Heinrich Heine
Universität
Düsseldorf

Bioinformatics in Allogeneic Stem Cell Transplantation: Microbiome Analysis, Risk Prediction, and Secure Data Transfer

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Philipp Spohr
aus Köln

Düsseldorf, Oktober 2025

Gedruckt mit der Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Gunnar W. Klau

2. Prof. Alexander T. Dilthey

Tag der mündlichen Prüfung: 07.05.2026

Acknowledgments

Professor Gunnar Klau initially encouraged me to pursue a doctorate in computer science and gave me this wonderful opportunity. As a supervisor, he was highly available, always positive, and a pleasure to work with. I do not take this for granted and am immensely grateful.

Professor Alexander Dilthey is not only an inexhaustible source of brilliant ideas but also genuinely interested in any challenging algorithmic problem. I want to extend my sincere thanks for the many in-depth sessions discussing and dissecting various algorithms and statistical methods.

I want to thank the scientific and non-scientific staff of all research groups I was involved in—particularly the Algorithmic Bioinformatics group at the HHU, as well as the groups of the Institute for Medical Microbiology and Hospital Hygiene—not only for being invaluable for my research through their scientific input, but also for providing a very friendly, inclusive, and engaging environment that made work a pleasure, not a chore. A special "Thank you" goes out to Khoa for extensively proofreading this thesis.

The Jürgen Manchot Foundation receives my heartfelt appreciation for giving me the opportunity to conduct my research in a free and unrestricted way; the exchange with fellow researchers involved in the AI use cases was exceptionally interesting.

Teaching is a great way of deepening one's own knowledge while sharing it with others. I want to express my gratefulness to all the students I taught, especially the ones whose final theses I supervised throughout the years, for the pleasure it was to work with them and their many intriguing questions and remarks. Hosting the Oberseminar at our chair was a great way to broaden my own horizons and get introduced to a multitude of fascinating topics.

Throughout my work on various projects I learned a substantial amount from all co-authors and collaborators; thus my deepest gratitude is given to them. While certainly stressful at times, our cooperation was always respectful, goal-oriented, and effective.

This thesis is something I could never have achieved without the support of my family and friends, which I am blessed to have in my life. Finally, I want to express my love and gratefulness to Maike, my wonderful wife, for proofreading, but also in general for being at my side and making every day fantastic.

Abstract

The availability of an increasing amount of machine-processable data sources in the clinical setting is enabling bioinformatics analyses with the goal of improving therapy- and diagnostics quality, leading ultimately to a better outcome for patients. Allogeneic stem cell transplantation is an essential treatment option for many diseases, especially malignant tumors of the hematopoietic system. In this thesis, we present three publications, covering applications for bioinformatics in the context of this procedure:

First, we show that Nanopore-based sequencing of patients' gut microbiomes prior to transplantation reveals three distinct clusters based on their composition. We also show that a significant fraction of the observed microbiome consists of fungal and viral components. The presence of specific bacterial strains is tracked throughout the procedure within individuals, showing a re-population of the same species by other strains in some instances. Overall, the disruptive effect on the microbiome is captured in a holistic way, enabled by the resolution obtained by the sequencing technology.

Second, we demonstrate that a risk-prediction system, using a random forest-based approach utilizing time-series based features, is able to provide daily stratification of patients into low, moderate, or high mortality risk, with AUROC scores continuously exceeding 0.8 from day 17. We are able to plot out individual patients' trajectories and contextualize their risk scores based on the clinical context, providing a system that is explainable and a valuable tool in assisting clinicians in their decision making process.

Third, we present the "Secure Whole-Genome Transfer System", SWGTS, a client-server software architecture that enables the privacy-preserving transfer of DNA data while providing scalability via a docker-compose architecture. We show that the system can scale well to multiple, simultaneously uploading clients. At the same time, the re-identification risk can be controlled to adhere to given ethical or legal requirements for which we give and discuss guidelines.

The papers are prefaced with background chapters explaining the concepts and methods required for a substantial understanding of the presented work. They are further expanded with additional motivation, a presentation of the research leading to the publication, extended discussion, and an outline of related future projects.

Publications

Major Contributions

- **Spohr P**, Scharf S, Rommerskirchen A, Henrich B, Jäger P, Klau GW, Haas R, Dilthey AT, and Pfeffer K
  Insights into gut microbiomes in stem cell transplantation
  by comprehensive shotgun long-read sequencing
  Scientific Reports, 2024
  https://doi.org/10.1038/s41598-024-53506-1
- **Spohr P**, Ried M, Kühle L, and Dilthey AT
  SWGTS-a platform for stream-based host DNA depletion
  Bioinformatics, 2024
  https://doi.org/10.1093/bioinformatics/btae332
- **Spohr P**, Fröhlich RC, Scharf S, Rommerskirchen A, Bobak J, Schweier S, Jäger P, Kobbe G, Dietrich S, Dilthey AT, Henrich B, Pfeffer K, Haas R, and Klau GW
  Dynamic Prediction of Mortality Risk Following Allogeneic Hematopoietic Stem Cell Transplantation
  Machine Learning Health, 2025
  https://doi.org/10.1088/3049-477X/adf74e

Additional Publications

- **Spohr P**, Dinkla K, Klau GW, and El-Kebir M
  eXamine: Visualizing annotated networks in Cytoscape
  F1000Research, 2018
  https://doi.org/10.12688/f1000research.14612.2
- Brand M, Tran NK, **Spohr P**, Schrinner S, and Klau GW
  The Homo-Edit Distance Problem
  bioRxiv, 2020
  https://doi.org/10.1101/2020.05.27.118273
- Schrinner S, Goel M, Wulfert M, **Spohr P**, Schneeberger K, and Klau GW
  Using the longest run subsequence problem within homology-based scaffolding
  Algorithms for Molecular Biology, 2021
  https://doi.org/10.1186/s13015-021-00191-8
- Bobak J, **Spohr P**, Richter S, Streuer A, Schulz F, Strupp C, Gerhards C, Schmitt N, Luft T, Dietrich S, Germing U, and Klau GW
  Dynamic Mortality Risk Prediction in Myelodysplastic Syndromes Using Longitudinal Clinical Data
  Preprint, Submitted to JCO Clinical Cancer, 2025
  https://doi.org/10.1101/2025.07.21.25331775

Contents

Chapter 1

Introduction

Allogeneic stem cell transplantation is the only curative treatment option for many hematological malignancies such as leukemia or myelodysplastic syndrome. While improvements in therapy and diagnostics have drastically increased outcomes, these diseases are still a major cause of death in developed countries. In addition, many mechanisms of tumorgenesis are still undiscovered and the interactions between these diseases and various exogenic and endogenic factors continue to be the subject of active research.

With the advent of digital medicine, new possibilities for this research are increasingly becoming available. By analyzing large datasets, we can uncover previously hidden patterns that may provide crucial information to inform future clinical decision-making. In this context, the Use-Case "Health" of the Manchot Research Group "Decision-making aided with artificial intelligence methods" began in 2019 with the primary goal of enabling personalized risk assessments for patients undergoing allogeneic stem cell transplantation at the University Hospital Düsseldorf and has since expanded to various related projects.

Here we will present three major contributions: First, a comprehensive analysis of the gut microbiome in patients undergoing allogeneic stem cell transplantation, which reveals three distinct clusters based on microbial compositions before the procedure. Second, a risk-prediction system that utilizes medical records to stratify patients into three risk categories. Third, a software suite for the privacy-preserving transfer of genomic data in a scalable fashion.

In the next section we provide essential background for the presented work. We then follow with the selected publications: Each publication is framed by a presentation of the initial motivation leading to the research as well as an extended discussion evaluating the research in retrospective. We conclude this thesis with a brief, general discussion of the presented work.

Chapter 2

Background

This chapter introduces fundamental concepts and background information to aid understanding of the presented papers. It does not aim to replace a topic-specific textbook; rather, it highlights and discusses the aspects that are especially relevant for the work presented in this thesis.

2.1 Allogeneic Stem Cell Transplantation

Cancers are a major cause of death, especially in highly developed countries. In Germany they were the second leading cause of death in 2023, resulting in a significant health and economic burden [1]. All diseases referred to as cancer share the same essential mechanism: cells proliferate abnormally. In the case of malign tumors, this eventually means that, due to invasion or spreading, physiological functions become negatively affected. Any type of cells can show this behavior, including the cells that regenerate blood cells, the hematopoietic stem cells. Cancers affecting them typically lead to a loss of functional blood cells as they are no longer generated or replaced by non-functional cells.

The procedure of allogeneic stem cell transplantation was developed during the Cold War era, under the threat of nuclear warfare, as an effort to counteract bone marrow function loss induced by radiation exposure. Hematopoietic stem cells from an external source are injected into the bloodstream, eventually settling in the bone marrow. The first transplantations treating a malignant hematological disease were performed in 1957 by E. Donnall Thomas [2] and have since become a crucial treatment option for cancers affecting the hematopoietic system. There are two aspects of this procedure that are particularly challenging and potentially dangerous for cancer patients. First, in many instances the existing cancerous stem cells need to be destroyed (this is referred to as myeloablative treatment), leaving the patient without an immune response and thus susceptible to infections. Second, immune cells generated by the transplanted stem cells can recognize the recipient as foreign and trigger a response. This is referred to as graft-versus-host disease. In order to counteract this, immunosuppressive medication is required, which in turn reinforces the susceptibility to infections. As a result, although the overall mortality has continuously decreased, it is still a dangerous procedure with approximately 18% of patients dying within 100 days [3].

Due to the myeloablative treatment, patients will initially have a significant reduction in all blood cells, which then eventually get replaced by new cells. It can therefore

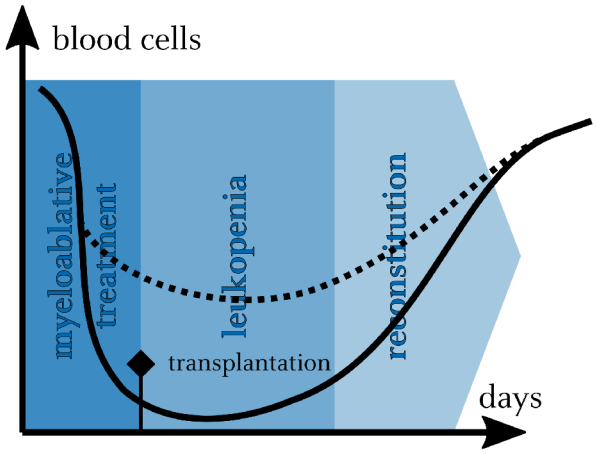


Figure 2.1: Blood cell counts throughout alloHSCT. The solid line represents a typical progression of blood cells while the dotted line represents a reduced myeloablative treatment course. The day of transplantation is marked with a diamond. Phases are labeled in the background.

make sense to subdivide the clinical course into three phases (see Figure 2.1): the pre-transplantation phase of myeloablative treatment, the subsequent phase of low blood cell count (specifically leukocytes), and the eventual phase of blood cell reconstitution. Notably, novel approaches using reduced or non-myeloablative treatment destroy fewer cells and rely on the graft-versus-tumor effect (the new immune cells removing the remaining original tumor cells).

2.2 The Human Microbiome

The existence of small organisms, which cannot be observed with the unaided eye, has already been hypothesized in antiquity. In his work on agriculture (published around 40 BC) Marcus Terentius Varro warns of "animalia quaedam minuta, quae non possunt oculi consequi, et per aera intus in corpus per os ac nares perveniunt atque efficiunt difficilis morbos"[1], even linking those hypothetical microorganism to human disease [4]. Eventually, the invention of the microscope was able to prove this hypothesis and to begin the exciting discovery of a hitherto unknown world.

Today, microbial life has been discovered everywhere throughout the human body to an extent where it is possible to think of humans as large symbiotic entities as opposed to just monolithic life forms. Human cells are constantly exchanging biological signals with those entities and the influence of this cross-talk on health and disease is significant [5]. As a result, the microbiome composition at a locale can be seen as both, a marker, and a cause for processes or states of interest. For instance, a reduction in gut microbiome diversity has been associated with acute-on-chronic liver failure, thus providing a possible marker [6] while colonization of mice with gut microbiomes from patients with inflammatory bowel diseases has led to severe colitis, thus showing a relevant causal re-

[1] particular small animals, which can't be seen by the eye, and which by air enter the body through mouth and nose causing grave diseases; author's translation

lationship [7]. The severity of interactions can be surprising: infections with *Toxoplasma gondii* have even been shown to alter human behavior [8].

In this thesis, we are specifically concerned with the gut microbiome, the collection of all microorganisms within the gastrointestinal tract, encompassing around ≈ 100 trillion individual organisms [9] and thereby being an exceptionally large one. While a healthy microbiome was found to be associated with a high species diversity as well as core microorganisms being temporally and spatially stable within one human [10], the microbiome makeup was found to be highly variable across geographic and cultural axes [11]. Nevertheless, in many instances, different groups of organisms perform similar functions and interactions due to shared functional genes [12]. It is therefore imperative to focus not only on taxonomic units but also on functional roles within a specific ecosystem to make meaningful comparisons between microbiome communities.

In the context of allogeneic stem cell transplantation the gut microbiome is naturally of particular interest since the antimicrobial therapy heavily disrupts the ecosystem. In this context it has previously been shown that a lower diversity within the gut microbiome prior to transplantation is indicative of a higher mortality risk [13].

Although an increasing amount of microorganisms are discovered within the gut microbiome, a large amount of organisms is likely not comprehensively described yet. Specifically, the assembly of metagenomes from sequencing data revealed a surprising amount of previously unknown organisms. The gut microbiomes of mice have even been reported to be dominated by unknown species that are essential for distinguishing between groups of interest [14]. Consequently, it is crucial to integrate unlabeled taxa into any comparative analysis of microbiomes.

2.3 Long Read Sequencing

Sequencing is the process of transforming DNA molecules present in a sample into digital string representations. It consists of wet-lab steps (e.g., preparing the DNA and making it available by breaking up cells) and dry-lab steps, which transform primary data types into the desired representation. Most methods are not able to read an entire genome contained in a cell. Instead, they rely on breaking apart the DNA into small fragments, which are then processed and transformed into a digital signal representing the most likely sequence of nucleotides. One estimated fragment sequence is commonly referred to as a "read". Since this thesis approaches the subject from the field of computer science, our goal is to abstract each read r as a string over the alphabet $\text{DNA} = \{\text{A}, \text{C}, \text{G}, \text{T}\}$. In practice, we usually track more information such as ambiguous characters (representing uncertainty between two or more bases), quality scores (the likelihood of positions in the string being correct), or methylation status. However, the simple representation of DNA reads as strings over the nucleotide alphabet is sufficient to explain many relevant problems and methods.

Throughout the years, different sequencing technologies have been developed and used. We differentiate between targeted sequencing and whole-genome sequencing: in the case of classical targeted sequencing we only observe a specific region of the genome adjacent to a primer sequence (thus requiring prior knowledge) while in the case of whole-genome sequencing we attempt to capture the entire DNA content of a cell. For

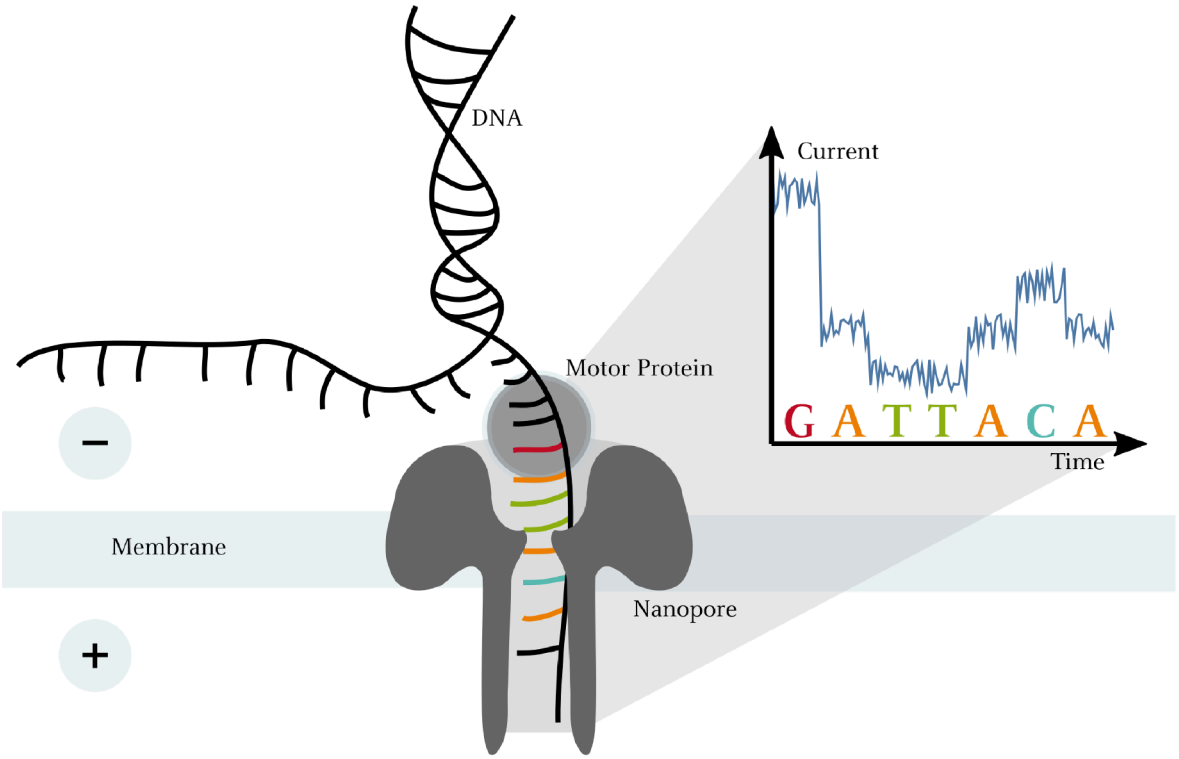


Figure 2.2: Nanopore sequencing. DNA fragments are pulled through a membrane along a voltage gradient from negative to positive through a pore duct, the "Nanopore". This is facilitated by motor proteins ligated to the DNA fragments. The pore induces a current, which is measured and translated into a sequence of bases. Note that in reality, multiple pore ducts are combined on a "flowcell" to allow parallel processing of fragments.

this thesis we are especially concerned with Nanopore sequencing (see Figure 2.2 for a schematic overview of the underlying mechanism).

This technology has a couple of particular properties that are worth noting. The main benefits of the technology are that individual reads can be very long and thereby span entire genes, that the relatively fast turnaround time allows for near-to real-time analysis, and that the sequencers themselves are small and thus portable. Initially, a major trade-off were relatively high base error rates of up to 15%, however current protocols using the R10.4 flowcells can reach a base accuracy of up to 99% [15].

Advantages of longer reads are clear: a small read sampled from a repetitive region does not allow to draw conclusions about the structure and number of repeats whereas a longer read stretching the entire region contains this information. In the context of strain resolution—the task of assigning reads to individual copies of genomes with a high similarity—longer reads allow linking more distant variation to the same copy. For an illustration of those advantages refer to Figure 2.3.

We want to also briefly discuss costs. The current cost for sequencing samples in our setting using Nanopore technology was around 200 € per sample while Illumina sequencing currently costs around 450 € per sample, generating a similar amount of around 10,000,000,000 bases. Those prices obviously fluctuate and depend on quantity of processed samples and other factors but highlight that, at the very least, Nanopore

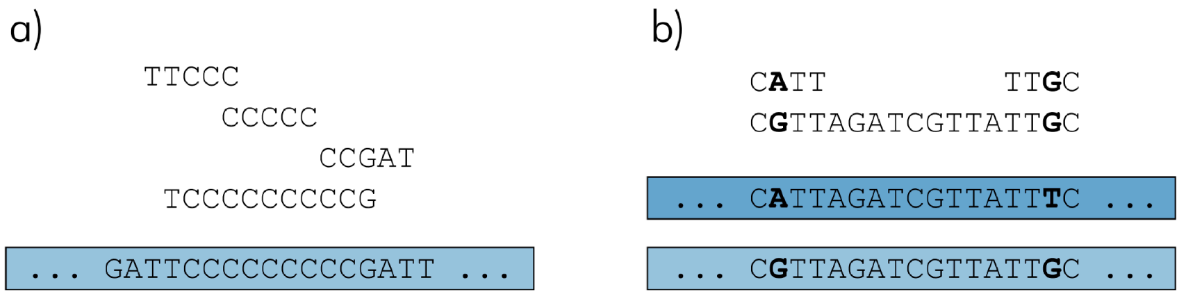


Figure 2.3: Advantages of long read sequencing. Panel a) shows three short reads and a longer read sampled from a fictive genome containing a large repetitive stretch of Cs. It is impossible to infer the amount of Cs in the stretch based on the short reads. However, a single long read spanning the repeat contains this information. Panel b) shows two short reads and a long read sampled from a fictive diploid genome at the bottom. The genomes differ at two positions with variation which are highlighted in bold. Based on the short reads, which are sampled randomly from the two copies, it is impossible to infer whether two variants occur on the same copy. However, a long read containing both positions contains this information.

prices are competitive. Prices in general are of course expected to become increasingly affordable based on the current trends. It should be noted that Nanopore sequencing also allows for a quick processing of samples. Walker et al. [16] have shown a rapid viral surveillance sequencing workflow that produced assembled genomes after 28 hours, clearly providing a proof that this technology can be a step towards real-time personalized medicine.

2.4 Abundance Estimation

One essential task when analyzing samples containing multiple organisms is to give an estimate about the organisms' abundances. In a bioinformatics context this usually means that we are interested in an abundance estimation matrix $M \in \mathbb{N}^{t \times s}$, where t is the number of taxonomic units and s the number of samples analyzed concurrently, and each entry represents the amount of DNA reads found in a specific sample assigned to the respective taxonomic unit.

Each abundance estimation workflow can be understood as a parameterized function that transforms a set of DNA reads into a matrix M. If we are concerned with concrete software implementations, we usually have two types of input: the DNA reads, and a database. A flat database would simply contain a genome sequence for every taxonomic unit. However, the units are often organized in trees as for example in the rank-based taxonomy frequently used in biology. Here for example, both *Escherichia albertii* and *Escherichia coli* are considered children of the parent node *Escherichia*. A non-flat database consists of a tree that contains all taxonomic units and reflects this complexity. Here the tree leaves contain actual genome sequences while the inner nodes do not. Assignment of a read to an inner node means that a read fits all genome sequences contained in the subtree rooted at the node equally well. A read containing a gene sequence conserved in all *Escherichia* would not be informative to make a distinction between *Escherichia alber-*

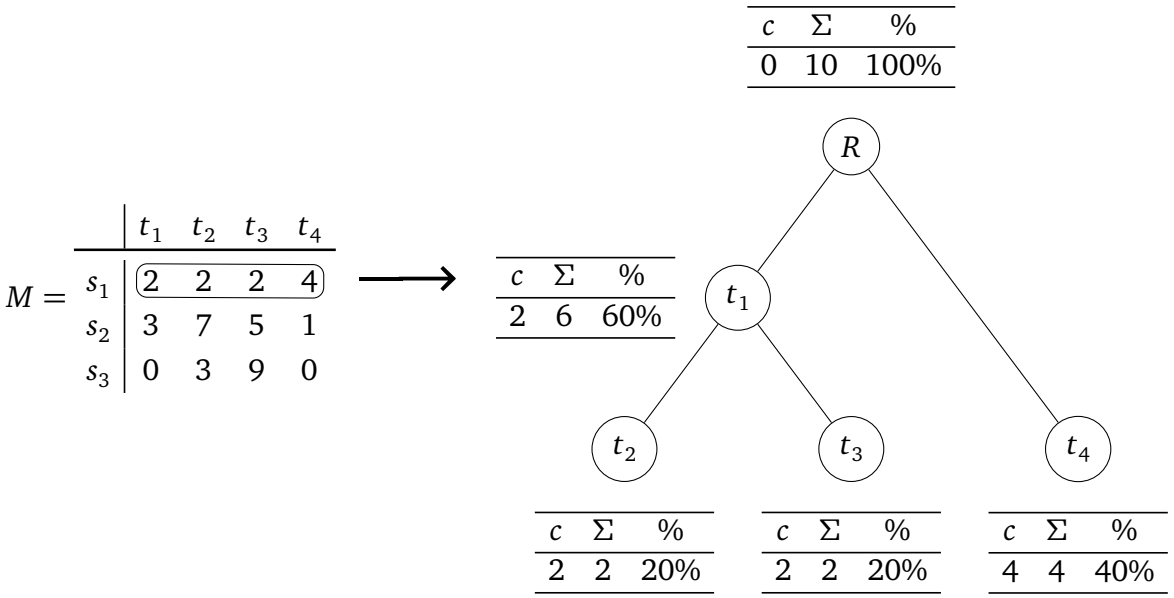


Figure 2.4: Example of an abundance estimation matrix M for a set of samples (s_1, s_2, s_3) and taxonomic units (t_1, t_2, t_3, t_4). The assignment of the reads belonging to sample s_1 is shown with respect to a taxonomic tree (holding all taxonomic units and an additional root node R) where each node contains: a) the assigned reads at the specific node (c), b) the total reads belonging to the subtree rooted at the node (Σ), and c) the fractional share of reads belonging to this subtree (%).

tii and *Escherichia coli*; it would therefore be assigned to *Escherichia*. Reports generated by abundance estimation software usually represent their abundance estimation matrix in the context of the tree, see Figure 2.4 for an example.

Real world workflows can have more complexity such as the ability to dynamically update the database by adding elements generated from the input reads. Another goal that modern workflows attempt to achieve is to estimate genome similarity between genomes contained in a sample and those in the database [17]. Notably, most methods also add virtual taxonomic units that represent technical outcomes such as a read not being assignable or failing quality control standards.

Abundance estimation is usually the precursor to subsequent comparative analysis, for example distinguishing between samples based on the resulting compositions, or linking different sample annotations to specific composition patterns based on significant correlation. The impact of biases or errors introduced in this step on hypotheses validation or falsification is therefore high and deserves special attention.

2.4.1 Classification Strategies

In the context of this thesis we will distinguish between two types of strategies for assigning DNA reads to taxonomic units: mapping-based approaches and k-mer based approaches. Generally speaking, mapping based approaches attempt to directly project a DNA read or a portion of it to the reference genome sequence–while usually allowing a certain amount of difference between read and reference. In contrast, k-mer based ap-

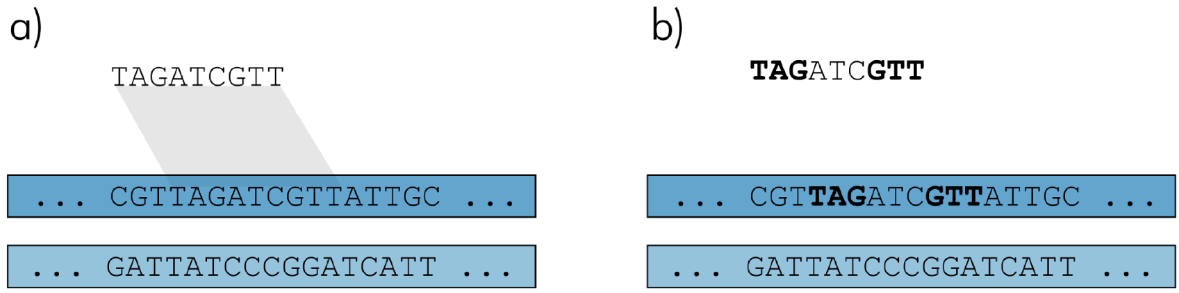


Figure 2.5: Key differences between mapping and *k*-mer based assignment. Panel a) shows a read being mapped to one of two possible candidate references. The read is projected into the genome space to the position it fits best. Panel b) shows a 3-mer approach: Here the read is classified based on containing two 3-mers that are unique to one of the references.

proaches look for small characteristic fragments in a DNA read (of size k) and then make a decision based on the k-mers found. As a consequence they do not directly yield information about a hypothetical location of origin on a reference genome. However, while helpful to understand conceptual differences between specific methods, this separation is not a strict one and individual methods may combine techniques from both approaches. See Figure 2.5 for a schematic comparison.

The tool minimap2 [18] utilizes a mapping-based approach: It begins by finding exact short matches which are referred to as anchors and then checks if the distances between the anchors in the read agree with the distances between the anchors in the reference. Finally, a dynamic programming alignment algorithm is used to "fill in the gaps", resulting in a full projection of the read into the genome space. Per default minimap2 allows that one read can be assigned to multiple possible candidate positions. In the context of abundance estimation, reads that achieve multiple mappings of comparable quality require additional post-processing. Strategies could include discarding of those reads (implying that they are not informative to distinguish between genomes of interest because they originate from a stretch of DNA that is present in both) or assigning them to a shared taxonomic ancestor if a taxonomic tree is available.

In contrast, Kraken2 [19] is an example for a k-mer based tool. The tool builds an index of characteristic k-mers (technically, for performance reasons, Kraken2 uses minimizers, which are the k-mers in a given window that minimize a hash function) out of a set of genome sequences. Notably, the k-mers are structured according to the underlying taxonomic tree: k-mers that are contained in multiple taxonomic units belonging to a shared ancestor but not found elsewhere become characteristic for that ancestor. The tool then scans a read for k-mers and uses a majority voting strategy to decide its placement.

2.4.2 Limitations and Challenges

The task of estimating the abundance of a specific organism in a sample, based on the number of uniquely assigned DNA reads, is not well-defined and usually non-trivial due to a multitude of reasons. We want to outline a selection of challenges here and discuss

their implications.

The lengths of existing genomes span a wide range, including short viral genomes such as the ≈ 30,000 bases long COVID-19 genome up to extremely long plant genomes like *Tmesipteris oblanceolata* consisting of 160 billion bases [20]. If the genomes of both are fragmented into pieces of length 100 and a fragment is randomly chosen (sequenced), the plant genome naturally has a higher likelihood of being observed. In addition, the number of cells and more specifically speaking the number of genome copies varies widely between organisms: A simple haploid bacteria only has a single copy of its genome, whereas a human consists of over 30 trillion cells, most containing two copies of the full genome. As a result, it is usually sensible to be aware of what a given method attempts to approximate and whether the quantity of interest is represented by this measure. The share of individuals in a sample belonging to a species is not necessarily approximated well by the amount of DNA belonging to that species–although both are clearly correlated. Similarly, in an analysis of food plants in gut samples, the estimated biomass is a natural quantification of interest whereas the amount of DNA reads, the estimated genome count, or the estimated count of individuals are usually not as informative.

Current sequencing technologies are not able to directly identify a full genome from an isolated cell, instead multiple wet-lab steps are required (such as cell membrane lysis, polymerase-chain reaction amplification, or the fragmentation of DNA). Naturally, every single procedure introduces its own bias, starting with the acquisition of the sample. Many metagenomic samples—for instance soil, water or stool samples—are not homogeneous and produce a variance that can only be counteracted by homogenization procedures to a certain extent. The effects can add up causing a significant impact, for example, the choice of DNA extraction method has been shown to cause a deviation of factor 10 in the observed abundance of DNA [21]. Potential subsequent steps with their own biases will naturally amplify this effect.

While a natural way of thinking about compositions is to look at a percentage share of DNA reads assigned to a taxonomic unit, relative to all DNA reads in the sample, the absolute number of observed reads carry a meaning of their own. Two samples with identical percentages but vastly deviating total numbers may represent a different presence of observed organisms (as well as differences in sampling or sample processing quality). Since all sequencing workflows are subject to the aforementioned biases, the analysis of low abundance components is especially challenging.

In addition, DNA reads that are not assignable to a specific organism pose a significant challenge: they might be artifacts produced by the sequencing technology, they might also belong to unknown yet present organisms existing in the analyzed ecosystem. When ignored, they lead to found species being overrepresented and a non-comprehensive representation of the analyzed microbiome. Recent large-scale metagenomic assembly projects have led to an increase in the assignable reads, however, significant fractions (15% and more) still remain unassignable [22], possibly reflecting low-abundance organisms.

Another large factor in making meaningful abundance estimations is the choice of database and therefore the set of possible assignments. This includes aspects like completeness, representation of variation, and quality. Public repositories like RefSeq grow at a rapid pace, containing 1,000,000 accessions in 2003 and already 500,000,000 in 2025 [23]. The growth is naturally not evenly distributed across all existing forms of life:

certain species such as typical lab model organisms are vastly overrepresented. Also, the quality of provided references varies greatly. However, even if a consistent quality and representation of the potential microbiomes were achievable, genomic DNA is also subject to mutations and changes, and thus the observed DNA of individuals in a sample will necessarily differ from a reference genome. In a more extreme way, the analysis of yet unknown, or less frequently analyzed microbiomes might require methods for detecting additional genomes not present in a database and the testing of their plausibility. The growth of available references has been shown to particularly affect species-level classification, as more closely related genomes being added lead to more reads being assigned at higher taxonomic levels and thus less specific assignments [24].

Taxonomy trees are not necessarily built based on genetic similarity and can be subject to changes in the scientific consensus, affecting a final reported outcome significantly. Depending on the application it is also not trivial to decide on the level of resolution required. Since some genetic information belongs to conserved regions that are shared at a high taxonomic level, some DNA reads can only be assigned to a high taxonomic unit. On the other end of the scale, it can be relevant in a clinical setting to make distinctions between individual strains of the same species, for instance, to detect toxicity or antibiotic resistance since both can be caused by the presence or absence of few genes.

As a result of the previously discussed effects, most data analysis focuses on overrepresented groups although individual small groups can be as impactful depending on the hypothesis. In a set of airway metagenomes researchers found low abundance species to be a separating factor between healthy children and children affected with cystic fibrosis [25].

When estimating DNA abundances it is imperative to be aware of the challenges described and have a clear understanding of the hypotheses that is analyzed to see if the limitations of a given method support the ability to accept or reject the hypothesis. In the context of the work presented in this thesis it makes sense to think of the abundance estimation as a profiling task, where a characteristic composition vector is assigned to a sample, which might not fully represent the unknown ground truth of all genomic DNA present in the sample, but rather a representation that serves as a foundation for comparative analysis between samples. We believe that, even if a set of DNA reads is assigned to a wrong species (due to flaws or limitations of the databases/methods), as long as the assignment and the errors are consistent across all samples, it is still possible to detect patterns and groups within a set of samples. Supplementary and subsequent analysis can then determine and quantify uncertainty about precise species makeup if that is relevant for the underlying research.

2.5 Genetic Privacy

The genome of a human represents a unique identifier, even more specific than a fingerprint. Not only can it be unequivocally assigned to an individual, but it also contains information about a person's sex, appearance, and susceptibility to certain diseases. With access to genomes of multiple people, it is possible to reconstruct phylogenies that reveal relationships and inheritance patterns. This is clearly highly sensitive information that is of interest to a variety of actors: Health insurance companies may be tempted to use the data for risk profiling, law enforcement agents for identifying the source of DNA

found at crime scenes, and researchers for studying genetic variation. Importantly, even if ones own genome is not available, information can still be inferred from the genomic data of relatives. Many high-profile cases have made headlines, including the identification of the so-called "Golden State Killer", who was identified by reconstructing a family phylogeny from a public DNA genealogy database [26]. In other cases the ethical controversies are more apparent: tissue samples taken from Henrietta Lacks, who was treated for a cervical cancer that eventually took her life, became a backbone for biomedical research around the world without her family ever consenting to this [27]. Subsequent publications, for example one that focused on assembling the whole genome [28], kept disclosing highly sensitive data about the family to the public.

The potential information coming from access to a full genome is enormous, but even small fragments of DNA can be highly revealing which is illustrated by panels using only 92 base positions, which can be used to uniquely identify people with a high certainty [29]. Some diseases are monogenic and triggered by a single base alteration, thus one piece of DNA can already be sufficient to identify an affected human. At the same time, many studies utilizing DNA benefit from a large amount of samples, as acquisition is often both time- and cost-intensive. The benefit of having access to DNA samples from different geographical sources further enhances those studies since this allows generalization of hypotheses that might otherwise be local to specific populations. As a result there are reasonable efforts to collect DNA from diverse sources in databases.

Beyond the ethical challenges of handling genetic data, there are also concrete legal frameworks regulating its use in many countries. In Germany, the "Datenschutz-Grundverordnung" (DSGVO) is of particular interest. Article 9, Paragraph 1 of the DSGVO explicitly forbids the processing of genetic data while Paragraph 2 lists possible exemptions including scientific research in letter (j). Other laws may apply as well such as the "Bundesdatenschutzgesetz" (BDSG), which in Article 1, Paragraph 27 allows usage of personal data for scientific research if the benefits outweigh the individual's interest in preventing its processing. The "Gendiagnostikgesetz" (GenDG) further emphasizes genetic data as particularly sensitive. These regulations have significant implications when it comes to the processing and storing of medical records, especially genetic sequencing data. Even temporary storage on a non-certified server without reasonable security safeguard may not provide sufficient legal certainty and could create legal risks for the server provider.

There is also legislation in other countries that in some instances specifies very concrete requirements for the handling of data: in the United States of America, the "Health Insurance Portability and Accountability Act" (HIPAA) imposes audit and safeguard requirements to ensure that protected information is not improperly accessed in Title 45, Part 164, Sections 308 and 312. Federally funded research is further regulated by the National Institute of Health's (NIH) "Genomic Data Sharing" (GDS) Policy, which requires investigators sharing genomic data to ensure that research subjects cannot be identified through the data and that repositories used are protected by appropriate measures. A common thread found in all these legal frameworks is the careful balancing of individual privacy with the benefits of data sharing for justified purposes such as health care and research.

2.6 Machine Learning on Clinical Datasets

Machine learning in general can be stratified into two approaches: unsupervised learning and supervised learning. The goal of unsupervised learning is to assign target properties (or labels) to an objects based on their already known feature properties. Examples would be the clustering of images based on their visual similarity. Unsupervised methods do not require prior knowledge of the prediction goal and can thus be applied to any dataset. In contrast, supervised learning is concerned with learning from a labeled dataset, where the target properties are known. A model is chosen and parameterized in a way that it assigns objects to the target properties well, in the hope that the assignments on unknown datasets are equally meaningful. Examples would be the prediction of efficiency for a particular drug based on its chemical structure. If we know for example that molecules containing a given functional group are more likely efficient for a specific type of cancer, we can learn this correlation and as a result our model can assign a higher likelihood to unknown drugs that contain the same functional group.

Evaluating supervised methods is usually done using a separate set of labeled data and checking for concordance between the predictions generated by the models and the actual labels. In general, predictions do not need to be binary, and can even be chosen from an infinite domain, for example, the prediction of a probability between 0 and 1 (assuming arbitrary precision) or a higher-dimensional object (e.g., a vector or a tensor).

If we want to formalize this in a very general and broad way, we are usually given a matrix $X \in P^{m \times n}$, where P is an arbitrary set of properties, m is the number of objects, and n the number of properties of said objects. We are looking for a matrix $Y = T^{m \times t}$, where t is the number of target properties (in the set T) that we want to predict. We do not constrain the domains P and T of any entry here but for many applications and methods the entries are assumed to be numeric. A trained machine learning model is now simply a function $f_\chi : P^{m \times n} \to T^{m \times t}$ that depends on a set of parameters χ. The goal is to find a combination of such a function and a set of parameters, usually by using existing pairs of input and output, X and Y.

Although an arbitrary complex model can always be parameterized in a way that any input matrix X gets transformed exactly into a set of labels Y (a trivial way to do so would be to simply have an explicit mapping function as the model), this is usually not desired since this function would not be generalizable to arbitrary datasets, but instead be highly specific to the one it was trained on. The underlying idea of applying machine learning techniques however, is that the models will learn existing patterns in the data, which are potentially highly complex and can not be modeled directly without prior knowledge of them. Ideally, those patterns are not specific to the concrete dataset the model was trained on, but can instead be found in all similar datasets.

While complex machine learning models have been proven to be very powerful, it is worth discussing that in many cases, if a lot of knowledge about the desired task is available, different options of lower-complexity are available and come with advantages of their own. We want to illustrate this briefly with the example of a very powerful machine learning technique, so-called neural networks. Without going into too much detail, the general idea is that our parameterized function models a network of neurons, which have likelihoods to activate other neurons depending on how strongly they themselves are activated. A subset of neurons is then used to determine the decision–their level of ac-

tivation represents the desired prediction. This emulates in a very basic fashion the way a human brain works, and leads—analogue to the human brain—to a general purpose machine that can be adapted to different tasks. If we are now tasked with an arbitrary task, it is possible to train a sufficiently large neural network and have a virtual brain that will eventually solve this task very well. However, we might no longer be able to easily identify how that decision was made in the same way that it is not always trivial for humans to reflect on how an instinctive or subconscious decision came to be. The parameters learned for the neurons can not be easily interpreted. Hypothetically, we could apply a very simple linear model with two parameters to solve the same task and by chance realize that this simple function is equally good. Now we have two models to solve the task but the second model provides the advantage of actually enabling an understanding how the task was solved.

There are also monetary incentives for having more lightweight models since computational power is expensive and large machine learning models take up a large amount of resources that might surpass the capabilities of many medical institutions. For instance, training of large language models can reach costs of 100 million US dollars and even inference using high-end models such as OpenAI's ChatGPT o1 exceeds 50 US dollars per 1,000,000 tokens [30]. As long as there are no substantial financial incentives to offset such expenditure, this level of spending remains unachievable for many medical institutions.

In addition, the issue of explainability is relevant for multiple reasons. It has been shown that the acceptance of machine learning systems in real-world settings is influenced by their perceived explainability [31]. Transparency and explainability can also be an explicit requirement of legal frameworks such as the "Right to explanation of individual decision-making", contained in Article 86 of the European Union's AI Act. However, it should also be noted that there is clearly value in systems that can make reliably accurate predictions even if their explainability is low.

Ethics are a fundamental part of machine learning, especially so on clinical datasets: Since we do not have the ability to see what the learned parameters represent for models with low explainability, it is possible that our model actually learns discriminatory patterns that, for example, over- or underrepresent people based on race, religion, or sexuality.

While the terms "artificial intelligence" and "machine learning" are heavily loaded with many associations and highly diverging definitions, in the last years they became almost synonymous with chatbots based on large language models, primarily due to the success stories of tools like ChatGPT, which is currently the 5th most visited website, recently surpassing services such as Wikipedia, WhatsApp, or Amazon [32]. Although we are not directly concerned with those models in the context of this thesis, we want to note promising applications in the clinical setting, such as the automated translation of doctor's letters and medical records into machine-readable data structures [33] since those developments will eventually impact the primary data sources used for the work presented here.

In this thesis, we are primarily concerned with clinical datasets, and it is therefore worth highlighting some of their properties here. All datasets are noisy, however clinical datasets are especially likely to contain irregularities and anomalies. This can be easily illustrated when looking at a hypothetical patient treated on an intensive care unit. Said

patient might be monitored with a 24-hour electrocardiogram. While receiving basic care, water can get into contact with the electrodes resulting in transient current spikes. As the patient is getting positioned to avoid pressure ulcers, an electrode might get disconnected leading to a temporary loss in signal for a set of channels until the disconnect is discovered. Throughout the day, the patient might also be required to get a magnetic resonance tomography scan which in turn requires all electrodes to be temporarily removed. The resulting signal will not record any heart activity during this time period. Naturally, the clinical routine adds variance to sampling time points. Manually collected data (such as blood tests) may be sampled at different hours on subsequent days, thereby hindering comparability.

Depending on the existing clinical infrastructure, data can be highly fragmented across different sources. Some measurement tools feed their data directly into one or multiple databases while others rely on manual entry (and are thus prone to human error). Additional measurements are only recorded on paper, lacking any digital record or standardization.

Summarizing those assumptions, we can derive frequently observed, shared properties of clinical datasets: They are sparse (they are missing data points, and they are temporally irregular), they are erroneous, and they require extensive pre-processing and curation to be utilized at all. As a result, methods need to be robust enough to address these challenges.

2.6.1 Ensemble Classifiers

Some of the methods presented here such as the XGBoost [34] or Random Forest methods are so-called ensemble classifiers and we want to give a short explanation of the idea behind this type of method as well as some of the important properties with context to the research here. Ensemble classifiers combine k functions ${f^1}_{\chi^1}, \ldots, {f^k}_{\chi^k}$ where each function ${f^i}_{\chi_i}$ maps an input matrix X to a target $Y_i \in T^{m\times t}$. Subsequently, a (possibly parameterized) aggregation function $A_\mu : (T^{m\times t})^k \to T^{m\times t}$ is used to produce the final output for a given input. While simple ensemble classifiers have independent classifiers, complex models can have inter-classifier dependencies so a function ${F^a}_{\chi^a}$ may have a parameter set χ^a that is itself a function of the output produced by either another function ${F^b}_{\chi^b}$ or a set of other functions. Typically this leads to "sequential" learning as the functions need to be evaluated sequentially (or iteratively) whereas, in the case of independent functions, they can be evaluated simultaneously. Common strategies for aggregation are majority voting in the case of binary label prediction or averaging in the case of continuous value prediction. The advantage of ensemble classifiers are usually that a highly heterogeneous dataset with many underlying hidden processes can be represented more easily as the individual classifiers may mirror parts of said processes. By keeping individual classifiers on a low complexity it is also possible to avoid overfitting while still having access to a complex prediction system.

