Jens Zentgraf, et al. Mibianto: ultra-efficient online microbiome analysis through k-mer based metagenomics.

[9] Georges P. Schmartz, Jacqueline Rehner, Miriam J. Schuff, Sören L. Becker, Marcin Krawczyk, et al. Between cages and wild: Unraveling the impact of captivity on animal microbiomes and antimicrobial resistance. .

[10] Georges P. Schmartz, Jacqueline Rehner, Madline Gund, Stefan Rupf, Matthias Hannig, et al. Decoding the diagnostic and therapeutic potential of microbiota using pan-body pan-disease microbiomics. .

[11] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.

[12] US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4, RIKEN Genomic Sciences Center: Sakaki Yoshiyuki 9 Fujiyama Asao 9 Hattori Masahira 9 Yada Tetsushi 9 Toyoda Atsushi 9 Itoh Takehiko 9 Kawagoe Chiharu 9 Watanabe Hidemi 9 Totoki Yasushi 9 Taylor Todd 9, Genoscope, CNRS UMR-8030: Weissenbach Jean 10 Heilig Roland 10 Saurin William 10 Artiguenave Francois 10 Brottier Philippe 10 Bruls Thomas 10 Pelletier Eric 10 Robert Catherine 10 Wincker Patrick 10, Institute of Molecular Biotechnology: Rosenthal André 12 Platzer Matthias 12 Nyakatura Gerald 12 Taudien Stefan 12 Rump Andreas 12 Department of Genome Analysis, GTC Sequencing Center: Smith Douglas R. 11 Doucette-Stamm Lynn 11 Rubenfield Marc 11 Weinstock Keith 11 Lee Hong Mei 11 Dubois JoAnn 11, Beijing Genomics Institute/Human Genome Center: Yang Huanming 13 Yu Jun 13 Wang Jian 13 Huang Guyang 14 Gu Jun 15, et al. Initial sequencing and analysis of the human genome. *nature*, 409(6822):860–921, 2001.

[13] Steven L Salzberg. Open questions: How many genes do we have? *BMC biology*, 16(1):1–3, 2018.

[14] Jack A Gilbert, Martin J Blaser, J Gregory Caporaso, Janet K Jansson, Susan V Lynch, and Rob Knight. Current understanding of the human microbiome. *Nature medicine*, 24(4):392–400, 2018.

[15] Christos A Ouzounis and Alfonso Valencia. Early bioinformatics: the birth of a discipline—a personal view. *Bioinformatics*, 19 (17):2176–2190, 2003.

[16] Jeff Gauthier, Antony T Vincent, Steve J Charette, and Nicolas Derome. A brief history of bioinformatics. *Briefings in bioinformatics*, 20(6):1981–1996, 2019.

[17] Davide Chicco and Giuseppe Jurman. Ten simple rules for providing bioinformatics support within a hospital. *BioData Mining*, 16(1):1–12, 2023.

[18] Sean Whalen, Jacob Schreiber, William S Noble, and Katherine S Pollard. Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3):169–181, 2022.

[19] Frank Konietschke, Karima Schwab, and Markus Pauly. Small sample sizes: A big data problem in high-dimensional data analysis. *Statistical Methods in Medical Research*, 30(3):687–701, 2021.

[20] Tomasz Miksa, Andreas Rauber, and Eleni Mina. Identifying impact of software dependencies on replicability of biomedical workflows. *Journal of Biomedical Informatics*, 64:232–254, 2016.

[21] Serghei Mangul, Lana S Martin, Eleazar Eskin, and Ran Blekhman. Improving the usability and archival stability of bioinformatics software, 2019.

[22] Pamela H Russell, Rachel L Johnson, Shreyas Ananthan, Benjamin Harnke, and Nichole E Carlson. A large-scale analysis of bioinformatics code on github. *PloS one*, 13(10):e0205898, 2018.

[23] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[24] Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[25] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[26] Kenneth S Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J Rodney Brister, and Christopher O'Sullivan. Stat: a fast, scalable, minhash-based k-mer tool to assess sequence read archive next-generation sequence submissions. *Genome Biology*, 22(1):1–15, 2021.

[27] Yue Meng, Yu Lei, Jianlong Gao, Yuxuan Liu, Enze Ma, Yunhong Ding, Yixin Bian, Hongquan Zu, Yucui Dong, and Xiao Zhu. Genome sequence assembly algorithms and misassembly identification methods. *Molecular Biology Reports*, 49(11): 11133–11148, 2022.

[28] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

[29] Zhenmiao Zhang, Chao Yang, Werner Pieter Veldsman, Xiaodong Fang, and Lu Zhang. Benchmarking genome assembly methods on metagenomic sequencing data. *Briefings in Bioinformatics*, 24(2):bbad087, 2023.

[30] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[31] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. metaspades: a new versatile metagenomic assembler. *Genome research*, 27(5):824–834, 2017.

[32] G David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[33] Jiefu Li, Jung-Youn Lee, and Li Liao. A new algorithm to train hidden markov models for biological sequences with partial labels. *BMC bioinformatics*, 22:1–21, 2021.

[34] Elisabeth Roesch, Joe G Greener, Adam L MacLean, Huda Nassar, Christopher Rackauckas, Timothy E Holy, and Michael PH Stumpf. Julia for biologists. *Nature Methods*, 20(5):655–664, 2023.

[35] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25 (14):1754–1760, 2009.

[36] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[37] Katalin Ferenc, Konrad Otto, Francisco Gomes de Oliveira Neto, Marcela Dávila López, Jennifer Horkoff, and Alexander Schliep. Empirical study on software and process quality in bioinformatics tools. *bioRxiv*, pages 2022–03, 2022.

[38] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.

[39] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, et al. Sustainable data analysis with snakemake. *F1000Research*, 10, 2021.

[40] Thomas Cokelaer, Sarah Cohen-Boulakia, and Frédéric Lemoine. Reprohackathons: promoting reproducibility in bioinformatics through training. *Bioinformatics*, 39(Supplement_1):i11–i20, 2023.

[41] Robyn S Lee and William P Hanage. Reproducibility in science: important or incremental? *The Lancet Microbe*, 1(2):e59–60, 2020.

[42] Mark Ziemann, Pierre Poulain, and Anusuiya Bora. The five pillars of computational reproducibility: bioinformatics and beyond. *Briefings in Bioinformatics*, 24(6):bbad375, 2023.

[43] Laura Wratten, Andreas Wilm, and Jonathan Göke. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature methods*, 18(10):1161–1168, 2021.

[44] Yanqing Wang, Fuhai Song, Junwei Zhu, Sisi Zhang, Yadong Yang, et al. Gsa: genome sequence archive. *Genomics, proteomics & bioinformatics*, 15(1):14–18, 2017.

[45] Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1): D368–D375, 2024.

[46] Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Shadab Ahmad, et al. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), 2022.

[47] Kenneth Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J Rodney Brister, and Christopher O'Sullivan. The sequence read archive: a decade more of explosive growth. *Nucleic acids research*, 50(D1):D387–D390, 2022.

[48] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[49] Nicole Ludwig, Meike Becker, Timo Schumann, Timo Speer, Tobias Fehlmann, Andreas Keller, and Eckart Meese. Bias in recent mirbase annotations potentially associated with rna quality issues. *Scientific reports*, 7(1):5162, 2017.

[50] Benjamin Goudey, Nicholas Geard, Karin Verspoor, and Justin Zobel. Propagation, detection and correction of errors using the sequence database network. *Briefings in bioinformatics*, 23(6): bbac416, 2022.

[51] Qingyu Chen, Ramona Britto, Ivan Erill, Constance J Jeffery, Arthur Liberzon, et al. Quality matters: Biocuration experts on the impact of duplication and other data quality issues in biological databases. *Genomics, proteomics & bioinformatics*, 18 (2):91, 2020.

[52] Edward S Dove, Yann Joly, Anne-Marie Tassé, and Bartha M Knoppers. Genomic cloud computing: legal and ethical points to consider. *European Journal of Human Genetics*, 23(10):1271–1278, 2015.

[53] Sushmita Basu, Jörg Gsponer, and Lukasz Kurgan. Depicter2: a comprehensive webserver for intrinsic disorder and disorder function prediction. *Nucleic Acids Research*, page gkad330, 2023.

[54] The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, 50(W1):W345–W351, 2022.

[55] Fabian Kern, Tobias Fehlmann, and Andreas Keller. On the lifetime of bioinformatics web services. *Nucleic acids research*, 48 (22):12523–12533, 2020.

[56] Julian E Sale and Barry L Stoddard. Fifty years of nucleic acids research<? mode editorial?>. *Nucleic Acids Research*, 52(1):1–3, 2024.

[57] Dominik Seelow. the 21st annual nucleic acids research web server issue 2023. *Nucleic Acids Research*, 51(W1):W1, 2023.

[58] Daniel J Rigden and Xosé M Fernández. The 2024 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 52(D1):D1–D9, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1173. URL `https://doi.org/10.1093/nar/gkad1173`.

[59] Dominik Seelow. the 19th annual nucleic acids research web server issue 2021. *Nucleic Acids Research*, 49(W1):W1–W4, 2021.

[60] Dominik Seelow. the 20th annual nucleic acids research web server issue 2022. *Nucleic Acids Research*, 50(W1):W1–W3, 2022.

[61] Wilson Wen Bin Goh and Limsoon Wong. The birth of biodata science: trends, expectations, and applications. *Genomics, proteomics & bioinformatics*, 18(1):5, 2020.

[62] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

[63] Matthew CB Tsilimigras and Anthony A Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of epidemiology*, 26(5):330–335, 2016.

[64] M Luz Calle. Statistical analysis of metagenomics data. *Genomics & informatics*, 17(1), 2019.

[65] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5:1–18, 2017.

[66] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[67] Eric Dexter, Gretchen Rollwagen-Bollens, and Stephen M Bollens. The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling. *Limnology and Oceanography: Methods*, 16(7):434–443, 2018.

[68] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome biology*, 17(1):1–14, 2016.

[69] C Titus Brown and Luiz Irber. sourmash: a library for minhash sketching of dna. *Journal of open source software*, 1(5):27, 2016.

[70] Biyuan Chen, Xueyi He, Bangquan Pan, Xiaobing Zou, and Na You. Comparison of beta diversity measures in clustering the high-dimensional microbial data. *PloS one*, 16(2):e0246893, 2021.

[71] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44 (2):139–160, 1982.

[72] Thomas P Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F Richardson, and Tamsyn M Crowley. A field guide for the compositional analysis of any-omics data. *GigaScience*, 8(9): giz107, 2019.

[73] Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

[74] Kasper U Kjeldsen, Lars Schreiber, Casper A Thorup, Thomas Boesen, Jesper T Bjerg, et al. On the evolution and physiology of cable bacteria. *Proceedings of the National Academy of Sciences*, 116(38):19116–19125, 2019.

[75] Chris Greening and Trevor Lithgow. Formation and function of bacterial organelles. *Nature Reviews Microbiology*, 18(12):677–689, 2020.

[76] Petra Anne Levin and Esther R Angert. Small but mighty: cell size and bacteria. *Cold Spring Harbor perspectives in biology*, 7(7): a019216, 2015.

[77] Rohit Ghai, Carolina Megumi Mizuno, Antonio Picazo, Antonio Camacho, and Francisco Rodriguez-Valera. Metagenomics uncovers a new group of low gc and ultra-small marine actinobacteria. *Scientific Reports*, 3(1):2471, Aug 2013. ISSN 2045-2322. doi: 10.1038/srep02471. URL `https://doi.org/10.1038/srep02471`.

[78] FRANCIS CRICK. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970. ISSN 1476-4687. doi: 10.1038/227561a0. URL `https://doi.org/10.1038/227561a0`.

[79] Lianet Noda-Garcia, Wolfram Liebermeister, and Dan S. Tawfik. Metaboliteenzyme coevolution hypothesis. our hypothesis addresses the origins of new metabolites and of new enzymes and the order of their recruitment. we aim to not only survey established knowledge but also present open questions and potential ways of addressing them.

[80] Ferric C Fang, Elaine R Frawley, Timothy Tapscott, and Andrés Vázquez-Torres. Bacterial stress responses during host infection. *Cell host & microbe*, 20(2):133–143, 2016.

[81] Nancy Merino, Heidi S Aronson, Diana P Bojanova, Jayme Feyhl-Buska, Michael L Wong, Shu Zhang, and Donato Giovannelli. Living at the extremes: extremophiles and the limits of life in a planetary context. *Frontiers in microbiology*, 10:780, 2019.

[82] Chiaki Kato, Lina Li, Yuichi Nogi, Yuka Nakamura, Jin Tamaoka, and Koki Horikoshi. Extremely barophilic bacteria isolated from the mariana trench, challenger deep, at a depth of 11,000 meters. *Applied and environmental microbiology*, 64(4):1510–1513, 1998.

[83] Jiwei Li, Zhiyan Chen, Xinxin Li, Shun Chen, Hengchao Xu, et al. The sources of organic carbon in the deepest ocean: implication from bacterial membrane lipids in the mariana trench zone. *Frontiers in Earth Science*, 9:653742, 2021.

[84] Susanne Schultze-Lam, D Fortin, BS Davis, and TJ Beveridge. Mineralization of bacterial surfaces. *Chemical Geology*, 132(1-4): 171–181, 1996.

[85] Florence Postollec, Anne-Gabrielle Mathot, Muriel Bernard, Marie-Laure Divanac'h, Sonia Pavan, and Danièle Sohier. Tracking spore-forming bacteria in food: from natural biodiversity to selection by processes. *International journal of food microbiology*, 158(1):1–8, 2012.

[86] Yu-Ming Cai. Non-surface attached bacterial aggregates: a ubiquitous third lifestyle. *Frontiers in Microbiology*, 11:557035, 2020.

[87] Hans-Curt Flemming and Stefan Wuertz. Bacteria and archaea on earth and their abundance in biofilms. *Nature Reviews Microbiology*, 17(4):247–260, Apr 2019. ISSN 1740-1534. doi: 10.1038/s41579-019-0158-9. URL `https://doi.org/10.1038/s41579-019-0158-9`.

[88] Axel Schippers, Lev N Neretin, Jens Kallmeyer, Timothy G Ferdelman, Barry A Cragg, R John Parkes, and Bo B Jørgensen. Prokaryotic cells of the deep sub-seafloor biosphere identified as living bacteria. *Nature*, 433(7028):861–864, 2005.

[89] Mara Madalina Mihai, Monica Beatrice Dima, Bogdan Dima, and Alina Maria Holban. Nanomaterials for wound healing and infection control. *Materials*, 12(13):2176, 2019.

[90] Florence Hoefler, Xavier Pouget-Abadie, Mariam Roncato-Saberan, Romain Lemarié, Eve-Marie Takoudju, et al. Clinical and epidemiologic characteristics and therapeutic management of patients with vibrio infections, bay of biscay, france, 2001–2019. *Emerging Infectious Diseases*, 28(12):2367, 2022.

[91] Christopher F Sharpley and Clemens Koehn. Frequency and content of the last fifty years of papers on aristotle's writings on biological phenomena. *Journal of the History of Biology*, 55(3): 585–607, 2022.

[92] Carolus Linnaeus. *Systema naturae*, volume 1. Stockholm Laurentii Salvii, 1758.

[93] Frederick A Matsen and Aaron Gallagher. Reconciling taxonomy and phylogenetic inference: formalism and algorithms for describing discord and inferring taxonomic roots. *Algorithms for Molecular Biology*, 7:1–11, 2012.

[94] Donald Hobern, Saroj K Barik, Les Christidis, Stephen T. Garnett, Paul Kirk, et al. Towards a global list of accepted species vi: The catalogue of life checklist. *Organisms Diversity & Evolution*, 21(4):677–690, 2021.

[95] José M Padial, Aurélien Miralles, Ignacio De la Riva, and Miguel Vences. The integrative future of taxonomy. *Frontiers in zoology*, 7(1):1–14, 2010.

[96] Carl R Woese, Otto Kandler, and Mark L Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, 1990.

[97] James R Brown and W Ford Doolittle. Archaea and the prokaryote-to-eukaryote transition. *Microbiology and Molecular Biology Reviews*, 61(4):456–502, 1997.

[98] Gareth A Coleman, Adrián A Davín, Tara A Mahendrarajah, Lénárd L Szánthó, Anja Spang, Philip Hugenholtz, Gergely J Szöllősi, and Tom A Williams. A rooted phylogeny resolves early bacterial evolution. *Science*, 372(6542):eabe0511, 2021.

[99] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[100] Mark J Pallen, Andrea Telatin, and Aharon Oren. The next million names for archaea and bacteria. *Trends in microbiology*, 29(4):289–298, 2021.

[101] Rudolf Amann and Ramon Rosselló-Móra. After all, only millions? *MBio*, 7(4):10–1128, 2016.

[102] Lee Sweetlove. Number of species on earth tagged at 8.7 million. *Nature*, Aug 2011. ISSN 1476-4687. doi: 10.1038/news.2011.498. URL `https://doi.org/10.1038/news.2011.498`.