An example for an ensemble classifier consisting of independent classifiers is the Random Forest classifier. Here, the individual classifiers are decision trees, where every node contains a single equality based on a single feature and every leaf contains a prediction. The prediction is now simply obtained by threading a sample through each tree by deciding if the respective feature is below or above the threshold (see Figure 2.6). For

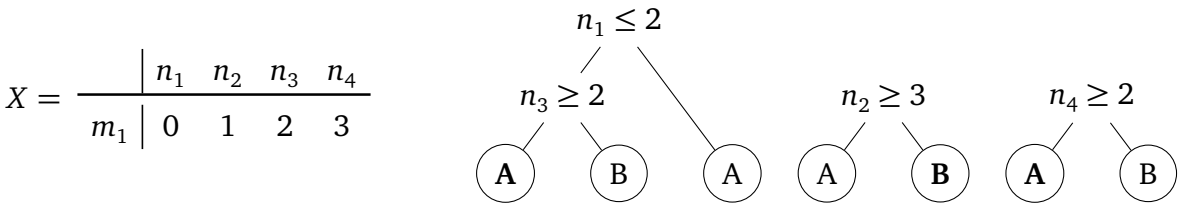


Figure 2.6: Random forest classifier. Here a sample consisting of four features is classified by an ensemble of three trees. If an equation is satisfied, the left path is followed, otherwise the right path. Two of the three trees and thereby the majority would classify the sample as "A".

aggregation, a simple majority voting is used, meaning the prediction that is output from most decision trees is used as the final output for a given row.

A commonly used, sequential ensemble classifier is the XGBoost algorithm, which differs in the way the training phase is conducted. It begins by assigning each sample an arbitrary output for each target. Then, a residual error term is calculated that serves as a distance metric between the current prediction and the true labels. A decision tree is now constructed with the goal of obtaining splits at each node that reduce the overall residual error. A new combined prediction is obtained by combining the output from the new tree with the previous predictions. The process is controlled with parameters that limit the impact of individual trees to prevent overfitting.

2.6.2 From Time Series to Feature Columns

A particularly interesting property of clinical datasets is that many of them represent time series, for example, subsequent measurements of vital signs. Intuitively, we want to be analyzing those measurements, not as independent data points, but as a unit. We illustrate this with the example of an electrocardiogram: a medical expert would not be able to make a meaningful diagnosis based on a set of unordered current values but could easily spot indicators for heart diseases, looking at a plotted line of the current values in their temporal order. Since many machine-learning methods need to work with numerical features, it becomes desirable to represent the essential properties of a time series (including those a human would utilize, when evaluating it) as numerical features. An alternative, straightforward approach would be to avoid modeling those properties and instead rely on the model itself being able to capture all the properties via appropriate parameter choice. A high-complexity model, such as a sufficiently large neural network, can perform image processing and analyze an EKG in a way that mirrors processes in the brain of a human analyst without an explicit modeling of time series properties.

By utilizing features based on time series characteristics, we shift complexity from the model and its parameters to an additional feature engineering step. We are thereby giving simpler models the option to catch this complexity, which they would have otherwise not been able to. For example, a linear regression model can not, by definition, catch any non-linear correlation in its data. However, if this correlation is already implicitly modeled in the features, it can be represented by the method. Analogous to the previously discussed dichotomy between complex and simpler machine learning models, where we sacrifice explainability for a possibly very powerful and accurate model, this approach comes with

a comparable advantage. Namely, it is possible to gain a more intuitive explainability since certain features, such as “the number of peaks present in a given series”, can be understood and visualized easily for a human observer. In contrast, this property might be impossible to derive from the parameters in a neural network.

Nevertheless, constructing and choosing suitable time series features may introduce significant bias into the method, since an excessive amount of manually chosen but uninformative features can lead to a separation of few data points by chance and thus to the learning of non-generalizable patterns. It is therefore advisable to pursuit a data-driven approach in choosing features by focusing on statistically significant separators within a given dataset [35]. The feature engineering step may be complex and can be computationally expensive. Convolutional methods have been shown to achieve good performance in capturing essential properties of time series [36] but do so at the cost of lesser explainability.

Chapter 3

Major Contributions

This thesis highlights different contributions that are prototypes for the possibilities of novel technology implementation in a clinical setting. Figure 3.1 illustrates the development of the related projects and their context. Starting with a cohort of 847 patients, undergoing the allogeneic stem cell transplantation procedure between 2004 and 2019, we began with the curation and processing of clinical data sources including paper charts, free-form text documents, and a multitude of database systems. In this context we created data extraction and transformation tools to combine the data sources and deliver machine-readable formats for further analysis. For the analysis of this complex and high-dimensional data, we created a web-based visualization and exploration tool, "MedReactor" in close cooperation with clinicians and their needs. Based on the clinical data sources, we introduce a machine-learning approach providing a live risk prediction system based on medical records and derived time series (Publication 1). For a smaller subset of 31 patients we collected stool samples throughout the clinical course of transplantation and used Nanopore sequencing to obtain DNA reads. A workflow we developed characterizes the microbiome compositions while providing additional confidence for the presence of individually reported species through a partial mapping. It also allows the comparison between observed strains of the same species at different time points (Publication 2). Extensive experimentation with the Nanopore barcoding technology (multiple samples are sequenced simultaneously and then reassigned to samples using markers) led us to the observation that barcode cross-talk is a serious issue which can distort reported compositions. As a result a study quantifying this effect and showing mitigation strategies is currently ongoing. Our microbiome analysis workflow was optimized to analyze sparse microbiomes and applied to different datasets such as pancreas tissue or wound microbiomes. Additionally, methylation analysis of the sequenced DNA is currently ongoing. Since a lot of the genetic data collected in the projects contains human DNA, it is highly sensitive. A side-project thus presented a highly scalable client-server architecture that allows for the privacy-preserving exchange of DNA data (Publication 3).

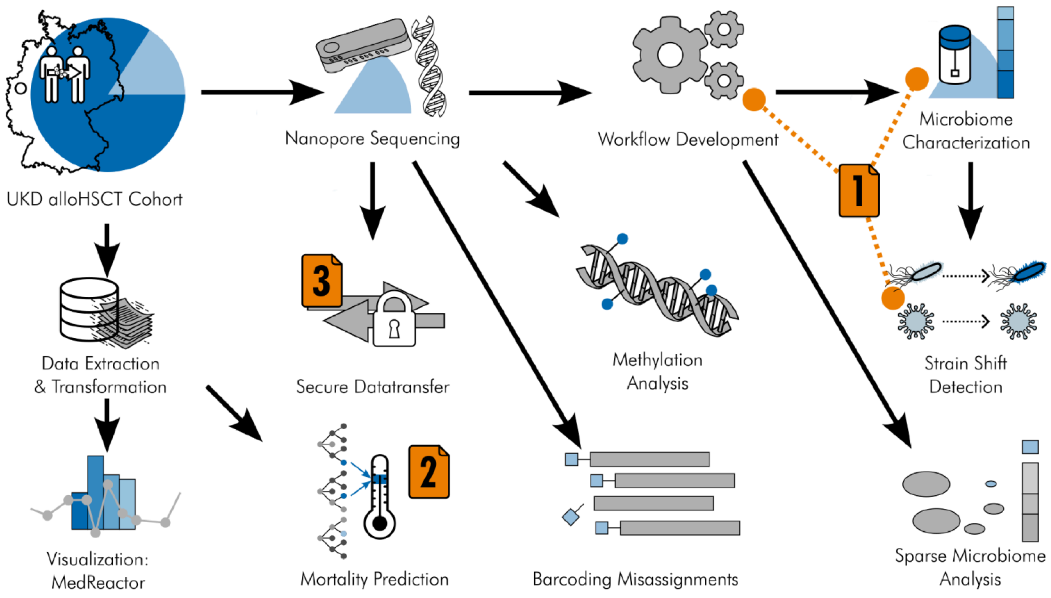


Figure 3.1: Outline of selected projects described in the thesis and their development. The major publications discussed in this thesis are numbered in order of their appearance and highlighted in orange.

3.1 Insights into gut microbiomes in stem cell transplantation by comprehensive shotgun long-read sequencing

3.1.1 Motivation: Towards personalized bedside medicine

The increasingly affordable prices for whole genome sequencing as well as the low processing times have made the goal of bedside medicine utilizing genomic medicine more achievable. As a step in this direction, we began with the collection of stool samples by patients undergoing allogeneic stem cell transplantation with a variety of hypotheses but also as a proof of feasibility. Was the Nanopore sequencing technology advanced enough that we could develop a workflow to build robust analysis and leverage the advantages such as the lower costs and the faster turnaround times? Eventually we settled on core questions to focus our efforts on. We were interested whether patients start the procedure with different microbiomes and if those compositions (or the presence of specific marker organisms) are associated with outcome (concrete: graft-versus-host disease and mortality). In addition, we wanted to identify if the primary source of the microbiome that reconstitutes after myeloablative treatment is exogenic or endogenic, meaning if the bacteria survive within the host in small quantities and then repopulate or if they are reintroduced via food or other sources. From a bioinformatics perspective, this translated to three specific goals our method needed to address simultaneously:

1. How do the compositions between samples differ?
2. Which organisms are present in a sample with high confidence?

3. Can organisms found in two associated samples be compared and identified as belonging to the same (or a different) strain?

In the following paper we present a reproducible workflow that is able to answer the questions and present results from its application to 32 stool samples from patients undergoing allogeneic stem cell transplantation. We found a three-cluster structure within the pre-transplantation samples characterized by *Bacteroides* domination, *Enterococcus* domination, and a mixed composition respectively.

scientific reports

Check for updates

OPEN

Insights into gut microbiomes in stem cell transplantation by comprehensive shotgun long-read sequencing

Philipp Spohr[1,4,5], Sebastian Scharf[2,5], Anna Rommerskirchen[2,5], Birgit Henrich[2], Paul Jäger[3], Gunnar W. Klau[1,4✉], Rainer Haas[3✉], Alexander Dilthey[2,4✉] & Klaus Pfeffer[2✉]

The gut microbiome is a diverse ecosystem, dominated by bacteria; however, fungi, phages/viruses, archaea, and protozoa are also important members of the gut microbiota. Exploration of taxonomic compositions beyond bacteria as well as an understanding of the interaction between the bacteriome with the other members is limited using 16S rDNA sequencing. Here, we developed a pipeline enabling the simultaneous interrogation of the gut microbiome (bacteriome, mycobiome, archaeome, eukaryome, DNA virome) and of antibiotic resistance genes based on optimized long-read shotgun metagenomics protocols and custom bioinformatics. Using our pipeline we investigated the longitudinal composition of the gut microbiome in an exploratory clinical study in patients undergoing allogeneic hematopoietic stem cell transplantation (alloHSCT; n = 31). Pre-transplantation microbiomes exhibited a 3-cluster structure, characterized by *Bacteroides* spp. /*Phocaeicola* spp., mixed composition and *Enterococcus* abundances. We revealed substantial inter-individual and temporal variabilities of microbial domain compositions, human DNA, and antibiotic resistance genes during the course of alloHSCT. Interestingly, viruses and fungi accounted for substantial proportions of microbiome content in individual samples. In the course of HSCT, bacterial strains were stable or newly acquired. Our results demonstrate the disruptive potential of alloHSCTon the gut microbiome and pave the way for future comprehensive microbiome studies based on long-read metagenomics.

Keywords Microbiome stability, Whole-genome sequencing (WGS), Metagenomics, Bacteriome, Mycobiome, Archaeome, Eukaryome, Virome, Leukemia, Bioinformatics pipeline, Oxford Nanopore, Hematopoietic stem cell transplantation (alloHSCT)

Allogeneic hematopoietic stem cell transplantation (alloHSCT) is a potentially curative treatment for patients with high-risk hematological malignancies encompassing acute myeloid leukemia, high-risk myelodysplastic syndromes (MDS), and lymphoid leukemia[1–4]. It is well-established that the gut bacteriome can be associated with specific outcomes of alloHSCT, including the occurrence of adverse events such as graft-versus-host-disease (GvHD) or life-threatening infections. Recently, adverse outcomes, including the risk of GvHD, have been linked to reduced bacterial diversity of the gut microbiome[5–8], while the risk of bloodstream infections has been linked to the domination of individual taxa within the gut bacteriome[5,6,8,9]. A set of microbial taxa, including *Enterobacteriaceae*, *Clostridiales*, and *Blautia* have been implicated in alloHSCT success[6,10–13]. While a number of explanations for the observed associations have been put forward, including modulation of the immune system by microbiota-derived components and microbiome-host crosstalk at the level of metabolites,

[1]Chair Algorithmic Bioinformatics, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [2]Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, University Hospital Düsseldorf, Düsseldorf, Germany. [3]Department of Hematology, Immunology, and Clinical Immunology, Heinrich Heine University Düsseldorf, University Hospital Düsseldorf, Düsseldorf, Germany. [4]Center for Digital Medicine, Düsseldorf, Germany. [5]These authors contributed equally: Philipp Spohr, Sebastian Scharf and Anna Rommerskirchen. ✉email: gunnar.klau@hhu.de; haas@med.uni-duesseldorf.de; alexander.dilthey@med.uni-duesseldorf.de; klaus.pfeffer@hhu.de

the observed associations—apart from an association of overall diversity with outcome, which has been replicated in international multi-center studies[14]—often remain inconsistent[5,15] and incompletely understood.

The vast majority of microbiome studies were performed using ribosomal DNA (rDNA) based sequencing methods designed for bacterial and fungal 16/23S, 18/28S or ITS region rDNA sequences, respectively, thus neglecting the non-fungal eukaryome, and the DNA virome[16–23]. rDNA sequencing is a well-established, scalable, and cost-effective technology; it has, however, important limitations as it bears potential amplification-induced biases in bacteriome/mycobiome composition estimates and challenges in reliably assigning accurate species—or genus-level labels[24,25]. 16S/18S rDNA sequencing is thus ill-suited to characterize the potential contributions to alloHSCT outcomes of other domains of microbial life. For example, in the context of infection prevention, fungi and protozoan parasites like *Cryptosporidium* spp. and *Toxoplasma gondii* can become relevant in the clinical management of alloHSCT[26,27]; and *Candida* has been associated with GvHD[14] and survival[28,29]. With regard to viruses, increases in persistent DNA viruses and reduced bacteriophage richness were observed to be associated with enteric GvHD[30].

A challenge in characterizing associations between microbiome and alloHSCT outcomes consists in the highly dynamic nature of the gut microbiome over the course of alloHSCT[31]. The gut microbiome undergoes substantial temporal variation even in healthy control individuals[32]; In the context of alloHSCT, patient microbiomes have often been impacted by multiple cycles of cytostatic and antiinfective therapeutic treatment prior to the initiation of alloHSCT[33], and continue to be biased by the effects of anti-bacterial, anti-fungal, and anti-viral prophylaxis, in addition to the effects of myeloablation and the subsequent establishment of a "new" immune system by the transplanted allograft[34].

Longitudinal studies taking the aforementioned aspects into account are sparse, thus, we developed a novel workflow to enable whole-microbiome profiling of patient microbiomes, covering all domains of microbial life (with the exception of RNA viruses)[35]. Interrogation of non-bacterial domains of microbial life from shotgun metagenomics is challenging[36–38]; reasons for this include contamination[39,40] and coverage gaps in the relevant reference databases, likely remaining despite significant recent expansion efforts[41,42]. We thus chose to implement the shotgun metagenomics step using long-read sequencing, based on the Oxford Nanopore technology, as the accuracy of taxonomic assignment generally increases with read length[43], and assembled a comprehensive 292 Gb reference database (MetaGut database v 1.0) as the basis of the bioinformatics workflow. Furthermore, k-mer-based classification is known to be sensitive to mis-classification in the presence of out-of-database genomes[43,44], which, the utilization of a large reference database notwithstanding, remains a relevant concern in the gut microbiome context. We thus propose to complement initial k-mer-based classification with a mapping-based verification approach to reduce the rate of false-positive taxonomic detections. Tailored bioinformatics increased the analytical accuracy and reduced the rate of false-positive taxonomic detections. Moreover, bioinformatic tools for the enumeration of antibiotic resistance genes were implemented. Based on this, we applied the pipeline to characterize microbiome dynamics over the course of alloHSCT in an explorative clinical study including patients (n = 31) before alloHSCT, during the phase of leukopenia, and hematological reconstitution. In addition, a comprehensive characterization of gut microbiomes was performed for a group of healthy volunteers (n = 11) and compared to patient microbiomes pre-Tx. Our pipeline proved to be a valuable tool for the comprehensive characterization of microbiomes. Furthermore, a 3-cluster structure of pre-Tx patient microbiomes was detected and the acquisition and replacement of bacterial strains during the course of alloHCST could be successfully monitored.

Results

A pipeline for the accurate and comprehensive characterization of gut microbiomes

To enable the characterization of the gut microbiome encompassing the bacteriome, mycobiome, archaeome, DNA virome (bacteriophages/viruses), and protozoa of hematological patients from stool samples at high-resolution, we developed an integrated method for robust microbiome characterization (Fig. 1). It comprises the following components: (i) protocols for stool sample processing and DNA extraction, based on a modified version of the Human Microbiome Project (HMP) protocol[45,46] for robust sample handling and DNA extraction and suitable for processing samples at different degrees of stool consistency (Methods); (ii) long-read sequencing and compositional analysis of microbiota, based on the Oxford Nanopore platform and a custom bioinformatics approach integrating Kraken2-based read assignments[47] with a newly developed mapping-based validation to ensure that sufficient-quality pairwise alignments exist between reads and the taxonomic entities they are assigned to; taxa with low rates of mapping-based read validation are flagged and not included in downstream analyses and visualizations (see Methods for details); (iii) targeted antibiotic resistance gene (ARG); and (iv) crAssphage analyses, based on read mapping to specific databases comprising (a) ARGs and (b) recently assembled crAssphage strain sequences[48]; (v) longitudinal tracking of strain dynamics: short-read Illumina sequencing data are used to detect single-nucleotide variants (SNVs) in bacterial reference genomes, and the temporal dynamics of the detected SNVs over multiple samples from the same individual are investigated to infer the maintenance, acquisition and loss of bacterial strains. Extraction of sufficient amounts of DNA for metagenomic sequencing was challenging during the period immediately following alloHSCT, when patient stool samples are known to vary in consistency, rendering the exhibit very low biomass. We thus incorporated a modified version of the HMP gut microbiome protocol[45,46], which proved to work well in our study across the course of alloHSCT. We initially carried out three experiments as a validation for the protocol and bioinformatic pipeline:

First, we applied the workflow to a well-defined microbial community standard (Zymo Gut Microbiome Standard) and observed good concordance at the genus level between the theoretical and the inferred composition and lesser concordance at the species level (Pearson's r = 0.82/0.67, for comparison a minimap2-based abundance

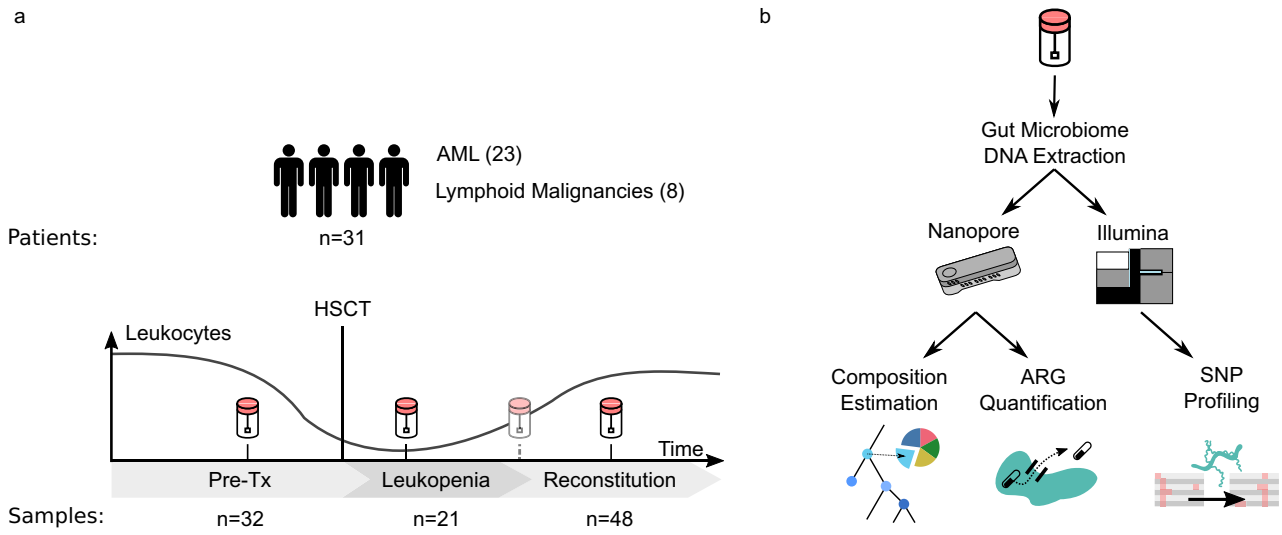


Figure 1. Overview of the exploratory clinical study (**a**) and the workflow (**b**). (**a**) 31 patients undergoing alloHSCT were recruited for an exploratory study and stool samples were collected longitudinally before transplantation (pre-Tx), during leukopenia (defined as white blood count of ≤ 1,000/µL), and during reconstitution for each patient. (**b**) The workflow comprises optimized protocols for sample preparation and DNA extraction, long-read metagenomics using the Oxford Nanopore technology, and custom bioinformatics for taxon validation, quantification of antibiotic resistance genes/crAssphage sequences, and tracking of strain dynamics based on short-read Illumina data.

estimation leads to r = 0.84 for both levels; Supplementary Fig. 1). However, an elevated proportion of false-positive species hits, accounting for 26.02% of total abundance were driven by reads from *Veillonella rogosae* and *Prevotella corporis* (which do not contain genomes in the database) being misassigned to different species of their respective genera (Supplementary Table 1).

Second, we applied the workflow to a cohort of 10 healthy volunteers. Here, we observed good agreement of high-level compositional metrics with the WGS-based component of LifeLines-DEEP[49] (Supplementary Fig. 2). Of note, at a median frequency of 25.47% (range 9.10%–51.72%), the inferred abundance of *Bacteroides spp.* was higher in the investigated cohort of healthy volunteers compared to the LifeLines-DEEP cohort with a median frequency of 14.99% (range 0.80%–48.94%).

Third, we assessed the agreement between Nanopore and Illumina based sequencing. We used Kraken2 with our default database on short-reads we had for a subset of samples. For each sample that generated at least 10.000 reads with both platforms we resampled to a fixed amount of 10.000 reads and then compared the resulting compositions (Supplementary Fig. 12). We observed an overall agreement (Pearson's r of 0.88) between the composition vectors.

Reliable characterization of gut microbiomes over the course of alloHSCT

To characterize the microbiomes of alloHSCT patients and to investigate the dynamic changes of the gut microbiome over the course of alloHSCT, we recruited a cohort of 31 patients, diagnosed with defined hematological malignancies (see Supplementary Note 1 and Methods for a description of the cohort and recruitment process, and Supplementary Table 2 for patient characteristics), undergoing alloHSCT at Düsseldorf University Hospital.

Based on 101 stool samples collected longitudinally at defined time points over the course of alloHSCT (pre-Tx, leukopenia, reconstitution, see Fig. 1), we assessed the ability of the workflow to enable microbiome characterization at different stages of the treatment cycle (Fig. 2). First, we observed that DNA yields in the hematological cohort differed significantly from the healthy cohort (patients 0.47 µg/g stool, healthy 5.95 µg/g stool; $p = 0.000011$) and also displayed significant variation between treatment time points (Kruskall–Wallis p-value: 0.0031 for the 3 treatment phases). DNA yields were typically lowest during leukopenia (median = 0.06 µg/g stool), compared to the pre-Tx and reconstitution periods (medians = 0.6 µg/g stool and 1.1 µg/g stool, respectively). Second, the utilized protocols enabled the generation of more than 100,000 reads per sample (89 samples with > = 100,000 reads). We observed that median read counts were lowest for samples taken during leukopenia (median = 128,330). Third, the median of the median sample read lengths was 653 base pairs (bp) with the longest read spanning 888,591 bp. DNA extraction and sequencing data statistics are summarized in Supplementary Table 3. Fourth, at median per-species validation rates of 77.79% (bacteria) and 50.00% (viruses) across treatment periods and in the healthy cohort showed generally high rates of mapping-based validation (Supplementary Fig. 3). Fungi exhibited lower validation rates (median = 5.33%); we observed pronounced differences in fungal validation rates between the healthy and alloHSCT samples, and also between the characterized time points, with the healthy samples generally exhibiting the lowest fungal validation rates (Supplementary Fig. 3).

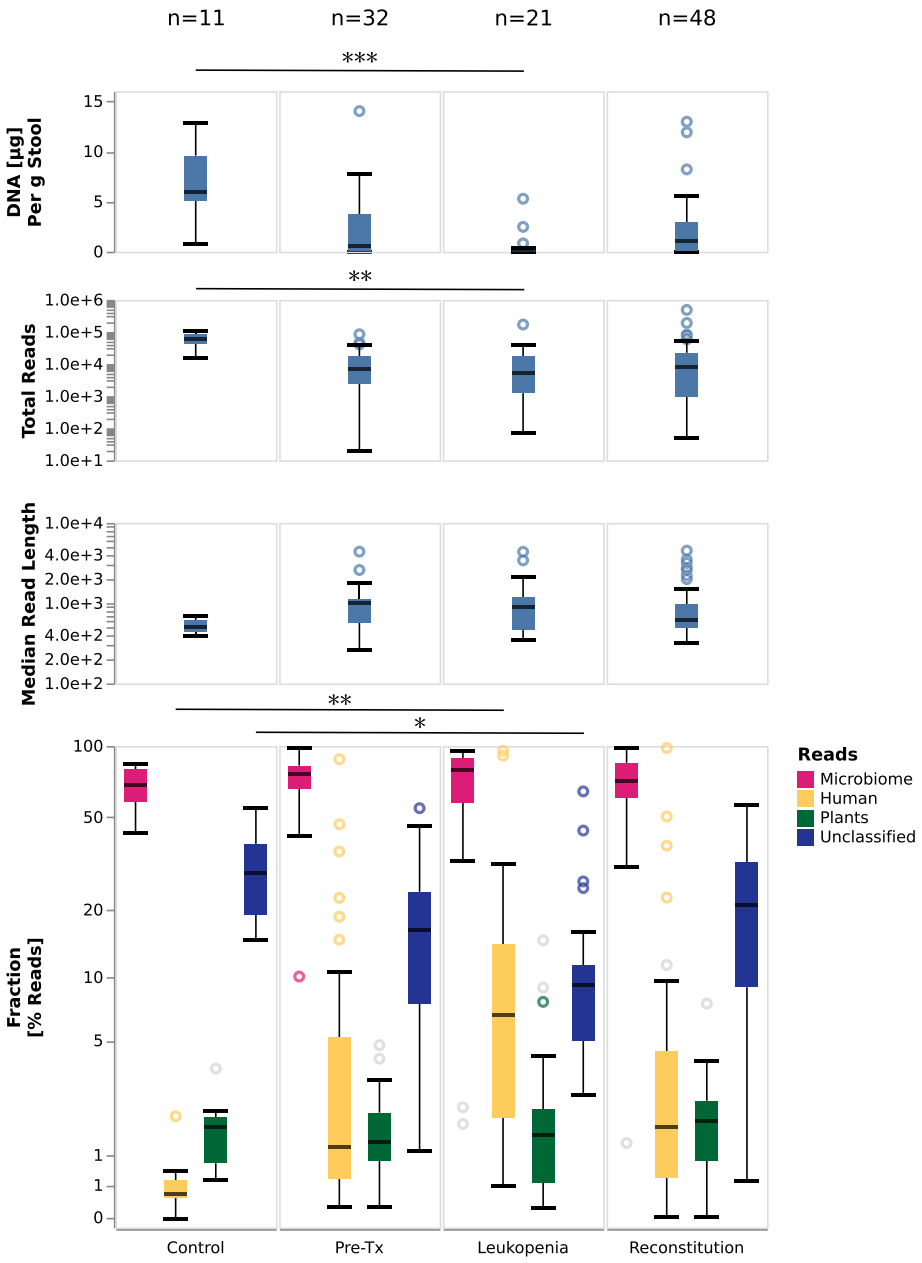


Figure 2. DNA amounts, sequencing metadata, and high-level stool sample compositions. Shown are the amounts of extracted DNA per gram of stool; the number of generated Nanopore sequencing reads; median read lengths; and the relative proportions of microbial-, human-, and plant-assigned reads, as well as the proportions of reads remaining unclassified by the initial Kraken2-based step. Gray circles in the bottom panel indicate outliers for which less than 80% of the reads are in taxa validated by the mapping-based validation step. For each panel, the most significant difference between included categories is indicated (Mann–Whitney U test; * [$p < 1.19 \times 10^{-3}$], ** [$p < 2.38 \times 10^{-4}$] or *** [$p < 2.38 \times 10^{-5}$]). Raw data is shown in Supplementary Table 3 and Supplementary Table 4.

Human-derived DNA is often not reported in microbiome studies; we however noted that the amount of human DNA differed significantly between the control group and the patient cohort, and also between the sampling time points within the hematological cohort. The relative proportion of human DNA was highest during the leukopenic period, possibly reflecting the effect of myeloablative therapy on the gut mucosa (Fig. 2). However, the increase in human DNA proportions might, at least in parts, reflect the decrease of microbiota after myeloablation and anti-infective therapy. We also observed that the per-read validation rates of the human reads were lowest for the control samples, i.e. indicating that a substantial proportion of the reads assigned by Kraken2 to the "Homo sapiens" species in these samples may reflect false-positive assignments (Supplementary Fig. 3).

Per-species validation rates for archaea and non-fungal eukaryotes were close to zero in most control and alloHSCT samples and also across the characterized time points. Analyses of the occurrence and abundance of

individual species were thus limited to the bacterial, viral and fungal components of the microbiome. Of note, the presence of plant-based DNA could also sometimes be validated, in particular during the leukopenic period, the most frequently validated plant DNA was *Cucumis sativus* (Cucumber) and different *Triticum* spp. (wheat), *Pisum sativum* (pea) and *Musa* spp. (banana) (Supplementary Fig. 3).

alloHSCT microbiomes exhibited dynamic and diverse compositions

Having established and validated our workflow as an informative method, we investigated high-level gut microbiome compositions (Fig. 3). DNA yield was lower in the patient cohort overall and specifically during leukopenia, possibly reflecting the impact of conditioning and antiinfective therapy as well as the stool consistency. The number of normalized species was highest in the samples from the control group and lowest for samples during leukopenia (Fig. 3, top panel). Besides, ARG-reads significantly differed between control samples (low) and alloHSCT samples (high) (Fig. 3, middle panel). The frequency of ARG reads of the control samples was significantly lower compared to alloHSCT samples ($p = 0.0006281$). Bacteria accounted for > 90% of microbial reads in 89/101 hematological samples and in 11/11 of control samples; bacteria thus dominated the large majority of the investigated samples. At median relative abundances of 0.53%, 0.11% and 0.02%, fungi, viruses, and archaea accounted for low, but non-negligible proportions of the characterized alloHSCT and

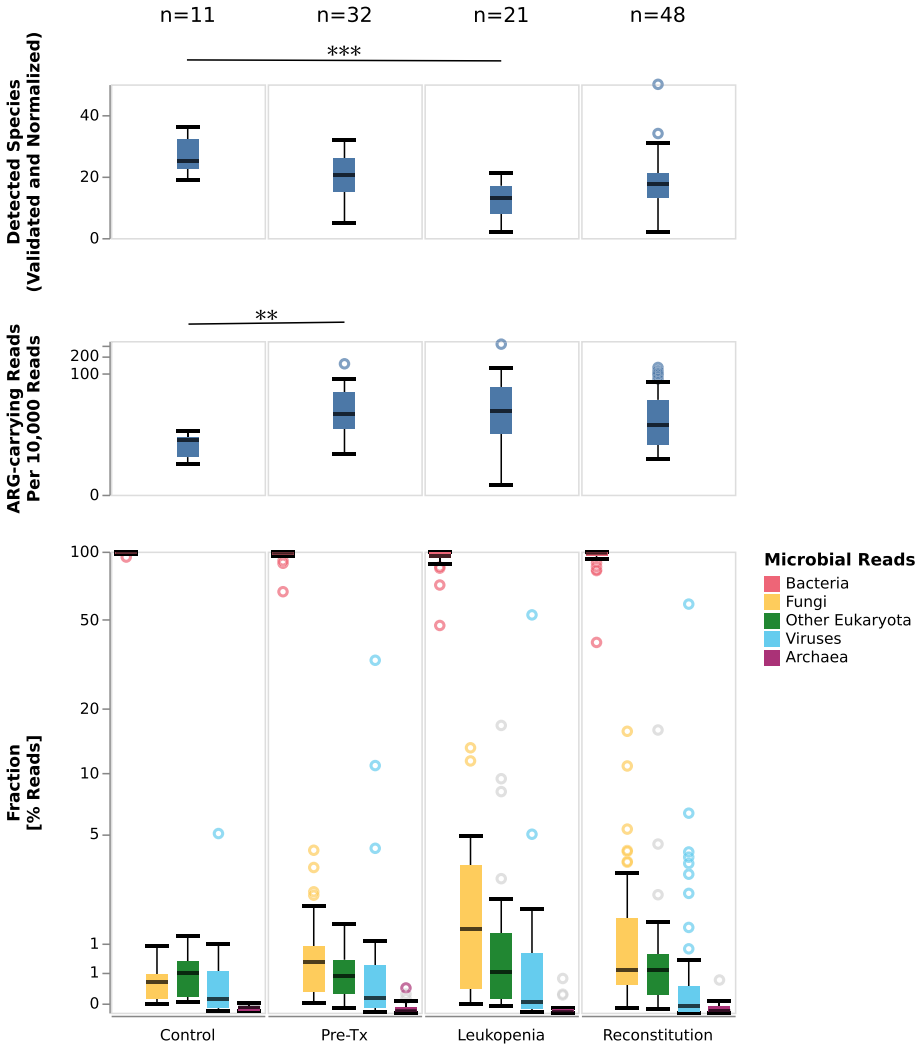


Figure 3. Diversity, ARG-carrying reads, and microbiome composition in stool samples from controls and alloHSCT patients at indicated phases. Shown are the numbers of detected genera, normalized to the sample with the lowest read count; the rate of reads that carry an ARG element; and the relative proportions of bacterial-, fungal-, archaeal-assigned reads as well as the proportion of reads assigned to DNA viruses (incl. bacteriophages), and non-fungal eukaryotes. Gray circles in the bottom panel indicate outliers for which less than 80% of the reads are in taxa validated by the mapping-based validation step. For each panel, the most significant difference between included categories is indicated (Mann–Whitney U test; * [$p < 1.19 \times 10^{-3}$], ** [$p < 2.38 \times 10^{-4}$] or *** [$p < 2.38 \times 10^{-5}$]). Raw data is shown in Supplementary Table 3 and Supplementary Table 4.

control microbiomes; median abundances of these groups were generally comparable between control and alloHSCT samples and between alloHSCT time points, with the exception of fungi, which showed an increased abundance in the alloHSCT samples in general and during leukopenia in particular (median = 1.32% vs 0.36% for control) (Fig. 3, bottom panel). Non-fungal eukaryotes were also estimated to account for small fractions of the characterized microbiomes (median = 0.49%).

The dominance of bacteria in most samples notwithstanding, individual microbiome samples were found to exhibit increased frequencies of non-bacterial taxa. For example, in two alloHSCT samples from one patient, the abundance of viral DNA was found to be ≥ 50%, accounted for by species belonging to the genus crAssphage; in two additional samples from different patients, it was ≥ 10% (Supplementary Table 14). Interestingly, in one control sample, viral abundance approached 5% (Fig. 3, bottom panel). Similarly, in five samples, fungal DNA was found at abundances ≥ 5%, accounted for mostly by the species *Saccharomyces cerevisiae*, *Candida albicans*, and *[Candida] glabrata* (Supplementary Table 6). In all of these instances, taxon presence was validated with read-mapping based verification.

Pre-Tx alloHSCT microbiomes could be grouped into 3 distinct clusters

To investigate the structure of the pre-transplantation gut microbiome in alloHSCT patients, we carried out a PCoA analysis and found that pre-Tx alloHSCT and control samples fell into 3 distinct clusters (Fig. 4).

Cluster 1, comprising 18 pre-Tx alloHSCT samples from distinct patients and all 11 control samples, was characterized by high abundances of species belonging to the genera of *Bacteroides* and *Phocaeicola,* accompanied by relatively high normalized numbers of detected species per sample; specifically, *Bacteroides uniformis, Phocaeicola vulgatus and Phocaeicola dorei* accounted for ≥ 25% of reads in Cluster 1 samples, while the median number of detected normalized and validated species per sample was 21.5, depicting a relatively high diversity. Cluster 3, comprising 6 pre-Tx alloHSCT samples, was characterized by *Enterococcus* spp. domination and exhibited the lowest diversity in terms of the number of detected species per sample; in 5/6 Cluster 3 samples, *Enterococcus faecium* accounted for ≥ 25% of sequencing reads, at a median number of 7.5 detected normalized and validated species per sample. Cluster 2, comprising 8 pre-Tx alloHSCT samples, was the most diverse in terms of the number of detected normalized species per sample (median = 26) and was also characterized by more uniform species abundance distributions. In Cluster 2, a median of 7.5 species were required to account for ≥ 50% of sequencing reads, compared to medians of 5.5 and 1 species in Clusters 1 and 3, respectively. Notably, Cluster 2 species contained in individual samples included *Citrobacter freundii* (max. abundance 12.6%), *Lactobacillus amylovorus* (max. abundance 41.4%), and *Escherichia coli* (max. abundance 27.5%); by contrast, the genera characteristic for Cluster 1 and 3 *(Bacteroides/Phocaeicola* and *Enterococcus*) were present at only low abundances in Cluster 2 (abundances ≤ 10% for 7/8 samples of Cluster 2) (Supplementary Table 15).

Suggestive structural differences between the 3 pre-Tx clusters were also apparent at the level of the fungal and viral microbiome components (Fig. 7). *Saccharomyces cerevisiae* (validated presence in 19 samples and > 10% relative within-fungal abundance in 15 samples) and *Candida albicans* (validated presence in 3 samples and > 10% within-fungal relative abundance in 3 samples) were the most abundant fungal species detected during the pre-Tx period; the presence of *Candida albicans*, however, was limited to Cluster 2 and 3 samples. At the level of the DNA virome, *Skunavirus* dominated (> 50% relative abundance) 8 of 14 Cluster 2 and Cluster 3 samples, but only 2 of 18 Cluster 1 samples. Conversely, crAssphage was found mostly in Cluster 1 samples (6 validated detections in total, 5 of which occurred in Cluster 1 samples); and 2 Cluster 1 samples were crAssphage-dominated, but no samples from Clusters 2 or 3.

Microbiomes in leukopenia exhibited decreased diversity and increased abundances of fungi and various bacterial genera

We proceeded to investigate the structure of microbiome samples collected during the leukopenic period, reflecting the combined effects of myeloablation, anti-infective prophylaxis, and recent alloHSCT. Consistent with an assumed bottleneck effect of anti-infective prophylaxis on the population of gut microbes, the leukopenic period exhibited the lowest median number of normalized detected and validated microbial species per sample (13; compared to 20.5 and 17.5 for the pre-Tx and reconstitution periods, across all clusters, respectively; Fig. 3). The proportion of the microbiome accounted for by fungal taxa (mycobiome), however, was substantially increased during leukopenia (median per sample: 1.32% of reads, compared to 0.65 and 0.54 for pre-Tx and reconstitution, respectively; Fig. 2); the main detected fungal genera were *Candida albicans* (2.06% mean absolute abundance across leukopenia samples) and *Saccharomyces cerevisiae* (1.97% mean absolute abundance across leukopenia samples).

Changes in microbiome composition compared to the pre-Tx period were also apparent at the level of individual bacterial taxa; *Bacteroides uniformis*, *Enterococcus faecium*, *Phocaeicola vulgatus*, and *Escherichia coli* were detected at abundances ≥ 5% in 4.76%, 47.62%, 9.52%, and 28.58% of samples during leukopenia, respectively, compared to 40.63%, 28.13%, 34.38%, and 18.75% for pre-Tx samples (Supplementary Table 5 and Figs. 4 and 5). Of note, the observed overall expansion of *Enterococcus faecium* during leukopenia was driven by samples from patients assigned to pre-Tx Cluster 1; patients assigned to the *Enterococcus*-associated pre-Tx Cluster 3, by contrast, generally showed a reduction in mean *Enterococcus faecium* abundance (44.99% pre-Tx, 11.30% leukopenia). Interestingly, changes at the level of individual species similar to bacteria were also detected for specific fungi: *Saccharomyces cerevisiae* was detected at an abundance of ≥ 5% in 53.13% of pre-Tx samples and 71.43% of leukopenia samples showing a slight increase. A decrease can be observed for uncultured crAssphage (18.75% of pre-Tx samples vs. 9.52% of leukopenia samples) (Supplementary Table 6, Supplementary Table 7, Supplementary Table 15).

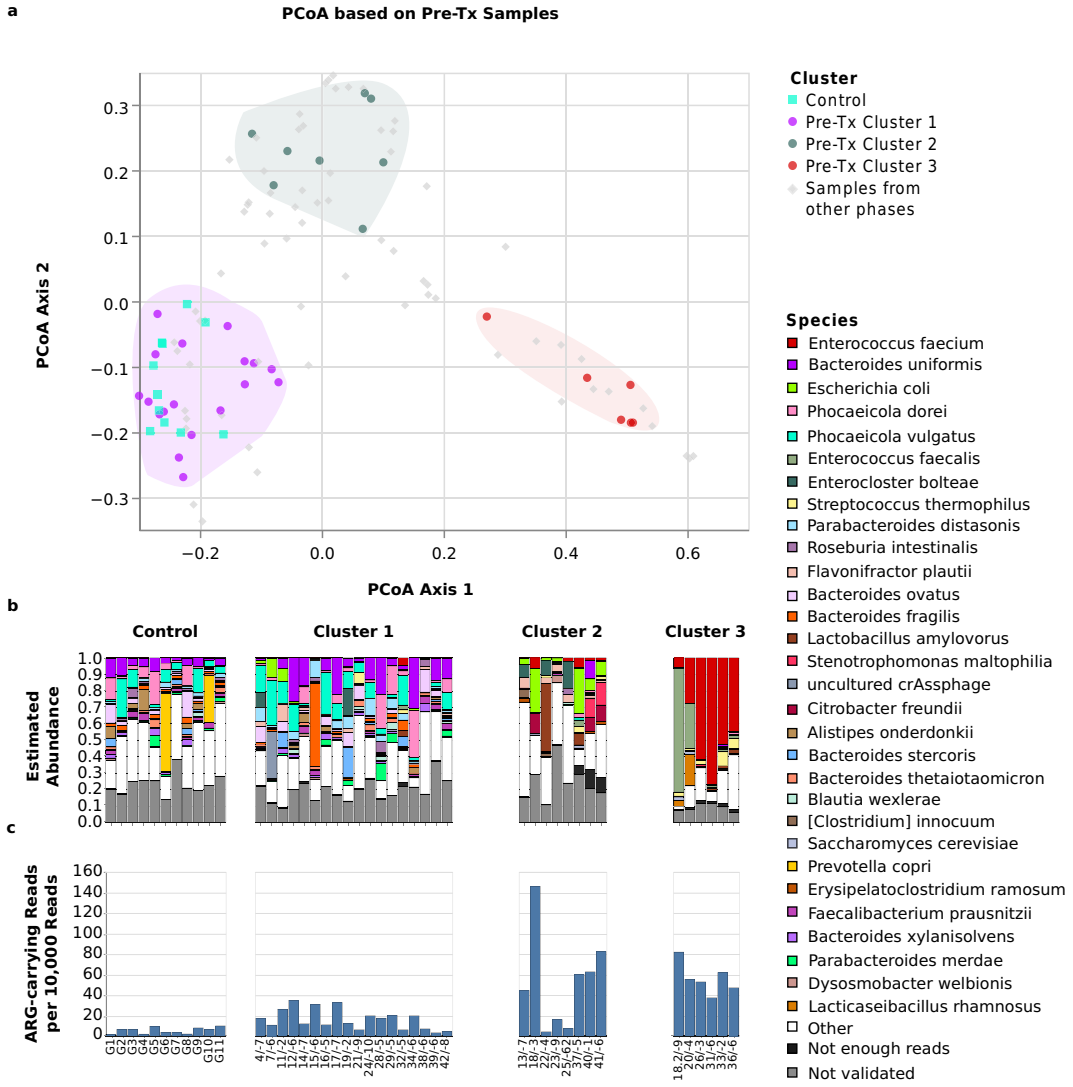


Figure 4. Microbiome structures and compositions of pre-Tx alloHSCT samples and healthy controls. The Figure shows (**a**) the positions of pre-Tx and control samples, displayed as colored dots, in the joint PCoA space of all samples, as well as the positions of pre-Tx Cluster 1, Cluster 2, and Cluster 3 in PCoA space (shaded areas), (**b**) bar plots visualizing the microbiome compositions of healthy control and pre-Tx alloHSCT samples, stratified by pre-Tx cluster membership, and (**c**) the rate of reads that carry an ARG element, separately for control and pre-Tx alloHSCT samples and stratified by pre-Tx cluster membership. The bar plots show the 30 species that attained the highest aggregated sum of frequency across all time points, and, of these, within each sample, only the species that 1.) were assigned at least 5 reads to and 2.) passed the mapping-based taxon validation step were depicted. The combined abundances of species with fewer than 5 reads is shown in the category "Not enough reads"; the category "Not validated" shows the combined abundances of species with more than 5 Kraken2-assigned reads that did not pass the mapping-based taxon validation step. Sample labels below the bar plots specify patient or control sample ID, followed, for alloHSCT samples, by the day of sampling relative to the stem cell transplantation time point.