[103] Patrick D Schloss, Rene A Girard, Thomas Martin, Joshua Edwards, and J Cameron Thrash. Status of the archaeal and bacterial census: an update. *MBio*, 7(3):10–1128, 2016.

[104] Patrick D Schloss and Jo Handelsman. Status of the microbial census. *Microbiology and molecular biology reviews*, 68(4):686–691, 2004.

[105] Microbiology by numbers. *Nature Reviews Microbiology*, 9(9): 628–628, Sep 2011. ISSN 1740-1534. doi: 10.1038/nrmicro2644. URL `https://doi.org/10.1038/nrmicro2644`.

[106] Stilianos Louca, Florent Mazel, Michael Doebeli, and Laura Wegener Parfrey. A census-based estimate of earth's bacterial and archaeal diversity. *PLoS biology*, 17(2):e3000106, 2019.

[107] John J Wiens. How many species are there on earth? progress and problems. *PLoS biology*, 21(11):e3002388, 2023.

[108] Ilias Lagkouvardos, Jörg Overmann, and Thomas Clavel. Cultured microbes represent a substantial fraction of the human and mouse gut microbiota. *Gut microbes*, 8(5):493–503, 2017.

[109] Jean-Christophe Lagier, Saber Khelaifia, Maryam Tidjani Alou, Sokhna Ndongo, Niokhor Dione, et al. Culture of previously uncultured members of the human gut microbiota by culturomics. *Nature Microbiology*, 1(12):16203, Nov 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.203. URL `https://doi.org/10.1038/nmicrobiol.2016.203`.

[110] Hilary P Browne, Samuel C Forster, Blessing O Anonye, Nitin Kumar, B Anne Neville, Mark D Stares, David Goulding, and Trevor D Lawley. Culturing of 'unculturable'human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604): 543–546, 2016.

[111] Alan W Walker and Lesley Hoyles. Human microbiome myths and misconceptions. *Nature Microbiology*, 8(8):1392–1396, 2023.

[112] Andrew D Steen, Alexander Crits-Christoph, Paul Carini, Kristen M DeAngelis, Noah Fierer, Karen G Lloyd, and J Cameron Thrash. High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal*, 13 (12):3126–3130, 2019.

[113] Sonia R Vartoukian. Cultivation strategies for growth of uncultivated bacteria. *Journal of oral biosciences*, 58(4):143–149, 2016.

[114] Eric J Stewart. Growing unculturable bacteria. *Journal of bacteriology*, 194(16):4151–4160, 2012.

[115] Jörg Overmann, Birte Abt, and Johannes Sikorski. Present and future of culturing bacteria. *Annual Review of Microbiology*, 71(1):711–730, 2017. doi: 10.1146/annurev-micro-090816-093449. URL `https://doi.org/10.1146/annurev-micro-090816-093449`. PMID: 28731846.

[116] Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, et al. A new view of the tree of life. *Nature microbiology*, 1(5):1–6, 2016.

[117] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, 09 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab776. URL `https://doi.org/10.1093/nar/gkab776`.

[118] Andrew Polaszek. A universal register for animal names. *Nature*, 437(7058):477–477, 2005.

[119] Robert M May and Paul H Harvey. Species uncertainties, 2009.

[120] Guillaume Tahon, Patricia Geesink, and Thijs JG Ettema. Expanding archaeal diversity and phylogeny: past, present, and future. *Annual Review of Microbiology*, 75:359–381, 2021.

[121] Kevin S Ikuta, Lucien R Swetschinski, Gisela Robles Aguilar, Fablina Sharara, Tomislav Mestrovic, et al. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 400(10369):2221–2248, 2022.

[122] Abigail Bartlett, Daniel Padfield, Luke Lear, Richard Bendall, and Michiel Vos. A comprehensive list of bacterial pathogens infecting humans. *Microbiology*, 168(12):001269, 2022.

[123] Sam P Brown, Daniel M Cornforth, and Nicole Mideo. Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control. *Trends in microbiology*, 20(7):336–342, 2012.

[124] Theodore H. Tulchinsky and Elena A. Varavikova. Chapter 1 - a history of public health. In Theodore H. Tulchinsky and Elena A. Varavikova, editors, *The New Public Health (Third Edition)*, pages 1–42. Academic Press, San Diego, third edition edition, 2014. ISBN 978-0-12-415766-8. doi: https://doi.org/10.1016/B978-0-12-415766-8.00001-X. URL `https://www.sciencedirect.com/science/article/pii/B978012415766800001X`.

[125] Michelle C Swick, Theresa M Koehler, and Adam Driks. Surviving between hosts: sporulation and transmission. *Virulence mechanisms of bacterial pathogens*, pages 567–591, 2016.

[126] Shira Doron and Sherwood L Gorbach. Bacterial infections: overview. *International Encyclopedia of Public Health*, page 273, 2008.

[127] B Joseph Hinnebusch, Clayton O Jarrett, and David M Bland. Molecular and genetic mechanisms that mediate transmission of yersinia pestis by fleas. *Biomolecules*, 11(2):210, 2021.

[128] R Barbieri, M Signoli, D Chevé, C Costedoat, S Tzortzis, G Aboudharam, D Raoult, and M Drancourt. Yersinia pestis: the natural history of plague. *Clinical microbiology reviews*, 34(1): 10–1128, 2020.

[129] Biao Tang, Abubakar Siddique, Chenhao Jia, Abdelaziz Ed-Dra, Jing Wu, Hui Lin, and Min Yue. Genome-based risk assessment for foodborne salmonella enterica from food animals in china: A one health perspective. *International Journal of Food Microbiology*, 390:110120, 2023. ISSN 0168-1605. doi: https://doi.org/10.1016/j.ijfoodmicro.2023.110120. URL `https://www.sciencedirect.com/science/article/pii/S0168160523000363`.

[130] Daniel F Monte, Nilton Lincopan, Paula J Fedorka-Cray, and Mariza Landgraf. Current insights on high priority antibiotic-resistant salmonella enterica in food and foodstuffs: A review. *Current opinion in food science*, 26:35–46, 2019.

[131] Nabanita Mukherjee, Vikki G Nolan, John R Dunn, and Pratik Banerjee. Sources of human infection by salmonella enterica serotype javiana: A systematic review. *PLoS One*, 14(9):e0222108, 2019.

[132] Won-Il Cho and Myong-Soo Chung. Bacillus spores: A review of their properties and inactivation processing technologies. *Food science and biotechnology*, 29:1447–1461, 2020.

[133] Ken Inweregbu, Jayshree Dave, and Alison Pittard. Nosocomial infections. *Continuing Education in Anaesthesia, Critical Care & Pain*, 5(1):14–17, 2005.

[134] Johannes Oosterom. The importance of hygiene in modern society. *International Biodeterioration & Biodegradation*, 41(3-4): 185–189, 1998.

[135] Olga de la Varga-Martínez, Esther Gómez-Sánchez, María Fe Muñoz, Mario Lorenzo, Estefanía Gómez-Pesquera, Rodrigo Poves-Álvarez, Eduardo Tamayo, and María Heredia-Rodríguez. Impact of nosocomial infections on patient mortality following cardiac surgery. *Journal of Clinical Anesthesia*, 69: 110104, 2021.

[136] Tania Nawfal Dagher, Charbel Al-Bayssari, Seydina M Diene, Eid Azar, and Jean-Marc Rolain. Bacterial infection during wars, conflicts and post-natural disasters in asia and the middle east: a narrative review. *Expert review of anti-infective therapy*, 18(6): 511–529, 2020.

[137] Sophia David, Sandra Reuter, Simon R Harris, Corinna Glasner, Theresa Feltwell, et al. Epidemic of carbapenem-resistant klebsiella pneumoniae in europe is driven by nosocomial spread. *Nature microbiology*, 4(11):1919–1929, 2019.

[138] Lindsey M Weiner-Lastinger, Sheila Abner, Jonathan R Edwards, Alexander J Kallen, Maria Karlsson, et al. Antimicrobial-resistant pathogens associated with adult healthcare-associated infections: summary of data reported to the national healthcare safety network, 2015–2017. *Infection Control & Hospital Epidemiology*, 41(1):1–18, 2020.

[139] Thibaud Vermeil, Alexandra Peters, C Kilpatrick, Daniela Pires, Benedetta Allegranzi, and Didier Pittet. Hand hygiene in hospitals: anatomy of a revolution. *Journal of Hospital Infection*, 101 (4):383–392, 2019.

[140] Aiqin Chen, Ze Yuan, Hanyan Chen, Xuehui Wang, Huan Li, and Xinyue Zhang. Investigation into the current status of cleaning, disinfection, and sterilization of da vinci surgical instruments—a cross-sectional survey. *Gland Surgery*, 12(4):487, 2023.

[141] William A Rutala and David J Weber. Disinfection, sterilization, and antisepsis: an overview. *American journal of infection control*, 47:A3–A9, 2019.

[142] Ehsan Ahmadi, Dale T Masel, Ashley Y Metcalf, and Kristin Schuller. Inventory management of surgical supplies and sterile instruments in hospitals: a literature review. *Health Systems*, 8 (2):134–151, 2019.

[143] Sharmila Devi. War driving cholera in syria. *The Lancet*, 400 (10357):986, 2022.

[144] Holly C Betts, Mark N Puttick, James W Clark, Tom A Williams, Philip CJ Donoghue, and Davide Pisani. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature ecology & evolution*, 2(10):1556–1562, 2018.

[145] J William Schopf and Bonnie M Packer. Early archean (3.3-billion to 3.5-billion-year-old) microfossils from warrawoona group, australia. *Science*, 237(4810):70–73, 1987.

[146] Catherine F Demoulin, Yannick J Lara, Luc Cornet, Camille François, Denis Baurain, Annick Wilmotte, and Emmanuelle J Javaux. Cyanobacteria evolution: Insight from the fossil record. *Free Radical Biology and Medicine*, 140:206–223, 2019.

[147] Matthew S Dodd, Dominic Papineau, Tor Grenne, John F Slack, Martin Rittner, Franco Pirajno, Jonathan O'Neil, and Crispin TS Little. Evidence for early life in earth's oldest hydrothermal vent precipitates. *Nature*, 543(7643):60–64, 2017.

[148] Ewan Callaway. Oldest homo sapiens fossil claim rewrites our species' history. *Nature*, 546:289–293, 2017.

[149] Ewan DS Wolff, Steven W Salisbury, John R Horner, and David J Varricchio. Common avian infection plagued the tyrant dinosaurs. *PLoS One*, 4(9):e7288, 2009.

[150] Ruggero D'Anastasio, Bernhard Zipfel, Jacopo Moggi-Cecchi, Roscoe Stanyon, and Luigi Capasso. Possible brucellosis in an early hominin skeleton from sterkfontein, south africa. *PloS one*, 4(7):e6439, 2009.

[151] Ewen Callaway. Bronze age skeletons were earliest plague victims. *Nature*, 22, 2015.

[152] Kathryn A Glatter and Paul Finkelman. History of the plague: An ancient pandemic for the age of covid-19. *The American journal of medicine*, 134(2):176–181, 2021.

[153] Mark Achtman. How old are bacterial pathogens? *Proceedings of the Royal Society B: Biological Sciences*, 283(1836):20160990, 2016.

[154] Andreas G Nerlich, Christian J Haas, Albert Zink, Ulrike Szeimies, and Hjalmar G Hagedorn. Molecular evidence for tuberculosis in an ancient egyptian mummy. *The Lancet*, 350 (9088):1404, 1997.

[155] Wilmar L Salo, Arthur C Aufderheide, Jane Buikstra, and Todd A Holcomb. Identification of mycobacterium tuberculosis dna in a pre-columbian peruvian mummy. *Proceedings of the National Academy of Sciences*, 91(6):2091–2094, 1994.

[156] Steve M Blevins and Michael S Bronze. Robert koch and the 'golden age'of bacteriology. *International Journal of Infectious Diseases*, 14(9):e744–e751, 2010.

[157] Elizabeth Craik. *The'Hippocratic'corpus: Content and context*. Routledge, 2014.

[158] Jordan A Kempker and Greg S Martin. The changing epidemiology and definitions of sepsis. *Clinics in chest medicine*, 37(2): 165–179, 2016.

[159] Kathryn A Glatter and Paul Finkelman. History of the plague: An ancient pandemic for the age of covid-19. *The American journal of medicine*, 134(2):176–181, 2021.

[160] Nick Lane. The unseen world: reflections on leeuwenhoek (1677)'concerning little animals'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1666):20140344, 2015.

[161] Steven M Opal. A brief history of microbiology and immunology. *Vaccines: A biography*, pages 31–56, 2010.

[162] Miklós Kásler. Ignaz semmelweis, the saviour of mothers. *POLGÁRI SZEMLE: GAZDASÁGI ÉS TÁRSADALMI FOLYÓIRAT*, 14(Spec.):385–410, 2018.

[163] Gerhart Drews. The roots of microbiology and the influence of ferdinand cohn on microbiology of the 19th century. *FEMS Microbiology Reviews*, 24(3):225–249, 2000.

[164] Russell W Currier and John A Widness. A brief history of milk hygiene and its impact on infant mortality from 1875 to 1925 and implications for today: a review. *Journal of food protection*, 81(10):1713–1722, 2018.

[165] Kendall A Smith. Louis pasteur, the father of immunology? *Frontiers in immunology*, 3:68, 2012.

[166] Tamara Giles-Vernick, Phaik Yeong Cheah, Gustavo Matta, and Nisia Trindade Lima. Louis pasteur, covid-19, and the social challenges of epidemics. *The Lancet*, 400(10369):2166–2168, 2022.

[167] Russell Currier. Pasteurisation: Pasteur's greatest contribution to health. *The Lancet Microbe*, 4(3):e129–e130, 2023.

[168] Lary Walker, Harry Levine, and Mathias Jucker. Koch's postulates and infectious proteins. *Acta neuropathologica*, 112:1–4, 2006.

[169] Christoph Gradmann. Robert koch and the pressures of scientific research: tuberculosis and tuberculin. *Medical history*, 45(1): 1–32, 2001.

[170] A Sakula. Robert koch: centenary of the discovery of the tubercle bacillus, 1882. *Thorax*, 37(4):246–251, 1982. ISSN 0040-6376. doi: 10.1136/thx.37.4.246. URL `https://thorax.bmj.com/content/37/4/246`.

[171] Donatella Lippi and Eduardo Gotuzzo. The greatest steps towards the discovery of vibrio cholerae. *Clinical Microbiology and Infection*, 20(3):191–195, 2014.

[172] Thomas J Silhavy, Daniel Kahne, and Suzanne Walker. The bacterial cell envelope. *Cold Spring Harbor perspectives in biology*, 2(5):a000414, 2010.

[173] Richard Coico. Gram staining. *Current Protocols in Microbiology*, 00(1):A.3C.1–A.3C.2, 2006. doi: https://doi.org/10.1002/9780471729259.mca03cs00. URL `https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/9780471729259.mca03cs00`.

[174] Manfred Rohde. The gram-positive bacterial cell wall. *Microbiology Spectrum*, 7(3):10–1128, 2019.

[175] Alexander JF Egan, Jeff Errington, and Waldemar Vollmer. Regulation of peptidoglycan synthesis and remodelling. *Nature Reviews Microbiology*, 18(8):446–460, 2020.

[176] Katherine R Hummels, Samuel P Berry, Zhaoqi Li, Atsushi Taguchi, Joseph K Min, Suzanne Walker, Debora S Marks, and Thomas G Bernhardt. Coordination of bacterial cell wall and outer membrane biosynthesis. *Nature*, 615(7951):300–304, 2023.

[177] Tharani Vijayakumar, Bose Divya, V Vasanthi, Madhu Narayan, Annasamy Ramesh Kumar, and Rajkumar Krishnan. Diagnostic utility of gram stain for oral smears–a review. *Journal of Microscopy and Ultrastructure*, 2023.

[178] Nicolas Jacquier, Patrick H Viollier, and Gilbert Greub. The role of peptidoglycan in chlamydial cell division: towards resolving the chlamydial anomaly. *FEMS microbiology reviews*, 39(2):262–275, 2015.

[179] Henrik Strahl and Jeff Errington. Bacterial membranes: structure, domains, and function. *Annual review of microbiology*, 71: 519–538, 2017.

[180] Yun Yang, Jiwei Liu, Bradley R Clarke, Laura Seidel, Jani R Bolla, Philip N Ward, Peijun Zhang, Carol V Robinson, Chris Whitfield, and James H Naismith. The molecular basis of regulation of bacterial capsule assembly by wzc. *Nature Communications*, 12(1):4349, 2021.

[181] Zhensong Wen and Jing-Ren Zhang. Chapter 3 - bacterial capsules. In Yi-Wei Tang, Max Sussman, Dongyou Liu, Ian Poxton, and Joseph Schwartzman, editors, *Molecular Medical Microbiology (Second Edition)*, pages 33–53. Academic Press, Boston, second edition edition, 2015. ISBN 978-0-12-397169-2. doi: https://doi.org/10.1016/B978-0-12-397169-2.00003-2. URL `https://www.sciencedirect.com/science/article/pii/B9780123971692000032`.