The reconstitution microbiomes were characterized by increased similarity to the pre-Tx microbiomes

Compared to samples collected during leukopenia, samples during the reconstitution period collectively showed similarity to pre-Tx microbiome samples. During the leukopenic period, approximately 52% (11/21) of collected samples were found to be within one of the 3 cluster spaces defined based on pre-Tx samples; during reconstitution, this fraction increased to 83% (40/48; Fig. 6). In addition, reconstitution microbiomes also resembled pre-Tx microbiomes at the level of observed overall abundances of bacteria, fungi, and DNA viruses (Fig. 3); at the level of bacterial species detected at ≥ 5% abundance (Supplementary Table 5); and at the level of the fungal and DNA viral species that exhibited the highest rates of validated detection at ≥ 5% relative abundance, i.e. *Saccharomyces cerevisiae* and *Candida albicans* for fungi, as well as as *Skunavirus* and crAssphages for DNA viruses (Supplementary Table 6, Supplementary Table 7, Fig. 7).

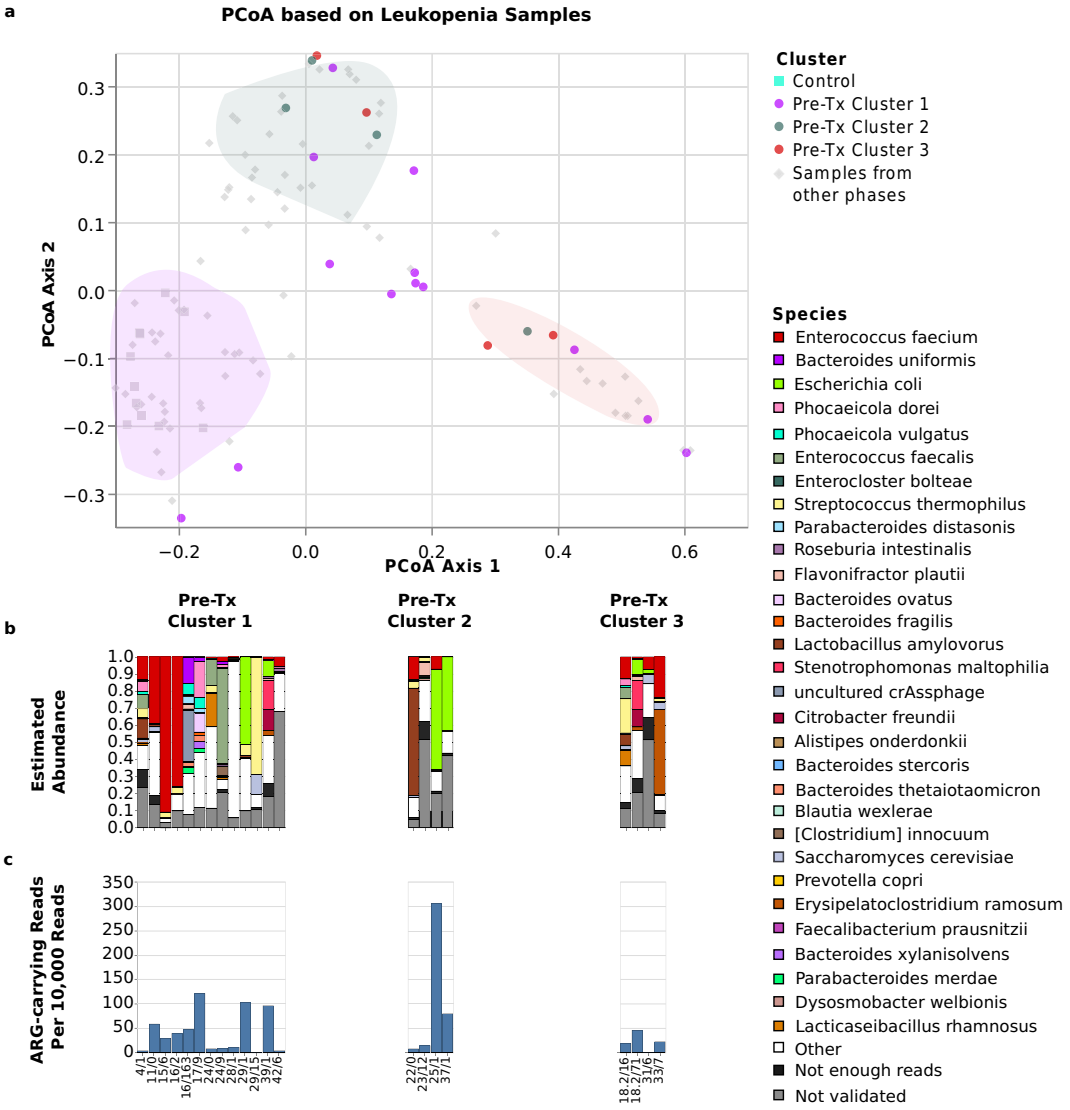


Figure 5. Microbiome structures and compositions during leukopenia. The Figure shows, for alloHSCT samples collected during the leukopenic period, (**a**) sample positions in the joint PCoA space of all samples, (**b**) bar plots visualizing sample microbiome compositions, and (**c**) rates of reads carrying an ARG element. All visualizations are stratified by pre-Tx cluster membership of the corresponding patients; in the top panel, pre-Tx cluster membership is indicated by dot color. The bar plots show the 30 species that attained the highest mean frequency across all time points, and, of these, within each sample, only the species that 1.) were assigned at least 5 reads and 2.) passed the mapping-based taxon validation step were depicted. The combined abundance of species with fewer than 5 reads is shown in the category "Not enough reads"; the category "Not validated" shows the combined abundance of species with more than 5 Kraken2-assigned reads that did not pass the mapping-based validation step. Sample labels below the bar plots specify patient ID, followed by the day of sampling relative to the stem cell transplantation time point.

Bacterial species exhibiting notable differences compared to the pre-Tx period included *Roseburia intestinalis*, which was detected in approximately twice as many samples at abundances ≥ 5% during reconstitution than during pre-Tx (13% compared to 6%), and which accounted for ≥ 50% of overall microbial abundance in individual reconstitution samples; *Parabacteroides distasonis*, which exhibited dominance (≥ 50% abundance) in one reconstitution sample (Supplementary Table 14), but overall decreased detection at ≥ 5% abundance (10% during reconstitution samples compared to 25% for pre-Tx samples); and *Phocaeicola vulgatus* and *Bacteroides uniformis*, which also exhibited decreased detection rates at the ≥ 5% abundance threshold (15% and 17% for reconstitution samples compared to 34% and 41% for pre-Tx samples, respectively; Supplementary Table 5). While similar differences were not detected for fungal species Supplementary Table 6), a few viral species exhibited exclusive detection at > 5% validated relative abundance in reconstitution samples compared to pre-Tx samples (Supplementary Table 7); while these detections were typically limited to individual samples, in some

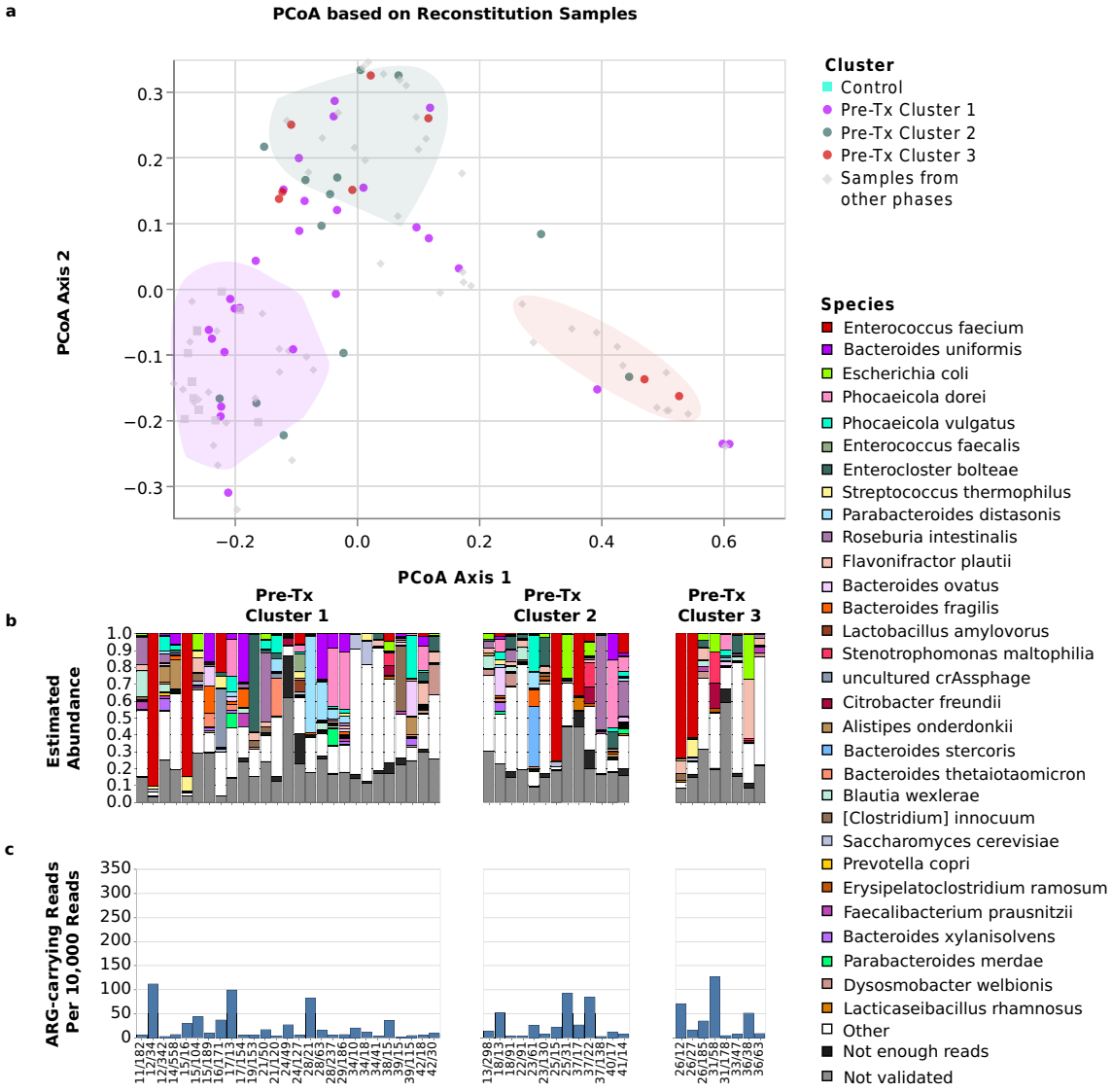


Figure 6. Microbiome structures and composition during reconstitution. The figure shows, for alloHSCT samples collected during the reconstitution period, (**a**) sample positions in the joint PCoA space of all samples, (**b**) bar plots visualizing sample microbiome compositions, and (**c**) rates of reads carrying an ARG element. All visualizations are stratified by pre-Tx cluster membership of the corresponding patients; in the top panel, pre-Tx cluster membership is indicated by dot color. The bar plots show the 30 species that attained the highest mean frequency across all time points, and, of these, within each sample, only the species that 1.) were assigned at least 5 reads and 2.) passed the mapping-based taxon validation step were depicted. The combined abundance of species with fewer than 5 reads is shown in the category "Not enough reads"; the category "Not validated" shows the combined abundance of species with more than 5 Kraken2-assigned reads that did not pass the mapping-based validation step. Sample labels below the bar plots specify patient ID, followed by the day of sampling relative to the stem cell transplantation time point.

instances the underlying viral species dominated the viral microbiome component (observed e.g., in the case of *Betapolyomavirus hominis* and *Afonbuvirus faecalis*; Fig. 7).

Pre-Tx clusters were not associated with longitudinal microbiome trajectories or clinical outcomes

Pre-Tx cluster assignment did not predict the cluster membership of samples from the same patient over the course of alloHSCT, neither during the leukopenic nor during the reconstitution period (Figs. 5, 6); that is, pre-Tx cluster membership was not associated with the microbiome trajectory at the level of individual patients. Similarly, the pre-Tx detection of *Candida albicans* or uncultured crAssphage was not associated with the detection of these genera during leukocytopenia or reconstitution (Fig. 7, Supplementary Fig. 7).

Furthermore, while the proportion of patients with adverse outcomes (relapse and death) or GvHD varied between the 3 clusters (44%, 25% and 67% for adverse outcomes in Cluster 1, 2 and 3, respectively; 11%, 12%,

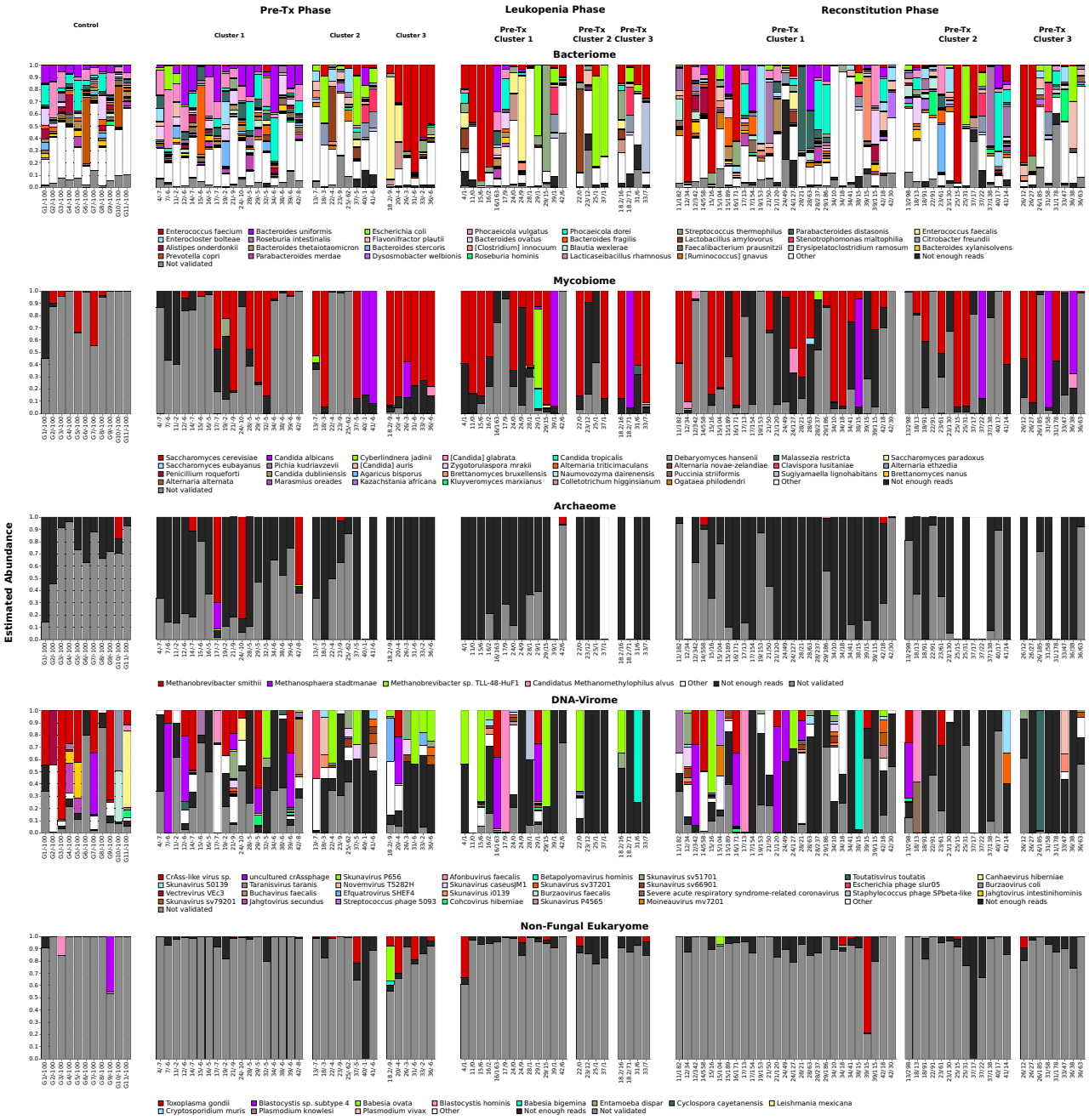


Figure 7. Integrated microbiome analysis of all characterized samples. Shown are, for each sample, bar plots that visualize the respective compositions of the bacterial, fungal, archaeal, DNA-viral, and non-fungal eukaryotic components of the sample microbiome. Normalization was applied independently to each bar plot. The bar plots only show the 30 (for bacteria, fungi and DNA viruses), 4 (for archaea), and 11 (for non-fungal eukaryotes) species that attained the highest mean relative frequency within the considered taxonomic category across all timepoints, and, of these, within each sample, only the species that (**a**) were assigned at least 5 reads by Kraken2, and (**b**) passed the mapping-based taxon validation step. The combined abundance of species within each taxonomic category with fewer than 5 reads is shown in the category "not enough reads"; the category "not validated" shows the combined abundance of species with more than 5 reads that did not pass the mapping-based validation step. Sample labels below the bar plots specify healthy control ID or the patient ID, followed by the day of sampling relative to the transplantation event.

and 17% for GvHD), these differences were not statistically significant ($p = 0.297$ for serious outcomes; $p = 0.938$ for GvHD; Pearsons' chi-squared test). Further, we did not observe an association between cluster membership and the type of hematological disease patients were treated for (Supplementary Table 8).

Analyses of antibiotic resistance, viral strain diversity and the tracking of individual marker taxa

We further investigated whether the whole-genome microbiome sequencing component could enable inferences about the presence of antibiotic resistance genes (ARGs), viral strain diversity, and the abundances of individual bacterial, fungal, or DNA-viral marker taxa which were described in previous publications (Supplementary Table 9. Literature List Marker taxa) or which were of particular interest. First, we investigated the detection of ARGs by quantifying the proportion of sequencing reads that aligned against known antibiotic resistance genes; the observed rates of ARG-carrying reads (Fig. 4) were relatively similar for pre-Tx and leukopenia alloHSCT samples (median = 20.52 ARG-carrying reads per 10.000 reads for pre-Tx samples; 23.00 for leukopenia samples), but substantially lower in the reconstitution (median = 13.39) and control (median = 7.01) samples. Of note, we observed lower rates of ARG-carrying reads in pre-Tx Cluster 1 samples than in Cluster 2 and Cluster 3 samples (Fig. 4; $p = 0.00102$); consistent with the observation that pre-Tx Clusters 2 and 3 contained microbiomes which were more divergent from the characterized control samples, possibly due to prior exposure to anti-infective medications. Next, prompted by the observation that crAssphages accounted for substantial proportions of total microbiome content in individual samples and by the availability of a large number of resolved crAssphage strain genomes from a recent publication[48], we investigated the detectability and relative abundances of different crAssphage strains using a mapping-based approach. This analysis showed that (i) mapping against a database containing only crAssphage genomes resulted in substantially higher estimated crAssphage abundances in many samples than classifying against the comprehensive MetaGut database (Supplementary Fig. 4), (ii) read mapping enabled sensitive detection and differentiation between different crAssphage strains (Supplementary Fig. 5), and (iii) many reads classified as crAssphage in the mapping-based analysis were assigned to other taxa by the Kraken2-based read assignment process (Supplementary Fig. 6). Of note, while these results confirmed the applicability of the generated long-read data to crAssphage-focused analyses, they also suggested that substantial crAssphage strain diversity is currently not represented in the comprehensive MetaGut database. Third, we investigated the longitudinal changes of the abundances of predefined marker taxa. To this end, we assembled a list of taxa reported to be associated with alloHSCT outcomes from the literature (Methods); and, to extend this analysis we added selected fungal and viral taxa accounting for substantial proportions of microbiome content (Methods, Supplementary Table 9). We then carried out a longitudinal analysis of the presence and abundance of the selected marker taxa over time (Supplementary Fig. 7), and found that the generated abundance estimates were informative for changes over time, based on largely non-overlapping abundance confidence intervals between time points, demonstrating, for example, fluctuations in the abundances of *Akkermansia muciniphilia* and *Blautia*.

Investigation of bacterial strain dynamics showed replacement of strains over the course of alloHSCT in some individuals

Finally, we investigated whether the significant impact of myeloablation, antiinfective therapy, and alloHSCT on the gut microbiome was associated with bacterial strain replacement, potentially indicating re-colonization of ecological niches in the microbiome. We developed a method to approximate species-level core genome average identities across samples based on short-read sequencing data, generated Illumina data for a subset of our samples with sufficient DNA yield (Methods), and applied the developed method for the detection of strain replacement events. Based on the generated data, we could characterize longitudinal strain dynamics in 31 cases, representing 6 bacterial species and 14 patients (Fig. 8). Of 44 considered intervals, 12 were classified as representing a strain-switching event; of these, 8 spanned an alloHSCT and 2 a relapse event; compared to 19 of the 29 alloHSCT or relapse intervals not representing a strain-switching event. While the rate of strain-switching events was thus increased for intervals spanning an alloHSCT or relapse event, the observed difference was not found to be statistically significant ($p = 0.17$; Fisher's exact test). This investigation demonstrates that individual strains can be replaced by other strains of the same species during the course of alloHSCT.

Discussion

Associations between the microbiome and alloHSCT outcomes are incompletely understood, in particular with respect to the role of non-bacterial domains of microbial life; studying these, however, is complicated by limitations of established technologies for microbiome characterization, such as 16S or 18S rDNA sequencing. We thus developed a method based on long-read shotgun metagenomics enabling interrogation of alloHSCT patient microbiomes across almost all domains of microbial life (bacteriome, archaeome, mycobiome, non-human/non-fungal eukaryome, and DNA-virome). It comprises robust protocols for sample preparation and DNA extraction as well as a specifically developed mapping-based bioinformatics approach to reduce the false-positive taxon detection rate associated with k-mer-based read classification.

We used the method to interrogate patient microbiomes in an explorative clinical study, and found that pre-Tx patient microbiomes fell into 3 clusters (see Fig. 4), which can be categorized as "*Bacteroides*- and *Phocaeicola*-dominated" (Cluster 1), "heterogeneous" (Cluster 2), and "*Enterococcus*-dominated" (Cluster 3). The included control samples clustered with the pre-Tx alloHSCT samples of Cluster 1. Interestingly, Cluster 1 exhibited high abundances of *Bacteroides*, often considered an important component of an "intact gut microbiome" and comprising many species known to be commensals and beneficial for human gut health[50]. Cluster 1 could thus be interpreted as least affected by alloHSCT-related microbiome dysbiosis; Clusters 2 and 3, by contrast, may be interpreted as more dysbiotic microbiome states[5]. Suggestive differences between the clusters were also detected at the level of viral and fungal microbiome components; for example, *Candida albicans* (of which an increase in HSCT is associated with adverse outcomes[28]) was detected more often in Cluster 2 and 3 samples , and *uncultured* crAssphage more often in Cluster 1 samples. Consistent with the interpretation of Cluster 1 as least dysbiotic, Cluster 1 also exhibited the lowest rates of antibiotic resistance genes, and patients in Cluster

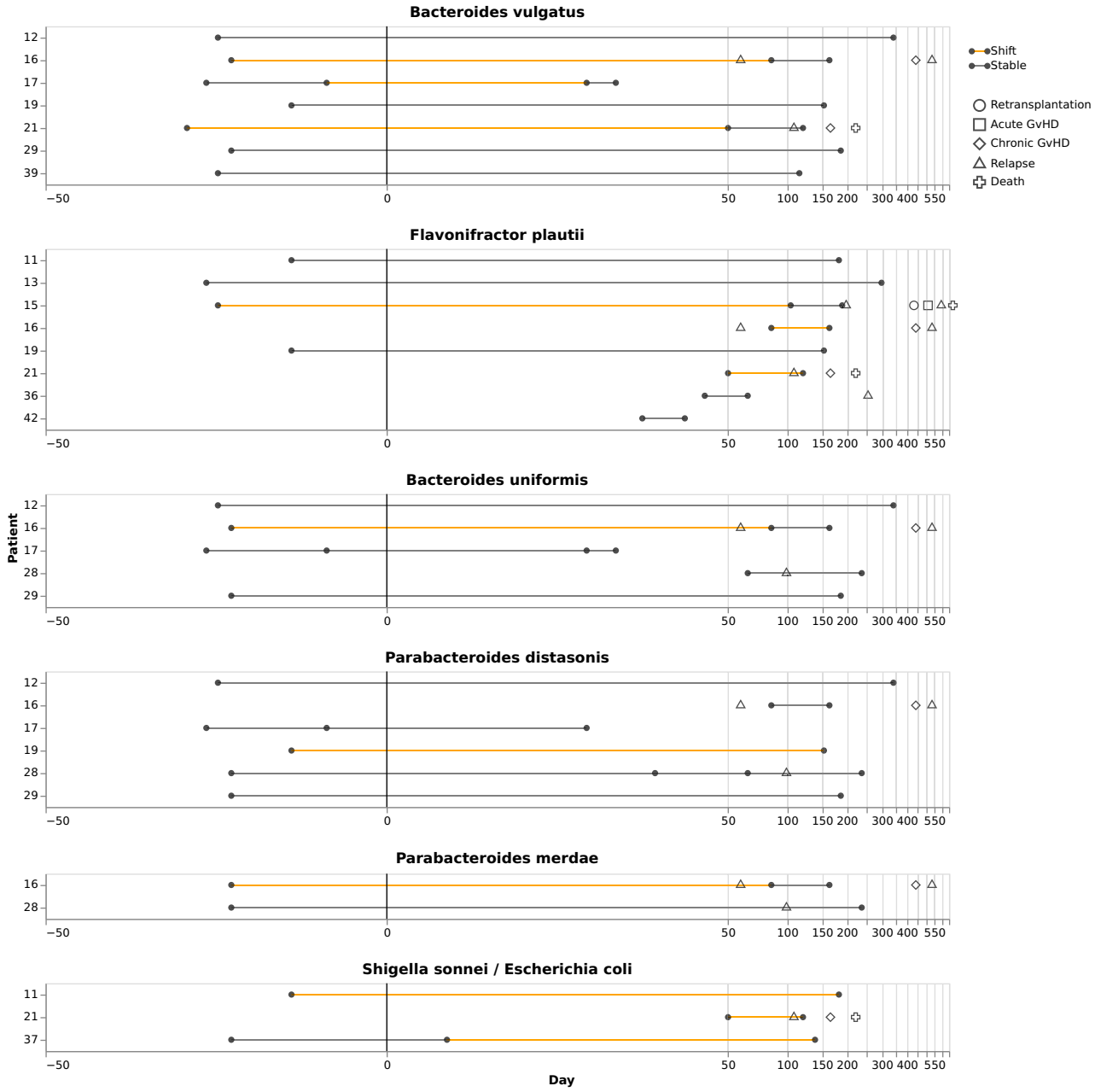


Figure 8. Strain replacement analysis. The Figure visualizes strain dynamics for combinations of bacterial species and sampling timepoints for which sufficient Illumina short-read data was available. Each species is represented by a main panel; main panels are labeled with patient IDs on the y-axis and sampling times on the x-axis. Each dot represents one sampling event, and each horizontal line represents the period between two sampling events. Horizontal lines are colored according to whether a strain replacement event was determined to have taken place between the two sampling events (orange) enclosing the line. Additional symbols indicate adverse events.

1 had a lower median number of pre-alloHSCT treatment cycles (median = 2 treatment cycles) compared to patients in Clusters 2 and 3 (combined median = 4.5 treatment cycles; $p = 0.01$; Mann–Whitney U test). A meta study from 2020 previously showed a significant relation between antibiotics administration and acute GvHD[51]. Future studies are required to investigate whether pre-Tx cluster membership is predictive of alloHSCT treatment success since the statistical power necessary to reliably detect such associations cannot be achieved with the 31 patients recruited for our study.

Further applying the pipeline to longitudinally collected samples, we found that patient microbiomes developed dynamically over the course of alloHSCT. Consistent with previous studies[12,31], we observed a reduction in microbiome diversity during leukopenia and partial recovery during the reconstitution period. Of note, the leukopenic period was also characterized by a marked increase in fungal abundances (see Fig. 3), and expansion of bacterial genera like *Enterococcus* (also observed in[5]) and *Streptococcus*. It should also be noted that the

acquisition of samples during leukopenia was challenging and sometimes outright not possible. Therefore the analyzed samples throughout this period might represent a biased selection. During the reconstitution period, patient microbiomes collectively showed a return in similarity to pre-Tx states; at the level of individual patient trajectories, however, pre-Tx cluster membership was not predictive of microbiome composition during leukocytopenia (see Fig. 5) or reconstitution (see Fig. 6). The analyzed reconstitution samples exhibited significant heterogeneity with respect to the time point of sampling (Supplementary Table 3), and individual patients were also represented with multiple samples in this analysis (Supplementary Table 3). It may also reflect the heterogeneity of alloHSCT patient journeys and the significant effects of these on the microbiome[33]. Consistently, the enriched incidence of bacterial strain replacement around stem cell transplantation and relapse events further emphasizes the potential microbiome disruption in treated patients.

Intriguingly, our study also confirmed the relevance of non-bacterial taxa in the context of alloHSCT. First, we demonstrated robust detection of non-bacterial microbiome components in alloHSCT microbiomes, including, in particular, fungal (*Saccharomyces and Candida*) and viral (crAssphage) taxa. Archaea and non-fungal eukaryotes were also detected, but at lower relative abundances (see Fig. 3). In this context, the mapping-based validation component was instrumental in enabling the distinction between confident hits and likely false-positives. Second, while the absolute proportion of non-bacterial taxa was small in most sampled patient microbiomes, viruses and fungi accounted for more than 50% and 10% of total microbial reads in individual microbiomes, respectively. Third, varying overall proportions of non-bacterial abundances, differences between the investigated treatment phases with respect to the validated detection of species at ≥ 5% abundance (in the case of viruses), as well as individual microbiome trajectories indicated that the non-bacterial components of the microbiome also exhibited high inter-patient and temporal plasticity. Interestingly, *Saccharomyces cerevisiae* and *Candida albicans* were detected at high relative abundances throughout the course of alloHSCT, potentially indicating relative stability of these microbiome components over the course of alloHSCT. Of note, in contrast to earlier studies based on 16S or 18S rDNA sequencing, employing an unbiased metagenomic approach, our workflow enabled the unbiased measurement of the relative quantitative abundances of different types of microbial life. To the best of our knowledge, our study is one of the first few studies to employ shotgun metagenomics in the context of alloHSCT[5,30,52], and, of these, the first to explicitly interrogate the mycobiome and DNA virome. Further studies are required to better characterize the relative stability of the fungal and viral microbiome components and the interplay between these and the bacterial microbiome, for instance at the level of bacteriophage-host relationships[53].

With respect to our pipeline, potential directions for future work include the further improvement of sample handling and DNA extraction protocols, potentially focused on extracting high-molecular weight DNA while retaining robustness, as well as improvements to the bioinformatics analysis components. In the current implementation, significant proportions of reads remained unclassified (Fig. 2); furthermore, the observed mapping-based validation rates for archaea, non-fungal eukaryotes, and (although to a lesser extent) fungi suggested substantial rates of residual mis-classification within these taxonomic groupings (Supplementary Fig. 3). Classification accuracy may benefit from incorporating gut-specific reference databases[54,55] or the usage of high-quality environment specific databases used in a taxonomy-agnostic way[56]. Moreover, the currently employed validation approach, based on using the proportion of reads with high-quality pairwise alignments as a proxy for distinguishing between likely true- and false-positive taxon detections, could be complemented with approaches based on horizontal genome coverage (which we explored with promising results for a likely false-positive eukaryotic hit, *Toxoplasma*; Supplementary Note 2), or analysis of k-mer distributions[57]. It is noteworthy that significant fractions of reads classified as human by Kraken2 did not validate in some samples (Supplementary Fig. 3). While we did not attempt an in-depth investigation, an improved understanding of this phenomenon may contribute to further reducing the rate of false-positive read assignments. An important feature of the current version of our workflow is that the functional analysis of microbiome samples covers the detection of ARGs; future versions could also incorporate analyses of metabolic pathways[58] and virulence factors[52]. Now, after successfully having established a workflow, we can attempt to map established microbiome biomarkers associated with alloHSCT outcomes—such as the quantitative definition of "dominance" by[31] as ≥ 30% abundance—onto the microbiome measurements produced by our method in future studies with sufficient patient numbers. As a prerequisite, a future calibration study should address factors such as the utilized extraction protocol, DNA sequencing technologies, and downstream bioinformatics, which influence the accuracy of microbiome measurements[59]). Importantly, in contrast to rDNA amplification-based approaches, our method enables the simultaneous, unbiased and quantitative measurement of human-derived DNA and the exploration of its biomarker potential, e.g., with respect to potential associations between human-derived DNA and mucosal integrity during leukopenia. Conceivably, such analyses could also include cell-of-origin analyses, leveraging the methylation detection capability of the Oxford Nanopore technology[60].

In summary, we have presented a method for characterizing microbiomes simultaneously for their bacteriome, archaeome, mycobiome, non-human/non-fungal eukaryome, and DNA virome composition. Applying it in an exploratory study to a cohort of alloHSCT patients, we carried out one of the first investigations of an all-kingdom microbiome in the context of alloHSCT based on shotgun metagenomics, confirmed the potential relevance of non-bacterial microbiome components, and identified a 3-cluster structure of pre-Tx patient microbiomes. Furthermore, we could demonstrate that microbial strains can be newly acquired or replaced during the course of alloHCST. These findings can serve as an important basis for future studies, aiming at a more comprehensive characterization of the role of the microbiome in alloHSCT-patients based on shotgun metagenomics.

Materials and methods

Patient recruitment and sampling scheme

Patient recruitment for collection of clinical alloHSCT samples was carried out between June 2018 and December 2019. Participation in the study was offered to all patients undergoing alloHSCT at the Department of Hematology, Immunology, and Clinical Immunology at University Hospital Düsseldorf, according to the following inclusion criteria: persons of legal age that are about to receive an allogeneic stem cell transplantation at the UKD and are able to consent. A full description of the recruited cohort is given in Supplementary Note 1; patient characteristics and treatment histories are summarized in Supplementary Table 2 and Supplementary Table 8. Fecal samples from alloHSCT patients were collected prior to transplantation, during the leukopenic period (defined as white blood count of ≤ 1,000/µL), and after reconstitution. For some patients more than one sample per time point was collected which is shown in Supplementary Fig. 11. Stool samples were collected in standard stool collection tubes (feces tubes 76 × 20 mm; Sarstedt), stored at 4 °C if possible and transported for further processing to the Institute for Medical Microbiology and Hospital Hygiene of Heinrich Heine University Düsseldorf as fast as possible.

A convenience control cohort of 10 individuals not diagnosed with hematological malignancies was recruited from employees of Düsseldorf University Hospital between February 2021 and June 2021. Control samples were collected from each individual after inclusion into the study and processed using the same protocols also applied to the alloHSCT samples.

This study was approved by the ethics committee of the Medical Faculty of Heinrich Heine University Düsseldorf (2019–509, Registration Number 2019055096) and all methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects.

Sample preparation and sequencing

Upon receipt, 1–4 aliquots from each stool sample of approximately 1 mL or 1 g of specimen each were placed into 1–4 labeled PowerBead Tubes containing 750 µL C1 buffer solution (Qiagen) each.

DNA extraction was carried out using a modified version of the HMP protocol[45]. Briefly, specimens were homogenized 30–40 s by vortexing and placed into a heating block at 65 °C and 95 °C for 10 min respectively. Samples were then stored at − 80 °C until DNA extraction. Aliquots for DNA extraction were thawed at 4 °C or room temperature (RT) and gently mixed by vortexing. 60 µL of C1 solution was added and briefly vortexed or inverted several times. Aliquots were vortexed at maximum speed for 10 min and centrifuged at 10,000 × g for 30 s at RT. From this step onwards, the DNeasy Power Soil Kit was used as per protocol. The concentration of the extracted DNA was determined by fluorometry using the Invitrogen Qubit 4 Fluorometer (Thermo Fisher Scientific) and aliquots were stored at -80 °C until the Nanopore and Illumina sequencing was performed.

Long-read Nanopore sequencing of all samples was carried out on the PromethION and MinION devices, using FLO-MIN106 R9-Version and FLO-PRO002 R9.4-Version flow cells, the SQK-LSK109 ligation sequencing kit, and the EXP-NBD104 or EXP-NBD114 barcoding kits for barcoded sequencing of native DNA, according to protocol NBE_9065_v109_revAA_14Aug2019. Basecalling was carried out using Guppy.

83 genomic DNA samples, selected to cover at least one pre-post pair per patient, were used for short-read Illumina sequencing. The samples were quantified by fluorometric assay (Qubit DNA HS Assay, Thermo Fisher Scientific) and their quality measured by capillary electrophoresis using the Fragment Analyzer and the 'DNF-488 HS Genomic DNA Kit' (Agilent Technologies, Inc.). Although the initial concentration of some samples was low, with some of them not measurable at all, we processed all of them, since the library preparation kit is also suitable for small input quantities. Library preparation was performed using the 'Nextera XT DNA Library Preparation Kit—Document # 15,031,942 v05 (Illumina, San Diego, USA.). Depending on whether the concentration of the sample could be determined, either 1 ng gDNA or 5 µL of the sample volume was used for the tagmentation step. Library amplification for final enrichment was performed with 13 cycles. Bead purified libraries were normalized and subsequently sequenced on the HiSeq3000 system (Illumina, San Diego, USA) with a read setup of 2 × 151 bp. The bcl2fastq2 tool was used to convert the bcl files to fastq files as well for adapter trimming and demultiplexing.

DNA extraction and sequencing data generation metrics are summarized in Supplementary Table 3.

Analysis of long-read sequencing data

Long-read sequencing data were analyzed using a comprehensive reference database ("MetaGut v. 1.0 database") comprising 292 giga base pairs of sequence and 1.5 million microbial genomes from all domains of microbial life (Supplementary Table 10). We constructed a custom Kraken2 database by using the Kraken2 build scripts for all domains, modified to download all assemblies marked as "representative genomes" (as opposed to just "Chromosome" or "Full Genome" assemblies) for fungi and protozoa. In addition, we included the *Mus musculus* reference genome (GCF_000001635.27_GRCm39). A full breakdown of database contents is shown in Supplementary File 1. Read classification and compositional analysis were carried out using a 2-step approach that combined an initial k-mer-based read assignment step with mapping-based validation. During the first step, all reads of a sample were classified using Kraken2[47], yielding an initial estimate of sample composition. During the second step, the presence of species (and genera) detected during the first step was validated using minimap2[61].

Briefly, to validate the presence of a species in a sample corresponding to node *n* in the reference database taxonomy, the following steps were carried out: (i) All sample reads assigned to node *n* and its descendants were collected; (ii) a combined reference genome for node *n* was created by linearly concatenating the reference genomes of all leaf-level descendants of node *n*; (iii) the collected reads for node *n* were mapped against the combined reference genome for node *n*; (iv) the proportion of mapping-validated reads was determined, where "mapping-validated" was defined as ≥ 70% of a read being covered by read-to-reference alignments with ≥ 70% identity; (v) finally, the species corresponding to node *n* was defined as "validated" in a sample if ≥ 20% of the

sample reads assigned to *n* achieved mapping-based validation. To increase computational efficiency, validation was carried out on a random subsample of 100,000 reads in each sample (or on the full sample if it contained fewer reads).

Of note, due to low per-species validation rates of archaea and non-fungal eukaryotes (see Results), analyses of the occurrence and abundance of individual species were limited to the bacterial, viral and fungal components of the microbiome.

Determination of read validation thresholds

The utilized read validation thresholds were defined empirically and based on an analysis of the Zymo Gut Microbiome Standard (https://www.zymoresearch.de/products/zymobiomics-gut-microbiome-standard). Briefly, the community standard was long-read-sequenced, employing the same protocols used for fecal samples. The generated reads were classified against the MetaGut database using Kraken2, and the assignments of individual reads were validated, using the 70%/70% criterion defined in the previous section, by mapping against (a) a database constructed from the Zymo-provided reference genomes of the species in the Zymo standard, and (b) the MetaGut reference database.

The analysis could not be conducted for the species *Veillonella rogosae* and *Prevotella corporis* since they were not present in the database.

When using the the Zymo-provided reference genomes for read validation (i.e. representing the assumed case that the genomes of all taxa present in the sample are perfectly represented in the reference database; but see below), per-genus read validation rates varied between 82.4% and 99.3%, with the exception of *Candida albicans, Salmonella enterica*, *Clostridium perfringens* and *Enterococcus faecalis* (Supplementary Table 11). 0 of the 2 and 3 of the 5 reads assigned to *Clostridium perfringens* and *Enterococcus faecalis*, respectively, could be validated, indicating that these represent false-positives; this is consistent with an expected absolute read count (relative theoretical abundance x number of generated reads) for these species of 1 and 0, respectively. *Candida albicans and Salmonella enterica* exhibited read validation rates of 50.2% and 69.1%; we hypothesized that these could be explained by a mismatch between the *Candida* and *Salmonella* genome present in the sequenced Zymo sample and the Zymo-provided reference genome. We tested and confirmed this hypothesis by de novo assembly of the Zymo sequencing data, followed by mapping of the assembled contigs against the database constructed from the Zymo-provided reference genomes using FastANI[62], weighted by assembled contig length; in this analysis, *Candida albicans* appeared as a clear outlier (FastANI estimates of 95.65%; Supplementary Table 12), indicating genetic divergence. No matching contigs were assembled for *Salmonella enterica*.

When using our comprehensive reference database for read validation (i.e. representing the case of potential divergence between the genomes in the sample and the next-closest database genomes) per-species read validation rates varied between 81 and 100% for species truly present in the Zymo standard (excluding *Enterococcus* and *Clostridium* due to low abundance), and between 0 and 100% for false-positive species (Supplementary Table 11). We hypothesized that per-taxon read validation rates would vary with the genetic distance between the sequenced and next-closest database genomes and confirmed this hypothesis by carrying out a correlation analysis (Pearson's $r = 0.484$ for the correlation between read validation rate and FastANI estimate for all species excluding *Enterococcus faecalis* and *Clostridium perfringens*; $p = 0.094$).

Based on these results, we decided that pragmatically employing a per-taxon read validation rate threshold of 20% would conservatively enable us to filter out false-positive classification results even in the presence of genetic divergence between the genomes present in the sequenced samples and the genomes present in our database.

Benchmarking of read validation method

To validate the validation method we generated a ground truth set consisting of reads and high-confidence assignments to references. We used the Zymo Gut Microbiome Standard and constructed a database by combining the Zymo-provided reference genomes with 185 additional contigs generated by the above mentioned de novo assembly (using Flye 2.9.1-b1780)[63]. To select suitable contigs we calculated the FastANI distances to the Zymo-provided references and selected references that had a) a similarity of 90% or higher to any reference b) a decrease in at least 1% to the next fitting reference and c) a length of at least 10,000 base pairs. The best match was used to label the contig.

We then mapped the reads exceeding a length of 500 base pairs using minimap2 requiring 80% of the read to align with at least 80% identity. This way we were able to assign 94% of the reads to a species in the Zymo Gut Microbiome Standard.

To quantify precision and recall of our validation we then applied Kraken2 with our full database on the readset (baseline) and performed the validation on species level. In addition, we removed each of the 17 species contained in the Zymo Gut Microbiome Standard from the database (rebuilding the entire index structure every time using the Kraken2 'fast-build' option). Refer to Supplementary Table 13 for a full presentation of the results.

The baseline experiment on species level already detects 1641 taxa i.e. the majority of detected Kraken2 species are false positive. Since many of those are discarded by default due to low read counts and not considered for the validation we focus exclusively on those taxa for which validation is performed.