[182] Farhana R Pinu, Ninna Granucci, James Daniell, Ting-Li Han, Sonia Carneiro, Isabel Rocha, Jens Nielsen, and Silas G Villas-Boas. Metabolite secretion in microorganisms: the theory of metabolic overflow put to the test. *Metabolomics*, 14:1–16, 2018.

[183] Matthew I Hutchings, Andrew W Truman, and Barrie Wilkinson. Antibiotics: past, present and future. *Current opinion in microbiology*, 51:72–80, 2019.

[184] Shuichi Nakamura and Tohru Minamino. Flagella-driven motility of bacteria. *Biomolecules*, 9(7):279, 2019.

[185] Akila Sridhar. The inner workings of the flagellar motor. *Nature Reviews Microbiology*, 18(12):673–673, 2020.

[186] Mikako Fujii, Satoshi Shibata, and Shin-Ichi Aizawa. Polar, peritrichous, and lateral flagella belong to three distinguishable flagellar families. *Journal of molecular biology*, 379(2):273–283, 2008.

[187] Stephen Melville and Lisa Craig. Type iv pili in gram-positive bacteria. *Microbiology and molecular biology reviews*, 77(3):323–341, 2013.

[188] Manuela K Hospenthal, Tiago RD Costa, and Gabriel Waksman. A comprehensive guide to pilus biogenesis in gram-negative bacteria. *Nature reviews microbiology*, 15(6):365–379, 2017.

[189] Nicholas A Ramirez and Hung Ton-That. *Bacterial Pili and Fimbriae*, pages 1–13. John Wiley & Sons, Ltd, 2020. ISBN 9780470015902. doi: https://doi.org/10.1002/9780470015902.a0000304.pub3. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000304.pub3`.

[190] Gang Ren, Xia Zhou, Ruimin Long, Maobin Xie, Ranjith Kumar Kankala, Shibin Wang, Yu Shrike Zhang, and Yuangang Liu. Biomedical applications of magnetosomes: State of the art and perspectives. *Bioactive Materials*, 28:27–49, 2023.

[191] D Muñoz, L Marcano, R Martín-Rodríguez, L Simonelli, A Serrano, A García-Prieto, ML Fdez-Gubieda, and A Muela. Magnetosomes could be protective shields against metal stress in magnetotactic bacteria. *Scientific Reports*, 10(1):11430, 2020.

[192] Christopher T Lefèvre and Dennis A Bazylinski. Ecology, diversity, and evolution of magnetotactic bacteria. *Microbiology and Molecular Biology Reviews*, 77(3):497–526, 2013.

[193] Bradley R Parry, Ivan V Surovtsev, Matthew T Cabeen, Corey S O'Hern, Eric R Dufresne, and Christine Jacobs-Wagner. The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity. *Cell*, 156(1):183–194, 2014.

[194] Akira Ishihama. The nucleoid: an overview. *EcoSal Plus*, 3(2): 10–1128, 2009.

[195] Mark Buchanan. Sizing up bacteria. *Nature Physics*, 10(11): 788–788, 2014.

[196] C Baril, C Richaud, G Baranton, and I Saint Girons. Linear chromosome of borrelia burgdorferi. *Research in microbiology*, 140(7):507–516, 1989.

[197] Aditi Kanhere and Manju Bansal. Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic acids research*, 33(10):3165–3175, 2005.

[198] Jacek Kominek, Drew T Doering, Dana A Opulente, Xing-Xing Shen, Xiaofan Zhou, et al. Eukaryotic acquisition of a bacterial operon. *Cell*, 176(6):1356–1366, 2019.

[199] Max A English, Raphaël V Gayet, and James J Collins. Designing biological circuits: synthetic biology within the operon model and beyond. *Annual Review of Biochemistry*, 90:221–244, 2021.

[200] Anna Ciok, Lukasz Dziewit, Jakub Grzesiak, Karol Budzik, Dorota Gorniak, Marek K Zdanowski, and Dariusz Bartosik. Identification of miniature plasmids in psychrophilic arctic bacteria of the genus variovorax. *FEMS Microbiology Ecology*, 92 (4):fiw043, 2016.

[201] James PJ Hall, João Botelho, Adrian Cazares, and David A Baltrus. What makes a megaplasmid? *Philosophical Transactions of the Royal Society B*, 377(1842):20200472, 2022.

[202] Donald R Helinski. A brief history of plasmids. *EcoSal Plus*, 10 (1):eESP–0028, 2022.

[203] Masaki Shintani, Zoe K Sanchez, and Kazuhide Kimbara. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Frontiers in microbiology*, 6:242, 2015.

[204] Oleg Reva and Burkhard Tümmler. Think big–giant genes in bacteria. *Environmental microbiology*, 10(3):768–777, 2008.

[205] T Ryan Gregory. Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics*, 6(9):699–708, 2005.

[206] Vishwa Patel and Nishad Matange. Adaptation and compensation in a bacterial gene regulatory network evolving under antibiotic selection. *Elife*, 10:e70931, 2021.

[207] Azra Ćutuk, Eda Sarić Hanjalić, Sibel Repuh, and Nuraiym Mamatnazarova. Gene regulation pathway modeling. In *2020 9th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–3, 2020. doi: 10.1109/MECO49872.2020.9134084.

[208] Hiten D Madhani. The frustrated gene: origins of eukaryotic gene expression. *Cell*, 155(4):744–749, 2013.

[209] Leise Riber and Lars Hestbjerg Hansen. Epigenetic memories: the hidden drivers of bacterial persistence? *Trends in Microbiology*, 29(3):190–194, 2021.

[210] Monica Rolando, Cristina Di Silvestre, Laura Gomez-Valero, and Carmen Buchrieser. Bacterial methyltransferases: from targeting bacterial genomes to host epigenetics. *Microlife*, 3: uqac014, 2022.

[211] Antoine Hocher, Shawn P. Laursen, Paul Radford, Jess Tyson, Carey Lambert, Kathryn M Stevens, Mathieu Picardeau, R. Elizabeth Sockett, Karolin Luger, and Tobias Warnecke. Histone-organized chromatin in bacteria. *bioRxiv*, 2023. doi: 10.1101/2023.01.26.525422. URL `https://www.biorxiv.org/content/early/2023/01/26/2023.01.26.525422`.

[212] Bibhusita Pani and Evgeny Nudler. Bacterial histones unveiled. *Nature Microbiology*, pages 1–3, 2023.

[213] Antoine Hocher, Shawn P Laursen, Paul Radford, Jess Tyson, Carey Lambert, et al. Histones with an unconventional dna-binding mode in vitro are major chromatin constituents in the bacterium bdellovibrio bacteriovorus. *Nature Microbiology*, pages 1–14, 2023.

[214] Josep Casadesús and David Low. Epigenetic gene regulation in the bacterial world. *Microbiology and molecular biology reviews*, 70(3):830–856, 2006.

[215] María A Sánchez-Romero, David R Olivenza, Gabriel Gutiérrez, and Josep Casadesús. Contribution of dna adenine methylation to gene expression heterogeneity in salmonella enterica. *Nucleic Acids Research*, 48(21):11857–11867, 2020.

[216] Brian P Anton and Richard J Roberts. Beyond restriction modification: epigenomic roles of dna methylation in prokaryotes. *Annual Review of Microbiology*, 75:129–149, 2021.

[217] Matthew J Blow, Tyson A Clark, Chris G Daum, Adam M Deutschbauer, Alexey Fomenkov, et al. The epigenomic landscape of prokaryotes. *PLoS genetics*, 12(2):e1005854, 2016.

[218] Satish Adhikari and Patrick D Curtis. Dna methyltransferases and epigenetic regulation in bacteria. *FEMS microbiology reviews*, 40(5):575–591, 2016.

[219] Amaury Payelleville and Julien Brillard. Novel identification of bacterial epigenetic regulations would benefit from a better exploitation of methylomic data. *Frontiers in microbiology*, 12: 685670, 2021.

[220] Douglas F Browning and Stephen JW Busby. Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, 14(10):638–650, 2016.

[221] Douglas F Browning and Stephen JW Busby. Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, 14(10):638–650, 2016.

[222] Citlalli Mejía-Almonte, Stephen JW Busby, Joseph T Wade, Jacques van Helden, Adam P Arkin, Gary D Stormo, Karen Eilbeck, Bernhard O Palsson, James E Galagan, and Julio Collado-Vides. Redefining fundamental concepts of transcription initiation in bacteria. *Nature Reviews Genetics*, 21(11):699–714, 2020.

[223] Patrick Cramer. Organization and regulation of gene transcription. *Nature*, 573(7772):45–54, 2019.

[224] Agnes Ullmann. Escherichia coli lactose operon. *eLS*, 2009.

[225] Marjan W Van Der Woude and Andreas J Bäumler. Phase and antigenic variation in bacteria. *Clinical microbiology reviews*, 17 (3):581–611, 2004.

[226] Marjan W Van der Woude. Phase variation: how to create and coordinate population diversity. *Current opinion in microbiology*, 14(2):205–211, 2011.

[227] James Chen, Hande Boyaci, and Elizabeth A Campbell. Diverse and unified mechanisms of transcription initiation in bacteria. *Nature Reviews Microbiology*, 19(2):95–109, 2021.

[228] Sofia Österberg, Teresa del Peso-Santos, and Victoria Shingler. Regulation of alternative sigma factor use. *Annual review of microbiology*, 65:37–55, 2011.

[229] Robert S Washburn and Max E Gottesman. Regulation of transcription elongation and termination. *Biomolecules*, 5(2): 1063–1078, 2015.

[230] Enrique Merino, Roy A Jensen, and Charles Yanofsky. Evolution of bacterial trp operons and their regulation. *Current opinion in microbiology*, 11(2):78–86, 2008.

[231] Michelle A Kriner, Anastasia Sevostyanova, and Eduardo A Groisman. Learning from the leaders: gene regulation by the transcription termination factor rho. *Trends in biochemical sciences*, 41(8):690–699, 2016.

[232] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355, 2010.

[233] Chengyuan Wang, Vadim Molodtsov, Emre Firlar, Jason T Kaelber, Gregor Blaha, Min Su, and Richard H Ebright. Structural basis of transcription-translation coupling. *Science*, 369(6509): 1359–1365, 2020.

[234] Kumari Kavita and Ronald R Breaker. Discovering riboswitches: the past and the future. *Trends in biochemical sciences*, 2023.

[235] Edmund Loh, Olivier Dussurget, Jonas Gripenland, Karolis Vaitkevicius, Teresa Tiensuu, Pierre Mandin, Francis Repoila, Carmen Buchrieser, Pascale Cossart, and Jörgen Johansson. A trans-acting riboswitch controls expression of the virulence regulator prfa in listeria monocytogenes. *Cell*, 139(4):770–779, 2009.

[236] Chelsea A Schiano and Wyndham W Lathem. Post-transcriptional regulation of gene expression in yersinia species. *Frontiers in cellular and infection microbiology*, 2:129, 2012.

[237] Luary C Martínez and Viveka Vadyvaloo. Mechanisms of post-transcriptional gene regulation in bacterial biofilms. *Frontiers in cellular and infection microbiology*, 4:38, 2014.

[238] Teresa Nogueira and Mathias Springer. Post-transcriptional control by global regulators of gene expression in bacteria. *Current opinion in microbiology*, 3(2):154–158, 2000.

[239] Elke Van Assche, Sandra Van Puyvelde, Jos Vanderleyden, and Hans P Steenackers. Rna-binding proteins involved in post-transcriptional regulation in bacteria. *Frontiers in microbiology*, 6:141, 2015.

[240] George A Mackie. Rnase e: at the interface of bacterial rna processing and decay. *Nature Reviews Microbiology*, 11(1):45–57, 2013.

[241] Susanne Huch, Lilit Nersisyan, Maria Ropat, Donal Barrett, Mengjun Wu, et al. Atlas of mrna translation and decay for bacteria. *Nature Microbiology*, pages 1–14, 2023.

[242] Liang Xue, Swantje Lenz, Maria Zimmermann-Kogadeeva, Dimitry Tegunov, Patrick Cramer, Peer Bork, Juri Rappsilber, and Julia Mahamid. Visualizing translation dynamics at atomic detail inside a bacterial cell. *Nature*, 610(7930):205–211, 2022.

[243] Sandip Kaledhonkar, Ziao Fu, Kelvin Caban, Wen Li, Bo Chen, Ming Sun, Ruben L Gonzalez Jr, and Joachim Frank. Late steps in bacterial translation initiation visualized using time-resolved cryo-em. *Nature*, 570(7761):400–404, 2019.

[244] Arlie J Rinaldi, Paul E Lund, Mario R Blanco, and Nils G Walter. The shine-dalgarno sequence of riboswitch-regulated single mrnas shows ligand-dependent accessibility bursts. *Nature communications*, 7(1):8976, 2016.

[245] Sujatha Thankeswaran Parvathy, Varatharajalu Udayasuriyan, and Vijaipal Bhadana. Codon usage bias. *Molecular biology reports*, 49(1):539–565, 2022.

[246] Yi Liu, Qian Yang, and Fangzhou Zhao. Synonymous but not silent: the codon usage code for gene expression and protein folding. *Annual review of biochemistry*, 90:375–401, 2021.

[247] Ekaterina Samatova, Jan Daberger, Marija Liutkute, and Marina V Rodnina. Translational control by ribosome pausing in bacteria: how a non-uniform pace of translation affects protein production and folding. *Frontiers in microbiology*, 11:619430, 2021.

[248] Charlotte Guyomar, Gaetano D'urso, Sophie Chat, Emmanuel Giudice, and Reynald Gillet. Structures of tmrna and smpb as they transit through the ribosome. *Nature communications*, 12 (1):4909, 2021.

[249] Boris Macek, Karl Forchhammer, Julie Hardouin, Eilika Weber-Ban, Christophe Grangeasse, and Ivan Mijakovic. Protein post-translational modifications in bacteria. *Nature Reviews Microbiology*, 17(11):651–664, 2019.

[250] Joaquín Sánchez and Jan Holmgren. Cholera toxin—a foe & a friend. *The Indian journal of medical research*, 133(2):153, 2011.

[251] Michael Taylor, David Curtis, and Ken Teter. A conformational shift in the dissociated cholera toxin a1 subunit prevents reassembly of the cholera holotoxin. *Toxins*, 7(7):2674–2684, 2015.

[252] Henry P Wood, F Aaron Cruz-Navarrete, Nicola J Baxter, Clare R Trevitt, Angus J Robertson, Samuel R Dix, Andrea M Hounslow, Matthew J Cliff, and Jonathan P Waltho. Allomorphy as a mechanism of post-translational control of enzyme activity. *Nature Communications*, 11(1):5538, 2020.

[253] D Thirumalai, Changbong Hyeon, Pavel I Zhuravlev, and George H Lorimer. Symmetry, rigidity, and allosteric signaling: from monomeric proteins to molecular machines. *Chemical reviews*, 119(12):6788–6821, 2019.

[254] Erin K Schrader, Kristine G Harstad, and Andreas Matouschek. Targeting proteins for degradation. *Nature chemical biology*, 5 (11):815–822, 2009.

[255] Matylda Anna Izert, Maria Magdalena Klimecka, and Maria Wiktoria Górna. Applications of bacterial degrons and degraders—toward targeted protein degradation in bacteria. *Frontiers in Molecular Biosciences*, 8:669762, 2021.

[256] Victor Chubukov, Luca Gerosa, Karl Kochanowski, and Uwe Sauer. Coordination of microbial metabolism. *Nature Reviews Microbiology*, 12(5):327–340, 2014.

[257] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[258] Christina R Bourne. Utility of the biosynthetic folate pathway for targets in antimicrobial discovery. *Antibiotics*, 3(1):1–28, 2014.

[259] Zheng Lu and James A Imlay. When anaerobes encounter oxygen: mechanisms of oxygen toxicity, tolerance and defence. *Nature reviews microbiology*, 19(12):774–785, 2021.

[260] Meina Neumann-Schaal, Dieter Jahn, and Kerstin Schmidt-Hohagen. Metabolism the difficile way: the key to the success of the pathogen clostridioides difficile. *Frontiers in microbiology*, 10:219, 2019.

[261] Maryam Khademian and James A Imlay. How microbes evolved to tolerate oxygen. *Trends in Microbiology*, 29(5):428–440, 2021.

[262] Drishya M George, Annette S Vincent, and Hamish R Mackey. An overview of anoxygenic phototrophic bacteria and their applications in environmental biotechnology for sustainable resource recovery. *Biotechnology reports*, 28:e00563, 2020.

[263] Patricia Sánchez-Baracaldo and Tanai Cardona. On the origin of oxygenic photosynthesis and cyanobacteria. *New Phytologist*, 225(4):1440–1446, 2020.

[264] Jason Olejarz, Yoh Iwasa, Andrew H Knoll, and Martin A Nowak. The great oxygenation event as a consequence of ecological dynamics modulated by planetary change. *Nature Communications*, 12(1):3985, 2021.