An ideal classifier would label reads from a species which is no longer present in the database as either "unclassified" or as belonging to a taxon that is an ancestor of the species. However, we observe that the majority of the reads gets reassigned to different orthogonal species producing numerous additional taxa that are not observed in the baseline experiment. Manual inspection reveals that many misassignments are to closely related species or entries that are only distant due to the underlying taxonomy and its labeling (for example *E. Coli* reads getting assigned to species of the genus *Shigella*).

Overall the recall for FP taxa is consistently high, varying between 75 and 99%. The recall for reads is significantly lower, varying between 5 and 10%. We hypothesize that this is due to the already high amount of false positive species assignments in the baseline experiment and the large amount of closely related species contained in the database that remains a challenge for LCA based approaches.

The precision of our filter is constantly 100% across all experiments, thus we never filter any real Zymo species.

"Kitome" and contamination analysis

To rule out a significant contribution of environmental contamination we also performed DNA extraction without stool followed by a sequencing run with three barcodes employing the default wetlab protocol. Here, the amount of DNA was below the detection limit of the Qubit Fluorometric analysis. Within the few obtained sequence reads a small amount of human reads and reads belonging to physiological skin microbiota such as Micrococcus luteus and Cutibacterium acnes could be validated. We thus conclude that the potential impact of environmental / Kitome contamination in our samples can be neglected.

Zymo and LifeLines-DEEP compositional analysis

Zymo community standard long-read sequencing data (see previous section) were mapped to Zymo-provided reference genomes using minimap2 and classified using the default workflow and the MetaGut database. Results were analyzed visually and by correlation analysis against the Zymo-provided expected abundance distribution. Whole-genome short-read sequencing data were obtained for 20 samples randomly selected from the Lifelines-DEEP cohort study (http://www.lifelines.nl), and classified using Kraken2 against the MetaGut v. 1.0 database using an absolute read cutoff of 5 reads and excluding reads classified as Viridiplantae.

Microbiome cluster analysis

To investigate high-level microbiome structures, a PCoA analysis of all samples (alloHSCT and control cohorts), based on the Bray–Curtis distance calculated on the fractional compositional estimates as reported by our pipeline post filtering, was carried out. Microbiome clusters were identified by performing k-means clustering on the PCoA components and scanning k for 2 through 6 for the best silhouette value which was achieved for $k = 3$.

Diversity analysis

Reported numbers of species were measured after downsampling each sample to the smallest number of reads observed in any sample ($n = 1029$ reads). Reads mapping to any taxon within Chordata or Viridiplantae were ignored.

ARG gene detection

Long-read sequencing data were mapped using minimap2 against the CARD database[64], comprising 2614 antibiotic resistance elements. A read was counted as carrying a specific ARG if a pairwise alignment with length ≥ 500 and < 50 undefined "N" characters was detected between the read and the ARG sequence. The ARG gene detection pipeline was applied to all reads from a sample, independent of other pipeline components.

crAssphage analysis

Long-read sequencing data were mapped using minimap2 against resolved crAssphage sequences[48] A read was counted as emanating from a crAssphage genome if a pairwise alignment with $\geq 70\%$ of query cover at 70% identity was detected between the read and a crAssphage reference sequence; the fraction of primary alignments divided by the number of query reads was used as an estimate for crAssphage abundance. To identify reads that could be confidently assigned to individual strains, filtering based on mapping quality ($MQ \geq 5$) was carried out, enabling assignment of 58% of reads. The crAssphage analysis component was applied to all reads from a sample, independent of other pipeline components.

Tracking of marker taxa

A comprehensive list of marker taxa reported to be associated with alloHSCT outcomes was assembled from the literature (Supplementary Table 9), and complemented with selected fungal and bacterial marker taxa detected in this study. The final list of taxa is summarized in Supplementary Table 9. Longitudinal abundance dynamics were assessed visually, with confidence intervals in the plots representing binomial fraction confidence intervals.

Detection of strain replacement events

Analysis of strain dynamics was based on (i) a novel measure ("aANI-lowFreq") robustly approximating average nucleotide identities between the majority bacterial strain of a species present in a sample *a* and bacterial strains of the same species present in another sample *b* from short-read sequencing data and robust even in the presence of low-frequency strains; (ii) empirical calibration of expected aANI-lowFreq values for samples carrying different dominant bacterial strains using short-read data from different patients, based on the assumptions that (a) bacterial species will vary in their aANI-lowFreq distributions and (b) that different patients will typically be colonized by different bacterial strains; (iii) visual assessment of aANI-lowFreq distances for within-patient longitudinal samples from the same patients; a strain replacement event was assumed to have taken place if the observed within-patient aANI-lowFreq distance falls within the observed between-patient distribution. We note that our approach was optimized for reducing the rate of false-positive strain replacement events, potentially trading off sensitivity against precision. We also note that we experimented with MetaSNV 2[65]; however, the subpopulation module failed due to insufficient substructure detection on each species.

aANI-lowFreq is based on counting, over shared regions of species-specific core genomes, the number of bacterial SNVs exclusive to sample *a*, i.e. not also found in sample *b*. As a strain replacement event is most clearly indicated by the presence of a novel majority bacterial strain in *a* that is not present in *b*, *a*-exclusive SNVs were defined as positions that carried an allele with ≥ 50% frequency supported by a minimum of 10 sequencing reads in sample *a* and ≥ 10% frequency in sample *b*; the 10% threshold was chosen to allow for the presence of sequencing errors and mis-aligned reads, and, at the utilized threshold on read coverage (see below), often translated to rejecting alleles that were observed on a single sequencing read on sample *b*. A set of species-specific core genome sequences was obtained from the proGenomes2 representative species level reference database[66]; a position in one of these core genome sequences was defined as "shared" if both the individual position as well the surrounding core genome reference genome contig across ≥ 80% of its positions achieved ≥ 10 × coverage in *a* and *b*. The aANI-lowFreq distance between *a* and *b* was defined as the count of *a*-exclusive SNVs divided by the length of the genomic regions classified as "shared" between *a* and *b*.

aANI-lowFreq was validated using simulation by obtaining 100 randomly selected NCBI RefSeq genomes for *Enterococcus faecium*, *Phocaeicola vulgatus*, *Escherichia coli*, and *Bacteroides uniformis*, as well as 72 genomes for *Lactobacillus gasseri*. For each genome, paired-end 2 × 100 bp short-read sequencing data at different coverage levels (5, 10, 20, 50, 200, and 1000) were simulated using dwgsim 0.1.14[67]. For each species and combination of coverage levels, we generated 50 random pairs of different genomes as well as 15 random pairs of identical genomes; for each pair, the aANI-lowFreq distance was calculated based on the simulated short reads, and a reference ANI (based on the NCBI RefSeq genome sequences) was obtained using FastANI[62]. The simulations demonstrated that from genome-wide coverages of ≥ 20x, substantial core genome proportions were classified as "shared" (Supplementary Fig. 8), and good approximation of FastANI by aANI-lowFreq, at slight levels of ANI underestimation (demonstrating conservativeness of aANI-lowFreq; Supplementary Fig. 9).

For detection of strain replacement events within the alloHSCT cohort, only species and sample pairs with at least 50% of the corresponding core genome were considered, and only species for which at least 10 between-patient aANI-lowFreq distances were available for calibration. For the visual assessment step (Supplementary Fig. 10), within-patient and between-patient aANI-lowFreq distances were plotted alongside each other, together with aANI-lowFreq distances from the RefSeq-based simulations, when available.

Data availability

The sequencing data—filtered for human DNA—is available at SRA under BioProject Identifier: PRJNA929328. Raw sequencing data and additional patient metadata are available upon reasonable request from the corresponding author.

Code availability

The bioinformatics workflow was made available under an Open Source license as a reproducible Snakemake[68] workflow with the DOI: https://doi.org/https://doi.org/10.5281/zenodo.8121116. Each step of the pipeline is executed within its own Conda environment, enabling finely-grained assignment of memory and CPU resources. Downstream analysis was carried out using a set of Jupyter Notebooks[69], also available at the DOI given above.

Received: 23 August 2023; Accepted: 1 February 2024
Published online: 19 February 2024

References

1. Xu, Z.-L. & Huang, X.-J. Optimizing allogeneic grafts in hematopoietic stem cell transplantation. *Stem Cells Transl. Med.* **10**(Suppl 2), S41–S47. https://doi.org/10.1002/sctm.20-0481 (2021).
2. Rafiee, M. *et al.* A concise review on factors influencing the hematopoietic stem cell transplantation main outcomes. *Health Sci. Rep.* **4**(2), e282. https://doi.org/10.1002/hsr2.282 (2021).
3. Kobbe, G., Schroeder, T., Haas, R. & Germing, U. The current and future role of stem cells in myelodysplastic syndrome therapies. *Expert Rev. Hematol.* **11**(5), 411–422. https://doi.org/10.1080/17474086.2018.1452611 (2018).
4. Rautenberg, C. *et al.* Prediction of Response and survival following treatment with azacitidine for relapse of acute myeloid leukemia and myelodysplastic syndromes after allogeneic hematopoietic stem cell transplantation. *Cancers (Basel)* https://doi.org/10.3390/cancers12082255 (2020).
5. Ilett, E. E. *et al.* Associations of the gut microbiome and clinical factors with acute GVHD in allogeneic HSCT recipients. *Blood Adv.* **4**(22), 5797–5809. https://doi.org/10.1182/bloodadvances.2020002677 (2020).
6. Taur, Y. *et al.* The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. *Blood* **124**(7), 1174–1182. https://doi.org/10.1182/blood-2014-02-554725 (2014).
7. Schluter, J. *et al.* The gut microbiota is associated with immune cell dynamics in humans. *Nature* **588**(7837), 303–307. https://doi.org/10.1038/s41586-020-2971-8 (2020).
8. Peled, J. U. *et al.* Microbiota as predictor of mortality in allogeneic hematopoietic-cell transplantation. *N. Engl. J. Med.* **382**(9), 822–834. https://doi.org/10.1056/NEJMoa1900623 (2020).
9. Montassier, E. *et al.* Chemotherapy-driven dysbiosis in the intestinal microbiome. *Aliment. Pharmacol. Ther.* **42**(5), 515–528. https://doi.org/10.1111/apt.13302 (2015).
10. Holler, E. *et al.* Metagenomic analysis of the stool microbiome in patients receiving allogeneic stem cell transplantation: Loss of diversity is associated with use of systemic antibiotics and more pronounced in gastrointestinal graft-versus-host disease. *Biol. Blood Marrow Transplant.* **20**(5), 640–645. https://doi.org/10.1016/j.bbmt.2014.01.030 (2014).
11. Jenq, R. R. *et al.* Intestinal blautia is associated with reduced death from graft-versus-host disease. *Biol. Blood Marrow Transplant.* **21**(8), 1373–1383. https://doi.org/10.1016/j.bbmt.2015.04.016 (2015).
12. Peled, J. U. *et al.* Intestinal microbiota and relapse after hematopoietic-cell transplantation. *J. Clin. Oncol.* **35**(15), 1650–1659. https://doi.org/10.1200/JCO.2016.70.3348 (2017).
13. Liu, X. *et al.* Blautia-a new functional genus with potential probiotic properties?. *Gut Microbes* **13**(1), 1–21. https://doi.org/10.1080/19490976.2021.1875796 (2021).
14. van der Velden, W. J. F. M. *et al.* Role of the mycobiome in human acute graft-versus-host disease. *Biol. Blood Marrow Transplant.* **19**(2), 329–332. https://doi.org/10.1016/j.bbmt.2012.11.008 (2013).

15. Shono, Y. *et al.* Increased GVHD-related mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in human patients and mice. *Sci. Transl. Med.* **8**(339), 339RA71. https://doi.org/10.1126/scitranslmed.aaf2311 (2016).
16. Sen, T. & Thummer, R. P. The impact of human microbiotas in hematopoietic stem cell and organ transplantation. *Front. Immunol.* **13**, 932228. https://doi.org/10.3389/fimmu.2022.932228 (2022).
17. Richard, M. L. & Sokol, H. The gut mycobiota: Insights into analysis, environmental interactions and role in gastrointestinal diseases. *Nat. Rev. Gastroenterol. Hepatol.* **16**(6), 331–345. https://doi.org/10.1038/s41575-019-0121-2 (2019).
18. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**(6285), 565–569. https://doi.org/10.1126/science.aad3369 (2016).
19. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**(7681), 507–511. https://doi.org/10.1038/nature24460 (2017).
20. Kurilshikov, A. *et al.* Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **53**(2), 156–165. https://doi.org/10.1038/s41588-020-00763-1 (2021).
21. Yalcin, S. S. *et al.* Intestinal mycobiota composition and changes in children with thalassemia who underwent allogeneic hematopoietic stem cell transplantation. *Pediatr. Blood Cancer* **69**(1), e29411. https://doi.org/10.1002/pbc.29411 (2022).
22. Borrel, G., Brugère, J.-F., Gribaldo, S., Schmitz, R. A. & Moissl-Eichinger, C. The host-associated archaeome. *Nat. Rev. Microbiol.* **18**(11), 622–636. https://doi.org/10.1038/s41579-020-0407-y (2020).
23. Stockdale, S. R. & Hill, C. Progress and prospects of the healthy human gut virome. *Curr. Opin. Virol.* **51**, 164–171. https://doi.org/10.1016/j.coviro.2021.10.001 (2021).
24. Edgar, R. C. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* **6**, e4652. https://doi.org/10.7717/peerj.4652 (2018).
25. Khachatryan, L. *et al.* Taxonomic classification and abundance estimation using 16S and WGS-A comparison using controlled reference samples. *Forensic Sci. Int. Genet.* **46**, 102257. https://doi.org/10.1016/j.fsigen.2020.102257 (2020).
26. Danziger-Isakov, L. Gastrointestinal infections after transplantation. *Curr. Opin. Gastroenterol.* **30**(1), 40–46. https://doi.org/10.1097/MOG.0000000000000016 (2014).
27. Rauwolf, K. K., Floeth, M., Kerl, K., Schaumburg, F. & Groll, A. H. Toxoplasmosis after allogeneic haematopoietic cell transplantation-disease burden and approaches to diagnosis, prevention and management in adults and children. *Clin. Microbiol. Infect.* **27**(3), 378–388. https://doi.org/10.1016/j.cmi.2020.10.009 (2021).
28. Malard, F. *et al.* Impact of gut fungal and bacterial communities on the outcome of allogeneic hematopoietic cell transplantation. *Mucosal Immunol.* **14**(5), 1127–1132. https://doi.org/10.1038/s41385-021-00429-z (2021).
29. Rolling, T. *et al.* Haematopoietic cell transplantation outcomes are linked to intestinal mycobiota dynamics and an expansion of Candida parapsilosis complex species. *Nat. Microbiol.* **6**(12), 1505–1515. https://doi.org/10.1038/s41564-021-00989-7 (2021).
30. Legoff, J. *et al.* The eukaryotic gut virome in hematopoietic stem cell transplantation: New clues in enteric graft-versus-host disease. *Nat. Med.* **23**(9), 1080–1085. https://doi.org/10.1038/nm.4380 (2017).
31. Taur, Y. *et al.* Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. *Clin. Infect. Dis.* **55**(7), 905–914. https://doi.org/10.1093/cid/cis580 (2012).
32. Vandeputte, D. *et al.* Temporal variability in quantitative human gut microbiome profiles and implications for clinical research. *Nat. Commun.* **12**(1), 6740. https://doi.org/10.1038/s41467-021-27098-7 (2021).
33. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. USA* **108**, 4554–4561. https://doi.org/10.1073/pnas.1000087107 (2011).
34. Le Bastard, Q., Chevallier, P. & Montassier, E. Gut microbiome in allogeneic hematopoietic stem cell transplantation and specific changes associated with acute graft vs host disease. *World J. Gastroenterol.* **27**(45), 7792–7800. https://doi.org/10.3748/wjg.v27.i45.7792 (2021).
35. Silverman, J. D., Shenhav, L., Halperin, E. A., Mukherjee, S. A. & David, L. A. Statistical considerations in the design and analysis of longitudinal microbiome studies. *BioRxiv* https://doi.org/10.1101/448332 (2018).
36. Lind, A. L. & Pollard, K. S. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome* **9**(1), 58. https://doi.org/10.1186/s40168-021-01015-y (2021).
37. "Unable to find information for 6350909." .
38. "Unable to find information for 4928522." .
39. Steinegger, M. & Salzberg, S. L. Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* **21**(1), 115. https://doi.org/10.1186/s13059-020-02023-1 (2020).
40. Lu, J. & Salzberg, S. L. Removing contaminants from databases of draft genomes. *PLoS Comput. Biol.* **14**(6), e1006277. https://doi.org/10.1371/journal.pcbi.1006277 (2018).
41. "Unable to find information for 10528610." .
42. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**(7), 960–970. https://doi.org/10.1038/s41564-021-00928-6 (2021).
43. Dilthey, A. T., Jain, C., Koren, S. & Phillippy, A. M. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat. Commun.* **10**(1), 3066. https://doi.org/10.1038/s41467-019-10934-2 (2019).
44. Meyer, F. *et al.* Critical assessment of metagenome interpretation: The second round of challenges. *Nat. Methods* **19**(4), 429–440. https://doi.org/10.1038/s41592-022-01431-4 (2022).
45. "HMP Protocol." https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/document.cgi?study_id=phs000228.v4.p1&phv=158680&phd=3190&pha=&pht=1184&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1#sec64 (accessed Dec. 05, 2022).
46. Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project. *Nature* **569**(7758), 641–648. https://doi.org/10.1038/s41586-019-1238-8 (2019).
47. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**(1), 257. https://doi.org/10.1186/s13059-019-1891-0 (2019).
48. Gulyaeva, A. *et al.* Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. *Cell Rep.* **38**(2), 110204. https://doi.org/10.1016/j.celrep.2021.110204 (2022).
49. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: Study design and baseline characteristics. *BMJ Open* **5**(8), e006772. https://doi.org/10.1136/bmjopen-2014-006772 (2015).
50. Zafar, H. & Saier, M. H. Gut Bacteroides species in health and disease. *Gut Microbes* **13**(1), 1–20. https://doi.org/10.1080/19490976.2020.1848158 (2021).
51. Gavriilaki, M., Sakellari, I., Anagnostopoulos, A. & Gavriilaki, E. The impact of antibiotic-mediated modification of the intestinal microbiome on outcomes of allogeneic hematopoietic cell transplantation: Systematic review and meta-analysis. *Biol. Blood Marrow Transplant.* **26**(9), 1738–1746. https://doi.org/10.1016/j.bbmt.2020.05.011 (2020).
52. Yan, J. *et al.* A compilation of fecal microbiome shotgun metagenomics from hematopoietic cell transplantation patients. *Sci. Data* **9**(1), 219. https://doi.org/10.1038/s41597-022-01302-9 (2022).
53. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498. https://doi.org/10.1038/ncomms5498 (2014).
54. Leviatan, S., Shoer, S., Rothschild, D., Gorodetski, M. & Segal, E. An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species. *Nat. Commun.* **13**(1), 3863. https://doi.org/10.1038/s41467-022-31502-1 (2022).

55. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**(1), 105–114. https://doi.org/10.1038/s41587-020-0603-3 (2021).
56. Pasolli, E. *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.e20. https://doi.org/10.1016/j.cell.2019.01.001 (2019).
57. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**(1), 198. https://doi.org/10.1186/s13059-018-1568-0 (2018).
58. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**(11), 962–968. https://doi.org/10.1038/s41592-018-0176-y (2018).
59. Sinha, R., Abnet, C. C., White, O., Knight, R. & Huttenhower, C. The microbiome quality control project: Baseline study design and future directions. *Genome Biol.* **16**, 276. https://doi.org/10.1186/s13059-015-0841-8 (2015).
60. Katsman, E. *et al.* Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. *Genome Biol.* **23**(1), 158. https://doi.org/10.1186/s13059-022-02710-1 (2022).
61. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191 (2018).
62. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**(1), 5114. https://doi.org/10.1038/s41467-018-07641-9 (2018).
63. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8 (2019).
64. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**(D1), D517–D525. https://doi.org/10.1093/nar/gkz935 (2020).
65. Van Rossum, T. *et al.* metaSNV v2: Detection of SNVs and subspecies in prokaryotic metagenomes. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab789 (2021).
66. Mende, D. R. *et al.* proGenomes2: An improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res.* **48**(D1), D621–D625. https://doi.org/10.1093/nar/gkz1002 (2020).
67. N. Homer, *DWGSIM*. github.com, 2022.
68. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res* **10**, 33. https://doi.org/10.12688/f1000research.29032.1 (2021).
69. Perez, F. & Granger, B. E. IPython: A system for interactive scientific computing. *Comput. Sci. Eng.* **9**(3), 21–29. https://doi.org/10.1109/MCSE.2007.53 (2007).

Acknowledgements

The study was funded and supported by the Jürgen Manchot Foundation and was part of the project "Decision-making with the help of Artificial Intelligence". We gratefully acknowledge the support of the medical and nursing staff of the hemato-oncological wards and units at the University Hospital Düsseldorf, without whom this study could not have been carried out. Support extracting and annotating clinical data provided by Sarah Schweier was appreciated. We thank Rebecca Fröhlich for development of the MedReactor tool for visualization of clinical data; Max Ried for computational infrastructure and support, Patrick Finzer, Colin MacKenzie and Tobias Wienemann for valuable discussions. For critical reading of the manuscript we want to acknowledge Sascha Dietrich and Jennifer Jaufmann. Sequencing as well as technical support regarding sequencing hardware were provided by the Genomics and Transcriptomics Laboratory (GTL) of the Biological and Medical Research Centre (BMFZ), Medical Faculty, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany. Computational support and infrastructure were provided by the "Centre for Information and Media Technology" (ZIM) at the University of Düsseldorf (Germany). We thank the participants and the staff of LifeLines-DEEP for their collaboration. Funding for the project was provided by the Top Institute Food and Nutrition Wageningen grant GH001. The sequencing was carried out in collaboration with the Broad Institute.

Author contributions

Project conceptualization: A.D., K.P., R.H. Wetlab protocol adaptation + validation: B.H., S.S. Acquisition, analysis, or interpretation of data: A.D., A.R., B.H., G.K., K.P., P.S., R.H., S.S. Bioinformatics method development: A.D., P.S. Software development and validation: P.S. Sample acquisition coordination and sequencing: A.R., S.S. Visualization of results: A.R., P.S., S.S. Supervision: A.D., B.H., G.K., K.P., R.H. Clinical coordination: P.J. Initial draft of manuscript: A.D., A.R., K.P., P.S., S.S. Draft revision: A.D., G.K., K.P., R.H. Funding acquisition: A.D., G.K., K.P., R.H. All authors read, edited and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. Manchot Foundation Project: Research Group "Decision-making with the help of Artificial Intelligence" to K.P., R.H, A.D., G.K. and research group Molecular Host–Pathogen-Interactions to K.P.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-53506-1.

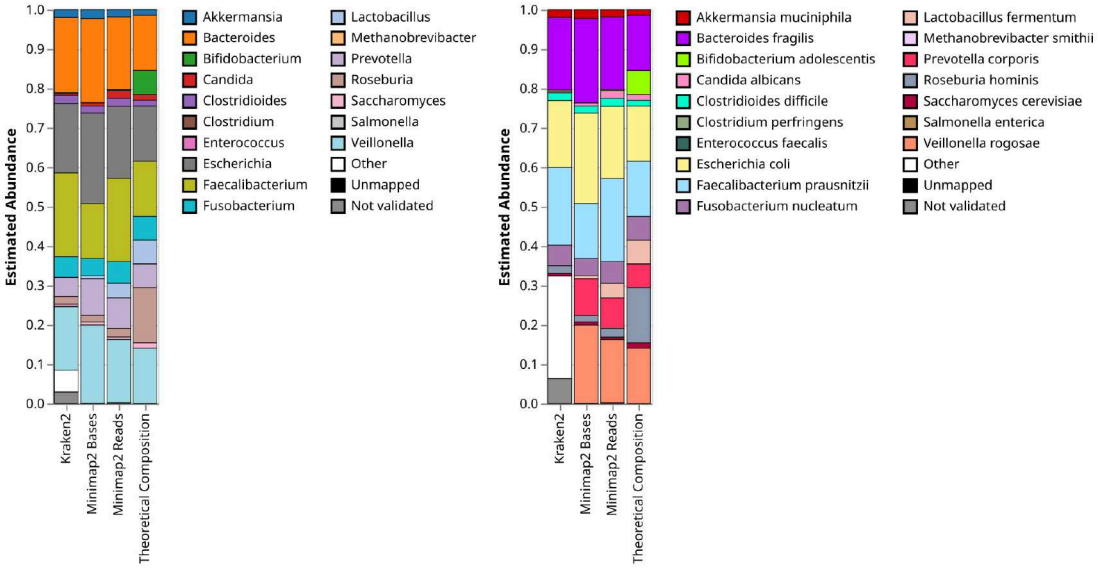
Correspondence and requests for materials should be addressed to G.W.K., R.H., A.D. or K.P.

Reprints and permissions information is available at www.nature.com/reprints.

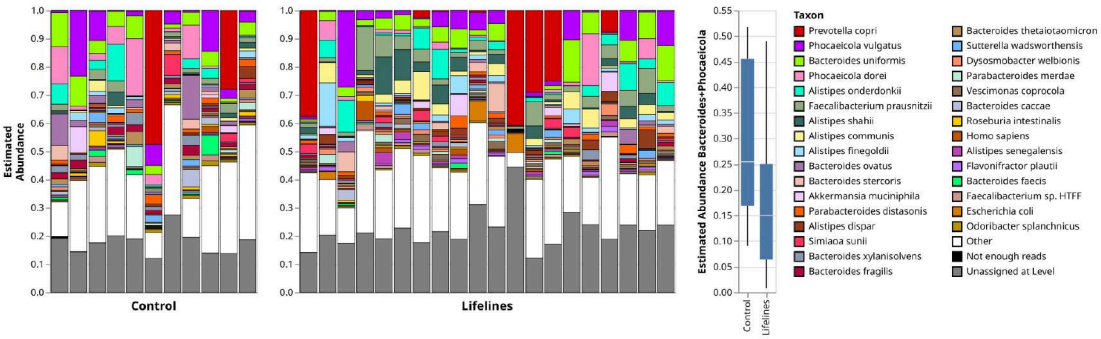
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.1.3 Supplementary Figures

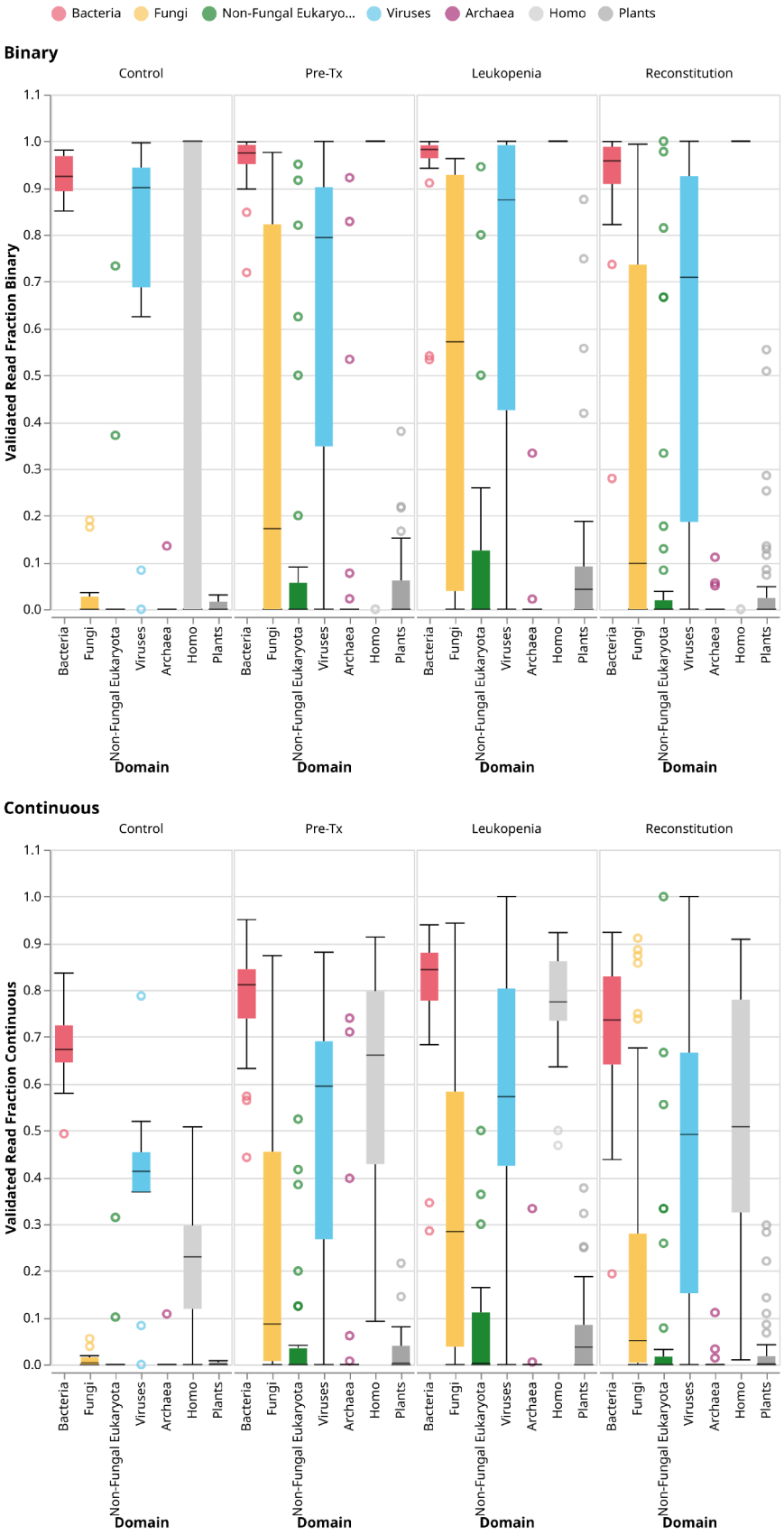
Additional supplementary files can be found online and are omitted due to being unsuitable for print.



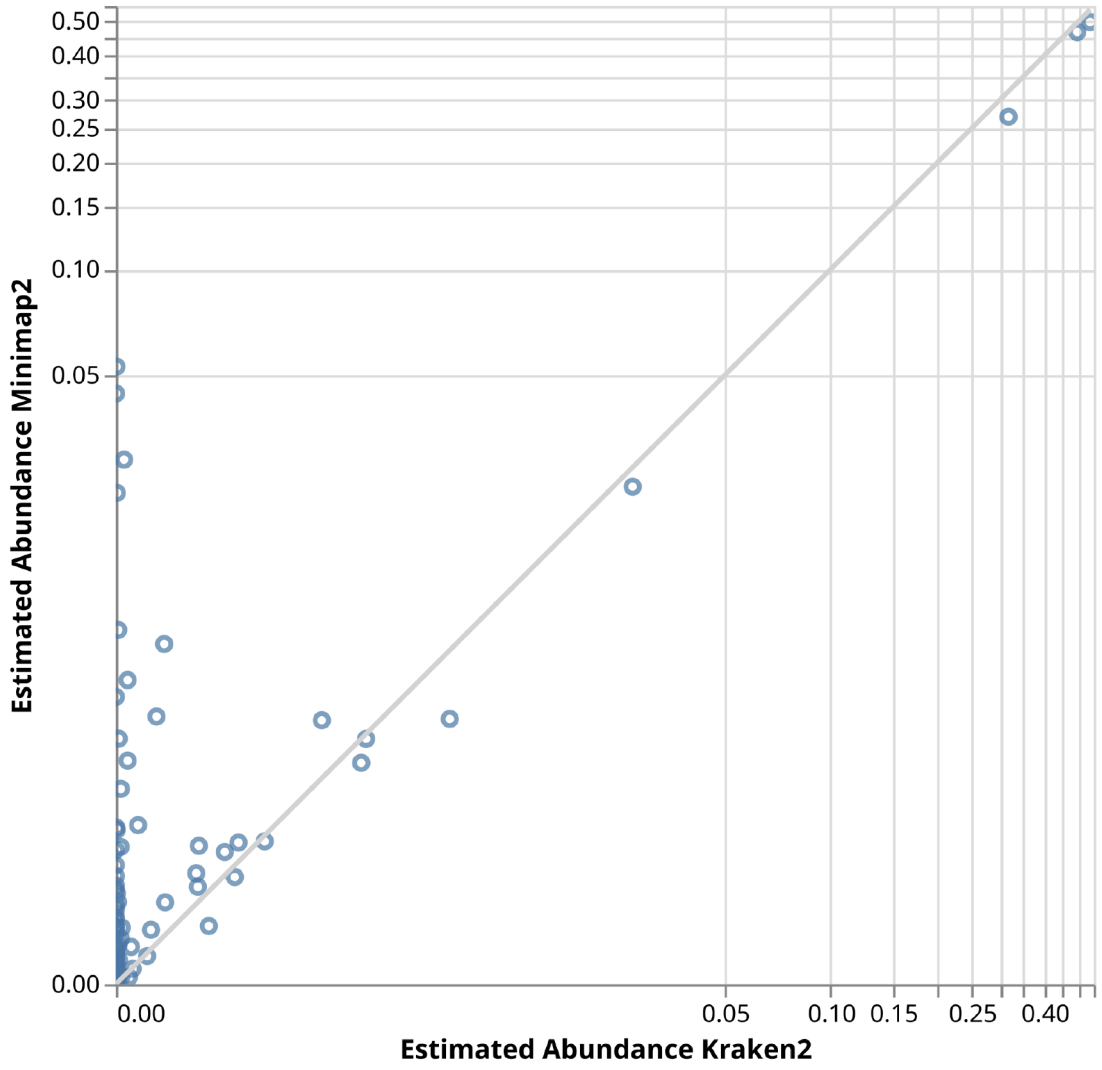
Supplementary Figure 1: Composition estimates for the Zymo Gut Microbiome Standard. Abundance estimation was carried out using Kraken2 and the MetaGut v1.0 database; for the minimap2 estimates we used either the fraction of reads or the fraction of bases while using only primary alignments to the Zymo-provided reference sequences. The theoretical composition, as reported by Zymo, is plotted in the last column.



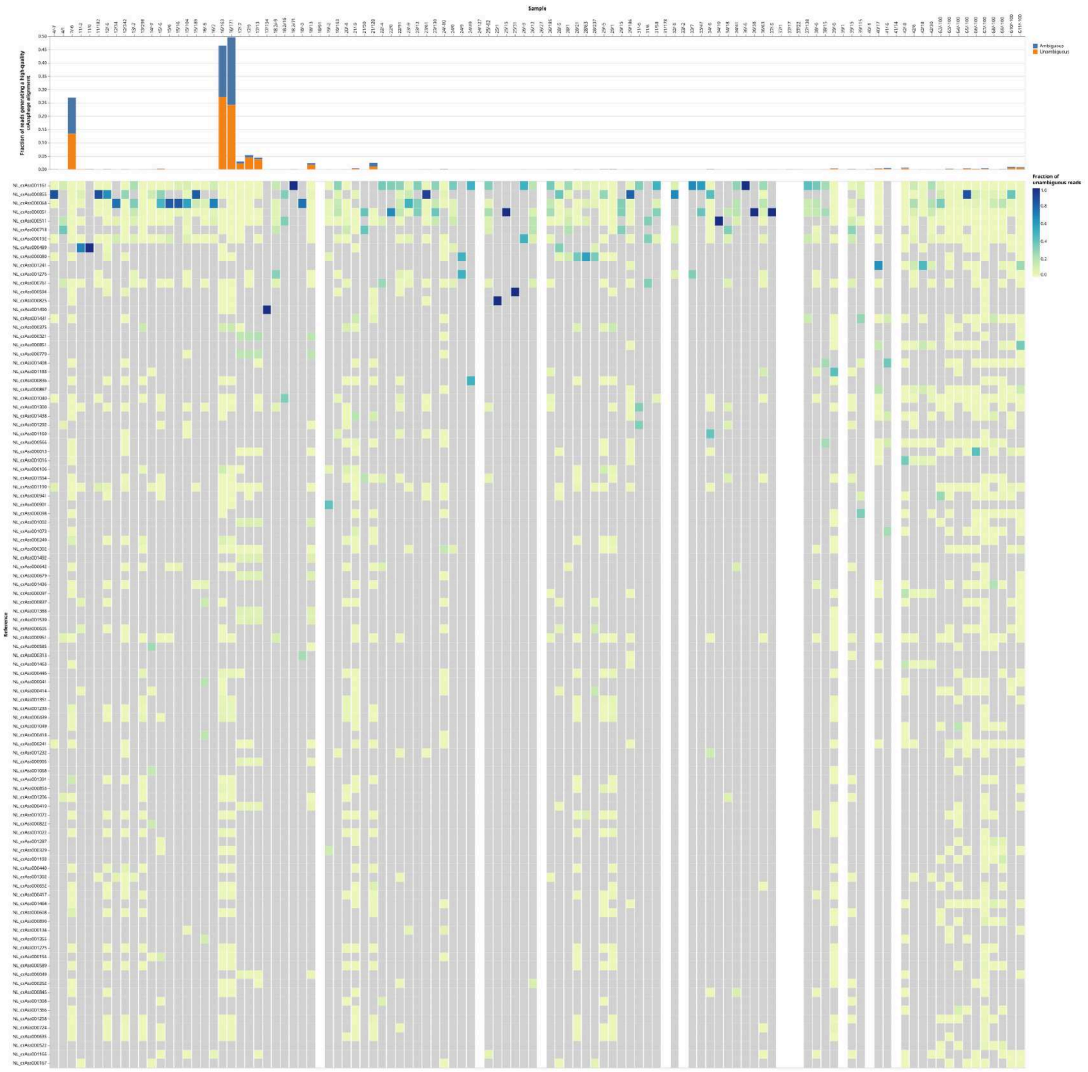
Supplementary Figure 2: Composition estimates for Lifelines Deep samples. Abundance estimation was carried out using Kraken2 with the MetaGut v1.0 database. The boxplot shows the fraction of species belonging to the *Bacteroides* and *Phocaeicola* genera for the Lifelines samples and for the control group samples of this study.



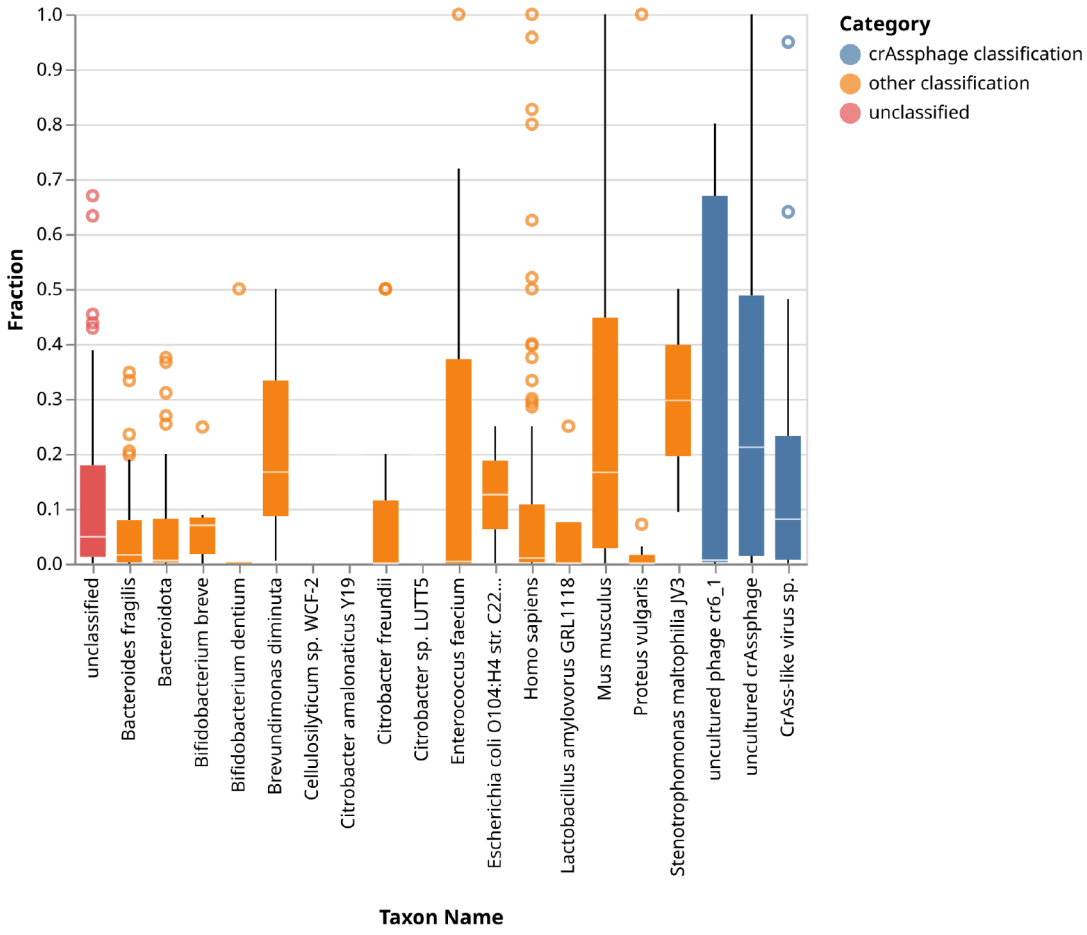
Supplementary Figure 3: Validation rates stratified by sampling time point. "Non-fungal eukaryota" excludes chordata and plants. For "binary" we considered a taxon to be validated if more than 20% of the associated reads could be validated. In this case, we counted all reads as validated. For "continuous" we report the actual (weighted) validation rates.



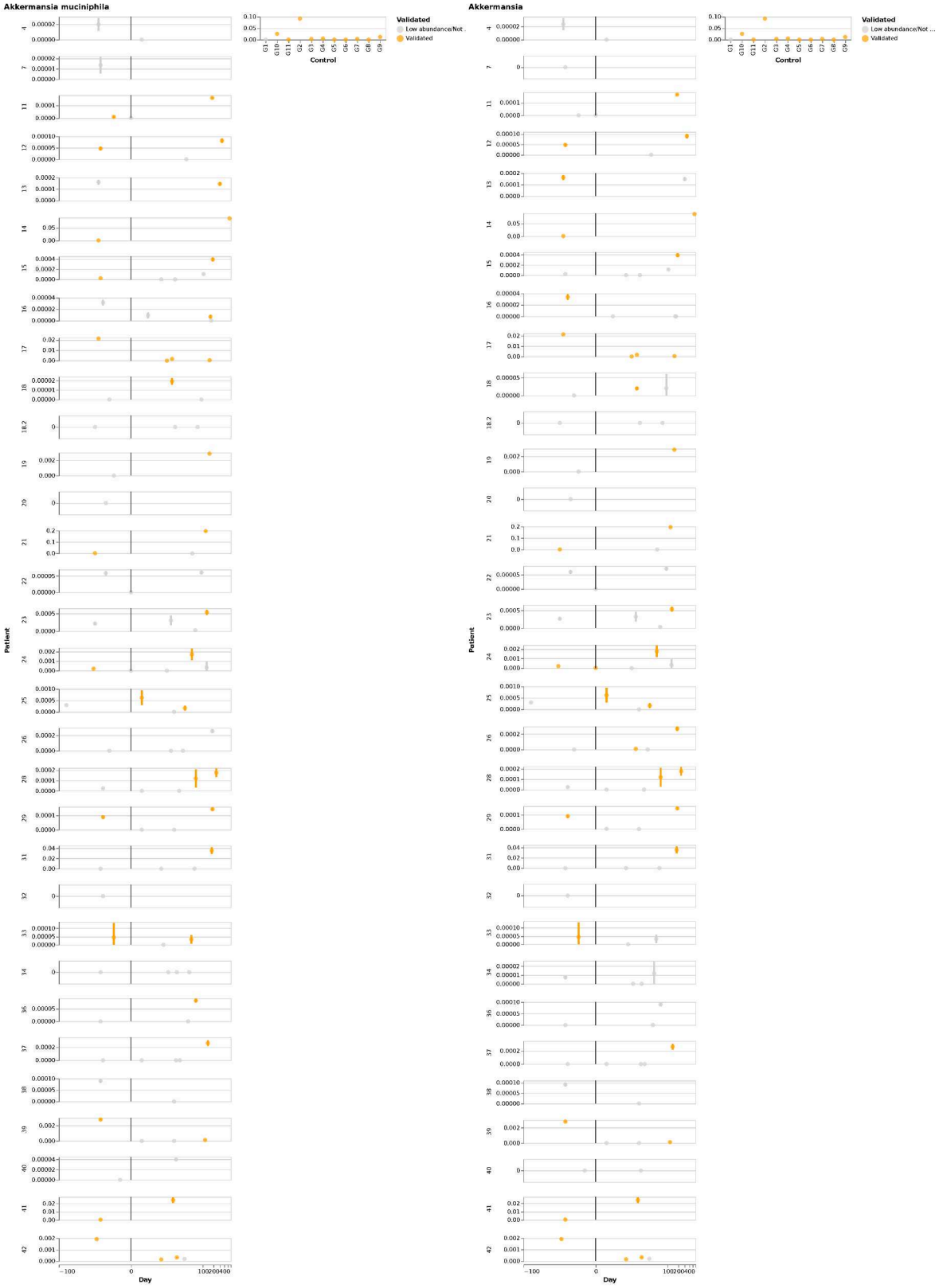
Supplementary Figure 4: Comparison of *crAssphage* abundance estimations. Reads were mapped against *crAssphage* references using minimap2 retaining only primary alignments that aligned at least 70% of the query with an identity of above 70%. We used the fraction of primary alignments divided by the number of query reads as a rough estimate for the abundance. For comparison the default abundances provided by Kraken2 and the MetaGut v1.0 database are shown.

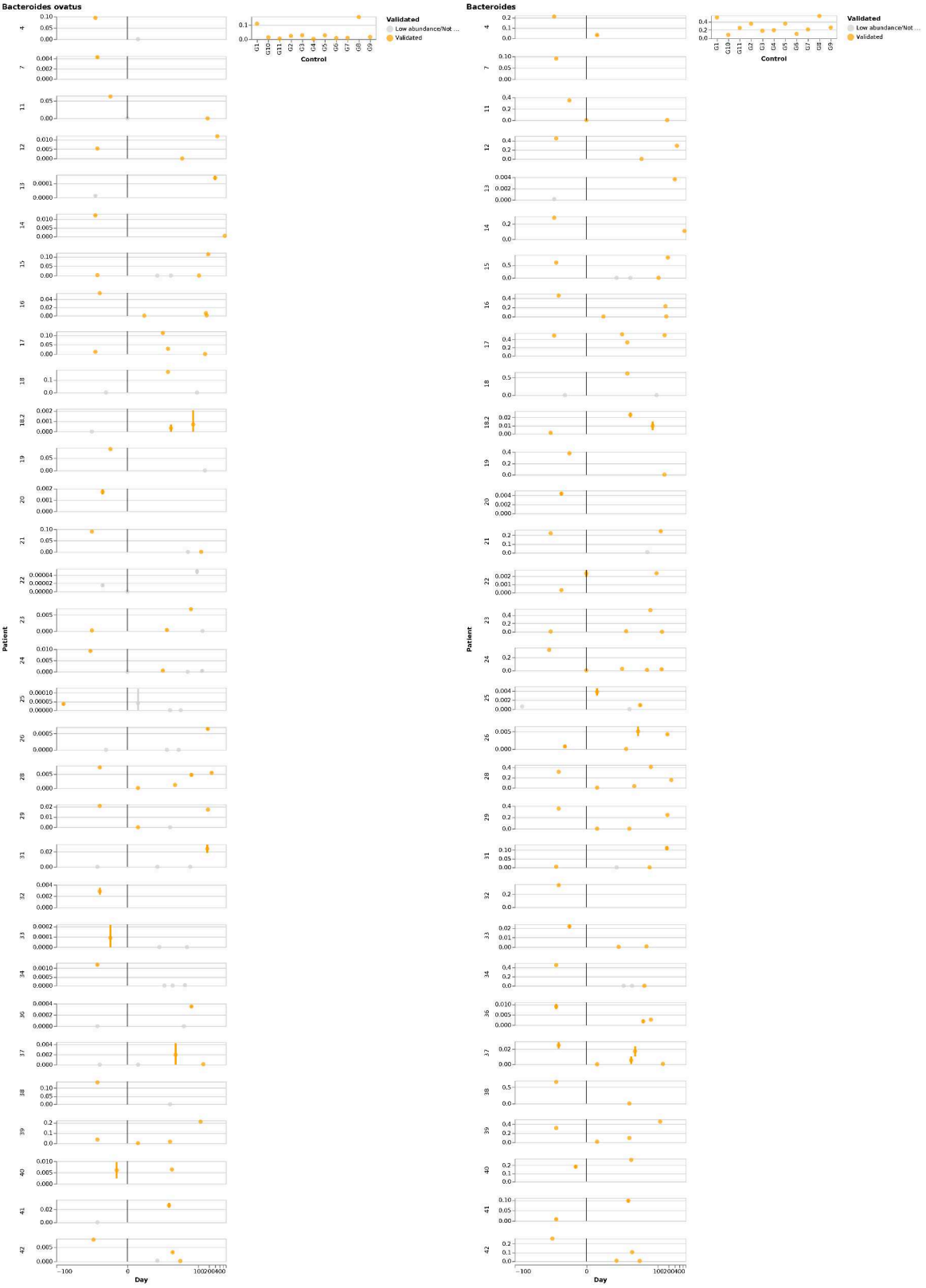


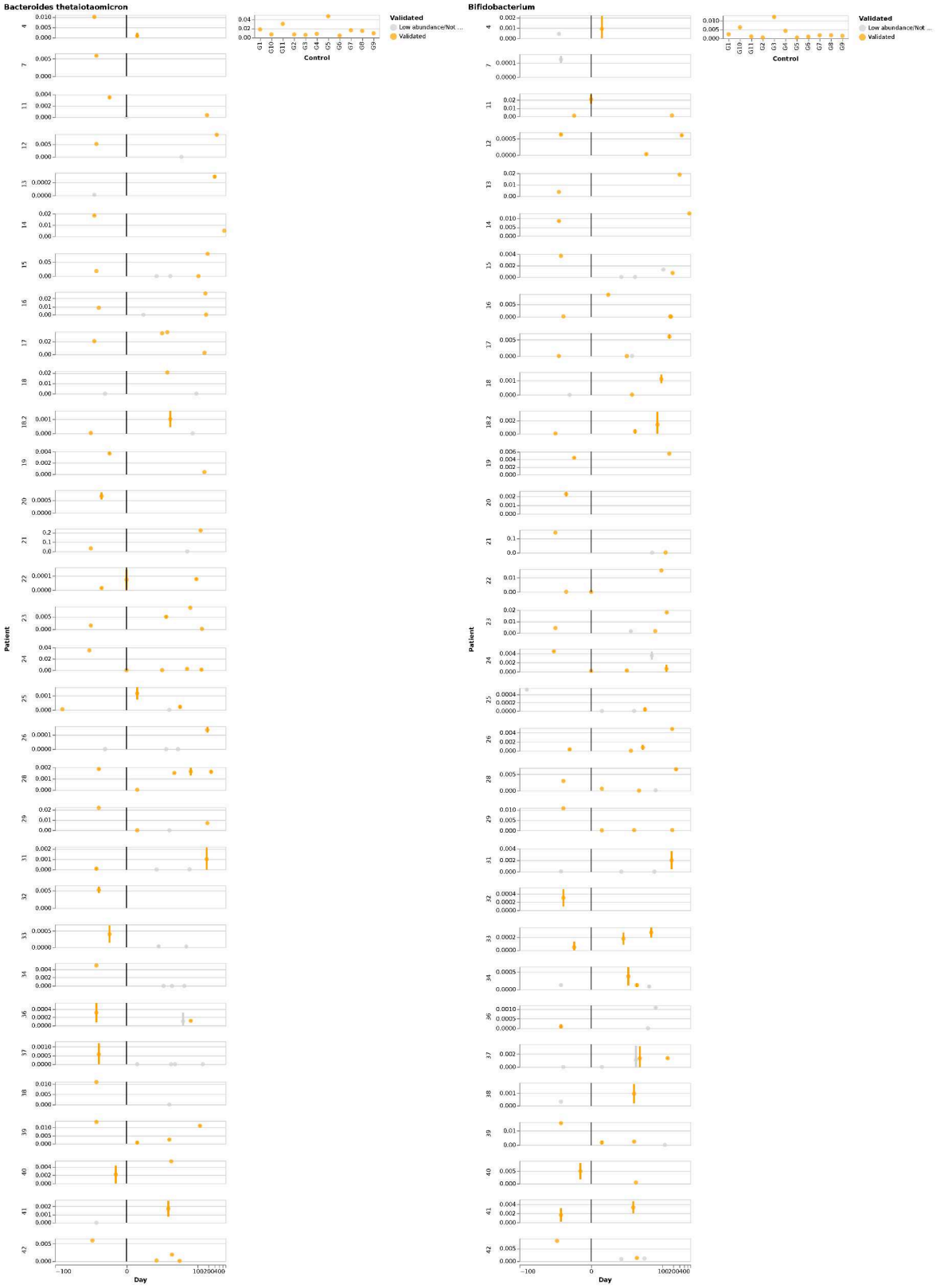
Supplementary Figure 5: Distribution of reads across specific *crAssphage* references. Barplots show the overall fraction of reads that are mapped to *crAssphage*. The heatmap is limited to the top 100 *crAssphage* sequences based on the mean of the assigned fractions and shows only ambiguous assignments (normalized to 100%). Sample-reference pairs with ambiguous assignments are colored gray. Since mapping quality directly corresponds to the uniqueness of each mapping, we used a threshold of mapping quality > 5 to determine reads that can be assigned to a specific reference.

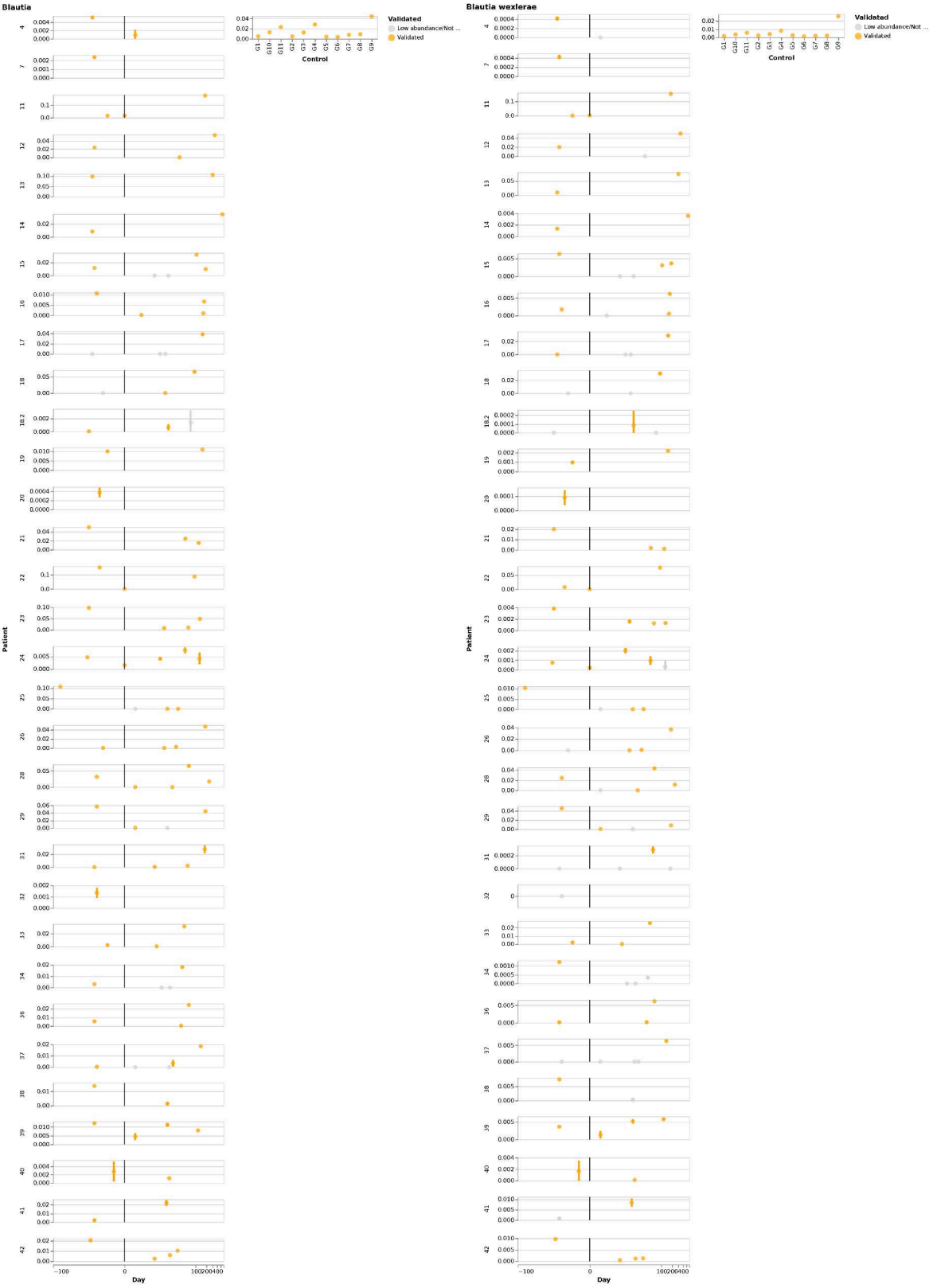


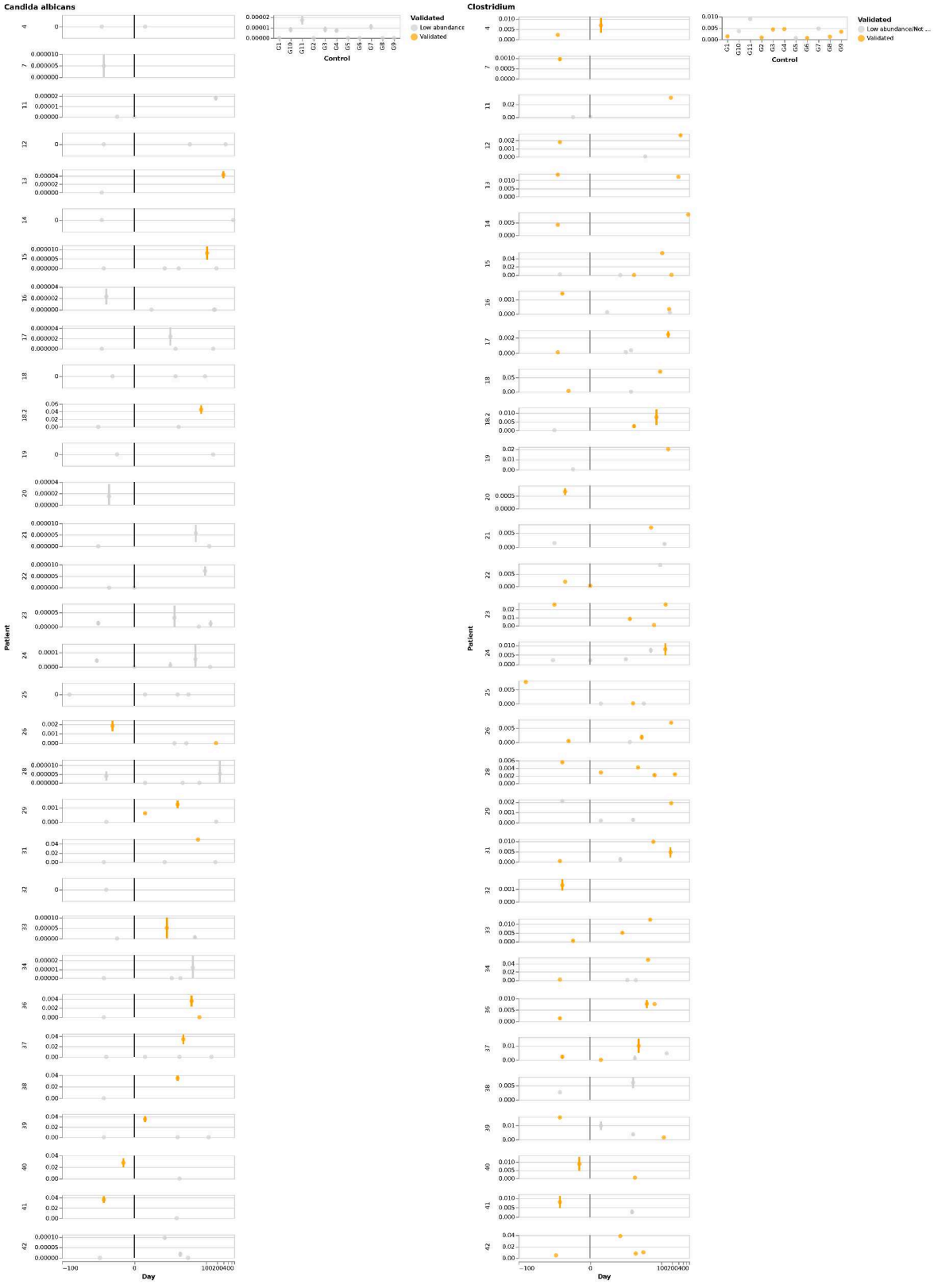
Supplementary Figure 6: Read migration from Kraken2 to minimap2. Fraction of Kraken2 taxon assignments for reads that were assigned to *crAssphage* using minimap2 and a custom database of curated *crAssphage* references. For visual clarity, only the top 20 taxa are shown.

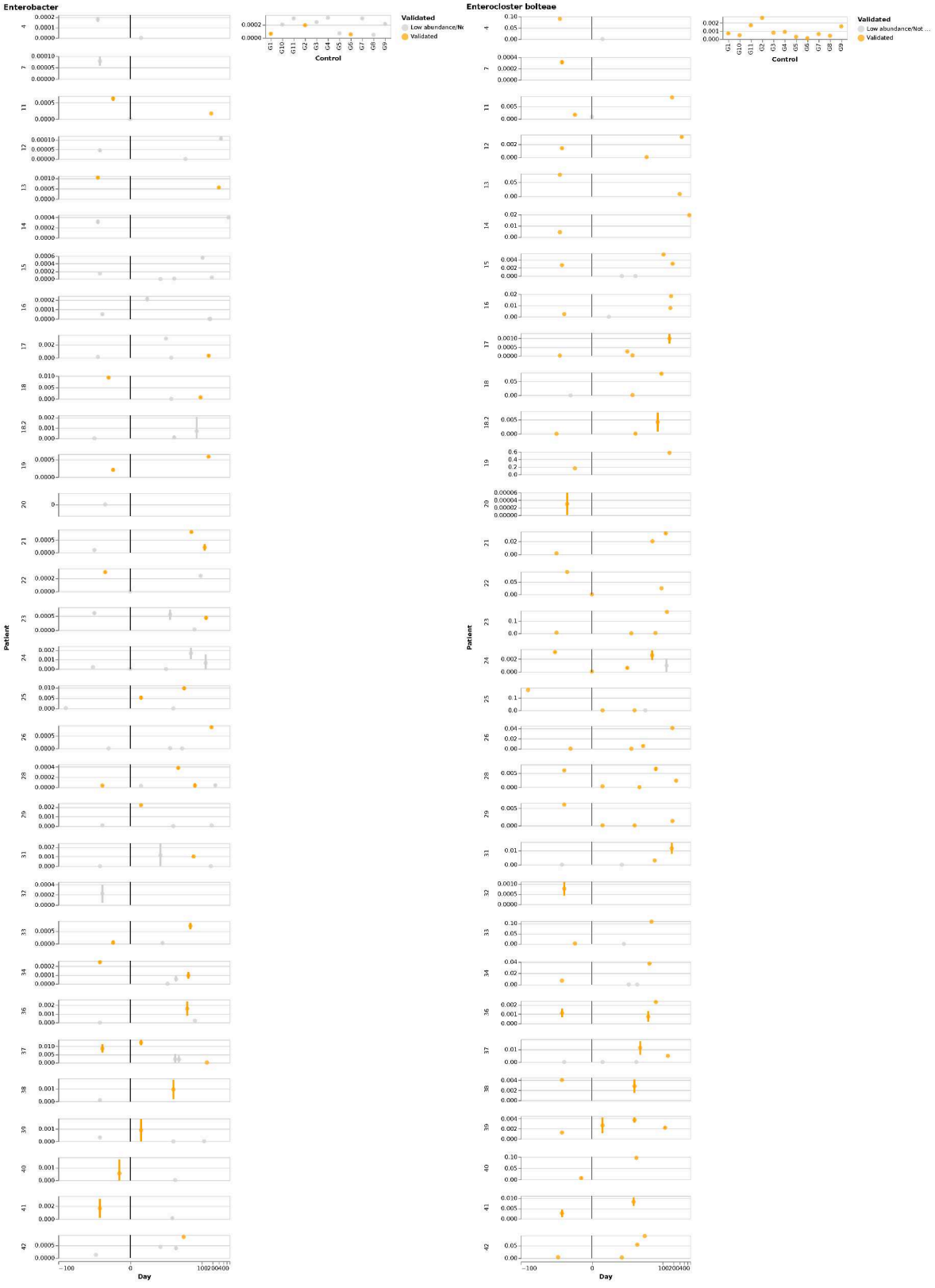

Akkermansia muciniphila
Akkermansia
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

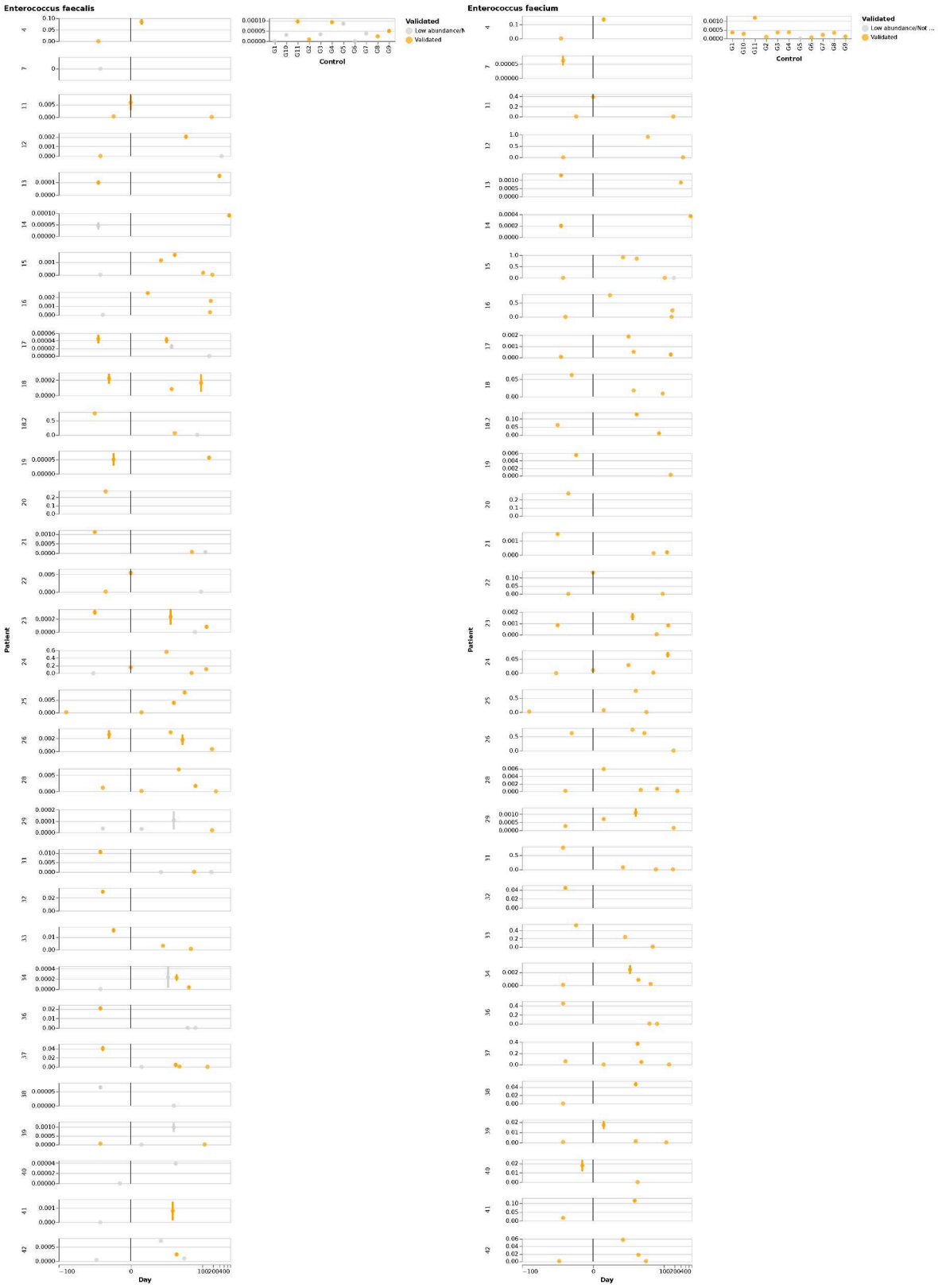

Bacteroides ovatus
Bacteroides
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

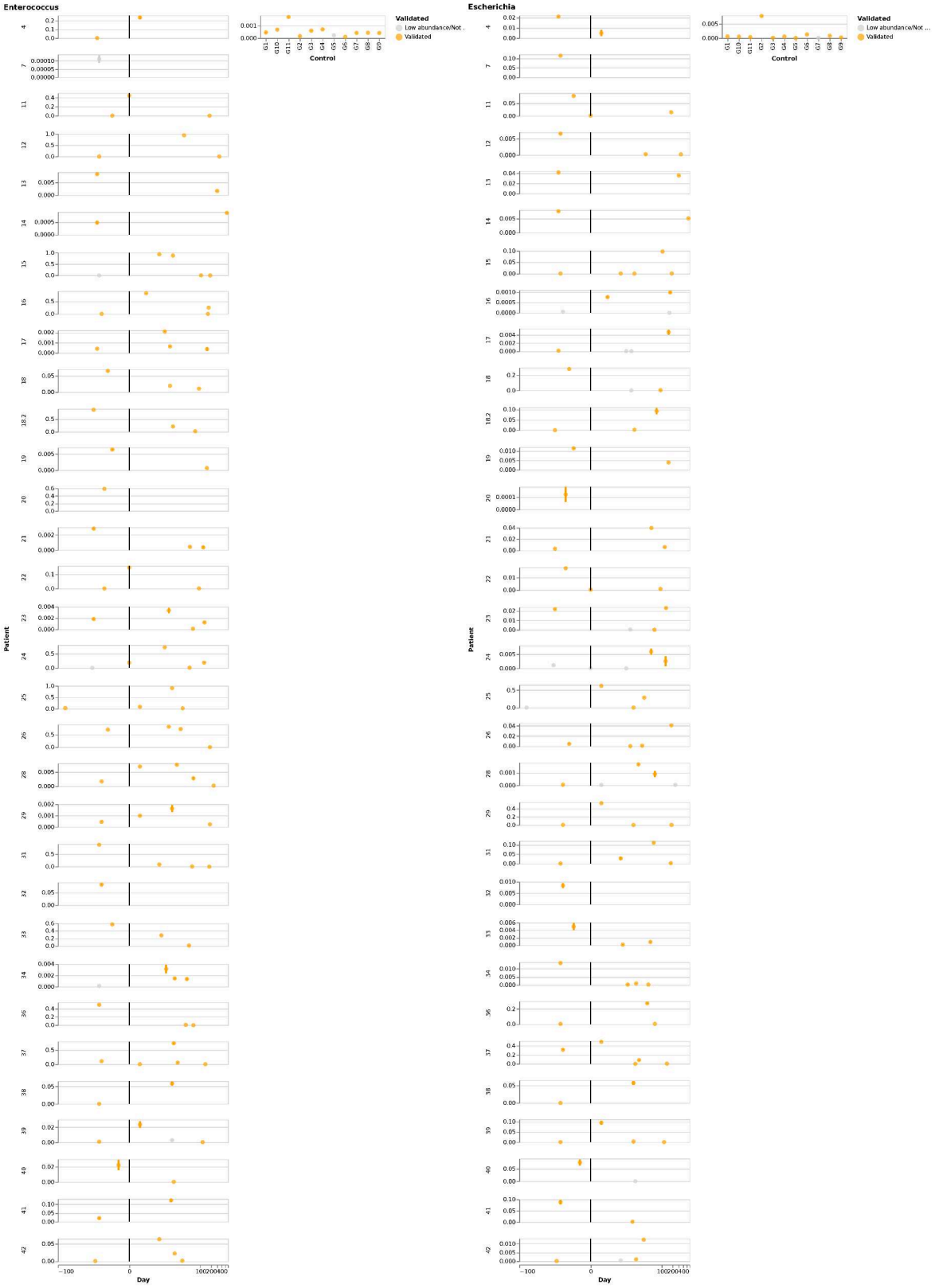

Bacteroides thetaiotaomicron
Bifidobacterium
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

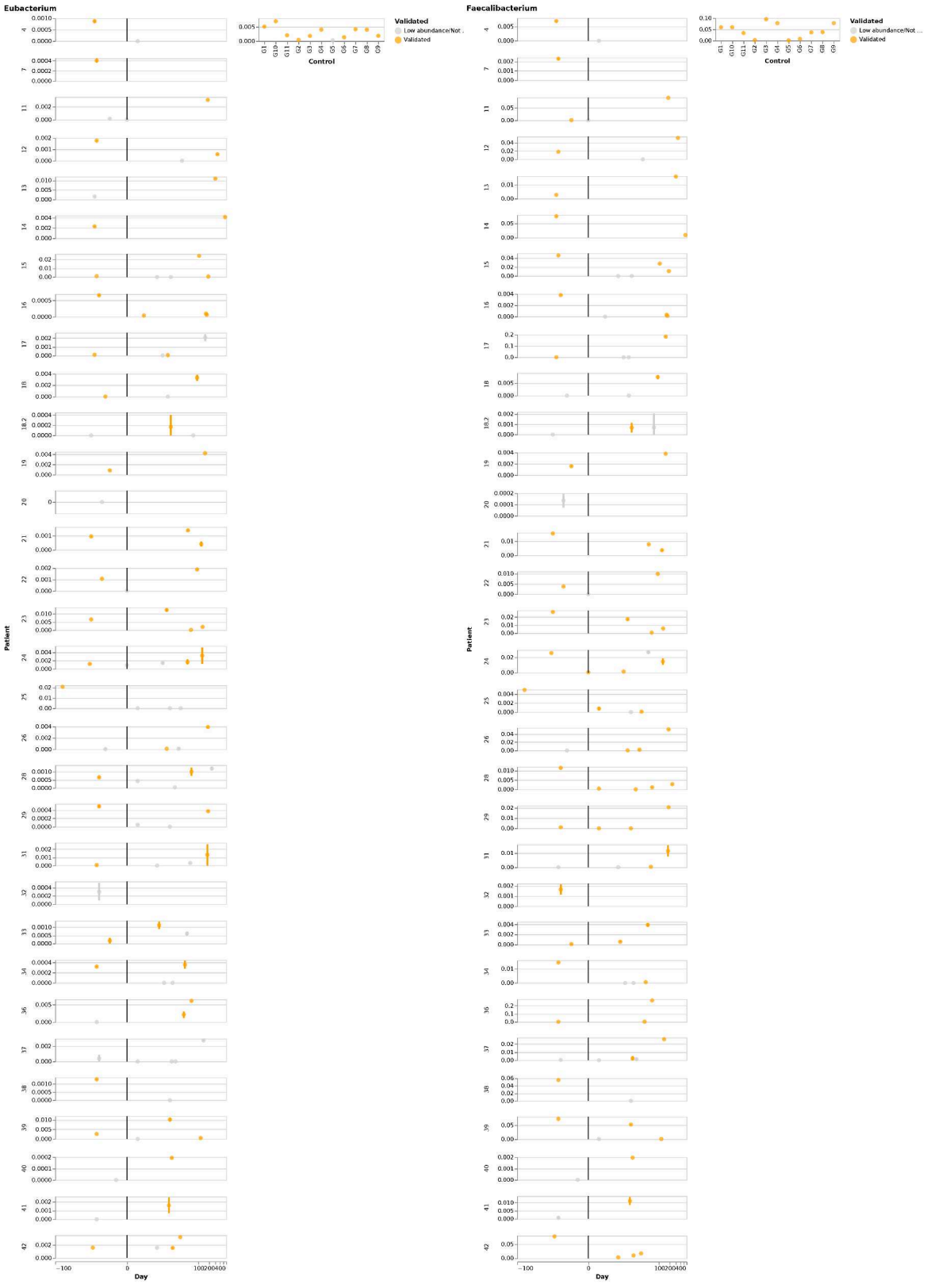

Blautia
Blautia wexlerae
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

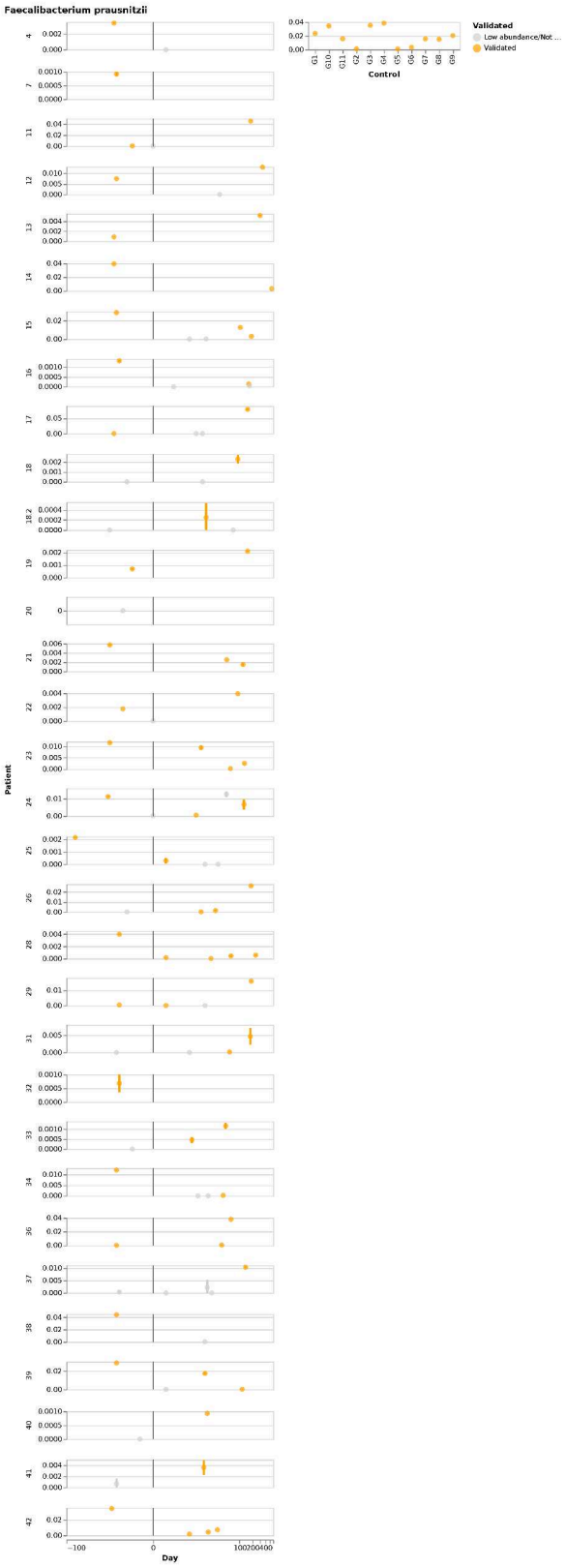

Candida albicans
Clostridium
Validated
Low abundance
Low abundance/Not ...
Validated
Control
Patient
Day

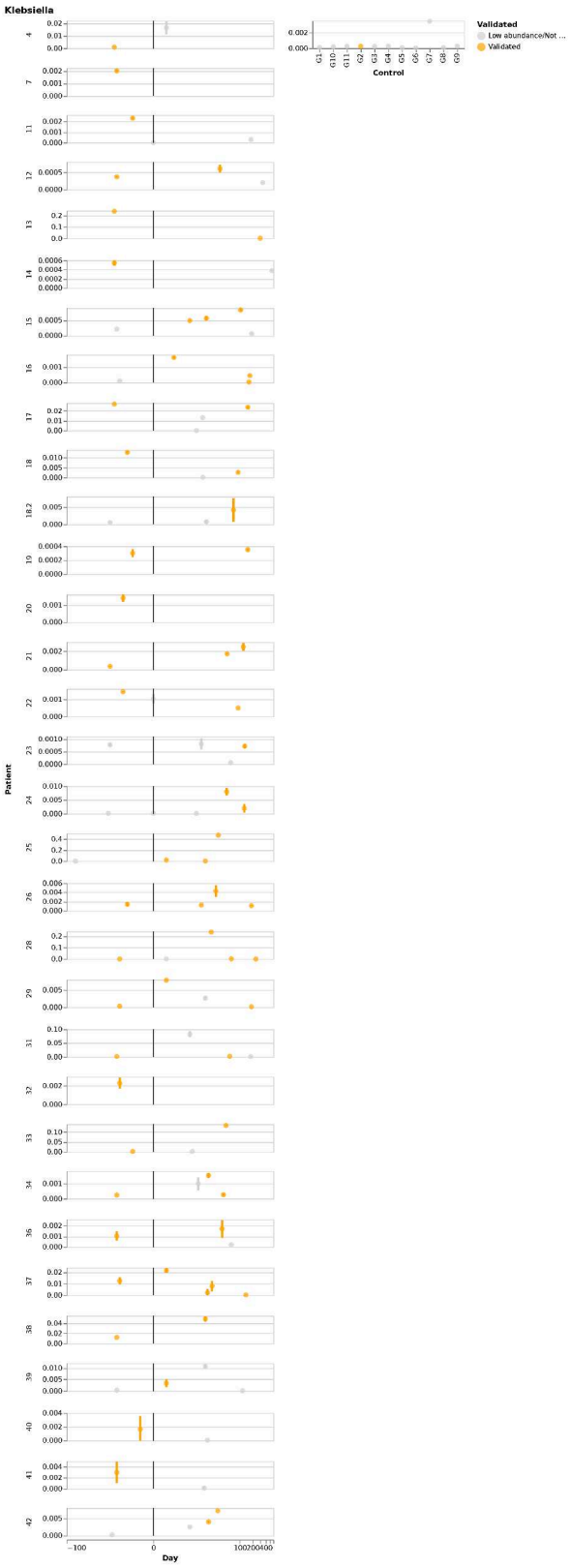

Enterobacter
Enterocloster bolteae
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

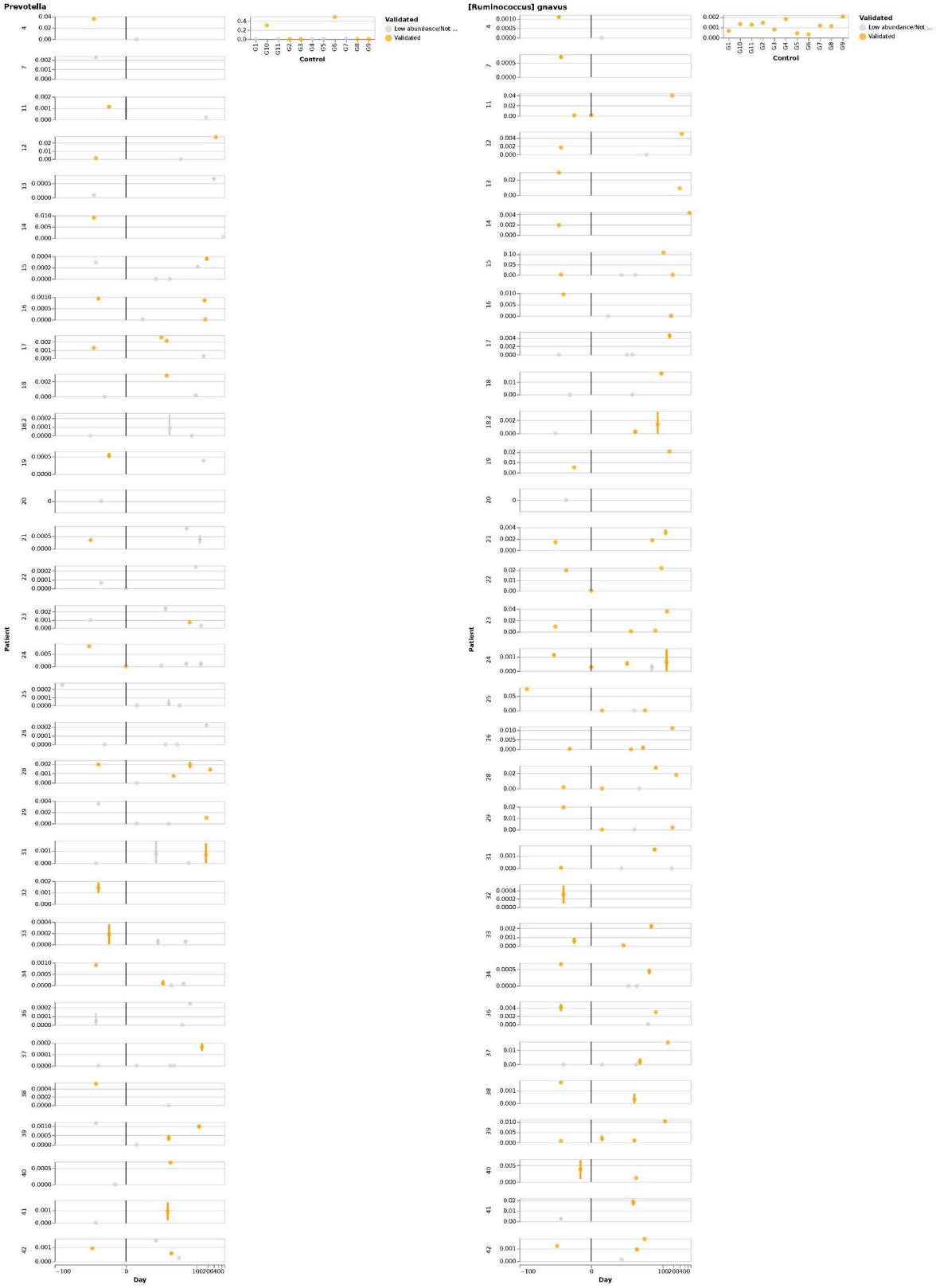

Enterococcus faecalis
Validated
Low abundance/N
Validated
Control
Patient
Day
Enterococcus faecium
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