[265] Hiroshi Kiyota, Yukiko Okuda, Michiho Ito, Masami Yokota Hirai, and Masahiko Ikeuchi. Engineering of cyanobacteria for the photosynthetic production of limonene from co2. *Journal of biotechnology*, 185:1–7, 2014.

[266] Beverly E Flood, Deon C Louw, Anja K Van der Plas, and Jake V Bailey. Giant sulfur bacteria (beggiatoaceae) from sediments underlying the benguela upwelling system host diverse microbiomes. *Plos one*, 16(11):e0258124, 2021.

[267] Fanuel Kawaka. Characterization of symbiotic and nitrogen fixing bacteria. *AMB Express*, 12(1):99, 2022.

[268] Jun Yang, Liying Lan, Yue Jin, Nan Yu, Dong Wang, and Ertao Wang. Mechanisms underlying legume–rhizobium symbioses. *Journal of Integrative Plant Biology*, 64(2):244–267, 2022.

[269] Esther R Angert. Alternatives to binary fission in bacteria. *Nature Reviews Microbiology*, 3(3):214–224, 2005.

[270] Alix Meunier, François Cornet, and Manuel Campos. Bacterial cell proliferation: from molecules to cells. *FEMS Microbiology Reviews*, 45(1):fuaa046, 2021.

[271] Damian Trojanowski, Joanna Hołówka, and Jolanta Zakrzewska-Czerwińska. Where and when bacterial chromosome replication starts: a single cell perspective. *Frontiers in Microbiology*, 9:2819, 2018.

[272] Zhi-Qiang Xu and Nicholas E Dixon. Bacterial replisomes. *Current opinion in structural biology*, 53:159–168, 2018.

[273] Jana Hirsch and Dagmar Klostermeier. What makes a type iia topoisomerase a gyrase or a topo iv? *Nucleic Acids Research*, 49 (11):6027–6042, 2021.

[274] Todd A Cameron and William Margolin. Insights into the assembly and regulation of the bacterial divisome. *Nature Reviews Microbiology*, pages 1–13, 2023.

[275] Suckjoon Jun, Fangwei Si, Rami Pugatch, and Matthew Scott. Fundamental principles in bacterial physiology—history, recent progress, and the future with focus on cell size control: a review. *Reports on Progress in Physics*, 81(5):056601, 2018.

[276] Elitza I Tocheva, Davi R Ortega, and Grant J Jensen. Sporulation, bacterial cell envelopes and the origin of life. *Nature Reviews Microbiology*, 14(8):535–542, 2016.

[277] Johannes M Keegstra, Francesco Carrara, and Roman Stocker. The ecological roles of bacterial chemotaxis. *Nature Reviews Microbiology*, 20(8):491–504, 2022.

[278] Hans-Curt Flemming, Jost Wingender, Ulrich Szewzyk, Peter Steinberg, Scott A Rice, and Staffan Kjelleberg. Biofilms: an emergent form of bacterial life. *Nature Reviews Microbiology*, 14 (9):563–575, 2016.

[279] Karin Sauer, Paul Stoodley, Darla M Goeres, Luanne Hall-Stoodley, Mette Burmølle, Philip S Stewart, and Thomas Bjarnsholt. The biofilm life cycle: Expanding the conceptual model of biofilm formation. *Nature Reviews Microbiology*, 20(10):608–620, 2022.

[280] N Allocati, M Masulli, C Di Ilio, and V De Laurenzi. Die for the community: an overview of programmed cell death in bacteria. *Cell death & disease*, 6(1):e1609–e1609, 2015.

[281] Georgia C Drew, Emily J Stevens, and Kayla C King. Microbial evolution and transitions along the parasite–mutualist continuum. *Nature Reviews Microbiology*, 19(10):623–638, 2021.

[282] Yves F Dufrêne and Albertus Viljoen. Binding strength of gram-positive bacterial adhesins. *Frontiers in microbiology*, 11:1457, 2020.

[283] Seema Patel, Nithya Mathivanan, and Arun Goyal. Bacterial adhesins, the pathogenic weapons to trick host defense arsenal. *Biomedicine & Pharmacotherapy*, 93:763–771, 2017.

[284] Mumtaz Virji. Pathogenic neisseriae: surface modulation, pathogenesis and infection control. *Nature Reviews Microbiology*, 7(4):274–286, 2009.

[285] Zineb Boumart, Philippe Velge, and Agnès Wiedemann. Multiple invasion mechanisms and different intracellular behaviors: a new vision of salmonella–host cell interaction. *FEMS microbiology letters*, 361(1):1–7, 2014.

[286] Sónia Castanheira and Francisco García-Del Portillo. Salmonella populations inside host cells. *Frontiers in cellular and infection microbiology*, 7:432, 2017.

[287] Preeti Malik-Kale, Carrie E Jolly, Stephanie Lathrop, Seth Winfree, Courtney Luterbach, and Olivia Steele-Mortimer. Salmonella–at home in the host cell. *Frontiers in microbiology*, 2: 125, 2011.

[288] B Brett Finlay and Grant McFadden. Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. *Cell*, 124(4):767–782, 2006.

[289] Pallavi Chandra, Steven J Grigsby, and Jennifer A Philips. Immune evasion and provocation by mycobacterium tuberculosis. *Nature Reviews Microbiology*, 20(12):750–766, 2022.

[290] Madeleine C van Dijk, Robin M de Kruijff, and Peter-Leon Hagedoorn. The role of iron in staphylococcus aureus infection and human disease: A metal tug of war at the host—microbe interface. *Frontiers in cell and developmental biology*, 10:857237, 2022.

[291] Min Dong, Geoffrey Masuyer, and Pål Stenmark. Botulinum and tetanus neurotoxins. *Annual review of biochemistry*, 88:811–837, 2019.

[292] Naresh Chand Sharma, Androulla Efstratiou, Igor Mokrousov, Ankur Mutreja, Bhabatosh Das, and Thandavarayan Ramamurthy. Diphtheria. *Nature Reviews Disease Primers*, 5(1):1–18, 2019.

[293] Olugbenga Ehuwa, Amit K Jaiswal, and Swarna Jaiswal. Salmonella, food safety and food handling practices. *Foods*, 10(5):907, 2021.

[294] Monica Gallo, Lydia Ferrara, Armando Calogero, Domenico Montesano, and Daniele Naviglio. Relationships between food and diseases: What to know to ensure food safety. *Food Research International*, 137:109414, 2020.

[295] Muhammad Tanveer Munir, Narjes Mtimet, Laurent Guillier, François Meurens, Phillipe Fravalo, Michel Federighi, and Pauline Kooh. Physical treatments to control clostridium botulinum hazards in food. *Foods*, 12(8):1580, 2023.

[296] Stéphane André, Tatiana Vallaeys, and Stella Planchon. Spore-forming bacteria responsible for food spoilage. *Research in Microbiology*, 168(4):379–387, 2017.

[297] Solveig Langsrud, Oddvin Sørheim, Silje Elisabeth Skuland, Valérie Lengard Almli, Merete Rusås Jensen, Magnhild Seim Grøvlen, Øydis Ueland, and Trond Møretrø. Cooking chicken at home: Common or recommended approaches to judge doneness may not assure sufficient inactivation of pathogens. *PLoS One*, 15(4):e0230928, 2020.

[298] James T Walker and Paul J McDermott. Confirming the presence of legionella pneumophila in your water system: a review of current legionella testing methods. *Journal of AOAC International*, 104(4):1135–1147, 2021.

[299] Chiqian Zhang and Jingrang Lu. Legionella: a promising supplementary indicator of microbial drinking water quality in municipal engineered water systems. *Frontiers in environmental science*, 9:684319, 2021.

[300] Lisa Paruch, Adam M Paruch, and Roald Sørheim. Dna-based faecal source tracking of contaminated drinking water causing a large campylobacter outbreak in norway 2019. *International Journal of Hygiene and Environmental Health*, 224:113420, 2020.

[301] Kubra F Naqvi, Stuart B Mazzone, and Michael U Shiloh. Infectious and inflammatory pathways to cough. *Annual review of physiology*, 85:71–91, 2023.

[302] Michelle E Wood, Rebecca E Stockwell, Graham R Johnson, Kay A Ramsay, Laura J Sherrard, et al. Face masks and cough etiquette reduce the cough aerosol concentration of pseudomonas aeruginosa in people with cystic fibrosis. *American journal of respiratory and critical care medicine*, 197(3):348–355, 2018.

[303] Michelle E Wood, Rebecca E Stockwell, Graham R Johnson, Kay A Ramsay, Laura J Sherrard, et al. Cystic fibrosis pathogens survive for extended periods within cough-generated droplet nuclei. *Thorax*, 74(1):87–90, 2019.

[304] Anuradha Sharma, Jitu M Kalita, and Vijaya L Nag. Screening for methicillin-resistant staphylococcus aureus carriage on the hands of healthcare workers: an assessment for hand hygiene practices. *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, 23 (12):590, 2019.

[305] Rajeshwari Nair, Eli N Perencevich, Michihiko Goto, Daniel J Livorsi, Erin Balkenende, Elizabeth Kiscaden, and Marin L Schweizer. Patient care experience with utilization of isolation precautions: systematic literature review and meta-analysis. *Clinical Microbiology and Infection*, 26(6):684–695, 2020.

[306] YS Marfatia, Ipsa Pandya, and Kajal Mehta. Condoms: Past, present, and future. *Indian journal of sexually transmitted diseases and AIDS*, 36(2):133, 2015.

[307] Anke Osterloh. Vaccination against bacterial infections: challenges, progress, and new approaches with a focus on intracellular bacteria. *Vaccines*, 10(5):751, 2022.

[308] Isabel Frost, Hatim Sati, Pilar Garcia-Vello, Mateusz Hasso-Agopsowicz, Christian Lienhardt, Valeria Gigante, and Peter

Beyer. The role of bacterial vaccines in the fight against antimicrobial resistance: an analysis of the preclinical and clinical development pipeline. *The Lancet Microbe*, 2022.

[309] Arezoo Asadi, Shabnam Razavi, Malihe Talebi, and Mehrdad Gholami. A review on anti-adhesion therapies of bacterial diseases. *Infection*, 47:13–23, 2019.

[310] Fernando L Gordillo Altamirano and Jeremy J Barr. Phage therapy in the postantibiotic era. *Clinical microbiology reviews*, 32(2):10–1128, 2019.

[311] Kotaro Kiga, Xin-Ee Tan, Rodrigo Ibarra-Chávez, Shinya Watanabe, Yoshifumi Aiba, et al. Development of crispr-cas13a-based antimicrobials capable of sequence-specific killing of target bacteria. *Nature communications*, 11(1):2934, 2020.

[312] Diana Garibo, Hugo A Borbón-Nuñez, Jorge N Díaz de León, Ernesto García Mendoza, Iván Estrada, et al. Green synthesis of silver nanoparticles using lysiloma acapulcensis exhibit high-antimicrobial activity. *Scientific reports*, 10(1):12805, 2020.

[313] Graham F Hatfull, Rebekah M Dedrick, and Robert T Schooley. Phage therapy for antibiotic-resistant bacterial infections. *Annual Review of Medicine*, 73:197–211, 2022.

[314] Lindsey Matthews, Jennifer S Goodrich, David J Weber, Nicholas H Bergman, and Melissa B Miller. The brief case: A fatal case of necrotizing fasciitis due to multidrug-resistant acinetobacter baumannii, 2019.

[315] Scott JC Pallett, Alex Trompeter, Marina Basarab, Luke SP Moore, and Sara E Boyd. Multidrug-resistant infections in war victims in ukraine. *The Lancet Infectious Diseases*, 23(8):e270–e271, 2023.

[316] Romney M Humphries and Andrea J Linscott. Practical guidance for clinical microbiology laboratories: diagnosis of bacterial gastroenteritis. *Clinical microbiology reviews*, 28(1):3–31, 2015.

[317] Jean-Christophe Lagier, Grégory Dubourg, Matthieu Million, Frédéric Cadoret, Melhem Bilen, Florence Fenollar, Anthony Levasseur, Jean-Marc Rolain, Pierre-Edouard Fournier, and Didier Raoult. Culturing the human microbiota and culturomics. *Nature Reviews Microbiology*, 16(9):540–550, 2018.

[318] Shi-Kai Yan, Run-Hui Liu, Hui-Zi Jin, Xin-Ru Liu, Ji Ye, Lei Shan, and Wei-Dong Zhang. " omics" in pharmaceutical research: overview, applications, challenges, and future perspectives. *Chinese journal of natural medicines*, 13(1):3–21, 2015.

[319] Jiajing Wang, Leping Jiang, and Hanlong Sun. Early evidence for beer drinking in a 9000-year-old platform mound in southern china. *PLoS One*, 16(8):e0255833, 2021.

[320] Klaus B Lengeler, Vratislav Stovicek, Ross T Fennessy, Michael Katz, and Jochen Förster. Never change a brewing yeast? why not, there are plenty to choose from. *Frontiers in Genetics*, 11: 582789, 2020.

[321] Mélanie Salque, Peter I Bogucki, Joanna Pyzel, Iwona Sobkowiak-Tabaka, Ryszard Grygiel, Marzena Szmyt, and Richard P Evershed. Earliest evidence for cheese making in the sixth millennium bc in northern europe. *Nature*, 493(7433): 522–525, 2013.

[322] Jan Steensels, Brigida Gallone, Karin Voordeckers, and Kevin J Verstrepen. Domestication of industrial microbes. *Current biology*, 29(10):R381–R393, 2019.

[323] Ross Kent and Neil Dixon. Contemporary tools for regulating gene expression in bacteria. *Trends in biotechnology*, 38(3):316–333, 2020.

[324] Yiting Liu, Jing Feng, Hangcheng Pan, Xiuwei Zhang, and Yunlei Zhang. Genetically engineered bacterium: Principles, practices, and prospects. *Frontiers in Microbiology*, 13:997587, 2022.

[325] Manish Kumar, Smita Sundaram, Edgard Gnansounou, Christian Larroche, and Indu Shekhar Thakur. Carbon dioxide capture, storage and production of biofuel and biomaterials by bacteria: A review. *Bioresource technology*, 247:1059–1068, 2018.

[326] Kanchan Samadhiya, Rimjhim Sangtani, Regina Nogueira, and Kiran Bala. Insightful advancement and opportunities for microbial bioplastic production. *Frontiers in Microbiology*, 12:674864, 2022.

[327] Inès Mnif and Dhouha Ghribi. Potential of bacterial derived biopesticides in pest management. *Crop Protection*, 77:52–64, 2015.

[328] Tithi Mehrotra, Subhabrata Dev, Aditi Banerjee, Abhijit Chatterjee, Rachana Singh, and Srijan Aggarwal. Use of immobilized bacteria for environmental bioremediation: A review. *Journal of Environmental Chemical Engineering*, 9(5):105920, 2021.

[329] Sharjeel Waqas, Muhammad Roil Bilad, Zakaria Man, Yusuf Wibisono, Juhana Jaafar, Teuku Meurah Indra Mahlia, Asim Laeeq Khan, and Muhammad Aslam. Recent progress in integrated fixed-film activated sludge process for wastewater treatment: A review. *Journal of environmental management*, 268: 110718, 2020.

[330] Timothy L Haskett, Andrzej Tkacz, and Philip S Poole. Engineering rhizobacteria for sustainable agriculture. *The ISME Journal*, 15(4):949–964, 2021.

[331] Maria Eugenia Inda and Timothy K Lu. Microbes as biosensors. *Annual Review of Microbiology*, 74:337–359, 2020.

[332] Jennifer A Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with crispr-cas9. *Science*, 346 (6213):1258096, 2014.

[333] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *science*, 337(6096):816–821, 2012.

[334] Marnix H Medema, Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B Biggins, et al. Minimum information about a biosynthetic gene cluster. *Nature chemical biology*, 11(9):625–631, 2015.

[335] Nehal Adel Abdelsalam, Mohamed Elhadidy, Nehal A Saif, Salma W Elsayed, Shaimaa F Mouftah, Ahmed A Sayed, and Laila Ziko. Biosynthetic gene cluster signature profiles of pathogenic gram-negative bacteria isolated from egyptian clinical settings. *Microbiology Spectrum*, 11(5):e01344–23, 2023.

[336] Jaycee Augusto Gumiran Paguirigan, Jung A Kim, Jae-Seoun Hur, and Wonyong Kim. Identification of a biosynthetic gene cluster for a red pigment cristazarin produced by a lichen-forming fungus cladonia metacorallifera. *Plos one*, 18(6): e0287559, 2023.

[337] Rahim Khan, Farinazleen Mohamad Ghazali, Nor Ainy Mahyudin, and Nik Iskandar Putra Samsudin. Aflatoxin biosynthesis, genetic regulation, toxicity, and control strategies: A review. *Journal of Fungi*, 7(8):606, 2021.

[338] Satria A Kautsar, Justin JJ van der Hooft, Dick de Ridder, and Marnix H Medema. Big-slice: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience*, 10(1):giaa154, 2021.