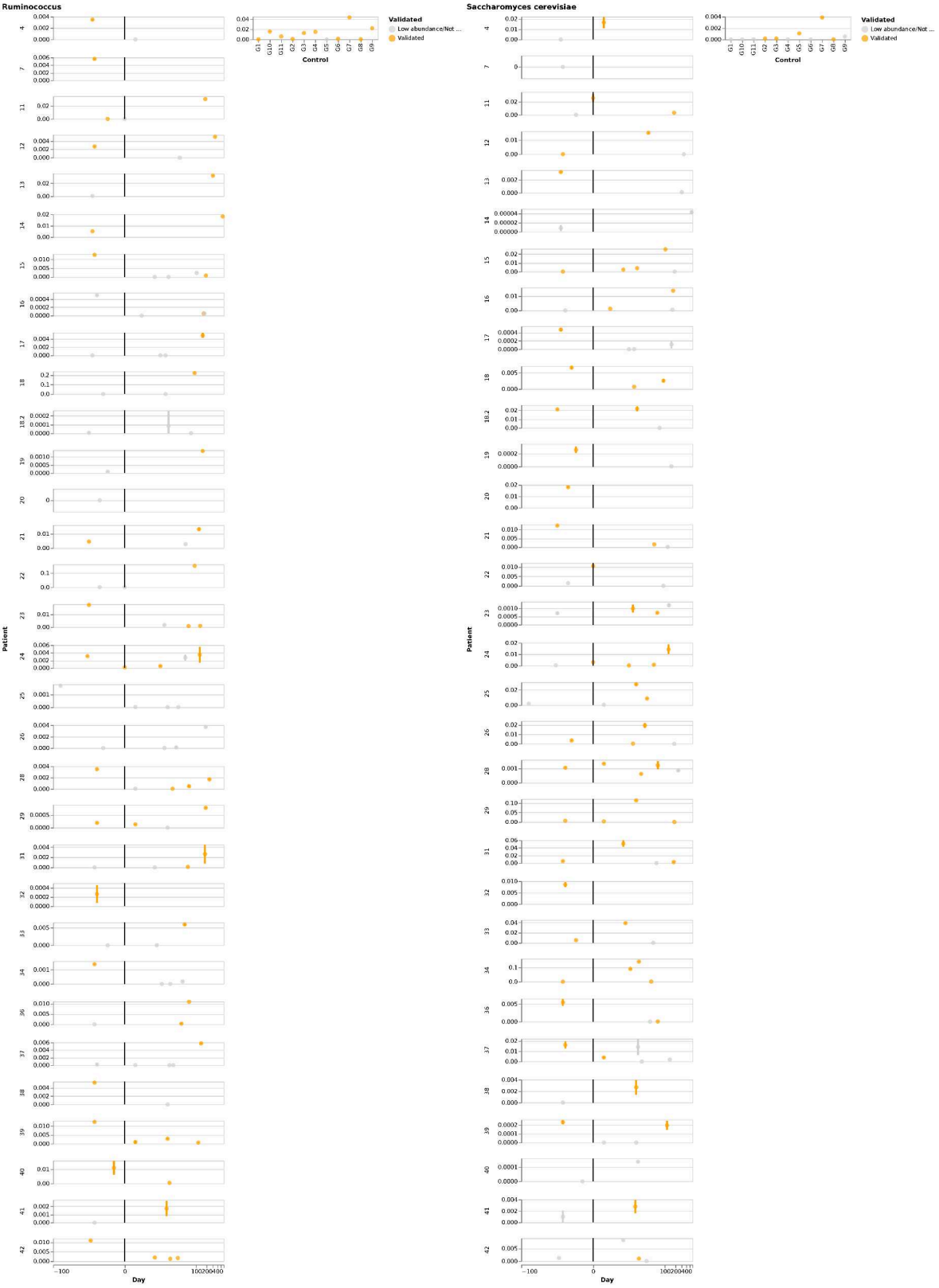

Enterococcus
Escherichia
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

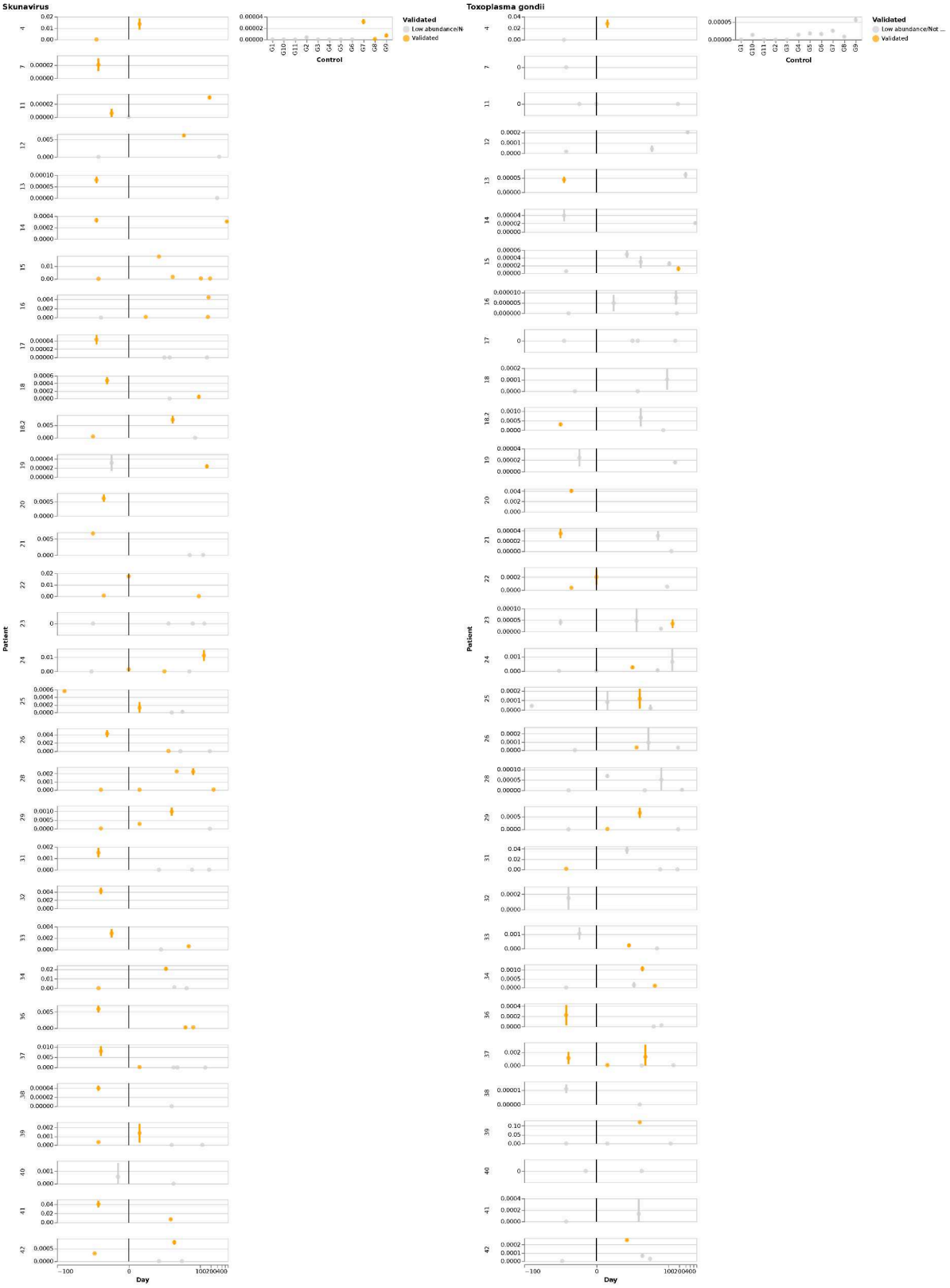

Eubacterium
Faecalibacterium
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

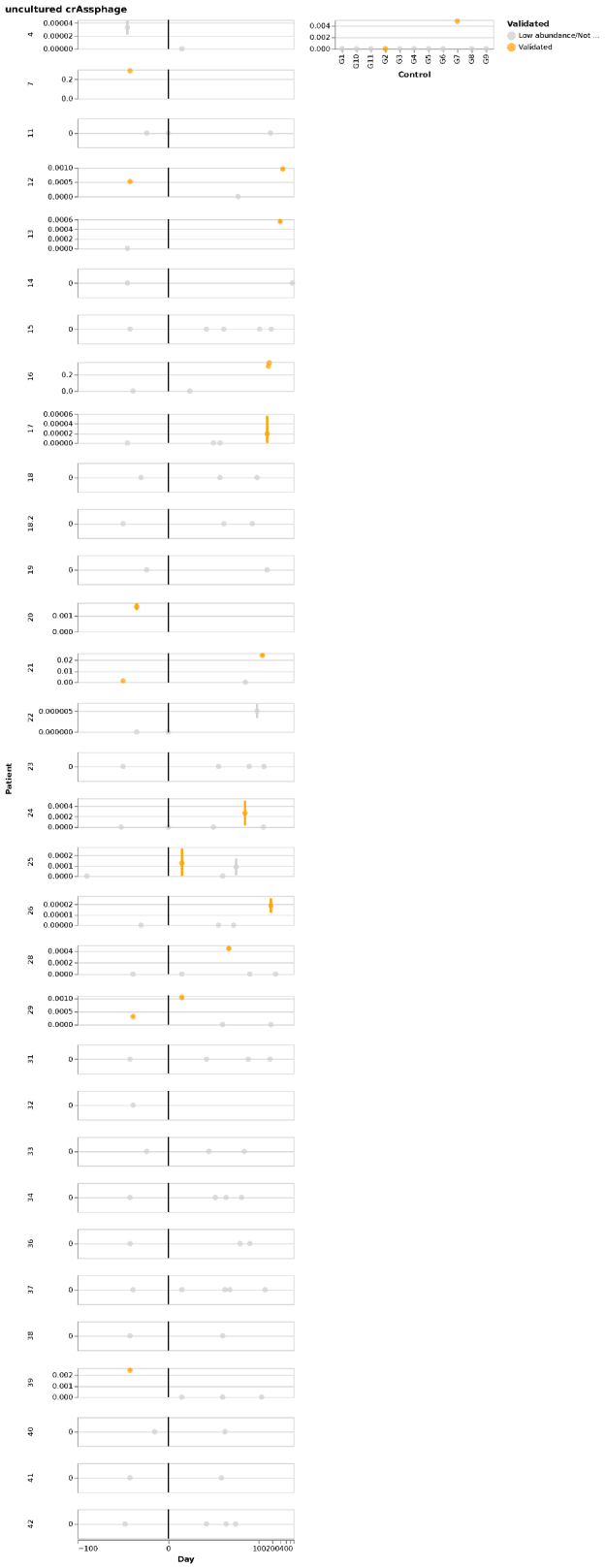

Faecalibacterium prausnitzii
Validated
Low abundance/Not ...
Validated
Control
Patient
Day


Klebsiella
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

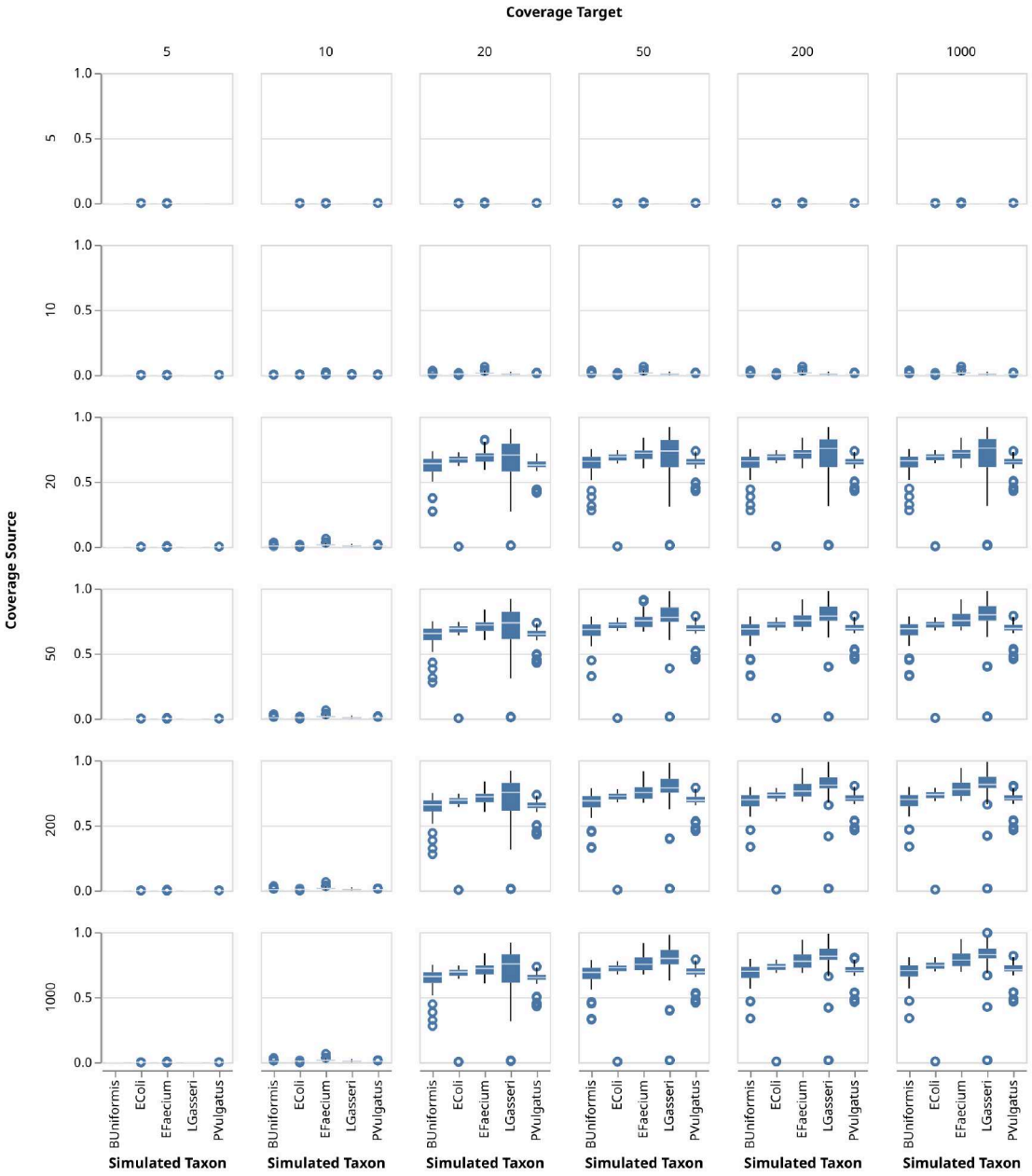

Prevotella
[Ruminococcus] gnavus
Validated
Low abundance/Not ...
Validated
Control
Patient
Day


Ruminococcus
Saccharomyces cerevisiae
Validated
Low abundance/Not ...
Validated
Control
Patient
Day

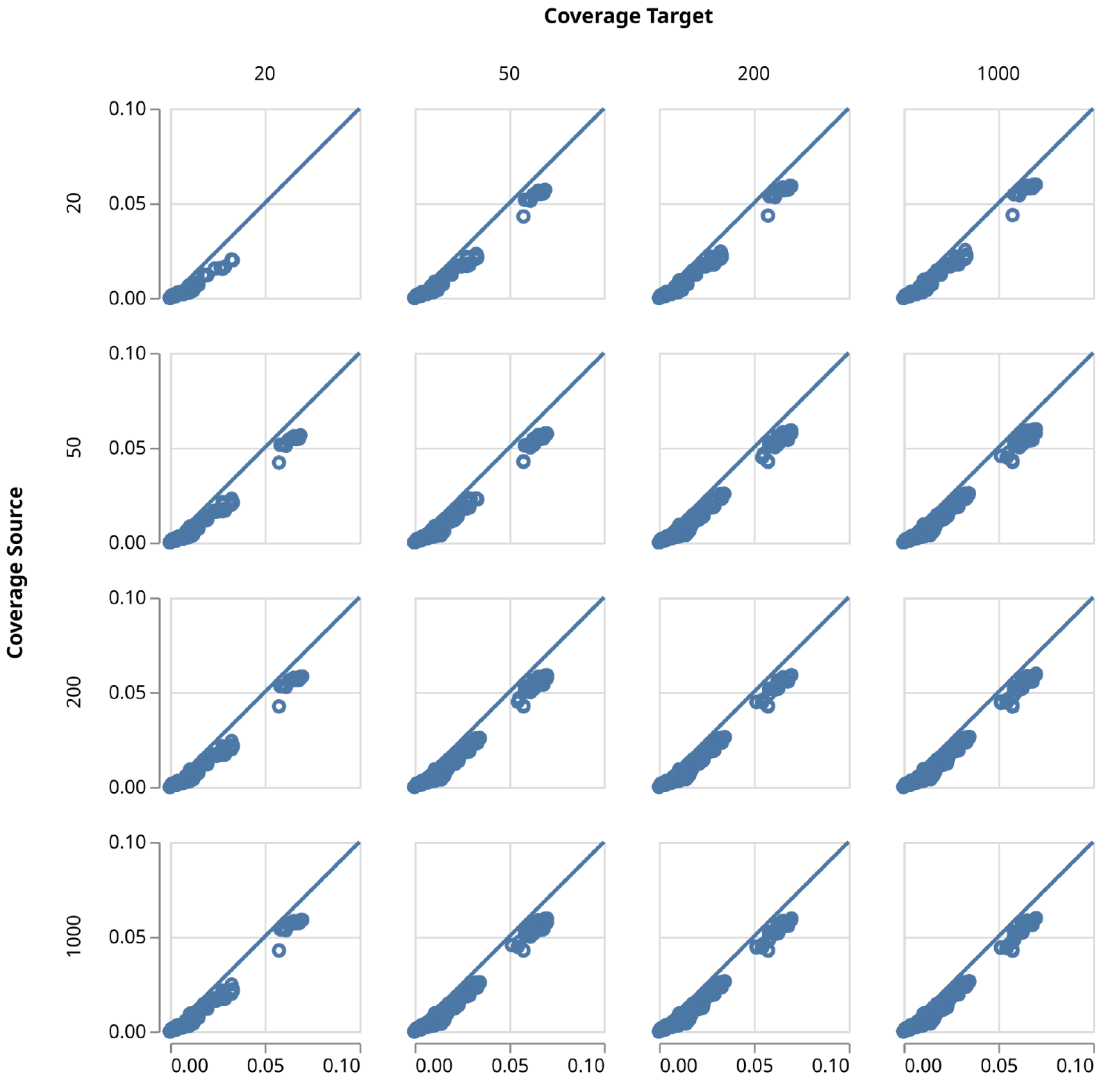

Skunavirus
Toxoplasma gondii
Validated
Low abundance/Not ...
Validated
Control
Patient
Day



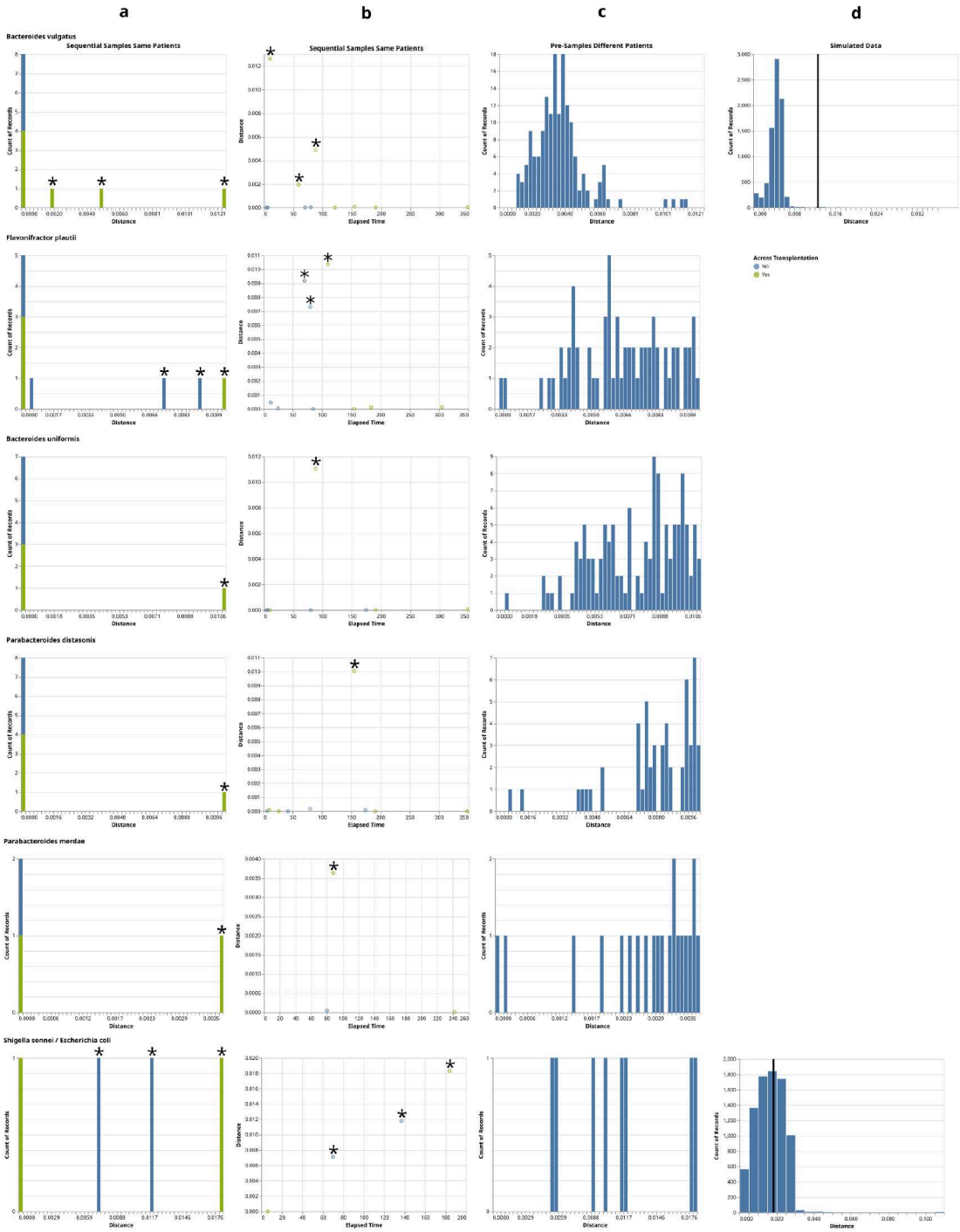
Supplementary Figure 7: Marker taxa. Abundance plots per patient for selected marker taxa using a binomial distribution to model confidence intervals based on $\alpha = 0.05$. Patients are shown with time (days) relative to alloHSCT on the x-axis and all control samples are collectively shown on the right side of the figure. Relative frequency and confidence intervals for $\alpha = 0.05$ of validated reads for the specified marker taxon are depicted on the y-axis. Each investigated taxon is shown on a separate page.



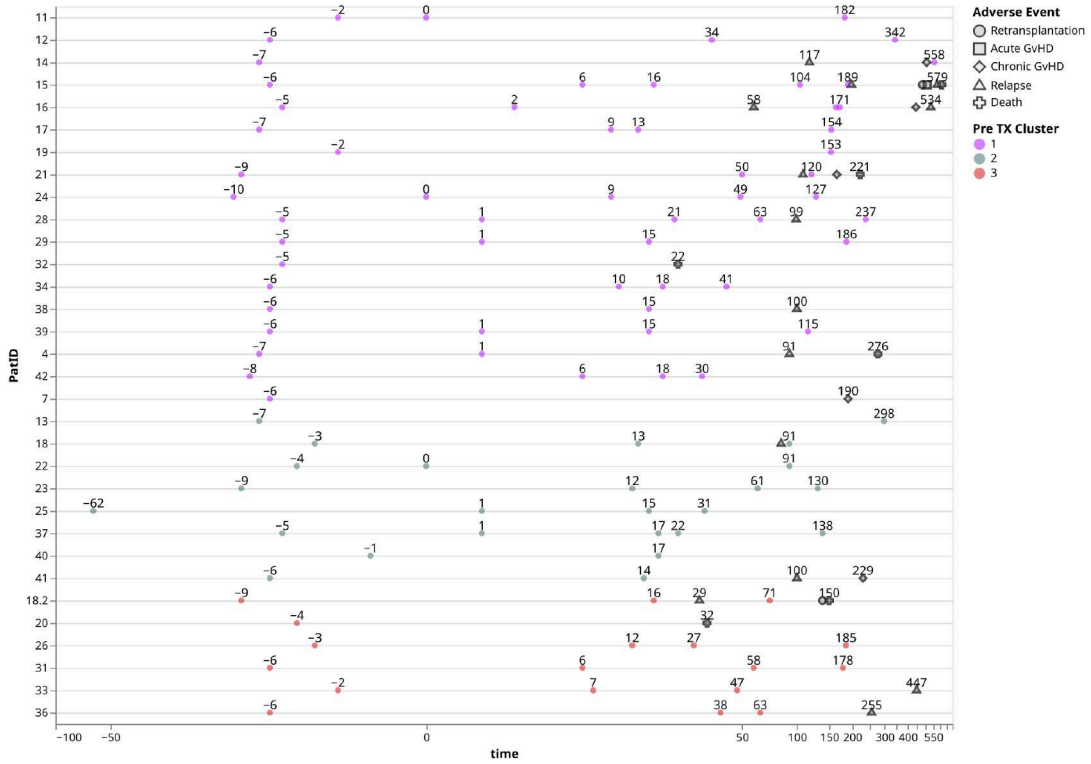
Supplementary Figure 8: Overlap for simulated genomes (ANI). The share of sequence overlap for pairs of simulated read sets for five bacterial taxa relative to different coverage values.



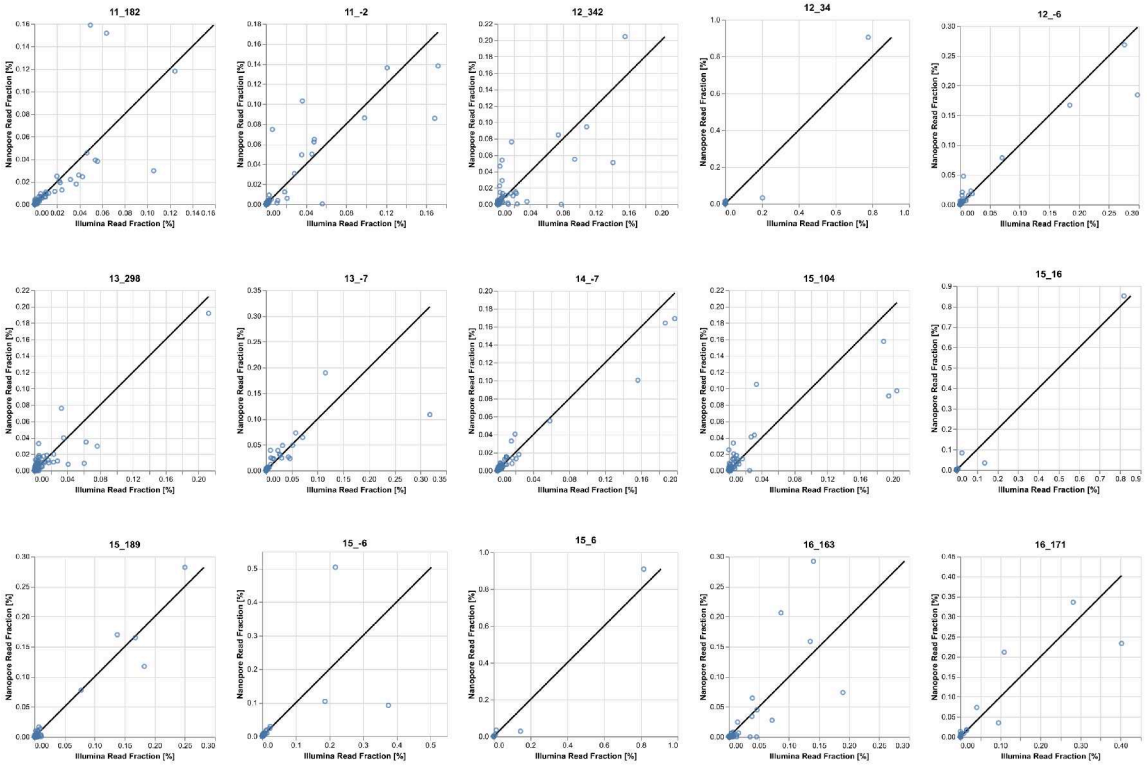
Supplementary Figure 9: aANI-lowFreq distance quality relative to coverage. The inner x-axis contains the FastANI genome distance and the inner y-axis the calculated aANI-lowFreq distance. Only distances with a shared overlap of the sequences above 70% are shown which removes all distances with simulated coverage 5 or 10.

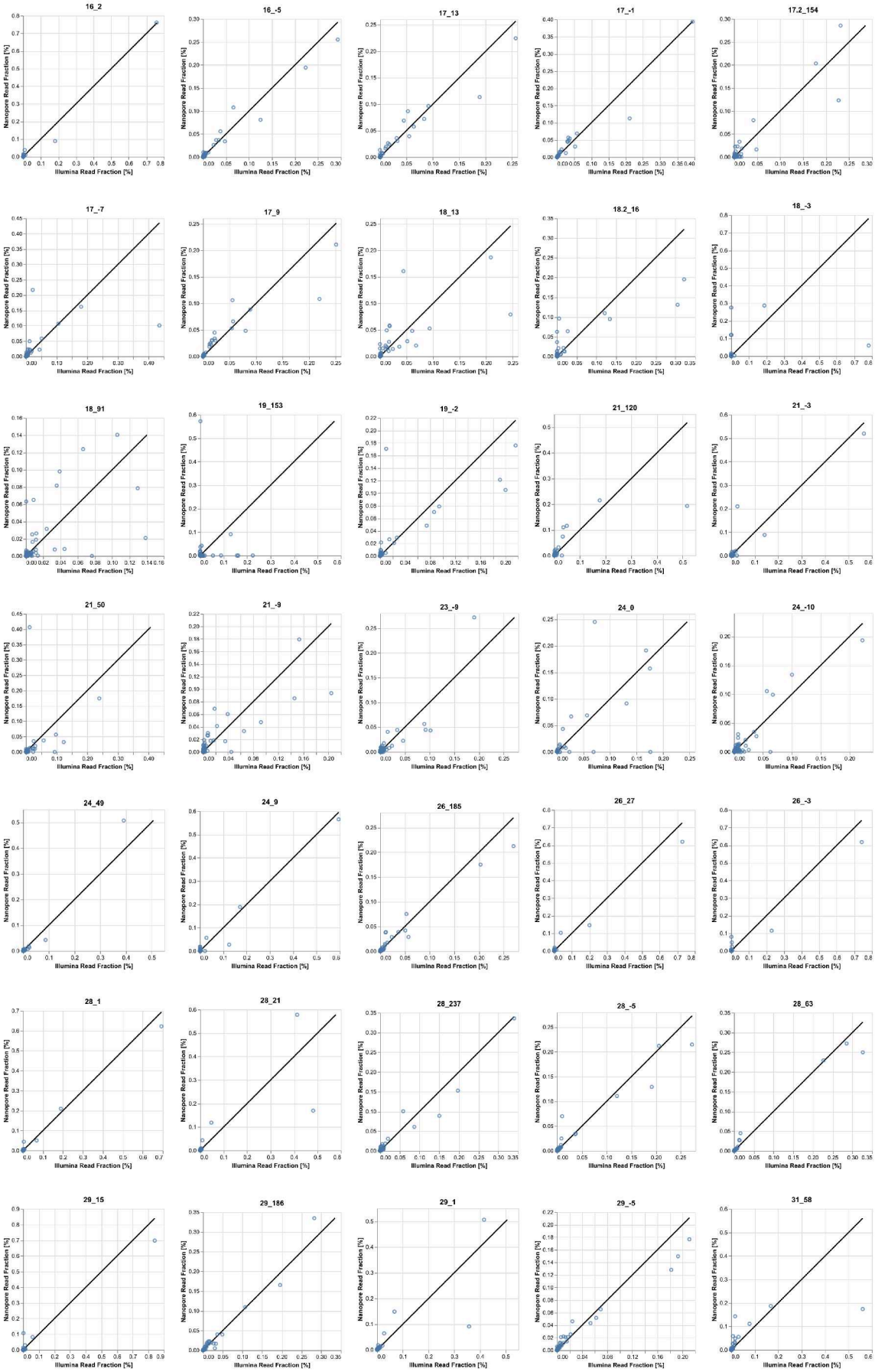


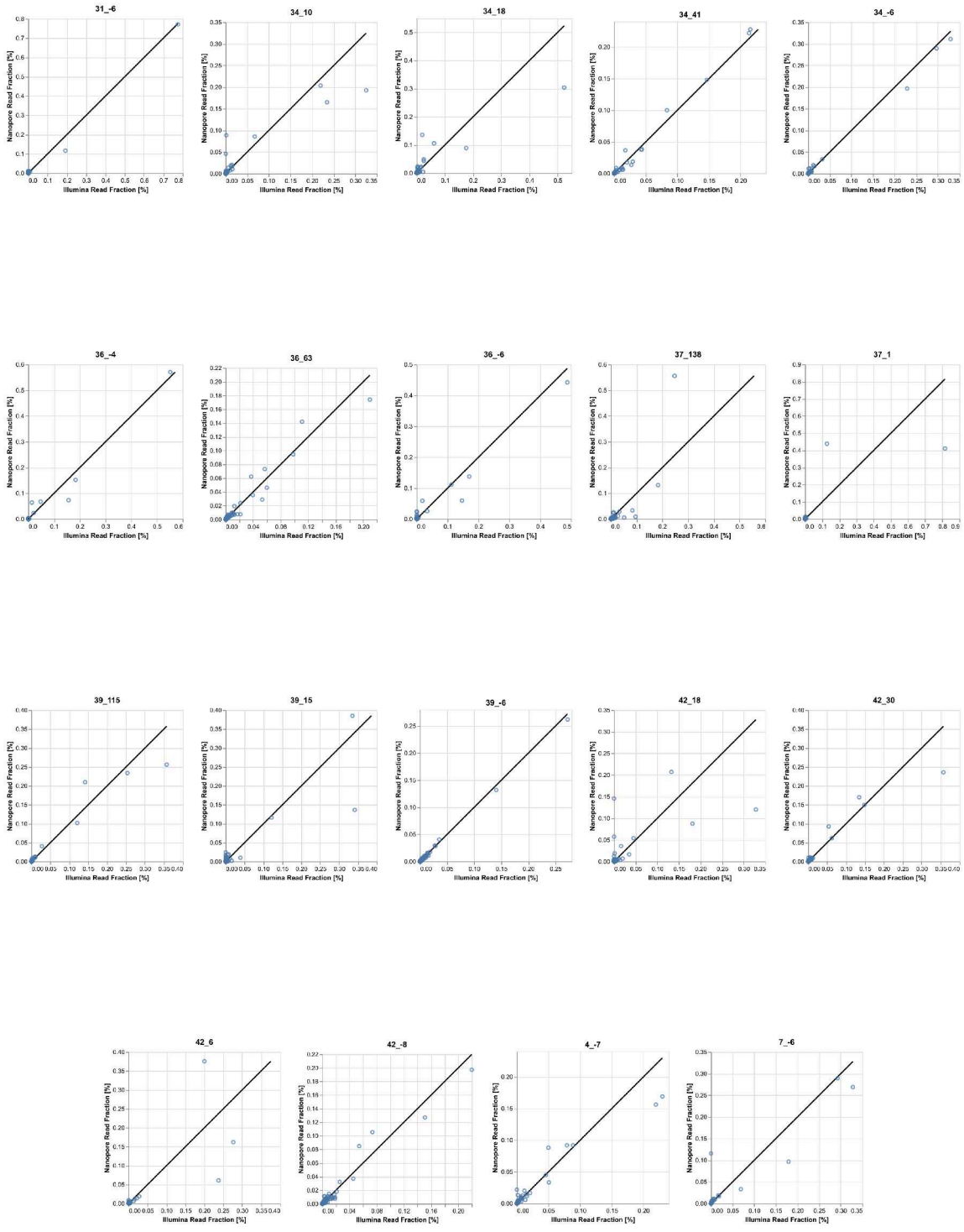
Supplementary Figure 10: Overview of high-confidence strain level distances. Distances for sequential samples within patients as histograms (Column a), aANI-lowFreq distances for sequential samples within patients relative to the elapsed time (Column b, in days), aANI-lowFreq distances between pre-Tx samples of different patients (Column c). In addition we plotted the respective distances from the simulation of *Bacteroides vulgatus* and *Escherichia coli* references (see Methods) and marked the highest observed aANI-lowFreq distance (right border) of the two preceding histograms in the row. Significant strain replacement events are marked with a star.



Supplementary Figure 11: Sampling timepoints. For visual clarity, time is shown on a symmetrical logarithmic scale and some overlapping labels were omitted. The patients are arranged depending on the composition of their pre-Tx sample (Cluster 1: lavender, Cluster 2: sage, Cluster 3: peach). Clinical adverse events are marked in gray with symbols (see legend).




16_2
16_-5
17_13
17_-1
17.2_154
17_-7
17_9
18_13
18.2_16
18_-3
18_91
19_153
19_-2
21_120
21_-3
21_50
21_-9
23_-9
24_0
24_-10
24_49
24_9
26_185
26_27
26_-3
28_1
28_21
28_237
28_-5
28_63
29_15
29_186
29_1
29_-5
31_58
Nanopore Read Fraction [%]
Illumina Read Fraction [%]



Supplementary Figure 12: Illumina-Nanopore agreement. X/Y plots showing the abundance for each sample that was processed with both Nanopore and Illumina and yielded at least 10,000 reads for each technology. Each sample was scaled down to 10,000 reads by random sampling. Mapping-based validation was not utilized here.

3.1.4 Discussion: A need for tailored methods

In addition to the discussion points mentioned in the paper we want to add additional perspective on the publication and the presented results.

Initially, one of the main goals was to analyze the pre-transplantation microbiome. However, in many cases the patients had already received therapy including antimicrobial drugs at the time of sampling. As a result, the observed compositions were not truly indicative of a starting condition prior to receiving any treatment. However, since many patients naturally received some form of therapy prior to receiving an allogeneic stem cell transplantation any comparison will have to account for those confounding factors.

In general, we observed that the quality of analysis is unsurprisingly highly dependent on the quality of the input data. In our case a comparison of samples was made more difficult since the time points of sampling varied significantly between patients. In addition, a large fraction of metadata was not available in machine-readable formats and needed manual pre-processing which is by nature more error-prone. While the data allowed us tackling our research question, the methods chosen to do so needed to be carefully adapted.

For a comparison of metagenomic profiles, we found that the actual labeling of taxonomic units is not as important and even "noise" can contribute to making a distinction between samples. As long as samples can be clearly subdivided into groups the constituting factors can instead be considered in subsequent analyses.

In contrast, when being concerned with the presence of organisms a signal needs to be clear and mapping is essential. This was especially relevant in our work when dealing with the samples collected during leukopenia, where the actual microbial DNA is sparse and thus a small amount of signal is used to infer biological meaning. Use cases like the analysis of cell-free DNA are the most extreme examples, here decisions need to be made based on a handful of DNA reads and decide about the plausibility of a taxonomic unit being present in our sample while excluding possible contamination sequencing artifacts and other confounding factors. The workflow "cfCop" [1] by Alona Tysha and Alexander Dilthey addresses those challenges, and we were in contact with the authors throughout our own research, adapting many ideas for our own analysis of sparse microbiomes.

We combined Kraken2-based profiling with a mapping-based validation to combine and while the approach was suitable we believe that there are clear advantages to using specific methods to address the questions. However, we also noted that the addition of the validation step did not change our analysis based on profiling, meaning that the same clustering of samples occurred regardless of whether we used the validation or not. This is due to the fact that "noise" of low-confidence or false-positive classifications was either evenly distributed across all samples or matched the existing groups.

A major realization from the paper affecting all potential research questions, was the disadvantage of using a full RefSeq database for taxonomic classification. While the impact on *k*-mer based approaches specifically had been shown [24], mapping-based strategies are confounded by multiple factors as well. First of all, the quantity of a species' sequenced individuals varies greatly with model organisms being vastly overrepresented. For example, *Escherichia coli* has a multitude of sequences and therefore a high representation of possible genomic configurations, transposable elements, and other sources

[1] `https://github.com/DiltheyLab/cfCOP`

of intraspecies variation. This results in the ability to map even unusual variants of the species. Nevertheless, it also means that a read with sequencing errors might have a perfect match in one of the numerous variants leading to a overestimation of intrasample diversity as reads can be spread across multiple references.

The quality of sequences also varies greatly with regards to completeness and accuracy. Some references have been shown to contain contamination: For *Toxoplasma gondii* we observed many false positive hits in our analysis that we could trace to human host contamination in the database. The contamination is plausible as *Toxoplasma gondii* is cultivated in human cells and we eventually found the respective sequences missing in newer RefSeq versions.

From our perspective there are superior approaches: We became aware of the classifier "MetaPhlAn" [37] in context of a cooperation with Nicola Segata who was designated Manchot Guest Professor at the HHU in 2023. At the core of the MetaPhlAn software is a database of unique marker genes derived from a large assembly of gut microbiomes. The assemblies are grouped based on their similarity and representatives are chosen for each group. This comes with several advantages: A large fraction of not yet described organisms is covered, and the taxonomy is based exclusively on genetic similarity. Notably, this leads to some species with a high diversity being represented as distinct groups while other organisms that are historically considered separate entities are being grouped together. We subsequently started a cooperation with the group and exchanged ideas about how to adapt the approach for long reads. In subsequent pipelines where we focused on sparse microbiomes it was feasible to use the representative genomes (and not just the marker genes) for mapping. Linda Cova has since expanded MetaPhlAn to support long reads, and we believe this will be a gold standard in metagenomic profiling moving forward.

3.2 Dynamic Mortality Risk Prediction in Myelodysplastic Syndromes Using Longitudinal Clinical Data

3.2.1 Motivation: Continuous re-evaluation of risk scores

While genomic data presents exciting novel possibilities to improve patient quality and generate foundational biomedical insights, generation of sequencing data is not trivial and comes with associated costs that can restrict an every-day implementation in the clinical setting. Nevertheless, with the ongoing digitalization of medicine a variety of other medical datasets are becoming available to be used for computational processing and analysis, enabling therapeutic decision support.

Although the vast amount of data available to clinicians can be immensely insightful, it can also be overwhelming. After in-depth conversations with practicing clinicians, we found that in many cases the majority of information available is rarely assessed simultaneously and therefore often lacks important context. Since clinical information systems typically only display a limited amount of measurements or show short time spans, an evaluation of trends is difficult for a human user. Even more challenging is the detection of multivariate patterns such as a slight increase in one value occurring concurrently with two subtle decreases in other values.

To address these limitations in how clinicians interact with data, we began developing a modern, interactive visualization tool for medical records that can display multiple data types and integrate them in intuitive ways. The tool, called MedReactor, got expanded in the context of a bachelor's thesis by Rebecca Fröhlich [38] and eventually provided a couple of decisive advantages: The web-based implementation enabled usage on any device with a browser, medication plans could be manually entered (as they are currently not tracked in a medical information system), and even microbiome compositions could be integrated into the overview. While this makes it easier for a clinician to be aware of clear, apparent effects and track known risk factors, an accurate risk prediction might be based on the interaction of numerous variables–patterns that are nearly impossible to identify for human observers.

Classical on-admission scores are still the state-of-the-art for clinical risk assessment and frequently used to assess the state of a patient on admission. However, they are usually based on a small set of parameters and inherently static. As a result they can not capture the state of an actual patient throughout the clinical course adequately. As a tool for population-based analysis they provide value retroactively by showing how a score can separate patients into two or more classes and how those correlate with specific outcomes. Since they do not indicate where a patient is within its class they offer very limited support for clinical decision making.

As a consequence, we identified the task of clinical risk assessment as natural expansion to the aforementioned visualization tool. In the following paper, we present a random forest-based risk prediction system that is able to identify multivariate patterns and can re-evaluate a patient's mortality risk on a daily basis, achieving high sensitivity and specificity scores. We use features that are derived from the clinical time series to capture the dynamics across the observation period. Risk scores are translated into a classification of low, moderate, or high risk, thus leading to a live monitoring system which can be understood intuitively and aid clinicians in their decision making.

MACHINE LEARNING
Health

PAPER

Dynamic prediction of mortality risk following allogeneic hematopoietic stem cell transplantation

Philipp Spohr[1,2,5,*], **Rebecca C Fröhlich**[1,5], **Sebastian Scharf**[3], **Anna Rommerskirchen**[3], **Jonathan Bobak**[4], **Sarah Schweier**[1,2], **Paul Jäger**[4], **Guido Kobbe**[4], **Sascha Dietrich**[4], **Alexander T Dilthey**[3], **Birgit Henrich**[3], **Klaus Pfeffer**[3,5], **Rainer Haas**[4,5] and **Gunnar W Klau**[1,2,5]

[1] Chair Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[2] Center for Digital Medicine, Düsseldorf, Germany
[3] Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, University Hospital Düsseldorf, Düsseldorf, Germany
[4] Department of Hematology, Immunology, and Clinical Immunology, Heinrich Heine University Düsseldorf, University Hospital Düsseldorf, Düsseldorf, Germany
[5] Shared First Author/Shared Last Author.
[*] Author to whom any correspondence should be addressed.

E-mail: philipp.spohr@hhu.de

Keywords: alloHSCT, mortality prediction, transplantation, time-series, clinical records, risk stratification

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
20 March 2025

REVISED
10 July 2025

ACCEPTED FOR PUBLICATION
4 August 2025

PUBLISHED
14 August 2025

Abstract

Allogeneic hematopoietic stem cell transplantation (alloHSCT) is a potentially curative treatment for high-risk hematological malignancies, but early mortality within 100 d remains a significant challenge, affecting approximately 18% of patients. Identifying high-risk patients early is critical for timely interventions. While static risk scores like the hematopoietic cell transplantation-specific comorbidity index and endothelial activation and stress index are valuable, they do not adapt to patients' clinical trajectories. Leveraging the increasing availability of digital medical datasets, we developed a dynamic risk stratification system based on a random forest approach to continuously predict mortality risk from day 0 to day 30 post-transplantation. Analyzing data from 847 patients undergoing alloHSCT between 2004 and 2019, our model achieved high classification performance, with AUROC scores continuously exceeding 0.8 from day 17, and effectively stratified patients into high, moderate, and low-risk groups. Feature importance analysis revealed a shift over time, with early predictions utilizing diverse variables and later stages relying more on inflammation-related bloodwork parameters. This dynamic approach surpasses static methods, demonstrating improved predictive performance and adaptability across treatment phases. To promote reproducibility and broader application, we provide open access to the software and accompanying bloodwork time-series dataset.

1. Introduction

Allogeneic hematopoietic stem cell transplantation (alloHSCT) is an efficient treatment for patients with high-risk hematological malignancies and for many of them the only option with curative potential. While the overall mortality of alloHSCT has steadily decreased due to reduced intensity conditioning regimens and advances in supportive therapy, a significant number of patients (approx. 18%) still die early within 100 d after transplantation, due to early relapse, infections, or graft-versus-host disease (GvHD) [1]. The 100 d mortality is a common indicator for the combined risk a patient is exposed to.

Identifying high risk patients is crucial to provide early extended diagnostic procedures and therapeutic intervention. Clinical risk scores such as the hematopoietic cell transplantation-specific comorbidity index (HCT-CI) [2] or the endothelial activation and stress index (EASIX) [3] are valuable predictors for mortality

risk, however, without taking into account the patient's clinical course, as they are typically determined only at admission and not continuously updated.

The increasing digitalization of medical datasets allows the utilization of a wide set of parameters and therefore enables the use of machine learning-based algorithms for risk prediction. Although final therapeutic decisions remain with physicians, such systems can be of valuable assistance. For their acceptance, explainability remains an important factor. While approaches based on complex machine-learning algorithms such as convolutional neural networks show promising results [4], the interpretability of their predictions is difficult. In contrast, tree-based models naturally facilitate transparency by providing interpretable feature importances. Shouval *et al* were able to train a decision tree-based model to predict 100 d mortality after alloHSCT that achieved an AUROC score of 0.701 using only 20 variables [5]. Arai *et al* achieved an AUROC score of 0.622 predicting acute GvHD using an alternating decision tree implementation [6]. Nevertheless, both models only provide a prediction at the time of admission that is not updated throughout the treatment. Models that are continuously updated have also been developed, specifically for high-frequency data ([7, 8]) but remain rare for more sparse data and are also challenging to easily evaluate by clinicians.

Thus, a dynamic risk stratification utilizing clinical and laboratory data sets would be a significant step forward to make predictions accessible.

Based on a group of 847 patients who underwent allogeneic HSCT between 2004 and 2019 at a single center, we present a risk stratification system that continuously evaluates each patient's mortality risk from day 0 to day 30. Our results show that the system achieves a high classification performance (AUROC scores continuously over 0.8 starting at day 17) and accurately stratifies patients into three risk groups: high, moderate, and low. We also demonstrate the method's increasing performance over time and observe a shift in feature importance. Early predictions utilize a broad range of features, while predictions between days 7–21 are mainly based on inflammation-related bloodwork parameters. This dynamic adaptation underscores the ability of our method to adjust its focus based on the patient's actual clinical course.

2. Results and discussion

2.1. A method for dynamic risk stratification

We present a machine learning-based risk assessment system that utilizes time-series data of bloodwork in addition to static parameters such as primary diagnosis or age to continuously evaluate an individual risk for each patient per day, starting from the day of transplantation as day 0 until day 30. Our system was trained and evaluated using a group of 847 patients receiving an allogeneic HSCT in the Department of Hematology, Oncology and Clinical Immunology of the University Hospital Düsseldorf between 2004 and 2019 (see figure 1). A random forest model classifying a patient as either 'deceased within 100 days' or 'alive after 100 days' was trained separately for every prediction day on the corresponding feature matrix and a subsequent risk stratification strategy assigns each patient to 'low', 'moderate' or 'high' risk so that physicians can be continuously provided with a risk assessment increasing their alertness for complications and permitting adjustments of the current therapy.

For a better understanding we publicly share the software alongside with the comprehensive dataset (static parameters: anamnesis, diagnosis, age, sex; dynamic parameters: microbiology data [pathogens/resistance testing]; bloodwork and biochemical parameters starting from day −7) for all patients undergoing alloHSCT, enabling full reproducibility of our work and its application to other cohorts (see materials and methods).

2.2. Group assignment reliable for short-term mortality

The assignment to the risk group showed good overall performance with deceased patients frequently getting assigned to high risk and likewise surviving patients getting assigned predominantly to low risk. Analyzing risk group assignments for the test set of 253 patients (24 deceased and 229 alive beyond day 100 post-transplantation), we found that, over the entire observation period of 30 d, on average 45.93% of deceased patients were classified as high-risk in comparison to only 12.36% within the surviving patients. On the other hand, an average of 63.12% of the surviving group were classified as low-risk compared to only 29.83% in the group of deceased patients (see figure 2, Panel (A)). In addition, the patients classified as low-risk within the group of deceased patients died primarily between days 76–100, suggesting better model performance for adverse events early after transplantation (see figure 2, Panel (B)). It was interesting to note a general increase in the proportion of high-risk patients in both, the surviving and deceased patients, up to day ten, which relates to the most critical period of severe neutropenia (which is considered a strong risk factor for bacterial and invasive fungal infection [9]). Thereafter, we observe a decreased risk in the surviving group of patients with none of them remaining in the high-risk category on day 16. Different from that

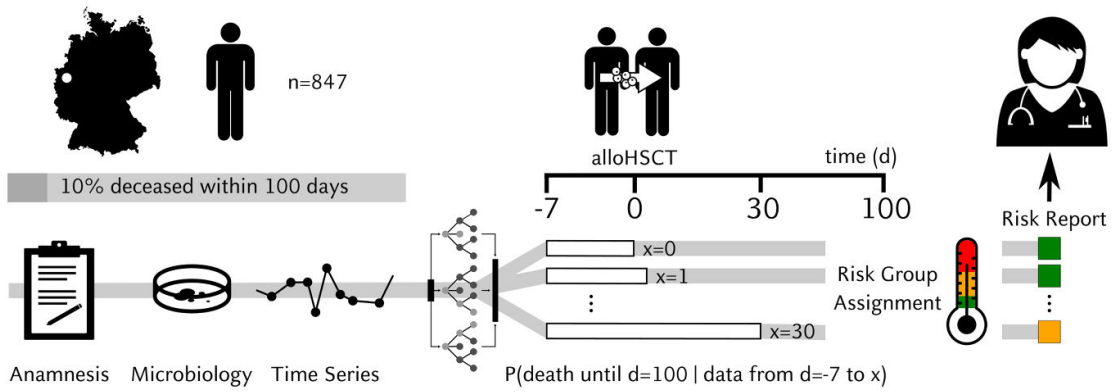


Figure 1. Overview of the study and method. Our cohort consists of 847 patients from the Department of Hematology, Oncology and Clinical Immunology of the University Hospital Düsseldorf who received an alloHSCT. Of those, 85 died within 100 d of transplantation. Our machine learning model predicts the patient's risk of dying within 100 d, starting from the day of transplantation up to 30 d afterwards. The data consist of static anamnesis parameters such as diagnosis, age/sex; microbiology data such as detected pathogens/resistance testing as well as continually measured parameters such as hemoglobin, blood cell count and biochemical parameters starting from day −7 to the day of prediction. In addition, derived features describing the series of aforementioned parameters are calculated. We utilize a random forest approach: Multiple decision trees assess the risk independently and vote for a decision. Finally, we translate the risk probability into a discrete risk classification, labeling each patient on every prediction day as either 'low', 'moderate' or 'high' risk, thus providing easily parable live assistance to clinicians.

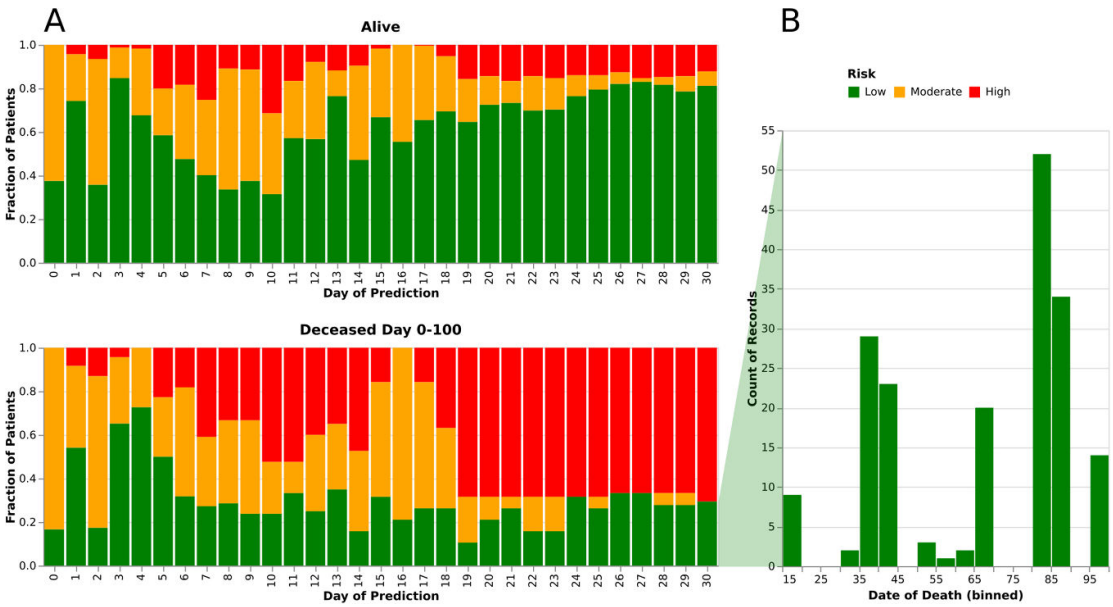


Figure 2. Overview of the risk group assignments for the test data set. (A) The fraction of patients assigned to a specific risk group is shown for each prediction day separately for the two outcomes. (B) The time point of death after alloHSCT (day 0) is shown for all 'low-risk' predictions made for deceased patients.

pattern, the deceased patients also show a tendency towards a lower risk, but during the following three days their proportion within the high-risk group is sharply raised to approximately 70%. This is of particular relevance as around day 17 the majority of patients should have a hematological (neutrophil) recovery which is associated with a decreased risk of complications [10]. The persistence of a patient in the high-risk group beyond day 17 is therefore a strong indicator and alarming sign that the mortality risk for the patient is high and potentially related to organ complications not necessarily related to the immunocompromised status. In general, patient risk trajectories were mostly constant progressions across risk levels without abrupt shifts. Over time, the moderate-risk group shrinks which reflects an increasing model certainty. Notably, individual misclassifications such as some high-risk patients surviving beyond 100 d still represented the status of patients well (See 'Misclassifications reflect clinical complexity'). Supplementary figure 1 illustrates all individual patient trajectories.

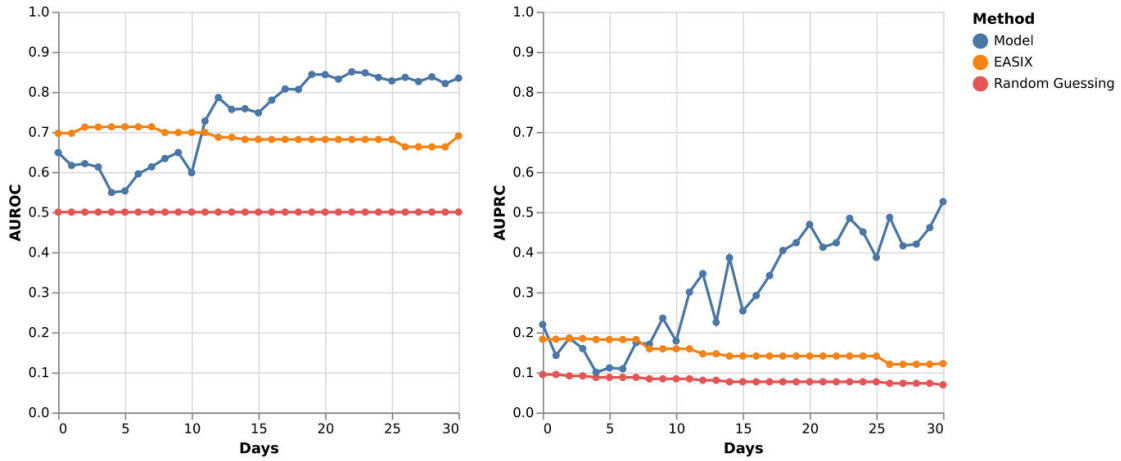


Figure 3. Overview of the model performance for individual days. Shown are (left panel) the area under the receiver operator curve (True Positive Rate vs. False Positive Rate) as well as (right panel) the area under the precision–recall curve for each prediction day. As a baseline we also plotted the performance of random guessing (assigning a random score between 0 and 1 to each row): for the AUROC this value is a constant 0.5 since FPR and TPR are symmetrical for random guessing; for the AUPRC the random guessing value depends on the class weights since the precision is affected by it.

2.3. Model predictive performance increases with time

As a basis for risk prediction, we trained a random forest model to predict 100 day mortality. Our model reaches a mean AUROC score of 0.74 (Std. Deviation 0.10, Min. 0.55, Max 0.85) and a mean AUPRC score of 0.31 (Std. Deviation 0.13, Min. 0.10, Max 0.53). We observed a steep increase in prediction quality starting at around day 10 (see figure 3). The overall lower stability of the AUPRC is the result of an overall small test set size and class imbalance; in the test set only 9.8% (Day 0)—6.9% (Day 30) of patients died. Notably, the model achieves an AUROC score of 0.648 for day 0 indicating an already usable risk prediction on the day of transplantation. As an additional baseline for comparison we included a classification using the EASIX score on day 0, which was found to be prognostic for overall survival [11], as a threshold. This was only possible for a subset of patients in the test set where the required measurements of LDH, Creatinine and Thrombocytes were available, limiting the direct comparability. Notably, our model does not have access to LDH as a feature since the value was not measured consistently for the majority of patients, possibly explaining the inferior performance in the early days. Nevertheless, the superiority of a dynamic score becomes apparent here, since the prognosis based on the EASIX does not take individual trajectories into account leading to an increasing disadvantage. Furthermore, the lack of separation between risk groups using only the EASIX score on day 0 also becomes clear when analyzing the Kaplan-Meyer survival curve (created on all patients from the training set where the calculation was possible) stratified into quartiles since there is no clear and consistent inter-group distinction (see supplementary figure 4).

2.4. Feature importance shifts throughout treatment course

The feature importance analysis highlights a strong emphasis on bloodwork values, particularly C reactive protein (CRP) and urea (HST), suggesting a correlation with inflammatory, renal function and protein catabolism (see figure 4). Variables such as age, sex, and other static parameters mostly achieve relevant feature importance values early on and late, whereas the bloodwork values dominate between days 7–24, indicating that the method prioritizes the patient's current status rather than static characteristics in this 'critical phase'. This is also reflected in the model's preference for a smaller amount of target features in the feature selection process in this timespan (see supplementary table 2). Clinically, we can correlate this to a delayed onset of inflammatory reactions, caused either by the therapy or infectious processes. The return to a broader set of features, starting at approximately day 24, reflects the neutrophil engraftment which at this point has completed in the majority of patients [12]. In general, we observed that the model gives measurements closer to the prediction day a higher importance (see supplementary figure 2). We noted that the resistance/pathogen features achieved non-zero importances and might therefore be worth exploring in risk assessment systems.

2.5. Misclassifications reflect clinical complexity

The model predictions generally align well with clinical risk assessments of physicians, while there are certainly mismatches in some patients, reflecting the multifactorial complexity of patient outcomes. For instance, there are six patients (B, D, F, G, H, I) predicted to be low risk (green) who died

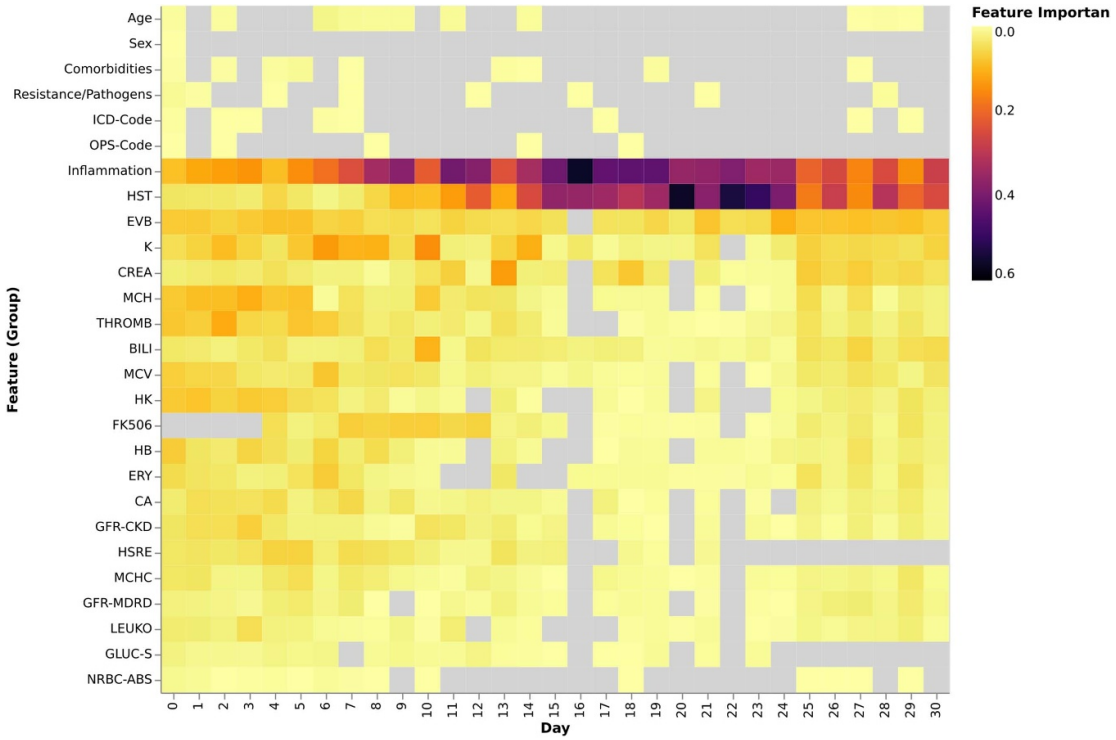


Figure 4. Overview of the feature importances. Features are grouped and their cumulative feature importance (Gini impurity decrease) is encoded via color. Age is a single feature describing the age at transplantation; sex describes the biological sex of the patient; comorbidities aggregates all known comorbidities prior to prediction; resistance/pathogens contains microbiological pathogen detections and resistance testing; ICD-code contains primary diagnosis information; OPS-code contains information about the transplantation; inflammation contains c-reactive protein and interleukin 6; HST urea; EVB red blood cell distribution width; K potassium; CREA creatinine in serum; MCH mean corpuscular hemoglobin; THROMB thrombocytes; BILI bilirubin; MCV mean corpuscular volume; HK hematocrit; FK506 tacrolimus; HB hemoglobin; ERY erythrocytes; CA calcium; GFR-CKD glomerular filtration rate (CKD); HSRE uric acid; MCHC mean corpuscular hemoglobin concentration; GFR-MDRD glomerular filtration rate (MDRD); LEUKO leukocytes; GLUC-S glucose in serum; NRBC-ABS nucleated red blood cells; see 'data preprocessing and feature engineering' for a detailed overview of all features. Features that are not used on a specific day for prediction are greyed out: this might be because (1) the feature was eliminated during feature selection, (2) the measurement was not available on that specific day or (3) the feature simply was not utilized by any decision tree.

post-transplantation of complications with 'sudden onset' such as sepsis, multi-organ failure, or pulmonary embolism. Another example is patient F who was discharged in good health on day 21 following allogeneic HSCT. The patient developed a lethal acute liver and kidney failure without leading anteceding symptoms resulting in death on day 84. Similarly, patient I succumbed to a sudden cardiac death following a lung embolism.

On the other hand, there are examples where the model predicted high risk (red), while the patients survived. For instance, patient K was correctly classified as high-risk because of severe complications like invasive aspergillosis and acute GvHD. Fortunately, the patient survived, reminding us that probability assessments do not permit us to forecast individual fates. We also found that predictions are more meaningful when patients are inpatients and thus clinical and laboratory parameters obtained regularly. For example, Patient D was followed as an outpatient between days 19 and 28 resulting in fewer measurements during this period. The risk scores are illustrated in figure 5, clinical annotations for patients A–K are provided in supplementary note 1. Supplementary figure 1 illustrates all individual patient trajectories. See supplementary note 2 for a mapping of letters to patient identification numbers used in the provided dataset.

2.6. High risk predictions can precede risk recognition by clinicians

Determining whether the risk prediction system can assist the judgment of experienced clinicians and is able to predict critical cases earlier remains a challenging task. To address this issue we used the judgment of our author RH, an experienced clinician who manually examined the features for 19 deceased patients of the test who were classified as high-risk by the algorithm at some point post-transplantation (see supplementary note 2 for a mapping to patient IDs used in the provided dataset as well as the full clinical assessment). Thereby, we identified patients for whom the system provided added value for the clinical reasoning.

For instance, the model predicted Patient X (deceased on day 19) to be at high risk for the first time on day 8. While both the algorithm and clinical assessment eventually considered the patient as high risk, this

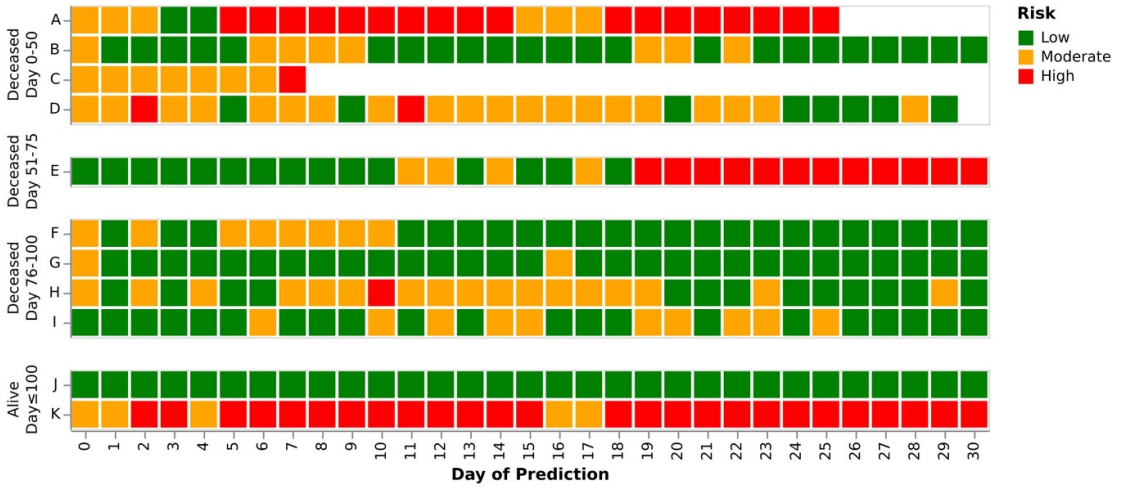


Figure 5. Exemplary risk prediction trajectories for a set of patients (with misclassifications overrepresented). **Patients A-K** are grouped by their outcome (day of death/or no death within 100 d) and their risk assignment is shown for every prediction day. 'White' missing squares indicate that the patient will be deceased within 5 d.

classification would not have been made that early based solely on clinical judgment since the sole results of the bloodwork did not show any obvious increased risk in comparison to the anteceding days.

Similarly, Patient Y was first flagged as high risk on day 7 and died on day 39. The shift in risk assessment from moderate to low risk occurring from day 1 to day 2 appeared implausible; but the upgrade from day 4 to day 5 to moderate risk would probably not have been made by clinicians. The same is true for the upgrade to high risk, as in the light of stable CRP values clinicians might have underestimated the continuous rise in urea levels observed since day 4. The course of this patient might serve as an example for the utility of the algorithm in recognizing complex patterns with subtle changes that may be missed by purely human reasoning.

Far from being a state of the art prospective comparison, these examples show the value of algorithmic systems to complement clinical decision-making. In total, we found that for 11 of the patients the algorithm made a high-risk classification earlier than the clinician would have done. On the other hand, for five patients and 3 patients the clinician's upgrade to high risk was somehow earlier or at the same time, respectively.

2.7. General response strategy to high risk classifications

In the case of a high-risk classification after allo-HSCT, various intensified diagnostic and therapeutic measures can be considered that go beyond the standard procedure. Together, they enable an individualized approach at an earlier stage in a risk-adapted manner.

In the case of a high risk classification in the time period between days 7 and 24 we suggest daily blood cultures even without persistent fevers, extended PCR-based viral detection for CMV, EBV, and HHV-6 at least two times per week as well as fungal detection strategies such as fuPCR [13], Galactomannan antigen test, or the 1,3-β-D-Glucan test. Additional infection markers such as IL-6 or PCT can be measured. We would also lower the thresholds to perform an early CT thorax, a preemptive bronchoalveolar lavage, or biopsy sampling in case of GvHD suspicion: The benefits of an earlier process detection outweighs potential harm caused by those interventions.

With the same justification, we could envision therapeutic intensification. The antibiotic treatment could be switched to broad-spectrum preparations such as Levofloxacin to pre-emptively treat undetected infections. In the same way the antifungal therapy can be escalated (From Fluconazole to Voriconazole, Posaconazole, or Amphotericin B). In case of an observable—although currently non-critical—PCR dynamic in combination with a high risk score, an early use of Letermovir, Ganciclovir or Foscarnet could anticipate a CMV infection. Similarly, steroids could already be given even at very early stages of GvHD (e.g. stage 1 skin GvHD). Additional immunosuppressants are another possibility in the case of a high risk score even before clinical manifestation.

Finally, a high score should lead to an increased attention to the patient's status by the therapeutic staff, intensifying the monitoring of all aspects.

Our risk score provides a valuable basis for taking pre-emptive measures before clinical complications become manifest. The combination of targeted diagnostics, escalation of prophylactic and therapeutic

measures and structured clinical decision-making holds the potential to significantly reduce morbidity and mortality.

2.8. Additional and improved data sources hold large potential

The next step should be an evaluation of our findings on an independent dataset of alloHSCT patients from other centers to see whether our AI model is of general value. We invite other researchers to apply our method on their datasets and use our dataset to validate their methods. Eventually, a prospective study where physicians assign patients to risk groups to compare their assessments to our model's predictions could show a manifest clinical benefit but is challenging taking into account the great number of patients needed to be recruited within a reasonable time.

With all these caveats and shortcomings of a retrospective study in mind we believe that our results are at least a proof-of-principle for the utility of real time AI based predictions to assist the physicians in their day to day clinical reasoning. Further refinements of the system can be achieved by incorporating more parameters and features such as time from diagnosis to transplantation, the number of prior transplantations, the graft source, and previous iron chelation therapy could be included [14]. This process could be further fostered by the growing digitalization of the patients data including vital parameters or the results of imaging examinations such as CT or MR scans.

3. Materials and methods

3.1. Software and data availability

The software is available as open source on github https://github.com/AlBi-HHU/dynamic_mortality. The software makes use of scikit-learn [15] and tsflex [16] and is implemented as a Snakemake [17] workflow, allowing ease-of-use and scalability to large datasets and cloud computing environments.

3.2. Patient cohort and characteristics

The main cohort was drawn from cases admitted to the UKD for alloHSCT in the years from 2004 until 2019 (see figure 6). For a detailed overview of the characteristics and the specific ICD-10 GM codes refer to supplementary table 1.

3.3. Data preprocessing and feature engineering

The raw data extracted from the system is transformed into one matrix per prediction day where every row is a case and every column is a feature. The following features are generated:

Demographic features sex is encoded as a binary feature, Age in days is encoded as an integer.

Comorbidities The risk groups are modeled after the HCT-CI, namely: Arrhythmia, Heart disease, Chronic inflammatory bowel diseases, Diabetes, Cerebrovascular diseases, Psychiatric disorders, Liver dysfunctions, Obesity, Infections, Rheumatic diseases, Gastric ulcer, Kidney dysfunctions, Lung dysfunctions, Tumor. For each group, the amount of associated diagnoses documented for a stay with a discharge date within 365 d prior to transplantation is encoded as an integer. See Supplementary Note 3 for a mapping of ICD-10-GM codes to the groups.

Resistances & pathogen detection microbiome resistance testing is encoded as a binary feature for each combination of pathogen, drug and both aerobic and anaerobic blood cultures.

Three additional binary feature columns are generated for the confirmed detection of MRSA, VRE, and MRGN within a year prior to the HSCT respectively.

ICD codes and OPS codes every OPS[1] Code (encoding donor/receiver match, blood vs. peripheral, and HLA match) and every main diagnosis ICD-10-GM[2] Code in the dataset is represented as a binary feature (meaning the patient has the code documented in the hospital stay across transplantation as a primary diagnosis).

Blood work measurements and time series we extracted all clinical blood work parameters starting from seven days before transplantation up to the day of prediction from the laboratory information system. Blood values are filtered if the median number of measurements across all measured patients is below 80% of all days between −7 and prediction. If multiple measurements exist on the same day the median is chosen, missing values are replaced with a constant value of −1000. For the values we calculate derivative time series

[1] www.bfarm.de/EN/Code-systems/Classifications/OPS-ICHI/OPS/_node.html, last accessed: 31 January 2025.
[2] www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-10-GM/_node.html, last accessed: 31 January 2025.

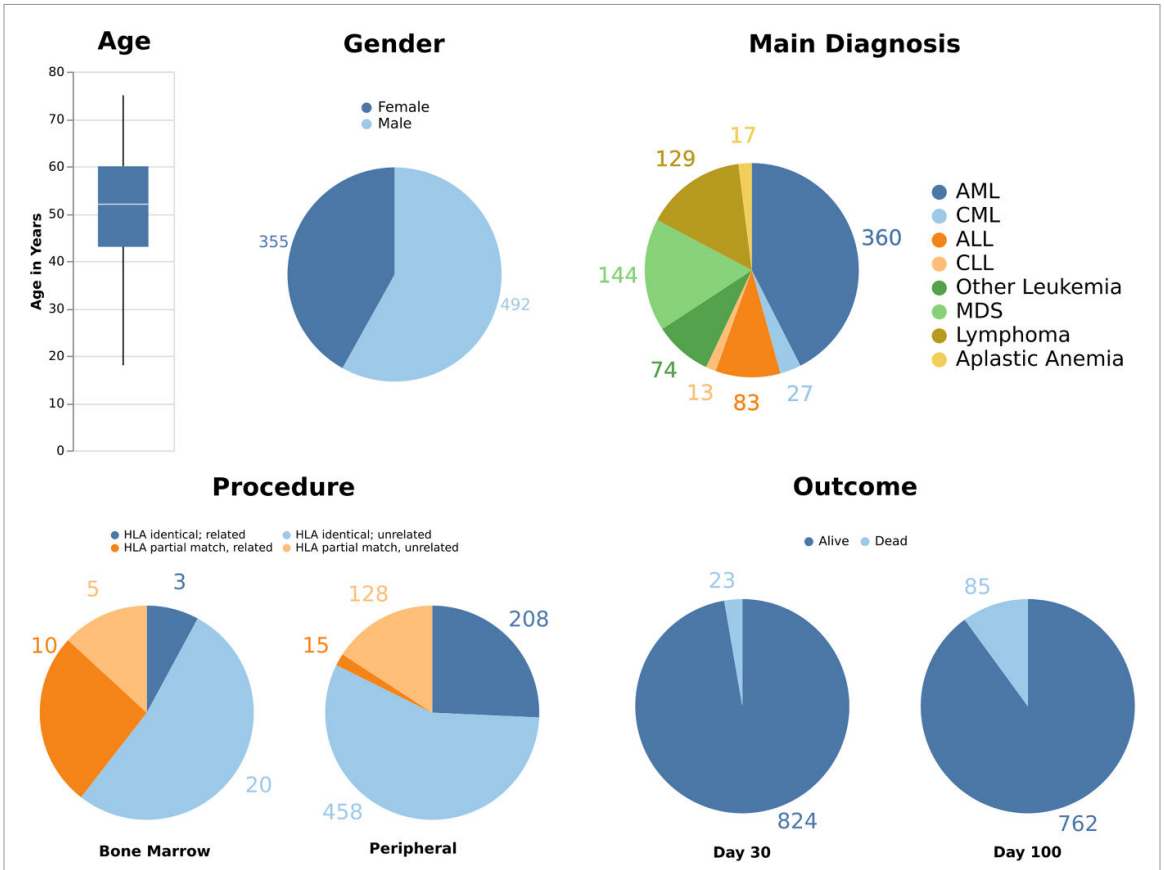


Figure 6. Overview of the main cohort from the University Hospital Düsseldorf and its characteristics.

features as feature columns in addition to the raw measurements, refer to Supplementary Note 4 for a full overview and explanations.

The time series features were calculated using tsflex which allows gaps within the time series; thus, the constant values of −1000 are only used for the raw measurement column.

Rows are only labeled as deceased within 100 d if a clear death event was recorded, and only labeled as alive if data indicating that is available after 100 d, otherwise the row is treated as 'lost to followup' and removed.

Rows that contain a death within 5 d after prediction are also removed to avoid prediction of imminent events (Note that this leads to a decrease of 33 rows over time).

3.4. Model training and evaluation

A random forest model classifying a patient as either 'deceased within 100 days' or 'alive after 100 days' is trained separately for every prediction day on the corresponding feature matrix. We want to note that we also tested XGBoost, Logistic Regression and Single Decision Trees as classifiers. XGBoost performed slightly but consistently worse while requiring more computational time for training, logistic regression and single decision trees performed significantly worse.

Since the number of rows decreases over time we aim to have a consistent test and training set (and thus ensuring comparability) for all prediction days by performing a stratified split on all rows that are contained on every prediction day into 30% test data and 70% training data.

The remaining 28 rows—representing patients deceased within 35 d—are then successively assigned to the sets such that for every day the split ratio stays as close to 30/70 as possible.

While we did not control for other aspects of the datasets, we notice that both sets are similar with respect to age, gender, performed procedure, and main diagnosis (see supplementary table 1).

To optimize the hyperparameters for each method, five-fold stratified cross validation with five repeats is used in combination with a full grid search: For every possible hyperparameter choice the training set is split five times into five disjunct folds, four of which are used to train the model and one of which is used for evaluation. In the end, the hyperparameter choice resulting in the highest average F1 score is used to train the model on the entire training data and then evaluate it on the test data.

3.5. Hyperparameter settings

For our random forest model we make use of trees with depth 2. For every decision in the tree we check all possible splits and take the one resulting in the highest impurity decrease. The other parameters were determined during the training using a full grid search with the following options (where m is the total number of available features):

Number of features per tree (maxFeatures): log2(m), sqrt(m)
Minimum amount of samples per leaf (minSamplesLeaf): 2, 4, 7
Number of decision trees (n_estimators): 100, 150, 200, 250
Target number of features for the feature selection: 20, 40, 100, 200, m

Refer to supplementary table 2 for the chosen hyperparameters for each prediction day.

3.6. Feature importance and selection

Feature selection is performed based on mutual information.

Reported feature importances were calculated using the rf.feature_importances_ function from scikit (mean of the Gini-impurity decrease across all trees).

3.7. Risk group assignment

To determine the assignments of patients to either the 'low', 'moderate' or 'high' risk groups on a specific day two thresholds θ_H and θ_L on the predicted scores are used. A patient achieving a prediction score of $\theta \geqslant \theta_H$ is assigned to the high risk group, a patient achieving a score of $\theta \leqslant \theta_L$ is assigned to the low risk group. Remaining patients are assigned to the moderate risk group. Using the training dataset, the threshold θ_H is chosen such that at most 20% of non-deceased patients fall into the high risk group. Analogously, the threshold θ_L is chosen such that at most 5% of deceased patients fall into the low risk group. In other words, we control the False Positive Rate to be less than 20% and the False Negative Rate to be less than 5%.

Thus we consider it worse to miss a patient with a high risk than to assign a patient a risk score that was higher than necessary; as a result the threshold for the low risk group is stricter. Although 'alert fatigue' [18] and subsequent desensitization should be considered when choosing the desired False Positive Rate, we believe a value of 20% is reasonable since our system is supposed to only perform a 'once-per-day' evaluation and the resulting overtreatment is primarily in the form of additional non-harmful diagnostics.

We believe that in the case of clinical implementation, the parameters need to be chosen based on the local capabilities and requirements for the system. Based on the threshold-independent performance of our method (AUROC/AUPRC) as well as the trajectories resulting from consecutive risk assessments for individual days, we believe our method to be robust regardless of a specific parameter choice.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.14779813.

Acknowledgment

We want to express our gratitude to the nursing and medical staff of the UKD for supporting our work, collecting samples and data and providing excellent care to the patients. For providing excellent technical assistance we want to thank Max Jakub Ried.

Funding

Manchot Foundation Project: Research Group 'Decision-making with the help of Artificial Intelligence' to AD, GKl, KP, PS, RH, SeS.

Ethics

The study was approved by the 'Ethikkommission an der Med. Fakultät der HHU Düsseldorf' (Study-Nr.: 2019-513).

Author contributions

RF and PS developed and evaluated the software. PS supervised the software development and evaluation. PS, RF and SeS wrote the initial manuscript draft; GKl and RH supervised manuscript preparation and

performed editorial revisions. PS, JB and GKl supervised the model development. SaS wrote software for data extraction from medical databases. SeS and PS programmed additional software for data extraction and processing of clinical records. RH and SD provided a detailed clinical perspective. AR coordinated ethics. PJ and GKo coordinated data acquisition. AD, BH, KP, GKl and RH initiated and supervised the project. All authors have contributed to the internal review process, read and approved the final manuscript.

ORCID iDs

Philipp Spohr 0000-0002-6039-377X
Sebastian Scharf 0000-0002-8299-0559
Anna Rommerskirchen 0009-0001-1524-0045
Jonathan Bobak 0009-0002-3022-5860
Sarah Schweier 0009-0001-7906-9881
Paul Jäger 0000-0001-9250-2064
Guido Kobbe 0000-0002-9745-8813
Sascha Dietrich 0000-0002-0648-1832
Alexander T Dilthey 0000-0002-6394-4581
Birgit Henrich 0000-0002-0565-5773
Klaus Pfeffer 0000-0002-5652-6330
Rainer Haas 0000-0002-3652-6595
Gunnar W Klau 0000-0002-6340-0090

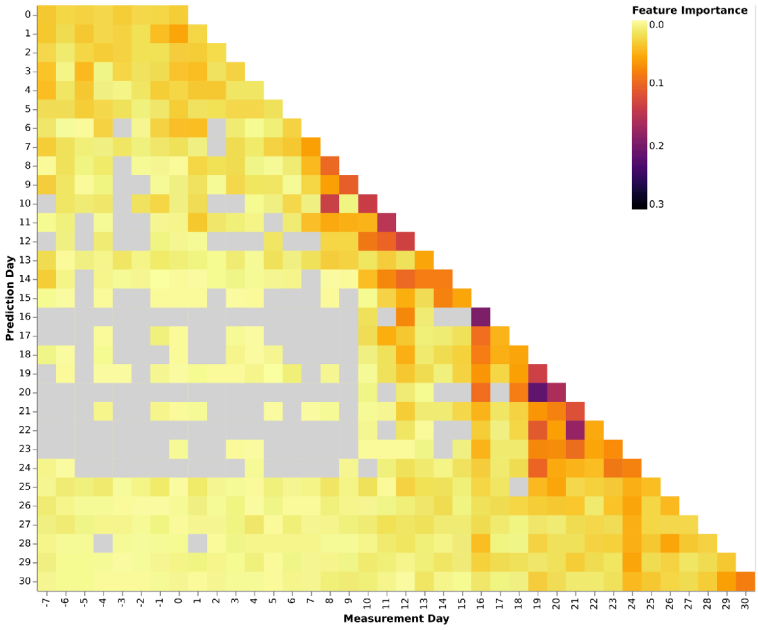
References

[1] Styczyński J *et al* 2020 Death after hematopoietic stem cell transplantation: changes over calendar year time, infections and associated factors *Bone Marrow Transplant.* **55** 126–36
[2] Sorror M L, Maris M B, Storb R, Baron F, Sandmaier B M, Maloney D G and Storer B 2005 Hematopoietic cell transplantation (HCT)-specific comorbidity index: a new tool for risk assessment before allogeneic HCT *Blood* **106** 2912–9
[3] Luft T *et al* 2020 EASIX and mortality after allogeneic stem cell transplantation *Bone Marrow Transplant.* **55** 553–61
[4] Jo T *et al* 2023 A convolutional neural network-based model that predicts acute graft-versus-host disease after allogeneic hematopoietic stem cell transplantation *Commun. Med.* **3** 67
[5] Shouval R *et al* 2015 Prediction of allogeneic hematopoietic stem-cell transplantation mortality 100 days after transplantation using a machine learning algorithm: a European Group for blood and marrow transplantation acute Leukemia working party retrospective data mining study *J. Clin. Oncol.* **33** 3144–51
[6] Arai Y *et al* 2019 Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation *Blood Adv.* **3** 3626–34
[7] Lapp L, Roper M, Kavanagh K, Bouamrane M-M and Schraag S 2023 Dynamic prediction of patient outcomes in the intensive care unit: a scoping review of the state-of-the-art *J. Intensive Care Med.* **38** 575–91
[8] Thorsen-Meyer H-C *et al* 2020 Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records *Lancet Digit. Health* **2** e179–91
[9] Bodey G P, Buckley M, Sathe Y S and Freireich E J 1966 Quantitative relationships between circulating leukocytes and infection in patients with acute leukemia *Ann. Intern. Med.* **64** 328–40
[10] Prébet T *et al* 2010 Platelet recovery and transfusion needs after reduced intensity conditioning allogeneic peripheral blood stem cell transplantation *Exp. Hematol.* **38** 55–60
[11] Jiang S *et al* 2021 Predicting sinusoidal obstruction syndrome after allogeneic stem cell transplantation with the EASIX biomarker panel *Haematologica* **106** 446–53
[12] Kim H T and Armand P 2013 Clinical endpoints in allogeneic hematopoietic stem cell transplantation studies: the cost of freedom *Biol. Blood Marrow Transplant.* **19** 860–6
[13] Scharf S, Bartels A, Kondakci M, Haas R, Pfeffer K and Henrich B 2021 fuPCR as diagnostic method for the detection of rare fungal pathogens, such as Trichosporon, Cryptococcus and Fusarium *Med. Mycol.* **59** 1101–13
[14] Kong S G, Jeong S, Lee S, Jeong J-Y, Kim D J and Lee H S 2021 Early transplantation-related mortality after allogeneic hematopoietic cell transplantation in patients with acute leukemia *BMC Cancer* **21** 177
[15] Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30
[16] Van Der Donckt J, Van Der Donckt J, Deprost E and Van Hoecke S 2022 tsflex: flexible time series processing & feature extraction *SoftwareX* **17** 100971
[17] Mölder F *et al* 2021 Sustainable data analysis with Snakemake *F1000Research* **10** 33
[18] Cash J J 2009 Alert fatigue *Am. J. Health-Syst. Pharm.* **66** 2098–101

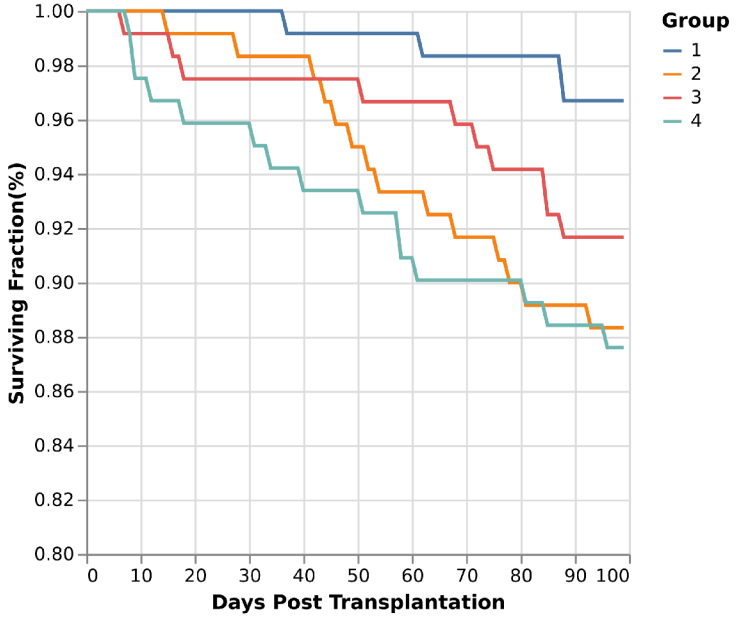
3.2.3 Supplementary Figures

Additional supplementary files can be found online and are omitted due to being unsuitable for print.

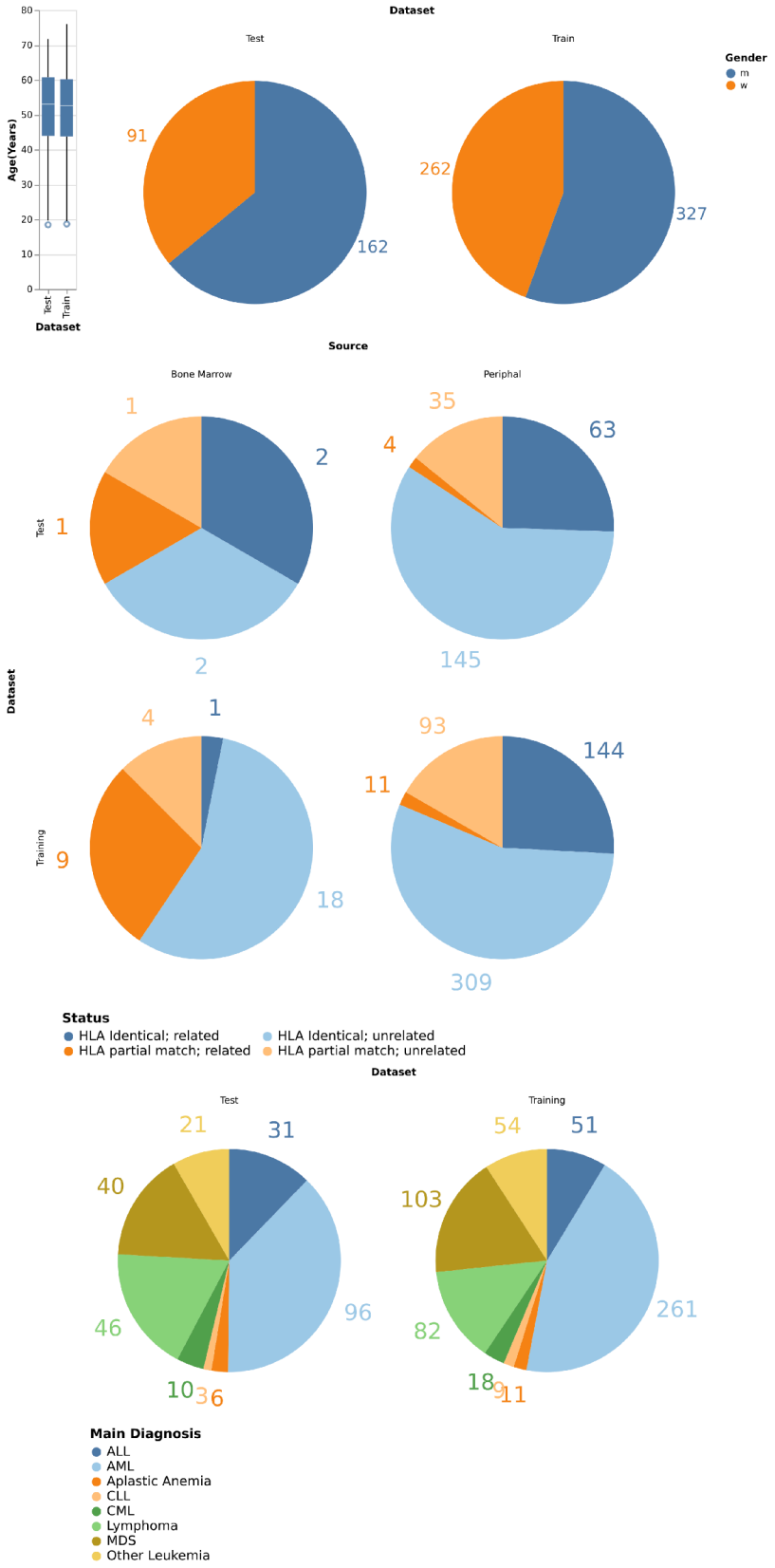
Supplementary Figure 1: Patient trajectories. Overview of all patient trajectories with their stratification into low (green), moderate (orange), and high (red) risk per day.