[339] Satria A Kautsar, Kai Blin, Simon Shaw, Tilmann Weber, and Marnix H Medema. Big-fam: the biosynthetic gene cluster families database. *Nucleic acids research*, 49(D1):D490–D497, 2021.

[340] Jorge C Navarro-Muñoz, Nelly Selem-Mojica, Michael W Mullowney, Satria A Kautsar, James H Tryon, et al. A computational framework to explore large-scale biosynthetic diversity. *Nature chemical biology*, 16(1):60–68, 2020.

[341] Kirstin Scherlach and Christian Hertweck. Mining and unearthing hidden biosynthetic potential. *Nature Communications*, 12(1):3864, 2021.

[342] Barbara R Terlouw, Kai Blin, Jorge C Navarro-Munoz, Nicole E Avalon, Marc G Chevrette, et al. Mibig 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic acids research*, 51(D1):D603–D610, 2023.

[343] Natsuko Ichikawa, Machi Sasagawa, Mika Yamamoto, Hisayuki Komaki, Yumi Yoshida, Shuji Yamazaki, and Nobuyuki Fujita. Dobiscuit: a database of secondary metabolite biosynthetic gene clusters. *Nucleic acids research*, 41(D1):D408–D414, 2012.

[344] Kai Blin, Hyun Uk Kim, Marnix H Medema, and Tilmann Weber. Recent development of antismash and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in Bioinformatics*, 20(4):1103–1113, 2019.

[345] Eric D Brown and Gerard D Wright. Antibacterial drug discovery in the resistance era. *Nature*, 529(7586):336–343, 2016.

[346] Kaitlyn C Belknap, Cooper J Park, Brian M Barth, and Cheryl P Andam. Genome mining of biosynthetic and chemotherapeutic gene clusters in streptomyces bacteria. *Scientific Reports*, 10(1): 2003, 2020.

[347] Marnix H Medema, Tristan de Rond, and Bradley S Moore. Mining genomes to illuminate the specialized chemistry of life. *Nature Reviews Genetics*, 22(9):553–571, 2021.

[348] Malek Zerikly and Gregory L Challis. Strategies for the discovery of new natural products by genome mining. *ChemBioChem*, 10(4):625–633, 2009.

[349] Namil Lee, Soonkyu Hwang, Jihun Kim, Suhyung Cho, Bernhard Palsson, and Byung-Kwan Cho. Mini review: Genome mining approaches for the identification of secondary metabolite biosynthetic gene clusters in streptomyces. *Computational and Structural Biotechnology Journal*, 18:1548–1556, 2020.

[350] Peter J Rutledge and Gregory L Challis. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature reviews microbiology*, 13(8):509–523, 2015.

[351] Edward Kalkreuter, Guohui Pan, Alexis J Cepeda, and Ben Shen. Targeting bacterial genomes for natural product discovery. *Trends in Pharmacological Sciences*, 41(1):13–26, 2020.

[352] Wenjun Li, Didier Raoult, and Pierre-Edouard Fournier. Bacterial strain typing in the genomic era. *FEMS microbiology reviews*, 33(5):892–916, 2009.

[353] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *science*, 269(5223):496–512, 1995.

[354] Bo Liu, Dandan Zheng, Siyu Zhou, Lihong Chen, and Jian Yang. Vfdb 2022: a general classification scheme for bacterial virulence factors. *Nucleic acids research*, 50(D1):D912–D917, 2022.

[355] Nicola De Maio, Liam P Shaw, Alasdair Hubbard, Sophie George, Nicholas D Sanderson, et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial genomics*, 5(9):e000294, 2019.

[356] Vivien Marx. Method of the year: long-read sequencing. *Nature Methods*, 20(1):6–11, 2023.

[357] Ryan R Wick, Louise M Judd, and Kathryn E Holt. Assembling the perfect bacterial genome using oxford nanopore and illumina sequencing. *PLOS Computational Biology*, 19(3):e1010905, 2023.

[358] Nathan D Olson, Steven P Lund, Rebecca E Colman, Jeffrey T Foster, Jason W Sahl, James M Schupp, Paul Keim, Jayne B Morrow, Marc L Salit, and Justin M Zook. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in genetics*, 6:235, 2015.

[359] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 2012.

[360] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[361] Tatiana Tatusova, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D Pruitt, Mark Borodovsky, and James Ostell. Ncbi prokaryotic genome annotation pipeline. *Nucleic acids research*, 44(14):6614–6624, 2016.

[362] Anshul Sharma, Sulhee Lee, and Young-Seo Park. Molecular typing tools for identifying and characterizing lactic acid bacteria: a review. *Food science and biotechnology*, 29:1301–1318, 2020.

[363] Paschalia Kapli, Ziheng Yang, and Maximilian J Telford. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7):428–444, 2020.

[364] Johannes Nemeth, Gabriela Oesch, and Stefan P Kuster. Bacteriostatic versus bactericidal antibiotics for patients with serious bacterial infections: systematic review and meta-analysis. *Journal of Antimicrobial Chemotherapy*, 70(2):382–395, 2015.

[365] Rustam I Aminov. A brief history of the antibiotic era: lessons learned and challenges for the future. *Frontiers in microbiology*, 1:134, 2010.

[366] Søren Brøgger Christensen. Drugs that changed society: History and current status of the early antibiotics: Salvarsan, sulfonamides, and $\beta$-lactams. *Molecules*, 26(19):6057, 2021.

[367] Alexander Fleming. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae. *Bulletin of the World Health Organization*, 79: 780–790, 2001.

[368] Ronald Bentley. Different roads to discovery; prontosil (hence sulfa drugs) and penicillin (hence $\beta$-lactams). *Journal of Industrial Microbiology and Biotechnology*, 36(6):775–786, 2009.

[369] Albert Schatz, Elizabeth Bugle, and Selman A Waksman. Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. *Proceedings of the Society for Experimental Biology and Medicine*, 55(1):66–69, 1944.

[370] John Ehrlich, Quentin R Bartz, Robert M Smith, Dwight A Joslyn, and Paul R Burkholder. Chloromycetin, a new antibiotic from a soil actinomycete. *Science*, 106(2757):417–417, 1947.

[371] BM Duggar. Aureomycin: a product of the continuing search for new antibiotics. *Annals of the New York Academy of Sciences*, 1241(1):163–169, 2011.

[372] Mariya Lobanovska and Giulia Pilla. Focus: drug development: Penicillin's discovery and antibiotic resistance: lessons for the future? *The Yale journal of biology and medicine*, 90(1):135, 2017.

[373] Dov Stekel. First report of antimicrobial resistance pre-dates penicillin. *Nature*, 562(7726), 2018.

[374] EP Abraham and E Chain. An enzyme from bacteria able to destroy penicillin. 1940. *Reviews of infectious diseases*, 10(4): 677–678, 1988.

[375] Charles H Rammelkamp and Thelma Maxon. Resistance of staphylococcus aureus to the action of penicillin. *Proceedings of the Society for Experimental Biology and Medicine*, 51(3):386–389, 1942.

[376] David Hansman, Lorraine Devitt, Helen Miles, and Ian Riley. Pneumococci relatively insensitive to penicillin in australia and new guinea. *Medical Journal of Australia*, 2(10):353–356, 1974.

[377] Mark C Enright, D Ashley Robinson, Gaynor Randle, Edward J Feil, Hajo Grundmann, and Brian G Spratt. The evolutionary history of methicillin-resistant staphylococcus aureus (mrsa). *Proceedings of the National Academy of Sciences*, 99(11):7687–7692, 2002.

[378] Andie S Lee, Hermínia De Lencastre, Javier Garau, Jan Kluytmans, Surbhi Malhotra-Kumar, Andreas Peschel, and Stephan

Harbarth. Methicillin-resistant staphylococcus aureus. *Nature reviews Disease primers*, 4(1):1–23, 2018.

[379] Mary Barber. Methicillin-resistant staphylococci. *Journal of clinical pathology*, 14(4):385, 1961.

[380] Teruyo Ito and Keiichi Hiramatsu. Acquisition of methicillin resistance and progression of multiantibiotic resistance in methicillin-resistant staphylococcus aureus. *Yonsei medical journal*, 39(6):526–533, 1998.

[381] Jessy Lallungawi Khawbung, Durbba Nath, and Supriyo Chakraborty. Drug resistant tuberculosis: A review. *Comparative immunology, microbiology and infectious diseases*, 74:101574, 2021.

[382] Katia Iskandar, Jayaseelan Murugaiyan, Dalal Hammoudi Halat, Said El Hage, Vindana Chibabhai, Saranya Adukkadukkam, Christine Roques, Laurent Molinier, Pascale Salameh, and Maarten Van Dongen. Antibiotic discovery and resistance: the chase and the race. *Antibiotics*, 11(2):182, 2022.

[383] Marcus Miethke, Marco Pieroni, Tilmann Weber, Mark Brönstrup, Peter Hammann, et al. Towards the sustainable discovery and development of new antibiotics. *Nature Reviews Chemistry*, 5(10):726–749, 2021.

[384] Evelina Tacconelli and Maria Diletta Pezzani. Public health burden of antimicrobial resistance in europe. *The Lancet Infectious Diseases*, 19(1):4–6, 2019.

[385] Alessandro Cassini, Liselotte Diaz Högberg, Diamantis Plachouras, Annalisa Quattrocchi, Ana Hoxha, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the eu and the european economic area in 2015: a population-level modelling analysis. *The Lancet infectious diseases*, 19(1):56–66, 2019.

[386] Zhigang Yu, Yue Wang, Ian R Henderson, and Jianhua Guo. Artificial sweeteners stimulate horizontal transfer of extracellular antibiotic resistance genes through natural transformation. *The ISME Journal*, 16(2):543–554, 2022.

[387] JIM O'neill. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. *Rev. Antimicrob. Resist.*, 2014.

[388] Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, 2022.

[389] Santiago Redondo-Salvo, Raúl Fernández-López, Raúl Ruiz, Luis Vielva, María de Toro, Eduardo PC Rocha, M Pilar Garcillán-Barcia, and Fernando de la Cruz. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nature communications*, 11(1):3602, 2020.

[390] Asad U Khan, Lubna Maryam, and Raffaele Zarrilli. Structure, genetics and worldwide spread of new delhi metallo-$\beta$-lactamase (ndm): a threat to public health. *BMC microbiology*, 17:1–12, 2017.

[391] JM Rolain, P Parola, and G Cornaglia. New delhi metallo-beta-lactamase (ndm-1): towards a new pandemia? *Clinical microbiology and infection*, 16(12):1699–1701, 2010.

[392] Fernando Baquero, Teresa M Coque, José-Luis Martínez, Sonia Aracil-Gisbert, and Val F Lanza. Gene transmission in the one health microbiosphere and the channels of antimicrobial resistance. *Frontiers in microbiology*, 10:2892, 2019.

[393] Brian J Arnold, I-Ting Huang, and William P Hanage. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, 20(4):206–218, 2022.

[394] Quentin J Leclerc, Jodi A Lindsay, and Gwenan M Knight. Mathematical modelling to study the horizontal transfer of antimicrobial resistance genes in bacteria: current state of the field and recommendations. *Journal of the royal society interface*, 16(157):20190260, 2019.

[395] Phillip Nazarian, Frances Tran, and James Q Boedicker. Modeling multispecies gene flow dynamics reveals the unique roles of different horizontal gene transfer mechanisms. *Frontiers in microbiology*, 9:2978, 2018.

[396] Elena Cabezón, Jorge Ripoll-Rozada, Alejandro Peña, Fernando De La Cruz, and Ignacio Arechaga. Towards an integrated model of bacterial conjugation. *FEMS microbiology reviews*, 39 (1):81–95, 2015.

[397] Kazuki Moriguchi, Noritaka Edahiro, Shinji Yamamoto, Katsuyuki Tanaka, Nori Kurata, and Katsunori Suzuki. Transkingdom genetic transfer from escherichia coli to saccharomyces cerevisiae as a simple gene introduction tool. *Applied and environmental microbiology*, 79(14):4393–4400, 2013.

[398] Calum Johnston, Bernard Martin, Gwennaele Fichant, Patrice Polard, and Jean-Pierre Claverys. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nature Reviews Microbiology*, 12(3):181–196, 2014.

[399] Trond Erik Vee Aune and Finn Lillelund Aachmann. Methodologies to increase the transformation efficiencies and the range of bacteria that can be transformed. *Applied microbiology and biotechnology*, 85:1301–1313, 2010.

[400] Anna Colavecchio, Brigitte Cadieux, Amanda Lo, and Lawrence D Goodridge. Bacteriophages contribute to the spread of antibiotic resistance genes among foodborne

pathogens of the enterobacteriaceae family–a review. *Frontiers in microbiology*, 8:1108, 2017.

[401] Jose Luis Balcazar. Bacteriophages as vehicles for antibiotic resistance genes in the environment. *PLoS pathogens*, 10(7): e1004219, 2014.

[402] Rafael Cantón, Murat Akova, Karen Langfeld, and Didem Torumkuney. Relevance of the consensus principles for appropriate antibiotic prescribing in 2022. *Journal of Antimicrobial Chemotherapy*, 77(Supplement_1):i2–i9, 2022.

[403] Steward Mudenda, Victor Daka, and Scott K Matafwali. World health organization aware framework for antibiotic stewardship: Where are we now and where do we need to go? an expert viewpoint. *Antimicrobial Stewardship & Healthcare Epidemiology*, 3(1):e84, 2023.

[404] Mike Sharland, Bernadette Cappello, Loice Achieng Ombajo, Joel Bazira, Ronald Chitatanga, et al. The who aware antibiotic book: providing guidance on optimal use and informing policy. *The Lancet Infectious Diseases*, 22(11):1528–1530, 2022.

[405] Veronica Zanichelli, Michael Sharland, Bernadette Cappello, Lorenzo Moja, Haileyesus Getahun, et al. The who aware (access, watch, reserve) antibiotic book and prevention of antimicrobial resistance. 2023.

[406] Yingfen Hsia, Brian R Lee, Ann Versporten, Yonghong Yang, Julia Bielicki, et al. Use of the who access, watch, and reserve classification to define patterns of hospital antibiotic use (aware): an analysis of paediatric survey data from 56 countries. *The Lancet Global Health*, 7(7):e861–e871, 2019.

[407] Daniel M Webber, Meghan A Wallace, and Carey-Ann D Burnham. Stop waiting for tomorrow: disk diffusion performed on early growth is an accurate method for antimicrobial susceptibility testing with reduced turnaround time. *Journal of clinical microbiology*, 60(5):e03007–20, 2022.

[408] Jan Hudzicki. Kirby-bauer disk diffusion susceptibility test protocol. *American society for microbiology*, 15:55–63, 2009.

[409] Lucien Barnes, Douglas M Heithoff, Scott P Mahan, John K House, and Michael J Mahan. Antimicrobial susceptibility testing to evaluate minimum inhibitory concentration values of clinically relevant antibiotics. *STAR protocols*, 4(3):102512, 2023.

[410] Henry L Stennett, Catherine R Back, and Paul R Race. Derivation of a precise and consistent timeline for antibiotic development. *Antibiotics*, 11(9):1237, 2022.

[411] Lidia Moreira Lima, Bianca Nascimento Monteiro da Silva, Gisele Barbosa, and Eliezer J Barreiro. $\beta$-lactam antibiotics: An

overview from a medicinal chemistry perspective. *European journal of medicinal chemistry*, 208:112829, 2020.

[412] Jed F Fisher and Shahriar Mobashery. Constructing and deconstructing the bacterial cell wall. *Protein science*, 29(3):629–646, 2020.

[413] Daina Zeng, Dmitri Debabov, Theresa L Hartsell, Raul J Cano, Stacy Adams, Jessica A Schuyler, Ronald McMillan, and John L Pace. Approved glycopeptide antibacterial drugs: mechanism of action and resistance. *Cold Spring Harbor perspectives in medicine*, 6(12), 2016.

[414] 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.

[415] Jaroslav Spížek and Tomáš Řezanka. Lincosamides: Chemical structure, biosynthesis, mechanism of action, resistance, and applications. *Biochemical pharmacology*, 133:20–28, 2017.

[416] Tanel Tenson, Martin Lovmar, and Måns Ehrenberg. The mechanism of action of macrolides, lincosamides and streptogramin b reveals the nascent peptide exit path in the ribosome. *Journal of molecular biology*, 330(5):1005–1014, 2003.

[417] Nora Vázquez-Laslop and Alexander S Mankin. How macrolide antibiotics work. *Trends in biochemical sciences*, 43(9):668–684, 2018.

[418] Jeffrey L Hansen, Joseph A Ippolito, Nenad Ban, Poul Nissen, Peter B Moore, and Thomas A Steitz. The structures of four macrolide antibiotics bound to the large ribosomal subunit. *Molecular cell*, 10(1):117–128, 2002.

[419] S Lohsen and DS Stephens. Current macrolide antibiotics and their mechanisms of action. *Antibiotic drug resistance*, pages 97–117, 2019.

[420] David C Hooper and George A Jacoby. Topoisomerase inhibitors: fluoroquinolone mechanisms of action and resistance. *Cold Spring Harbor perspectives in medicine*, 6(9), 2016.

[421] Piotr Wierzbiński, Joanna Hubska, Michał Henzler, Bartłomiej Kucharski, Rafał Bieś, and Marek Krzystanek. Depressive and other adverse cns effects of fluoroquinolones. *Pharmaceuticals*, 16(8):1105, 2023.

[422] Aben Ovung and Jhimli Bhattacharyya. Sulfonamide drugs: Structure, antibacterial property, toxicity, and biophysical interactions. *Biophysical reviews*, 13(2):259–272, 2021.

[423] William PJ Smith, Benjamin R Wucher, Carey D Nadell, and Kevin R Foster. Bacterial defences: mechanisms, evolution and antimicrobial resistance. *Nature Reviews Microbiology*, pages 1–16, 2023.

[424] Wanda C Reygaert. An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS microbiology*, 4(3):482, 2018.

[425] Jose M Munita and Cesar A Arias. Mechanisms of antibiotic resistance. *Virulence mechanisms of bacterial pathogens*, pages 481–511, 2016.

[426] Elizabeth M Darby, Eleftheria Trampari, Pauline Siasat, Maria Solsona Gaya, Ilyas Alav, Mark A Webber, and Jessica MA Blair. Molecular mechanisms of antibiotic resistance revisited. *Nature Reviews Microbiology*, 21(5):280–295, 2023.

[427] Manar Ali Abushaheen, Amal Jamil Fatani, Mohammed Alosaimi, Wael Mansy, Merin George, et al. Antimicrobial resistance, mechanisms and its clinical significance. *Disease-a-Month*, 66(6):100971, 2020.

[428] William PJ Smith, Benjamin R Wucher, Carey D Nadell, and Kevin R Foster. Bacterial defences: mechanisms, evolution and antimicrobial resistance. *Nature Reviews Microbiology*, pages 1–16, 2023.

[429] Idan Yelin and Roy Kishony. Antibiotic resistance. *Cell*, 172(5): 1136–1136, 2018.

[430] Elizabeth M Darby, Eleftheria Trampari, Pauline Siasat, Maria Solsona Gaya, Ilyas Alav, Mark A Webber, and Jessica MA Blair. Molecular mechanisms of antibiotic resistance revisited. *Nature Reviews Microbiology*, 21(5):280–295, 2023.

[431] Catherine L Tooke, Philip Hinchliffe, Eilis C Bragginton, Charlotte K Colenso, Viivi HA Hirvonen, Yuiko Takebayashi, and James Spencer. $\beta$-lactamases and $\beta$-lactamase inhibitors in the 21st century. *Journal of molecular biology*, 431(18):3472–3500, 2019.

[432] Maria S Ramirez and Marcelo E Tolmasky. Aminoglycoside modifying enzymes. *Drug resistance updates*, 13(6):151–171, 2010.

[433] Adam J Schaenzer and Gerard D Wright. Antibiotic resistance by enzymatic modification of antibiotic targets. *Trends in molecular medicine*, 26(8):768–782, 2020.

[434] Ichrak Benamri, Maryame Azzouzi, Kholoud Sanak, Ahmed Moussa, and Fouzia Radouani. An overview of genes and mutations associated with chlamydiae species' resistance to antibiotics. *Annals of Clinical Microbiology and Antimicrobials*, 20: 1–11, 2021.

[435]

[436] Theresa C Barrett, Wendy WK Mok, Allison M Murawski, and Mark P Brynildsen. Enhanced antibiotic resistance development from fluoroquinolone persisters after a single exposure to antibiotic. *Nature communications*, 10(1):1177, 2019.

[437] Kim Lewis. Persister cells. *Annual review of microbiology*, 64: 357–372, 2010.

[438] Adam C Palmer and Roy Kishony. Opposing effects of target overexpression reveal drug mechanisms. *Nature communications*, 5(1):4296, 2014.

[439] Meenakshi Venkatesan, Michael Fruci, Lou Ann Verellen, Tatiana Skarina, Nathalie Mesa, Robert Flick, Chester Pham, Radhakrishnan Mahadevan, Peter J Stogios, and Alexei Savchenko. Molecular mechanism of plasmid-borne resistance to sulfonamide antibiotics. *Nature Communications*, 14(1):4031, 2023.

[440] World Health Organization et al. Antimicrobial resistance surveillance in europe 2022–2020 data. 2022.

[441] J-M Rolain, C Abat, M-T Jimeno, P-E Fournier, and Didier Raoult. Do we need new antibiotics? *Clinical Microbiology and Infection*, 22(5):408–415, 2016.

[442] Kasim Allel, Lucy Day, Alisa Hamilton, Leesa Lin, Luis Furuya-Kanamori, Catrin E Moore, Thomas Van Boeckel, Ramanan Laxminarayan, and Laith Yakob. Global antimicrobial-resistance drivers: an ecological country-level study at the human–animal interface. *The Lancet Planetary Health*, 7(4):e291–e303, 2023.

[443] Bradley J Langford, Miranda So, Marina Simeonova, Valerie Leung, Jennifer Lo, et al. Antimicrobial resistance in patients with covid-19: a systematic review and meta-analysis. *The Lancet Microbe*, 2023.

[444] Clark Donald Russell. Eradicating infectious disease: can we and should we? *Frontiers in Immunology*, 2:53, 2011.

[445] Jay Patel, Anne Harant, Genevie Fernandes, Ambele Judith Mwamelo, Wolfgang Hein, Denise Dekker, and Devi Sridhar. Measuring the global response to antimicrobial resistance, 2020–21: a systematic governance analysis of 114 countries. *The Lancet Infectious Diseases*, 23(6):706–718, 2023.

[446] Angela Willemsen, Simon Reid, and Yibeltal Assefa. A review of national action plans on antimicrobial resistance: strengths and weaknesses. *Antimicrobial Resistance & Infection Control*, 11 (1):90, 2022.

[447] Darija Kuruc Poje, Vesna Mađarić, Vlatka Janeš Poje, Domagoj Kifer, Philip Howard, and Srećko Marušić. Antimicrobial stewardship effectiveness on rationalizing the use of last line of

antibiotics in a short period with limited human resources: a single centre cohort study. *BMC Research Notes*, 12(1):1–6, 2019.

[448] Samiran Banerjee and Marcel GA van der Heijden. Soil microbiomes and one health. *Nature Reviews Microbiology*, 21(1):6–20, 2023.

[449] Salvador Castañeda-Barba, Eva M Top, and Thibault Stalder. Plasmids, a molecular cornerstone of antimicrobial resistance in the one health era. *Nature Reviews Microbiology*, 22(1):18–32, 2024.

[450] Fernando Pérez-Rodríguez and Birce Mercanoglu Taban. A state-of-art review on multi-drug resistant pathogens in foods of animal origin: risk factors and mitigation strategies. *Frontiers in Microbiology*, 10:2091, 2019.

[451] Shabbir Simjee and Gabriella Ippolito. European regulations on prevention use of antimicrobials from january 2022. *Brazilian Journal of Veterinary Medicine*, 44, 2022.

[452] Odion O Ikhimiukor, Erkison Ewomazino Odih, Pilar Donado-Godoy, and Iruka N Okeke. A bottom-up view of antimicrobial resistance transmission in developing countries. *Nature microbiology*, 7(6):757–765, 2022.

[453] Didier Wernli, Peter S Jørgensen, Stephan Harbarth, Scott P Carroll, Ramanan Laxminarayan, Nicolas Levrat, John-Arne Røttingen, and Didier Pittet. Antimicrobial resistance: the complex challenge of measurement to inform policy and the public. *PLoS medicine*, 14(8):e1002378, 2017.

[454] World Health Organization et al. Global research agenda for antimicrobial resistance in human health. *World Health Organization. https://www. who. int/publications/m/item/global-research-agenda-forantimicrobial-resistance-in-human-health*, 2023.

[455] Salvador Castañeda-Barba, Eva M Top, and Thibault Stalder. Plasmids, a molecular cornerstone of antimicrobial resistance in the one health era. *Nature Reviews Microbiology*, pages 1–15, 2023.

[456] Avoxa – Mediengruppe Deutscher Apotheker GmbH. Eravacyclin: Xerava®: 10: 2022. URL `https://www.pharmazeutische-zeitung.de/arzneistoffe/daten/2022/eravacyclinxeravar102022/`.

[457] Khalid Eljaaly, Jessica K Ortwine, Mohammed Shaikhomer, Thamer A Almangour, and Matteo Bassetti. Efficacy and safety of eravacycline: A meta-analysis. *Journal of Global Antimicrobial Resistance*, 24:424–428, 2021.

[458] Congjuan Xu, Xiaoya Wei, Yongxin Jin, Fang Bai, Zhihui Cheng, Shuiping Chen, Xiaolei Pan, and Weihui Wu. Development

of resistance to eravacycline by klebsiella pneumoniae and collateral sensitivity-guided design of combination therapies. *Microbiology Spectrum*, 10(5):e01390–22, 2022.

[459] Zewen Wen, Yongpeng Shang, Guangjian Xu, Zhangya Pu, Zhiwei Lin, Bing Bai, Zhong Chen, Jinxin Zheng, Qiwen Deng, and Zhijian Yu. Mechanism of eravacycline resistance in clinical enterococcus faecalis isolates from china. *Frontiers in Microbiology*, 11:916, 2020.

[460] Carmelo Biondo. Bacterial antibiotic resistance: The most critical pathogens, 2023.

[461] Mark S Butler, Ian R Henderson, Robert J Capon, and Mark AT Blaskovich. Antibiotics in the clinical pipeline as of december 2022. *The Journal of Antibiotics*, pages 1–43, 2023.

[462] D Thomas and C Wessel. The state of innovation in antibacterial therapeutics. *Reports BIA, editor*, 2022.

[463] Federico Serral, Florencia A Castello, Ezequiel J Sosa, Agustín M Pardo, Miranda Clara Palumbo, et al. From genome to drugs: New approaches in antimicrobial discovery. *Frontiers in Pharmacology*, 12:647060, 2021.

[464] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[465] Paul S Hoffman. Antibacterial discovery: 21st century challenges. *Antibiotics*, 9(5):213, 2020.

[466] Kim Lewis. The science of antibiotic discovery. *Cell*, 181(1): 29–45, 2020.

[467] Kai Blin, Hyun Uk Kim, Marnix H Medema, and Tilmann Weber. Recent development of antismash and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in Bioinformatics*, 20(4):1103–1113, 2019.

[468] Kai Blin, Simon Shaw, Alexander M Kloosterman, Zach Charlop-Powers, Gilles P Van Wezel, Marnix H Medema, and Tilmann Weber. antismash 6.0: improving cluster detection and comparison capabilities. *Nucleic acids research*, 49(W1):W29–W35, 2021.

[469] Dmitry Meleshko, Hosein Mohimani, Vittorio Tracanna, Iman Hajirasouliha, Marnix H Medema, Anton Korobeynikov, and Pavel A Pevzner. Biosyntheticspades: reconstructing biosynthetic gene clusters from assembly graphs. *Genome research*, 29 (8):1352–1362, 2019.

[470] Geoffrey D Hannigan, David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, et al. A deep learning genome-mining

strategy for biosynthetic gene cluster prediction. *Nucleic acids research*, 47(18):e110–e110, 2019.

[471] Michael A Skinnider, Chad W Johnston, Mathusan Gunabalasingam, Nishanth J Merwin, Agata M Kieliszek, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nature communications*, 11(1):6058, 2020.

[472] Manish Boolchandani, Alaric W D'Souza, and Gautam Dantas. Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics*, 20(6):356–370, 2019.

[473] Bo Liu and Mihai Pop. Ardb—antibiotic resistance genes database. *Nucleic acids research*, 37(suppl_1):D443–D447, 2009.

[474] Michael Feldgarden, Vyacheslav Brover, Narjol Gonzalez-Escalona, Jonathan G Frye, Julie Haendiges, et al. Amrfinderplus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific reports*, 11(1):12728, 2021.

[475] Sushim Kumar Gupta, Babu Roshan Padmanabhan, Seydina M Diene, Rafael Lopez-Rojas, Marie Kempf, Luce Landraud, and Jean-Marc Rolain. Arg-annot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial agents and chemotherapy*, 58(1):212–220, 2014.

[476] Nathalie Bonin, Enrique Doster, Hannah Worley, Lee J Pinnell, Jonathan E Bravo, et al. Megares and amr++, v3. 0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic acids research*, 51(D1):D744–D752, 2023.

[477] James J Davis, Alice R Wattam, Ramy K Aziz, Thomas Brettin, Ralph Butler, et al. The patric bioinformatics resource center: expanding data and analysis capabilities. *Nucleic acids research*, 48(D1):D606–D612, 2020.

[478] Martin Hunt, Alison E Mather, Leonor Sánchez-Busó, Andrew J Page, Julian Parkhill, Jacqueline A Keane, and Simon R Harris. Ariba: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial genomics*, 3(10), 2017.

[479] Michael Inouye, Harriet Dashnow, Lesley-Ann Raven, Mark B Schultz, Bernard J Pope, Takehiro Tomita, Justin Zobel, and Kathryn E Holt. Srst2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine*, 6(11): 1–16, 2014.

[480] Emily F Wissel, Brooke M Talbot, Noriko AB Toyosato, Robert A Petit III, Vicki Hertzberg, Anne Dunlop, and Timothy D Read.

hamroaster: a tool for comparing performance of amr gene detection software. *bioRxiv*, pages 2022–01, 2022.

[481] Julian R. Marchesi and Jacques Ravel. The vocabulary of microbiome research: a proposal. *Microbiome*, 3(1):31, Jul 2015. ISSN 2049-2618. doi: 10.1186/s40168-015-0094-5. URL `https://doi.org/10.1186/s40168-015-0094-5`.

[482] Elisa T Granato, Thomas A Meiller-Legrand, and Kevin R Foster. The evolution and ecology of bacterial warfare. *Current biology*, 29(11):R521–R537, 2019.

[483] Gabriele Berg, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, et al. Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, 8(1):1–22, 2020.

[484] Song Hong, Yanlei Sun, Dapeng Sun, and Chengshu Wang. Microbiome assembly on drosophila body surfaces benefits the flies to combat fungal infections. *Iscience*, 25(6), 2022.

[485] Niccoló Alfano, Alexandre Courtiol, Hanna Vielgrader, Peter Timms, Alfred L. Roca, and Alex D. Greenwood. Variation in koala microbiomes within and between individuals: effect of body region and captivity status. *Scientific Reports*, 5(1):10189, May 2015. ISSN 2045-2322. doi: 10.1038/srep10189. URL `https://doi.org/10.1038/srep10189`.

[486] Alexandria M Palaferri Schieber, Yujung Michelle Lee, Max W Chang, Mathias Leblanc, Brett Collins, Michael Downes, Ronald M Evans, and Janelle S Ayres. Disease tolerance mediated by microbiome e. coli involves inflammasome and igf-1 signaling. *Science*, 350(6260):558–563, 2015.

[487] Charlie G Buffie, Vanni Bucci, Richard R Stein, Peter T McKenney, Lilan Ling, et al. Precision microbiome reconstitution restores bile acid mediated resistance to clostridium difficile. *Nature*, 517(7533):205–208, 2015.

[488] Allyson L Byrd and Julia A Segre. Adapting koch's postulates. *Science*, 351(6270):224–226, 2016.

[489] Takumi Murakami, Nozomu Takeuchi, Hiroshi Mori, Yuu Hirose, Arwyn Edwards, Tristram Irvine-Fynn, Zhongqin Li, Satoshi Ishii, and Takahiro Segawa. Metagenomics reveals global-scale contrasts in nitrogen cycling and cyanobacterial light-harvesting mechanisms in glacier cryoconite. *Microbiome*, 10(1):1–14, 2022.

[490] Yongqin Liu, Mukan Ji, Tao Yu, Julian Zaugg, Alexandre M Anesio, et al. A genome and gene catalog of glacier microbiomes. *Nature Biotechnology*, 40(9):1341–1348, 2022.

[491] Guillem Salazar, Lucas Paoli, Adriana Alberti, Jaime Huerta-Cepas, Hans-Joachim Ruscheweyh, et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell*, 179(5):1068–1083, 2019.

[492] Massimo Bourquin, Susheel Bhanu Busi, Stilianos Fodelianakis, Hannes Peter, Alex Washburne, Tyler J Kohler, Leïla Ezzat, Grégoire Michoud, Paul Wilmes, and Tom J Battin. The microbiome of cryospheric ecosystems. *Nature communications*, 13(1):3087, 2022.

[493] Himel Mallick, Siyuan Ma, Eric A Franzosa, Tommi Vatanen, Xochitl C Morgan, and Curtis Huttenhower. Experimental design and quantitative analysis of microbial community multiomics. *Genome biology*, 18(1):1–16, 2017.

[494] Bianca De Saedeleer, Antoine Malabirade, Javier Ramiro-Garcia, Janine Habier, Jean-Pierre Trezzi, et al. Systematic characterization of human gut microbiome-secreted molecules by integrated multi-omics. *ISME communications*, 1(1):82, 2021.

[495] Anna Heintz-Buschart, Patrick May, Cédric C Laczny, Laura A Lebrun, Camille Bellora, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature microbiology*, 2(1):1–13, 2016.

[496] Francesco Delogu, Benoit J Kunath, Pedro M Queirós, Rashi Halder, Laura A Lebrun, Phillip B Pope, Patrick May, Stefanie Widder, Emilie EL Muller, and Paul Wilmes. Forecasting the dynamics of a complex microbial community using integrated meta-omics. *Nature Ecology & Evolution*, pages 1–13, 2023.

[497] Sean M Gibbons, Thomas Gurry, Johanna W Lampe, Anirikh Chakrabarti, Veerle Dam, et al. Perspective: leveraging the gut microbiota to predict personalized responses to dietary, prebiotic, and probiotic interventions. *Advances in Nutrition*, 13 (5):1450–1461, 2022.

[498] Ron Sender, Shai Fuchs, and Ron Milo. Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans. *Cell*, 164(3):337–340, 2016.

[499] Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, 14(8):e1002533, 2016.

[500] Mireia Valles-Colomer, Aitor Blanco-Míguez, Paolo Manghi, Francesco Asnicar, Leonard Dubois, et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature*, 614(7946):125–135, 2023.

[501] Structure, function and diversity of the healthy human microbiome. *nature*, 486(7402):207–214, 2012.

[502] Richard L Gallo. Human skin is the largest epithelial surface for interaction with microbes. *Journal of Investigative Dermatology*, 137(6):1213–1214, 2017.

[503] Allyson L Byrd, Yasmine Belkaid, and Julia A Segre. The human skin microbiome. *Nature Reviews Microbiology*, 16(3):143–155, 2018.

[504] Elizabeth Thursby and Nathalie Juge. Introduction to the human gut microbiota. *Biochemical journal*, 474(11):1823–1836, 2017.

[505] Floyd E. Dewhirst, Tuste Chen, Jacques Izard, Bruce J. Paster, Anne C. R. Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G. Wade. The human oral microbiome. *Journal of Bacteriology*, 192(19):5002–5017, 2010. doi: 10.1128/jb.00542-10. URL `https://journals.asm.org/doi/abs/10.1128/jb.00542-10`.

[506] Ana Elena Pérez-Cobas, Jerónimo Rodríguez-Beltrán, Fernando Baquero, and Teresa M Coque. Ecology of the respiratory tract microbiome. *Trends in Microbiology*, 2023.

[507] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[508] Richard J Abdill, Elizabeth M Adamowicz, and Ran Blekhman. Public human microbiome data are dominated by highly developed countries. *PLoS biology*, 20(2):e3001536, 2022.

[509] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176 (3):649–662, 2019.

[510] Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature biotechnology*, 39(1):105–114, 2021.

[511] Chan Yeong Kim, Muyoung Lee, Sunmo Yang, Kyungnam Kim, Dongeun Yong, Hye Ryun Kim, and Insuk Lee. Human reference gut microbiome catalog including newly assembled genomes from under-represented asian metagenomes. *Genome Medicine*, 13(1):1–20, 2021.

[512] Pranvera Hiseni, Knut Rudi, Robert C Wilson, Finn Terje Hegge, and Lars Snipen. Humgut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data. *Microbiome*, 9(1):1–12, 2021.

[513] Wenxi Li, Hewei Liang, Xiaoqian Lin, Tongyuan Hu, Zhinan Wu, et al. A catalog of bacterial reference genomes from cultivated human oral bacteria. *npj Biofilms and Microbiomes*, 9(1):45, 2023.

[514] Jie Zhu, Liu Tian, Peishan Chen, Mo Han, Liju Song, et al. Over 50,000 metagenomically assembled draft genomes for the human oral microbiome reveal new taxa. *Genomics, Proteomics & Bioinformatics*, 20(2):246–259, 2022.

[515] Cynthia Maria Chibani, Alexander Mahnert, Guillaume Borrel, Alexandre Almeida, Almut Werner, Jean-François Brugère, Simonetta Gribaldo, Robert D Finn, Ruth A Schmitz, and Christine Moissl-Eichinger. A catalogue of 1,167 genomes from the human gut archaeome. *Nature Microbiology*, 7(1):48–61, 2022.

[516] Shuqin Zeng, Dhrati Patangia, Alexandre Almeida, Zhemin Zhou, Dezhi Mu, R Paul Ross, Catherine Stanton, and Shaopu Wang. A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nature Communications*, 13(1):5139, 2022.

[517] Alexandre Almeida, Alex L Mitchell, Miguel Boland, Samuel C Forster, Gregory B Gloor, Aleksandra Tarkowska, Trevor D Lawley, and Robert D Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, 2019.

[518] Stephen Nayfach, Zhou Jason Shi, Rekha Seshadri, Katherine S Pollard, and Nikos C Kyrpides. New insights from uncultivated genomes of the global human gut microbiome. *Nature*, 568 (7753):505–510, 2019.

[519] Mathilde Poyet, Mathieu Groussin, Sean M Gibbons, Julian Avila-Pacheco, Xiaofang Jiang, et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nature medicine*, 25(9): 1442–1452, 2019.

[520] Lisa M Olsson, Fredrik Boulund, Staffan Nilsson, Muhammad Tanweer Khan, Anders Gummesson, et al. Dynamics of the normal gut microbiota: A longitudinal one-year population study in sweden. *Cell Host & Microbe*, 30(5):726–739, 2022.

[521] The integrative human microbiome project. *Nature*, 569(7758): 641–648, 2019.

[522] Chloe X Yap, Anjali K Henders, Gail A Alvares, David LA Wood, Lutz Krause, et al. Autism-related dietary preferences mediate autism-gut microbiome associations. *Cell*, 184(24):5916–5931, 2021.

[523] Cedric CS Tan, Karrie KK Ko, Hui Chen, Jianjun Liu, Marie Loh, SG10K_Health Consortium, Minghao Chia, and Niranjan

Nagarajan. No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. *Nature Microbiology*, 8(5):973–985, 2023.

[524] Velma TE Aho, Marek Ostaszewski, Camille Martin-Gallausiaux, Cédric C Laczny, Jochen G Schneider, and Paul Wilmes. Snapshot: the expobiome map. *Cell Host & Microbe*, 30 (9):1340–1340, 2022.

[525] Paul Wilmes, Camille Martin-Gallausiaux, Marek Ostaszewski, Velma TE Aho, Polina V Novikova, Cédric C Laczny, and Jochen G Schneider. The gut microbiome molecular complex in human health and disease. *Cell Host & Microbe*, 30(9):1201–1206, 2022.

[526] Harry J Flint, Karen P Scott, Petra Louis, and Sylvia H Duncan. The role of the gut microbiota in nutrition and health. *Nature reviews Gastroenterology & hepatology*, 9(10):577–589, 2012.

[527] Danping Zheng, Timur Liwinski, and Eran Elinav. Interaction between microbiota and immunity in health and disease. *Cell research*, 30(6):492–506, 2020.

[528] Skye RS Fishbein, Bejan Mahmud, and Gautam Dantas. Antibiotic perturbations to the gut microbiome. *Nature Reviews Microbiology*, pages 1–17, 2023.

[529] Lisa Maier, Mihaela Pruteanu, Michael Kuhn, Georg Zeller, Anja Telzerow, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*, 555(7698):623–628, 2018.

[530] Lisa Maier, Camille V Goemans, Jakob Wirbel, Michael Kuhn, Claudia Eberl, et al. Unravelling the collateral damage of antibiotics on gut bacteria. *Nature*, 599(7883):120–124, 2021.

[531] Matthew A Jackson, Serena Verdi, Maria-Emanuela Maxan, Cheol Min Shin, Jonas Zierer, et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nature communications*, 9(1):2655, 2018.

[532] Jasminka Talapko, Aleksandar Včev, Tomislav Meštrović, Emina Pustijanac, Melita Jukić, and Ivana Škrlec. Homeostasis and dysbiosis of the intestinal microbiota: Comparing hallmarks of a healthy state with changes in inflammatory bowel disease. *Microorganisms*, 10(12):2405, 2022.

[533] Melanie Schirmer, Ashley Garner, Hera Vlamakis, and Ramnik J Xavier. Microbial genes and pathways in inflammatory bowel disease. *Nature Reviews Microbiology*, 17(8):497–511, 2019.

[534] Clélia Coutzac, Jean-Mehdi Jouniaux, Angelo Paci, Julien Schmidt, Domenico Mallardo, et al. Systemic short chain fatty acids limit antitumor effect of ctla-4 blockade in hosts with cancer. *Nature communications*, 11(1):2168, 2020.

[535] Emanuel E Canfora, Johan W Jocken, and Ellen E Blaak. Short-chain fatty acids in control of body weight and insulin sensitivity. *Nature Reviews Endocrinology*, 11(10):577–591, 2015.

[536] Damian R Plichta, Daniel B Graham, Sathish Subramanian, and Ramnik J Xavier. Therapeutic opportunities in inflammatory bowel disease: mechanistic dissection of host-microbiome relationships. *Cell*, 178(5):1041–1056, 2019.

[537] Stefano Romano, George M Savva, Janis R Bedarf, Ian G Charles, Falk Hildebrand, and Arjan Narbad. Meta-analysis of the parkinson's disease gut microbiome suggests alterations linked to intestinal inflammation. *npj Parkinson's Disease*, 7(1):27, 2021.

[538] Lieve Desbonnet, Gerard Clarke, F Shanahan, Timothy G Dinan, and JF3903109 Cryan. Microbiota is essential for social development in the mouse. *Molecular psychiatry*, 19(2):146–148, 2014.

[539] Djawad Radjabzadeh, Jos A Bosch, André G Uitterlinden, Aeilko H Zwinderman, M Arfan Ikram, et al. Gut microbiome-wide association study of depressive symptoms. *Nature Communications*, 13(1):7128, 2022.

[540] Jane A Foster and Karen-Anne McVey Neufeld. Gut–brain axis: how the microbiome influences anxiety and depression. *Trends in neurosciences*, 36(5):305–312, 2013.

[541] Yong Fan, René Klinkby Støving, Samar Berreira Ibraim, Tuulia Hyötyläinen, Florence Thirion, et al. The gut microbiota contributes to the pathogenesis of anorexia nervosa in humans and mice. *Nature Microbiology*, pages 1–16, 2023.

[542] Hilmar P Sigurdsson, Heather Hunter, Lisa Alcock, Ross Wilson, Ilse Pienaar, Elizabeth Want, Mark R Baker, John-Paul Taylor, Lynn Rochester, and Alison J Yarnall. Safety and tolerability of adjunct non-invasive vagus nerve stimulation in people with parkinson's: a study protocol. *BMC neurology*, 23(1):58, 2023.

[543] Masuma Afrin Taniya, Hea-Jong Chung, Abdullah Al Mamun, Safaet Alam, Md Abdul Aziz, et al. Role of gut microbiome in autism spectrum disorder and its therapeutic regulation. *Frontiers in Cellular and Infection Microbiology*, 12:998, 2022.

[544] M. J. Bull and N. T. Plummer. Part 1: The Human Gut Microbiome in Health and Disease. *Integr Med (Encinitas)*, 13, 2014.

[545] Sigrid Breit, Aleksandra Kupferberg, Gerhard Rogler, and Gregor Hasler. Vagus nerve as modulator of the brain–gut axis in psychiatric and inflammatory disorders. *Frontiers in psychiatry*, page 44, 2018.

[546] Bruno Bonaz, Thomas Bazin, and Sonia Pellissier. The vagus nerve at the interface of the microbiota-gut-brain axis. *Frontiers in neuroscience*, 12:49, 2018.

[547] A Murat Eren, Gary G Borisy, Susan M Huse, and Jessica L Mark Welch. Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences*, 111(28): E2875–E2884, 2014.

[548] Ziyang Min, Lei Yang, Yu Hu, and Ruijie Huang. Oral microbiota dysbiosis accelerates the development and onset of mucositis and oral ulcers. *Frontiers in Microbiology*, 14:1061032, 2023.

[549] Jinfeng Wang, Ji Qi, Hui Zhao, Shu He, Yifei Zhang, Shicheng Wei, and Fangqing Zhao. Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Scientific reports*, 3(1):1843, 2013.

[550] Melissa Agnello, J Marques, L Cen, B Mittermuller, A Huang, N Chaichanasakul Tran, W Shi, X He, and RJ Schroth. Microbiome associated with severe caries in canadian first nations children. *Journal of Dental Research*, 96(12):1378–1385, 2017.

[551] Jonathon L Baker, Jessica L Mark Welch, Kathryn M Kauffman, Jeffrey S McLean, and Xuesong He. The oral microbiome: Diversity, biogeography and human health. *Nature Reviews Microbiology*, pages 1–16, 2023.

[552] Shirley Greenbaum, Gili Greenbaum, Jacob Moran-Gilad, and Adi Y Weintraub. Ecological dynamics of the vaginal microbiome in relation to health and disease. *American Journal of Obstetrics and Gynecology*, 220(4):324–335, 2019.

[553] Silvia Carmona-Cruz, Luz Orozco-Covarrubias, and Marimar Sáez-de Ocariz. The human skin microbiome in selected cutaneous diseases. *Frontiers in cellular and infection microbiology*, 12: 834135, 2022.

[554] Suyun Yu, Huiping Zhang, Liping Wan, Min Xue, Yunfeng Zhang, and Xiwen Gao. The association between the respiratory tract microbiome and clinical outcomes in patients with copd. *Microbiological Research*, 266:127244, 2023.

[555] Alejandra Escobar-Zepeda, Arturo Vera-Ponce de León, and Alejandro Sanchez-Flores. The road to metagenomics: from microbiology to dna sequencing technologies and bioinformatics. *Frontiers in genetics*, 6:348, 2015.

[556] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[557] Luisa W Hugerth and Anders F Andersson. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Frontiers in microbiology*, 8: 1561, 2017.

[558] Michael Eisenstein. Microbiology: making the best of pcr bias. *Nature Methods*, 15(5):317–320, 2018.

[559] Ana Elena Pérez-Cobas, Laura Gomez-Valero, and Carmen Buchrieser. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microbial genomics*, 6(8), 2020.

[560] J L Stein, T L Marsh, K Y Wu, H Shizuya, and E F DeLong. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology*, 178(3):591–599, 1996. doi: 10.1128/jb.178.3.591-599.1996. URL `https://journals.asm.org/doi/abs/10.1128/jb.178.3.591-599.1996`.

[561] Michael DJ Lynch and Josh D Neufeld. Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*, 13(4):217–229, 2015.

[562] Francesco Durazzi, Claudia Sala, Gastone Castellani, Gerardo Manfreda, Daniel Remondini, and Alessandra De Cesare. Comparison between 16s rrna and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific reports*, 11(1):3030, 2021.

[563] Dmitry Antipov, Mikhail Raiko, Alla Lapidus, and Pavel A Pevzner. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome research*, 29(6):961–968, 2019.

[564] Xavier Didelot, A Sarah Walker, Tim E Peto, Derrick W Crook, and Daniel J Wilson. Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology*, 14(3):150–162, 2016.

[565] Joachim Johansen, Damian R Plichta, Jakob Nybo Nissen, Marie Louise Jespersen, Shiraz A Shah, et al. Genome binning of viral entities from bulk metagenomics data. *Nature Communications*, 13(1):965, 2022.

[566] Philipp Rausch, Malte Rühlemann, Britt M Hermes, Shauni Doms, Tal Dagan, et al. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome*, 7(1):1–19, 2019.

[567] Mykhaylo Usyk, Brandilyn A Peters, Smruthi Karthikeyan, Daniel McDonald, Christopher C Sollecito, et al. Comprehensive evaluation of shotgun metagenomics, amplicon sequencing,

and harmonization of these platforms for epidemiological studies. *Cell Reports Methods*, 3(1), 2023.

[568] Philip A Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, 38(3):276–278, 2020.

[569] Ludmila Chistoserdova. Functional metagenomics: recent advances and future challenges. *Biotechnology and Genetic Engineering Reviews*, 26(1):335–352, 2009.

[570] Jacob T Nearing, Gavin M Douglas, Molly G Hayes, Jocelyn MacDonald, Dhwani K Desai, et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, 13(1):342, 2022.

[571] Carine Poussin, Lusine Khachatryan, Nicolas Sierro, Vijay Kumar Narsapuram, Fernando Meyer, et al. Crowdsourced benchmarking of taxonomic metagenome profilers: lessons learned from the sbv improver microbiomics challenge. *BMC genomics*, 23(1):1–19, 2022.

[572] Ruijie Xu, Sreekumari Rajeev, and Liliana CM Salvador. The selection of software and database for metagenomics sequence analysis impacts the outcome of microbial profiling and pathogen detection. *Plos one*, 18(4):e0284031, 2023.

[573] Bin Hu, Shane Canon, Emiley A Eloe-Fadrosh, Michal Babinski, Yuri Corilo, et al. Challenges in bioinformatics workflows for processing microbiome omics data at scale. *Frontiers in Bioinformatics*, 1:826370, 2022.

[574] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11):1063–1071, 2017.

[575] Fernando Meyer, Peter Hofmann, Peter Belmann, Ruben Garrido-Oter, Adrian Fritz, Alexander Sczyrba, and Alice C McHardy. Amber: assessment of metagenome binners. *Gigascience*, 7(6):giy069, 2018.

[576] Fernando Meyer, Till-Robin Lesker, David Koslicki, Adrian Fritz, Alexey Gurevich, Aaron E Darling, Alexander Sczyrba, Andreas Bremges, and Alice C McHardy. Tutorial: assessing metagenomics software with the cami benchmarking toolkit. *Nature protocols*, 16(4):1785–1801, 2021.

[577] Matteo Calgaro, Chiara Romualdi, Davide Risso, and Nicola Vitulo. benchdamic: benchmarking of differential abundance methods for microbiome data. *Bioinformatics*, 39(1):btac778, 2023.

[578] Fernando Meyer, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nature methods*, 19(4):429–440, 2022.

[579] Bei Gao, Liang Chi, Yixin Zhu, Xiaochun Shi, Pengcheng Tu, Bing Li, Jun Yin, Nan Gao, Weishou Shen, and Bernd Schnabl. An introduction to next generation sequencing bioinformatic analysis in gut microbiome studies. *Biomolecules*, 11(4):530, 2021.

[580] Chao Yang, Debajyoti Chowdhury, Zhenmiao Zhang, William K Cheung, Aiping Lu, Zhaoxiang Bian, and Lu Zhang. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, 19:6301–6314, 2021.

[581] Richa Bharti and Dominik G Grimm. Current challenges and best-practice protocols for microbiome analysis. *Briefings in bioinformatics*, 22(1):178–193, 2021.

[582] Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*, 35(9):833–844, 2017.

[583] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

[584] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17): i884–i890, 2018.

[585] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[586] Francesco Beghini, Lauren J McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *elife*, 10:e65088, 2021.

[587] Yoshihiko Tomofuji, Kyuto Sonehara, Toshihiro Kishikawa, Yuichi Maeda, Kotaro Ogawa, et al. Reconstruction of the personal information from human genome reads in gut metagenome sequencing data. *Nature Microbiology*, pages 1–16, 2023.

[588] Sara Saheb Kashaf, Alexandre Almeida, Julia A Segre, and Robert D Finn. Recovering prokaryotic genomes from host-associated, short-read shotgun metagenomic sequencing data. *Nature protocols*, 16(5):2520–2541, 2021.

[589] Cedric C Laczny, Tomasz Sternal, Valentin Plugaru, Piotr Gawron, Arash Atashpendar, Houry Hera Margossian, Sergio Coronado, Laurens van der Maaten, Nikos Vlassis, and Paul Wilmes. Vizbin-an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, 3:1–7, 2015.

[590] Oskar Hickl, Pedro Queirós, Paul Wilmes, Patrick May, and Anna Heintz-Buschart. binny: an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets. *Briefings in Bioinformatics*, 23(6):bbac431, 2022.

[591] Yu-Wei Wu, Blake A Simmons, and Steven W Singer. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, 2016.

[592] Dongwan D Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, 2019.

[593] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

[594] Till LV Bornemann, Sarah P Esser, Tom L Stach, Tim Burg, and Alexander J Probst. ubin: A manual refining tool for genomes from metagenomes. *Environmental Microbiology*, 2023.

[595] Matthew R Olm, Christopher T Brown, Brandon Brooks, and Jillian F Banfield. drep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME journal*, 11(12): 2864–2868, 2017.

[596] John Vollmers, Sandra Wiegand, Florian Lenk, and Anne-Kristin Kaster. How clear is our current view on microbial dark matter?(re-) assessing public mag & sag datasets with mdmcleaner. *Nucleic Acids Research*, 50(13):e76–e76, 2022.

[597] Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. Gtdb-tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, 38(23): 5315–5316, 2022.

[598] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic acids research*, 50(D1):D785–D794, 2022.

[599] Alessio Milanese, Daniel R Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, et al. Microbial abundance, activity and population genomic profiling with motus2. *Nature communications*, 10(1):1014, 2019.

[600] Aitor Blanco-Míguez, Francesco Beghini, Fabio Cumbo, Lauren J McIver, Kelsey N Thompson, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using metaphlan 4. *Nature Biotechnology*, pages 1–12, 2023.

[601] Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome biology*, 20:1–13, 2019.

[602] Duy Tin Truong, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome research*, 27(4):626–638, 2017.

[603] Francesco Beghini, Lauren J McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *elife*, 10:e65088, 2021.

[604] Will PM Rowe and Martyn D Winn. Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics*, 34 (21):3601–3608, 2018.

[605] Michael Inouye, Harriet Dashnow, Lesley-Ann Raven, Mark B Schultz, Bernard J Pope, Takehiro Tomita, Justin Zobel, and Kathryn E Holt. Srst2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine*, 6(11): 1–16, 2014.

[606] Amy D Willis. Rarefaction, alpha diversity, and statistics. *Frontiers in microbiology*, 10:2407, 2019.

[607] Sven Kleine Bardenhorst, Tom Berger, Frank Klawonn, Marius Vital, André Karch, and Nicole Rübsamen. Data analysis strategies for microbiome studies in human populations—a systematic review of current practice. *Msystems*, 6(1):10–1128, 2021.

[608] Nicola Segata, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett, and Curtis Huttenhower. Metagenomic biomarker discovery and explanation. *Genome biology*, 12:1–18, 2011.

[609] Dmitry Antipov, Nolan Hartwick, Max Shen, Mikhail Raiko, Alla Lapidus, and Pavel A Pevzner. plasmidspades: assembling plasmids from whole genome sequencing data. *Bioinformatics*, 32(22):3380–3387, 2016.

[610] David Pellow, Itzik Mizrahi, and Ron Shamir. Plasclass improves plasmid sequence classification. *PLoS computational biology*, 16(4):e1007781, 2020.

[611] Chamika Nandasiri, Sasindu Alahakoon, Gayal Dassanayake, Anuradha Wickramarachchi, and Indika Perera. Metapcbin: Plasmid/chromosome classification for metagenomic contigs using machine learning techniques. In *2022 Moratuwa Engineering Research Conference (MERCon)*, pages 1–6. IEEE, 2022.

[612] Anuradha Wickramarachchi and Yu Lin. Graphplas: refined classification of plasmid sequences using assembly graphs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):57–67, 2021.

[613] Qianhui Zhu, Shenghan Gao, Binghan Xiao, Zilong He, and Songnian Hu. Plasmer: an accurate and sensitive bacterial plasmid prediction tool based on machine learning of shared k-mers and genomic features. *Microbiology Spectrum*, 11(3): e04645–22, 2023.

[614] Léa Pradier, Tazzio Tissot, Anna-Sophie Fiston-Lavier, and Stéphanie Bedhomme. Plasforest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC bioinformatics*, 22(1):1–17, 2021.

[615] Xiaohui Zou, Marcus Nguyen, Jamie Overbeek, Bin Cao, and James J Davis. Classification of bacterial plasmid and chromosome derived sequences using machine learning. *Plos one*, 17 (12):e0279280, 2022.

[616] Xubo Tang, Jiayu Shang, Yongxin Ji, and Yanni Sun. Plasme: a tool to identify plasmid contigs from short-read assemblies using transformer. *Nucleic Acids Research*, 51(15):e83–e83, 2023.

[617] Elodie Drula, Marie-Line Garron, Suzan Dogan, Vincent Lombard, Bernard Henrissat, and Nicolas Terrapon. The carbohydrate-active enzyme database: functions and literature. *Nucleic acids research*, 50(D1):D571–D577, 2022.

[618] Clémence Joseph, Haris Zafeiropoulos, Kristel Bernaerts, and Karoline Faust. Predicting microbial interactions with approaches based on flux balance analysis: an evaluation. *BMC bioinformatics*, 25(1):36, 2024.

[619] World Health Organization et al. Parkinson disease: a public health approach: technical brief. 2022.

[620] Sigurlaug Sveinbjornsdottir. The clinical symptoms of parkinson's disease. *Journal of neurochemistry*, 139:318–324, 2016.

[621] Ai Huey Tan, Shen Yang Lim, and Anthony E Lang. The microbiome–gut–brain axis in parkinson disease—from basic research to the clinic. *Nature Reviews Neurology*, 18(8):476–495, 2022.

[622] Walter Willett, Johan Rockström, Brent Loken, Marco Springmann, Tim Lang, et al. Food in the anthropocene: the eat–lancet commission on healthy diets from sustainable food systems. *The lancet*, 393(10170):447–492, 2019.

[623] AF Yassin, H Hupfer, C Siering, H-P Klenk, and P Schumann. Auritidibacter ignavus gen. nov., sp. nov., of the family micrococcaceae isolated from an ear swab of a man with otitis externa, transfer of the members of the family yaniellaceae li et al. 2008 to the family micrococcaceae and emended description of the suborder micrococcineae. *International Journal of Systematic and EvolutionaryMicrobiology*, 61(2):223–230, 2011.

[624] Andrew C Yang, Fabian Kern, Patricia M Losada, Maayan R Agam, Christina A Maat, et al. Dysregulation of brain and choroid plexus cell types in severe covid-19. *Nature*, 595(7868): 565–571, 2021.

[625] Andreas Keller, Laura Gröger, Thomas Tschernig, Jeffrey Solomon, Omar Laham, et al. mirnatissueatlas2: an update to the human mirna tissue atlas. *Nucleic Acids Research*, 50(D1): D211–D221, 2022.

[626] Anne Hecksteden, Georges Pierre Schmartz, Yanni Egyptien, Karen Aus der Fünten, Andreas Keller, and Tim Meyer. Forecasting football injuries by combining screening, monitoring and machine learning. *Science and medicine in football*, 7(3): 214–228, 2023.

[627] Ernesto Aparicio-Puerta, Pascal Hirsch, Georges P Schmartz, Tobias Fehlmann, Verena Keller, Annika Engel, Fabian Kern, Michael Hackenberg, and Andreas Keller. isomirdb: microrna expression at isoform resolution. *Nucleic Acids Research*, 51(D1): D179–D185, 2023.

[628] Ernesto Aparicio-Puerta, Pascal Hirsch, Georges P Schmartz, Fabian Kern, Tobias Fehlmann, and Andreas Keller. mieaa 2023: updates, new functional microrna sets and improved enrichment visualizations. *Nucleic Acids Research*, page gkad392, 2023.

[629] Anastasis Oulas, Margarita Zachariou, Christos T Chasapis, Marios Tomazou, Umer Z Ijaz, Georges Pierre Schmartz, George M Spyrou, and Alexios Vlamis-Gardikas. Putative antimicrobial peptides within bacterial proteomes affect bacterial predominance: a network analysis perspective. *Frontiers in Microbiology*, 12:752674, 2021.

[630] Georges Pierre Schmartz, Fabian Kern, Tobias Fehlmann, Viktoria Wagner, Bastian Fromm, and Andreas Keller. Encyclopedia of tools for the analysis of mirna isoforms. *Briefings in Bioinformatics*, 22(4):bbaa346, 2021.

[631] Pascal Hirsch, Azat Tagirdzhanov, Aleksandra Kushnareva, Ilia Olkhovskii, Simon Graf, et al. Abc-humi: the atlas of biosynthetic gene clusters in the human microbiome. *Nucleic Acids Research*, page gkad1086, 2023.

[632] Nicola J Müller, Daniel Porawski, Lukas Wilde, Dennis Fink, Guillaume Trap, Annika Engel, and Georges P Schmartz. Neuro-explicit semantic segmentation of the diffusion cloud chamber. *Review of Scientific Instruments*, 94(6), 2023.

[633] H Lin and SD Peddada. Analysis of compositions of microbiomes with bias correction. nat commun 11: 3514, 2020.

[634] Andrew D Fernandes, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing rna-seq, 16s rrna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:1–13, 2014.

[635] Marco Cappellato, Giacomo Baruzzo, and Barbara Di Camillo. Investigating differential abundance methods in microbiome data: A benchmark study. *PLoS Computational Biology*, 18(9): e1010467, 2022.

[636] Lu Yang and Jun Chen. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome*, 10(1):130, 2022.

[637] Tobias Bähr, Ariane Baumhögger, Gabriele Geis, and Sören Gatermann. Complete genome sequences of eight auritidibacter ignavus strains isolated from ear infections in germany. *Microbiology Resource Announcements*, 12(11):e00666–23, 2023.

[638] Brian Bushnell, Jonathan Rood, and Esther Singer. Bbmerge–accurate paired shotgun read merging via overlap. *PloS one*, 12 (10):e0185056, 2017.

[639] Xubo Tang, Jiayu Shang, Yongxin Ji, and Yanni Sun. Plasme: a tool to identify plasmid contigs from short-read assemblies using transformer. *Nucleic Acids Research*, 51(15):e83–e83, 2023.

[640] Samuel C Forster, Junyan Liu, Nitin Kumar, Emily L Gulliver, Jodee A Gould, et al. Strain-level characterization of broad host range mobile genetic elements transferring antibiotic resistance from the human microbiome. *Nature Communications*, 13(1): 1445, 2022.

[641] Antonio Pedro Camargo, Lee Call, Simon Roux, Stephen Nayfach, Marcel Huntemann, et al. Img/pr: a database of plasmids from genomes and metagenomes with rich annotations and metadata. *Nucleic acids research*, 52(D1):D164–D173, 2024.

[642] Cedric C Laczny, Valentina Galata, Achim Plum, Andreas E Posch, and Andreas Keller. Assessing the heterogeneity of in silico plasmid predictions based on whole-genome-sequenced clinical isolates. *Briefings in bioinformatics*, 20(3):857–865, 2019.

[643] Mantas Sereika, Rasmus Hansen Kirkegaard, Søren Michael Karst, Thomas Yssing Michaelsen, Emil Aarre Sørensen, Rasmus Dam Wollenberg, and Mads Albertsen. Oxford nanopore r10. 4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature methods*, 19(7):823–826, 2022.

[644] Zhaohui Cao, Wenlong Zuo, Lanxiang Wang, Junyu Chen, Zepeng Qu, Fan Jin, and Lei Dai. Spatial profiling of microbial communities by sequential fish with error-robust encoding. *Nature Communications*, 14(1):1477, 2023.

[645] Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen AMM Elshaier, Mamdouh M Gomaa, and Aboul Ella Hassanien. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, 56(7):5975–6037, 2023.

[646] Hai-Cheng Yi, Zhu-Hong You, De-Shuang Huang, and Chee Keong Kwoh. Graph representation learning in bioinformatics: trends, methods and applications. *Briefings in Bioinformatics*, 23(1):bbab340, 2022.

[647] Pei-Hua Wang, Jen-Hao Chen, Yu-Yuan Yang, Chien Lee, and Yufeng Jane Tseng. Recent advances in quantum computing for drug discovery and development. *IEEE Nanotechnology Magazine*, 2023.

# *Acknowledgement*

First, I wish to sincerely thank my supervisor Prof. Dr. Andreas Keller for displaying inspiring leadership qualities as well as for trusting, guiding, and of course, welcoming me at his Chair for Clinical Bioinformatics (CCB) at Saarland University for the duration of this Ph.D..
Moreover, I would like to thank the remaining committee members for their valuable time they invested in the evaluation of this thesis.
I further want to thank the Saarbrücken Graduate School of Computer Science for their financial support.

I want to thank all authors, co-authors, and unmentioned scientific staff for their contributions to the individual projects contained within this thesis. Among these I want to explicitly highlight: Prof. Dr. Alexey Gurevich, Prof. Dr. Sören Becker, Prof. Dr. Rolf Müller, Dr. Verena Keller, Dr. Jan Dastbaz, Prof. Dr. Marcus Unger, Prof Dr. Marcin, and Prof. Dr. Robert Bals. Furthermore, I am grateful for support offered by Jacqueline Rehner who not only performed most of the experiments but also greatly helped me with error tracing during data analysis.

I am grateful for the pleasant project handovers as well as constructive discussions with former members of the CCB: Dr. Valentina Galata, Nadja Grammes, Dr. Tobias Fehlmann, and Dr. Ernesto Aparicio-Puerta. Further, I thank both, the students and HIWIs, Anna Hartung, Kijin Kim, Saurabh Pandey, and Emma Hoffmann for their contributions to these works as well as unpublished ones.
Many thanks also go to the current CCB team, Jérémy Amand, Matthias Flotho, Anna Priester, Dr. Fabian Kern, Pascal Hirsch, Annika Engel, Viktoria Wagner, and Friederike Grandke for various kinds of during the years.
Moreover, I want to use this opportunity to express my sincere gratitude to Sabine Lessel for protecting me against German bureaucracy to the best of her capabilities.

Villmools Merci, u meng Frënn, Verloobten, a Famill vir hire Versteesdemech, hire Support, an dofir dass ech mech wärend dëser intensiver Zäit zu all Moment konnt blann op si verloossen.