Supplementary Figure 2: Feature recency. Feature importances for the time series features only. The x-axis shows the day of measurement and the y-axis the day of prediction.



Supplementary Figure 4: EASIX-based survival curves. Stratification into four groups is performed based on the quartiles on admission.



Supplementary Figure 3: Specific overview of the cohort characteristics. All relevant characteristics are shown for both, the training and the test data set.

3.2.4 Discussion: Advocating for increased digitalization efforts

While the results are overall positive, the first obvious observation that warrants scrutiny is the relatively low improvement in solution quality that is attained by providing the time series features. In our use case the time series are rather "sparse" with only one data point per day being used. In practice, it is possible that fluctuations on a smaller scale might be meaningful. Additional parameters such as heart rate, blood pressure, temperature were not available but are also likely valuable features that unveil information when viewed as time series.

In a related project, we applied similar methods to a dataset containing patients suffering from aneurysmal subarachnoid hemorrhage treated on a neurosurgical intensive care unit. In contrast to the paper presented here, the data is collected from the monitoring devices and tracks the vital parameters with a high frequency. In this setting we notice a higher feature importance for the time series features and thus conclude that their benefit is higher in similar datasets.

The general robustness of this longitudinal approach has since also been shown on another dataset covering patients with myelodysplastic syndrome [39]. Here we were also able to obtain independent validation cohorts and could demonstrate generalizability. Contrasting with the other papers, the time series in this project cover a longer span of up to 32 quarters. We have therefore shown applicability on both larger and smaller timescales.

As of now, all methods are omitting features that are non-trivial to translate into machine-readable format such as the visual appearance of a patient or their spoken self-anamnesis. In an actual clinical setting, those features can have a significant impact on decision-making. As we view algorithmic risk assessment systems as aides for clinical practitioners who retain the ultimate responsibility, the omission of those features is less critical as those factors are naturally taken into consideration.

While not necessarily evident from the paper, a majority of the work that led to the publication consisted of obtaining, curating, and combining the data from multiple primary sources which in some cases where not digital but analog documents that needed to be transcribed. Many data points of interest such as medication regimens could not be reliably reconstructed at all. We believe that this was a major limiting factor and it reflects on the general state of digital maturity in Germany where many hospitals increasingly possess hardware and infrastructure but are still lacking in the implementation and utilization of information technology in the actual every-day processes [40]. A cultural component might be contributing to this since data privacy is valued highly among the German populace. This is not necessarily something we would rate negatively since an eventual digital infrastructure created in this climate has the potential to meet high ethical standards. Another reason why implementations are challenging, is the federal structure of the German national health care system, which makes creation of a unified digital infrastructure particularly challenging [41].

However, there are also projects like the ePA (electronic patient record) which, although slow in adaptation, explicitly mentions opportunities for scientific research as a goal [42]. Caused by the Krankenhauszukunftsgesetz (enacted on October 29, 2020) providing significant monetary incentives, a recent evaluation shows that, while major issues are still persisting, the overall trend is a positive one [43]. We believe that the ongoing efforts to increase digitalization in medicine are absolutely imperative to remain

competitive and provide the highest quality of patient care possible.

3.3 SWGTS–a platform for stream-based host DNA depletion

3.3.1 Motivation: Sample sharing in a connected world

During the COVID-19 pandemic we were collectively faced with unique challenges that also profoundly affected the way how we conducted research. At the same time the pandemic presented us with an opportunity to observe a real-life pandemic's infection dynamics and utilize modern research to help understand transmission pathways and inform decision making.

At the core of research projects conducted at the HHU and UKD was the genotyping of collected samples, which was then subsequently primarily used for phylogenetic inference but also other projects such as the detection of viral quasispecies (multiple COVID-19 strains present in the same species) and more base level research such as the PanCoV workflow for Nanopore technology error correction [44].

The collection of DNA samples was performed with nasal swabs that naturally contained host DNA of the respective humans in addition to the target DNA of interest. As this data is highly sensitive many researchers were confronted with conflicting objectives: Collecting and sharing as much data as possible would enable large-scale reconstructions of phylogenies with a high resolution so it appeared desirable to pool sequencing efforts by collecting data at centralized servers. However, it also came at the risk of exposing unwanted personal information to parties that should not have access to the data.

Following the fundamentals outlined in Article 5 of the DSGVO a number of critical goals were identified:

The fourth principle, minimization of data (Datenminimierung), dictates that data may only be collected to the degree that is required for the scope and purpose of the task. Since human data is not required for a study as described above, it should not be part of the collected data and therefore be excluded as quickly as possible. It is also crucial in order to adhere to the seventh principle, integrity and confidentiality (Integrität und Vertraulichkeit), that data needs to be handled safely without the risk of exposing it to a third party. Since subsequent releases of the pathogen DNA could be public, it is essential that this data is depleted of host DNA with high sensitivity.

Mitigation strategies are diverse and come with several disadvantages such as requiring the sample-collecting party to assure that collected DNA was free of human contamination. In this case guaranteeing a consistent quality level is difficult since different strategies for removing host DNA might be used by the respective data providers.

In the following paper, we present a system that is able to manage the exchange of pathogen DNA data, and perform the host depletion step (the removing of any human DNA) simultaneously with the upload, thus making sure that the receiving party will at no point be in possession of a critical mass of human DNA reads that would pose a risk for identification of an individual. At the same time we achieved high scalability through the Docker Compose architecture and good transmission speeds that allow productive use in real-life settings.

Bioinformatics, 2024, **40(6)**, btae332
https://doi.org/10.1093/bioinformatics/btae332
Advance Access Publication Date: 24 May 2024
Applications Note

OXFORD

Sequence analysis

SWGTS—a platform for stream-based host DNA depletion

Philipp Spohr [1,2,*], **Max Ried** [1,2], **Laura Kühle** [1,2], **Alexander Dilthey** [2,3,*]

[1]Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf, 40225, Germany
[2]Center for Digital Medicine, Düsseldorf, 40225, Germany
[3]Institute of Medical Microbiology and Hospital Hygiene, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, 40225, Germany

*Corresponding author. Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf, 40225 Germany. E-mail: philipp.spohr@hhu.de (P.S.); Institute of Medical Microbiology and Hospital Hygiene, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, 40225 Germany. E-mail: alexander.dilthey@med.uni-duesseldorf.de (A.D.)

Associate Editor: Can Alkan

Abstract

Motivation: Microbial sequencing data from clinical samples is often contaminated with human sequences, which have to be removed prior to sharing. Existing methods for human read removal, however, are applicable only after the target dataset has been retrieved in its entirety, putting the recipient at least temporarily in control of a potentially identifiable genetic dataset with potential implications under regulatory frameworks such as the GDPR. In some instances, the ability to carry out stream-based host depletion as part of the data transfer process may be preferable.

Results: We present SWGTS, a client–server application for the transfer and stream-based host depletion of sequencing reads. SWGTS enforces a robust upper bound on the maximum amount of human genetic data from any one client held in memory at any point in time by storing all incoming sequencing data in a limited-size, client-specific intermediate processing buffer, and by throttling the rate of incoming data if it exceeds the speed of host depletion carried out on the SWGTS server in the background. SWGTS exposes a HTTP–REST interface, is implemented using docker-compose, Redis and traefik, and requires less than 8 Gb of RAM for deployment. We demonstrate high filtering accuracy of SWGTS; incoming data transfer rates of up to 1.65 megabases per second in a conservative configuration; and mitigation of re-identification risks by the ability to limit the number of SNPs present on a popular population-scale genotyping array covered by reads in the SWGTS buffer to a low user-defined number, such as 10 or 100.

Availability and implementation: SWGTS is available on GitHub: https://github.com/AIBi-HHU/swgts (https://doi.org/10.5281/zenodo.10891052). The repository also contains a jupyter notebook that can be used to reproduce all the benchmarks used in this article. All datasets used for benchmarking are publicly available.

1 Introduction

The sharing of raw sequencing reads within the scientific community, e.g. publicly *via* SRA or ENA or across institutions as part of a multi-center study, is an important best practice in microbial genomics and microbiome research. Sharing of raw sequencing reads typically involves the identification and removal of "contaminant" human reads ("host depletion"); these are produced as a by-product of many sequencing workflows (Bush *et al.* 2020) and carry the risk of individual re-identification (Tomofuji *et al.* 2023). Of note, even small amounts of human genetic data are sufficient to raise privacy and identifiability concerns, as <100 SNPs may be sufficient for subject re-identification (Lin *et al.* 2004, Pakstis *et al.* 2010). Multiple effective host depletion methods have been developed, including Hostile (Constantinides *et al.* 2023), dehumanizer (https://github.com/SamStudio8/dehumanizer) or ReadItAndKeep (Hunt *et al.* 2022); in addition, general mapping or classification tools like minimap2 (Li 2018) or Kraken 2 (Wood *et al.* 2019) can also be used. Existing methods, however, only support the "bulk depletion" use case, i.e. removal of human data from a locally stored sequencing data file.

Here, we present the Secure Whole-Genome Transfer System (SWGTS), the first approach for the depletion of host reads in an incoming stream of sequencing data; by relying on a defined-size buffer for the temporary storage of incoming and as-of-yet uncontaminated sequencing data. SWGTS enforces a robust upper bound on the amount of human read data stored at any point during the transfer process. To motivate the use case for SWGTS, consider the case of a sequencing data exchange between Bob and Alice, with data flowing from Bob to Alice. Alice wants to ensure that the data received from Bob does not contain an identifiable amount of human genetic sequencing data at any point in time. For example, Alice may operate a public sequencing data archive; provide a cloud-based bioinformatics data analysis service; function as the centralized sequencing data collection node for a multi-center pathogen genome sequencing study; or want to avoid re-identifiability and legal risks associated with human genetic data potentially arising under the GDPR (Shabani and Marelli 2019). Alice could ask Bob to carry out human decontamination prior to transmitting any data, but whether and how Bob actually implements decontamination is out of Alice's control. Alice may therefore implement her own decontamination process, but existing "bulk"

Received: 10 November 2023; **Revised:** 7 May 2024; **Editorial Decision:** 17 May 2024; **Accepted:** 23 May 2024

approaches only become applicable after receipt of the complete dataset. If Bob's data are sufficiently contaminated, Alice will thus—at least for the period of time until her own decontamination process is successfully executed, which may vary depending on available system resources or be affected by unforeseen technical failures—be in control of identifiable human genetic data. Using SWGTS, however, Alice can set the size of the buffer for incoming sequencing data to appropriately minimize re-identification risks even if the un-decontaminated data held in memory was exclusively human; using SWGTS, Alice thus ensures that she is never in control of a personally identifiable human genetic sequencing dataset.

SWGTS implements a minimap2-based approach for the identification of human reads and carries out depletion by removal of sequences that align against the human genome ("Host Subtraction"); by exclusive retention of sequences that align against a target pathogen genome ("Simple Pathogen Retention"); or by exclusive retention of sequences that align to a target pathogen genome where the reference being aligned to also includes the human genome ("Host-Competitive Pathogen Retention").

2 Materials and methods

2.1 Implementation

SWGTS is based on a client–server architecture, implemented using docker-compose. The SWGTS client splits a locally stored sequencing read dataset into equally-sized packets of a defined size ("chunks") and sequentially transmits these to the SWGTS server. For each ongoing upload, the SWGTS server maintains an in-memory "buffer" of a configured maximum size of reads that have not yet undergone host depletion; an incoming chunk of reads is accepted if and only if the contained reads can be added to the buffer without the buffer exceeding its specified maximum size, and otherwise the chunk is rejected. Host depletion is implemented as a continuously running background process on the server, based on mapping the chunks sent by the client with the in-memory "mappy" version of minimap2 against the human reference genome ("Host Subtraction"); the target pathogen genome ("Simple Pathogen Retention"); or against the human reference genome plus the target pathogen genome ("Host-Competitive Pathogen Retention"). Filtering criteria in different modes are summarized in Supplementary Table S2. Once a chunk has been sent to mappy, it is removed from the buffer; filtered reads are always deleted; non-filtered reads can be stored on the SWGTS server and the IDs of these reads can also be transmitted back to the sending client. The communication between client and server is implemented using a traefik (https://traefik.io/)-based HTTP/REST interface; redis (https://redis.io/) is used for in-memory storage on the SWGTS server. SWGTS comes with a Python-based CLI client and with a React-based browser demo client application. The most important configurable parameters include the maximum size of the buffer on the server side and the number of threads used for host depletion. A full specification of all parameters and of the REST API is given in the SWGTS repository.

Evaluation

Four "contaminated" pathogen datasets were created representing the possible combinations of SARS-CoV-2 and multi-resistant *Staphylococcus aureus* (MRSA) as well as of Illumina (1000 000 reads per dataset) and Oxford Nanopore Technologies (ONT; 250 000 reads per dataset). For each dataset, 99% pathogen reads sampled from three different pathogen isolates (Walker *et al.* 2022) were mixed with 1% human reads sampled from three 1000 Genomes Project (1000 Genomes Project Consortium *et al.* 2015) samples; see Supplementary Note S2 for datasets and accessions. Hostile was used with its human-t2t-hla reference which was also utilized for SWGTS in "Host-Competitive Pathogen Retention" and "Host Subtraction" mode. For SARS-CoV-2 we used the MN90
8947.3 reference without the poly-A tail provided by ReadItAndKeep and for MRSA the assembly GCF_000 013425.1 of *Staphylococcus aureus* subsp. NCTC 8325. For measuring data transfer rates, two systems were used. System A is a server system with an AMD EPYC 7742 64-Core Processor with 128 Threads and 1 TiB RAM. System B is a VM configured with 10 virtual cores using an Intel® Xeon® CPU E5-2618L v4 @ 2.20 GHz with 20 GB RAM and 1*GbE uplink.

Relationship between SWGTS buffer size and re-identifiability

The number of SNP positions present on a common population-scale SNP genotyping array (Illumina Infinium Omni2.5–8) covered by reads in the SWGTS buffer under worst-case assumptions (only human data submitted) was pragmatically treated as a proxy for re-identifiability risk (see Supplementary Note S1). To empirically characterize the relationship between the number of such SNPs and buffer size, we randomly selected reads from three 1000 Genomes Project samples (Supplementary Note S2) until the cumulative length of the selected reads was equal to the selected buffer size, trimming the last selected read if necessary. The collected reads were mapped to the human reference genome (GRCh38), counting the number of Infinium SNP positions covered by primary alignments (Supplementary Table S1, Fig. 1B, left panel). To complement experimental results and assist the determination of an appropriate buffer size without having to carry out extensive simulations, we developed statistical models for the expected number of such SNPs (Supplementary Note S1).

3 Results

First, we confirmed the accuracy of SWGTS with respect to discriminating between human and pathogen reads. At a buffer size large enough to process all reads and depending on the depletion mode, SWGTS retained between 96.5% and 100.0% of target pathogen reads and removed between 96.0% and 100% of human reads on the four "contaminated" datasets (Supplementary Fig. S1, Supplementary Table S1), consistent with the host depletion accuracy of minimap2 reported in the literature (Bush *et al.* 2020, Hunt *et al.* 2022) and very similar to the host retention and pathogen removal reads of Hostile (v.0.4.0, matching the "Host Subtraction" mode of SWGTS; the slight difference results from SWGTS requiring a mapping quality of at least 20 using default settings, this can be configured) and ReadItAndKeep (v.0.3.0, matching the "Regular Retention" mode of SWGTS). A decrease in buffer size resulted in a decrease of retained pathogen reads (and a slight increase in filtered human reads) since all reads exceeding the buffer size are immediately rejected; from a buffer size of ≥10 000 bases, both rates were within 3% of rates observed for the largest buffer size (Supplementary Fig. S1).

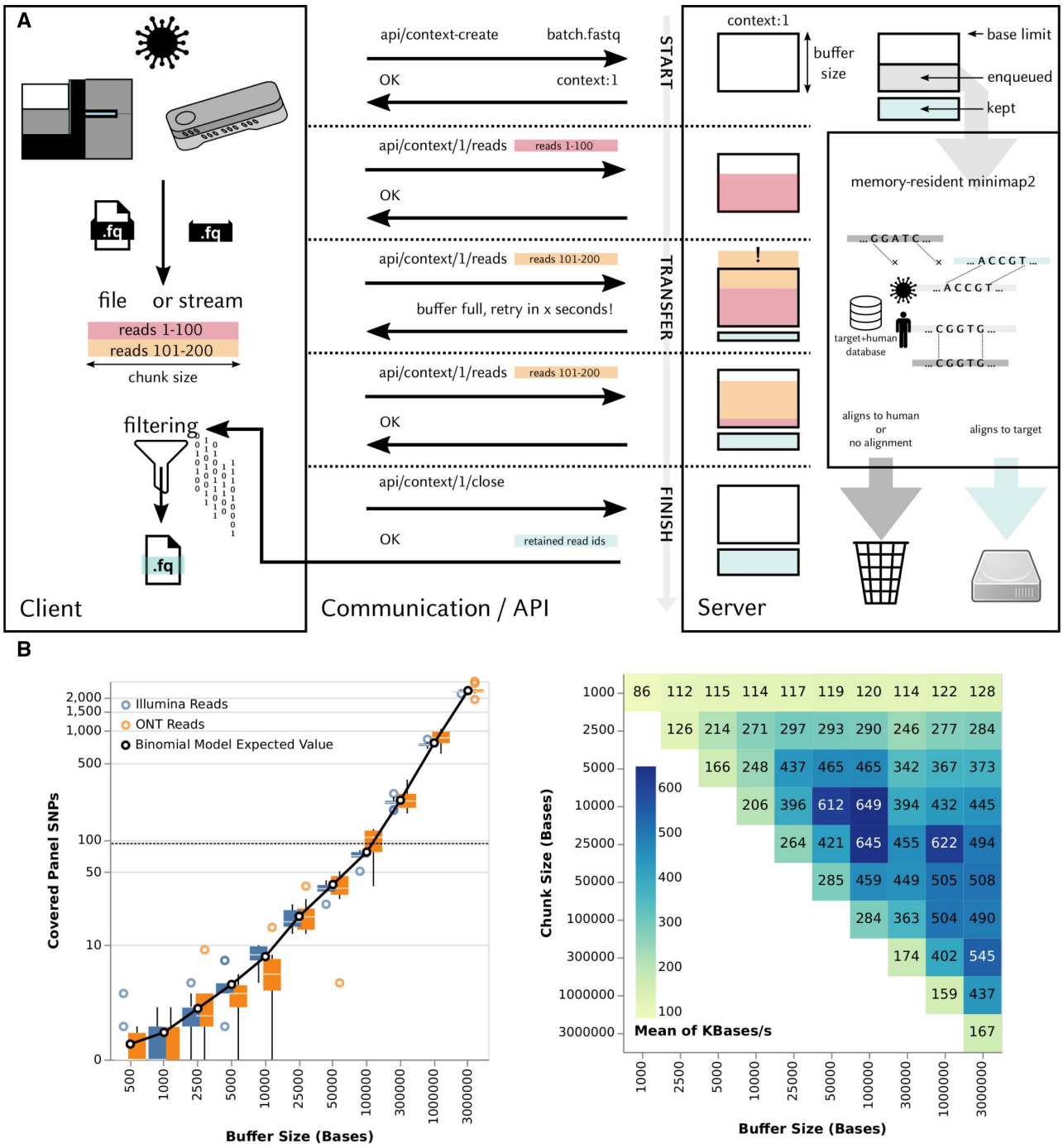


Figure 1. (A) Overview of SWGTS. SWGTS is a client–server application for the transfer and stream-based depletion of sequencing data from human reads. An upload to the SWGTS server is initiated by a context creation request; on the SWGTS server, each context corresponds to a unique buffer of unprocessed sequencing data of a defined maximum size (buffer size). After creation of the context, the client sequentially transmits equally-sized packets of sequencing data (chunk size) to the server, including the ID of the context just created; the server accepts chunks and adds the reads contained therein to the buffer corresponding to the specified context ID, for as long as doing so would not lead to the buffer exceeding its specified maximum size; otherwise, a "buffer full, retry in × seconds" notification is sent to the client. Single reads that are larger than the buffer are sent but immediately rejected and treated as filtered. A (multi-threaded) background process on the server continuously pulls reads from the buffer and carries out host depletion by aligning the received reads against a combined reference of the human genome and the target pathogen (in "Host-Competitive Pathogen Retention" mode). Reads that map to the pathogen are saved to disk; reads mapping to either the human reference or not at all are discarded. When a context is closed, the server sends a summary containing the IDs of the reads that were not discarded to the client. (B) Impact of the "buffer size" parameter on potential re-identifiability and performance. (Left panel) Relationship between buffer size and the number of Illumina Infinium Omni2.5-8 Kit Panel SNPs, which is pragmatically treated as a proxy for re-identifiability, covered by reads in the SWGTS buffer under worst-case assumptions (only human reads submitted); shown are empirical distributions based on real Illumina and ONT reads from three human samples (see Section 2; 10 replicates), as well as the expected value of the number of such SNPs under an approximate statistical model ("Binomial Model Expected Value", see Supplementary Note S1). (Right panel) Relationship between buffer size and mean transfer rate for sequencing datasets between two systems, averaged over the four "contaminated" datasets representing SARS-CoV-2 and MRSA as well as Illumina and ONT (see Section 2) and in "host subtraction" mode.

Second, we investigated how different buffer and transfer chunk sizes influence the performance of SWGTS by measuring how these variables affected transfer rates from System A to System B for the four "contaminated" datasets in "host subtraction" mode (Fig. 1B, Supplementary Fig. S2, Supplementary Table S1). Optimal performance across pathogens and sequencing data types was achieved (a) when chunk size was a fraction of buffer size, enabling parallel processing of chunks on the server side and, based on more frequently updated mapping rate estimates than for larger chunk sizes, a more accurate calibration of the "retry-after" messages sent to the client in case the buffer was full; (b) when the buffer was sufficiently large so that few retries were necessary; and (c) for chunk sizes $\geq$5000 bp and $<$50 000 bp.

Third, we characterized the relationship between the buffer size parameter and the number of Illumina Infinium Omni2.5–8 SNPs covered by reads in the SWGTS buffer under the worst-case assumption of only human data being submitted by the client; the number of such SNPs was pragmatically treated as a proxy for re-identifiability risk. Simulations (see Section 2) and statistical modeling (Supplementary Note S1) showed that the number of covered SNPs remained at $\leq$10 for a buffer size of 10 000 and $\leq$100 for a buffer size of 100 000 in almost all cases (Fig. 1B, Supplementary Table S1); furthermore, experimental results were highly consistent with the developed statistical model (Fig. 1B, Supplementary Note S1).

Finally, we evaluated which absolute data transfer speeds may be expected when using SWGTS, deploying an SWGTS server on System B (limited to 10 cores; two containers used for handling API requests; seven worker threads) and between 1 and 30 SWGTS clients on System A sending data simultaneously. Using a buffer size of 10 000 bases and transmitting the simulated SARS-CoV-2-containing Illumina read datasets, we observed data transfer rates (incoming data processed by System B) of up to approximately 0.4 megabases/s; using a buffer size of 100 000 bases and transmitting the simulated MRSA-containing Nanopore read datasets, we observed data transfer rates of up to approximately 1.8 megabases/s (see Supplementary Table S1 for full results). We note that the performance of SWGTS depends on multiple factors including network speed, the number of concurrent uploads, and the properties of the submitted data. Since the bulk of the required RAM is made up of the reference index, held in shared memory, and as each connection only adds a small fixed amount of memory, memory usage is generally consistent across buffer and chunk sizes and the number of incoming data streams and SWGTS can be deployed using relatively small amounts of RAM ($<$8 GB).

4 Discussion

We have presented SWGTS, the first approach for the transfer and stream-based human depletion of sequencing data. By combining a configurable-size buffer with a background host depletion process, SWGTS can enforce a robust upper bound on the maximum amount of human genetic data from any one client held in memory at any point in time.

Using conservative experiments, we have demonstrated that SWGTS can reliably keep the number of SNP positions present on a popular population-scale SNP genotyping array, which we have treated pragmatically as a proxy for re-identifiability risk, covered by reads in the SWGTS buffer below 100 or even 10. While the question of genetic identifiability is subtle, the risk of re-identification from so few (arbitrarily selected) SNPs has to be classified as low; in particular, to the best of our knowledge, no approaches have been presented that would allow for individual re-identification from 10 SNPs. Furthermore, we have shown consistency between the empirical results and a simple statistical model (Supplementary Note S1), which users may consult to identify a suitable application-dependent buffer size setting.

Using SWGTS can lead to significantly reduced data transfer rates; improving these, e.g. based on k-mer-based read prescreening, is therefore an important goal for future work. Additional potential improvements to SWGTS include the incorporation of a specialized short-read mapper; the improved handling of reads that exceed the buffer size, e.g. by splitting, chunk-wise processing and reconciliation; decoupling of transmission chunk size and mapping chunk size; and support for authentication and metadata handling mechanisms. These limitations notwithstanding, however, potential users of SWGTS may find that the benefits of a privacy-ensuring data transfer mechanism—such as only having to demonstrate conformity with data protection rules for demonstrably non-identifiable sequencing data—may outweigh the costs of a decreased data transfer rate.

Acknowledgements

The authors would like to thank Nguyen Khoa Tran for discussions regarding statistical models.

Author contributions

A.D., M.R., and P.S. conceptualized the idea. M.R. and P.S. implemented the software. P.S. drafted the initial manuscript and performed the benchmarking. M.R. designed the docker-compose architecture. A.D. supervised the project. L.K. designed the advanced statistical model for the base limit. All authors have read, contributed to, and approved of the final manuscript.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

We acknowledge funding by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung; Netzwerk Universitätsmedizin, GenSurv/MolTraX), award number 01KX2021, and by the Jürgen Manchot Foundation.

References

1000 Genomes Project Consortium, Auton A, Brooks LD *et al.* A global reference for human genetic variation. *Nature* 2015;**526**:68–74.

Bush SJ, Connor TR, Peto TEA *et al.* Evaluation of methods for detecting human reads in microbial sequencing datasets. *Microb Genom* 2020;**6**:mgen000393.

Constantinides B, Hunt M, Crook DW *et al.* Hostile: accurate decontamination of microbial host sequences. *Bioinformatics* 2023;**39**:btad728.

Hunt M, Swann J, Constantinides B *et al.* ReadItAndKeep: rapid decontamination of SARS-CoV-2 sequencing reads. *Bioinformatics* 2022; **38**:3291–3.

Lin Z, Owen AB, Altman RB *et al.* Genomic research and human subject privacy. *Science* 2004;**305**:183.

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.

Pakstis AJ, Speed WC, Fang R *et al.* SNPs for a universal individual identification panel. *Hum Genet* 2010;**127**:315–24.

Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR: assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. *EMBO Rep* 2019; **20**:e48316.

Tomofuji Y, Sonehara K, Kishikawa T *et al.* Reconstruction of the personal information from human genome reads in gut metagenome sequencing data. *Nat Microbiol* 2023;**8**:1079–94.

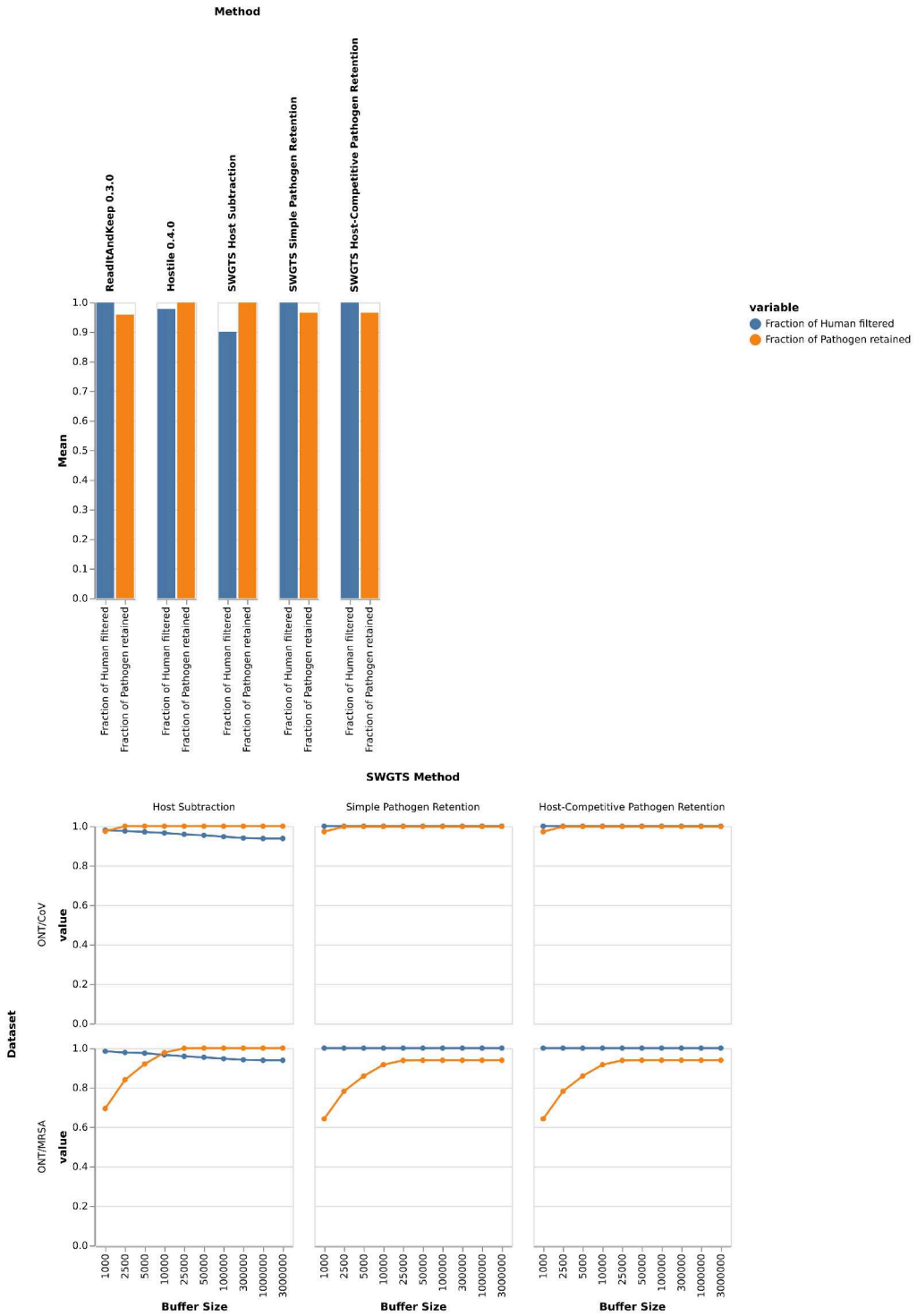
Walker A, Houwaart T, Finzer P, *et al.* Characterization of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection clusters based on integrated genomic surveillance, outbreak analysis and contact tracing in an urban setting. *Clin Infect Dis* 2022;**74**:1039–46.

Wood DE, Lu J, Langmead B *et al.* Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;**20**:257.

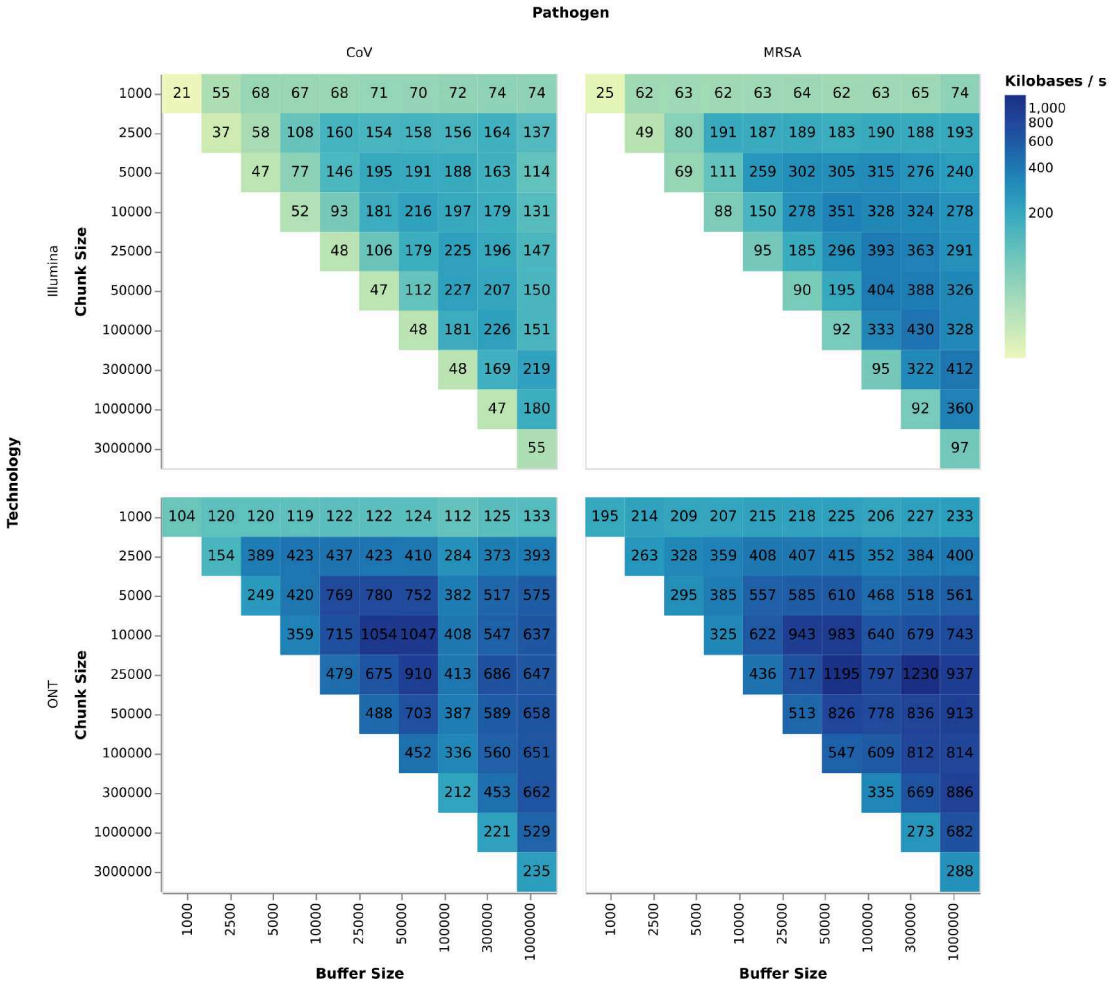
Bioinformatics, 2024, 40, 1–5
https://doi.org/10.1093/bioinformatics/btae332
Applications Note

3.3.3 Supplementary Figures

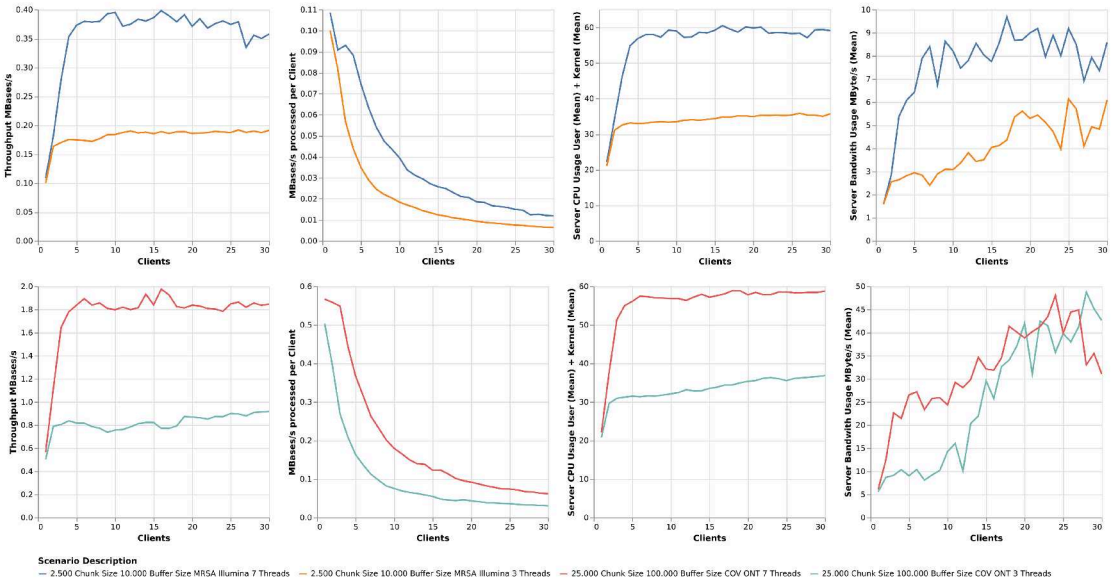
Additional supplementary files can be found online and are omitted due to being unsuitable for print.



Supplementary Figure 1: Filter quality. Fraction of both human reads filtered (sensitivity) and pathogen reads retained (specificity) with respect to different operation modes and datasets.



Supplementary Figure 2: Impact of chunk and buffer size. Transmission speeds for two different datasets and two different sequencing technologies are shown with respect to chunk and buffer size.



Supplementary Figure 3: Transmission rates and client scaling. Throughput for two different configurations of the software in combination with two different datasets with respect to the number of transmitting clients.

3.3.4 Discussion: Chances for real-world implementation

While we are satisfied with the software as a proof-of-concept, there are still obstacles preventing a widespread real-world adaptation. The released version is relatively easy to set up and can be used through a web interface for potential uploading clients; however, a minimum of expert knowledge is required to configure the server side software. In addition, real-life applications require secondary functions such as the labeling of samples with metadata and user management that were not provided with the initial release.

When thinking about scalability for large-scale data collection, network usage and transmission speeds become restrictive factors. Since HTTP is not designed for high-speed, high-throughput transactions a natural expansion of the algorithm is to provide different transmission protocol backend options. While this sacrifices the stateless nature of the provided REST implementation the higher speeds can be used to process larger data quantities. An experimental implementation using WebSockets for the file transfer showed promising—although not fully conclusive—results [45].

Naturally, the system is also limited when it comes to bad actors. If a party wants to upload critical data to a server running the SWGTS software in order to plant sensitive genetic information there it is easily possible to do so by writing the data in the metadata or read-id fields. While we could implement additional checks for the software this is a limitation we believe needs to be appropriately addressed by the respective legal frameworks that would regulate storage of genetic data.

Another strategy that is currently evaluated consists of transmitting the files split into encrypted chunks using a regular transfer protocol and then using a parallel transmission of encryption keys to control the potential amount of human DNA on the server. This way the main data transfer can be performed at maximum speed while an additional control flow handles the data privacy.

Chapter 4

Conclusions

It is clear that the broad focus of this thesis has led to a selection of rather loosely connected topics. A narrower approach could have potentially enabled more depth for any individual field of research. However, we believe there is also value to be found in a broader approach. Throughout the work on this thesis synergies were found in many unexpected places: Among others, network analysis tools and algorithms, originally designed for gene set enrichment, were surprisingly helpful in visualizing community interactions in microbiomes while fuzzy clustering proved to be a viable approach to quantify strain shifts. Generally speaking, strategies and information could be re-contextualized, leading to solutions that would have otherwise not been easy to obtain.

This effect was also evident in the composition of our project teams. In today's world, where many experts are highly specialized, establishing effective communication channels across disciplines and developing a shared language is a non-trivial process. In our case common ground had to be built between microbiologists, clinicians, computer scientists, and statisticians. The rewards of such a cooperation with well-established communication is substantial: We were able to profit from a broad set of perspectives and align our efforts so that they were not only meaningful for our respective fields but also advanced a shared objective, fostering synergy and knowledge transfer.

Especially in medicine, reproducibility of research remains challenging. While reviewing related work, we frequently encountered many instances of unavailable data or software. Provided datasets were often incomplete and software was rarely usable without major effort and modification. There is justified reluctance to share medical datasets due to their highly sensitive nature and the difficulties of effective anonymization. However, when it comes to software in particular, we believe that the standards need to be continuously raised as reproducibility of results is crucial for reliable and meaningful research. Strategies like federated learning (sharing only models or their parameters as opposed to the data) can be solutions to privacy concerns. Based on our own experience, we hope that the community keeps moving towards a culture, where reproducible research, within reasonable ethical and legal boundaries, becomes the norm.

While the three papers selected here highlight research that culminated in a publication, it is also worth noting that various other projects yielding smaller or no publications at all have been conducted as well. In many cases this is not necessarily due to a lack of quality but chance and opportunity factor in it as well. Within the scientific ecosystem negative or inconclusive results rarely receive the same attention even though the

underlying research behind it may be highly relevant. We would therefore encourage other researchers to be brave and pursuit interesting research even if a clear payoff is not immediately apparent.

Looking ahead, we are excited to continue the numerous ongoing projects–both by tying up loose ends and by exploring new directions.

References

[1] Statistisches Bundesamt. *Causes of Death*. Available at: https://www-genesis.destatis.de/datenbank/online/statistic/23211/table/23211-0001 Accessed May 28, 2025.

[2] Edward D. Thomas et al. "Intravenous infusion of bone marrow in patients receiving radiation and chemotherapy." In: *N Engl J Med* 257.11 (Sept. 1957), pp. 491–496. URL: https://doi.org/10.1056/nejm195709122571102.

[3] Jan Styczyski et al. "Death after hematopoietic stem cell transplantation: changes over calendar year time, infections and associated factors." In: *Bone Marrow Transplantation* 55.1 (Jan. 2020), pp. 126–136. ISSN: 1476-5365. URL: https://doi.org/10.1038/s41409-019-0624-z.

[4] Varro. *On Agriculture*. Ed. by William D. Hooper and Harrison Boyd Ash. 1934. DOI: 10.4159/dlcl.varro-agriculture.1934. URL: http://dx.doi.org/10.4159/DLCL.varro-agriculture.1934.

[5] Joel Doré and Sandra Ortega Ugalde. "Human-microbes symbiosis in health and disease, on earth and beyond planetary boundaries." In: *Frontiers in Astronomy and Space Sciences* Volume 10 - 2023 (2023). ISSN: 2296-987X. DOI: 10.3389/fspas.2023.1180522. URL: https://www.frontiersin.org/journals/astronomy-and-space-sciences/articles/10.3389/fspas.2023.1180522.

[6] Cristina Solé et al. "Gut microbiome is profoundly altered in acute-on-chronic liver failure as evaluated by quantitative metagenomics. Relationship with liver cirrhosis severity." In: *Journal of Hepatology* 68 (Apr. 2018), S11–S12. ISSN: 0168-8278. DOI: 10.1016/S0168-8278(18)30240-X. URL: https://doi.org/10.1016/S0168-8278(18)30240-X.

[7] Graham J. Britton et al. "Microbiotas from Humans with Inflammatory Bowel Disease Alter the Balance of Gut Th17 and $ROR_{\gamma}t^{+}$ Regulatory T Cells and Exacerbate Colitis in Mice." In: *Immunity* 50.1 (Jan. 2019), 212–224.e4. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2018.12.015. URL: https://doi.org/10.1016/j.immuni.2018.12.015.

[8] Jaroslav Flegr. "Effects of toxoplasma on human behavior." en. In: *Schizophr Bull* 33.3 (Jan. 2007), pp. 757–760.

[9] Ruth E. Ley et al. "Human gut microbes associated with obesity." In: *Nature* 444.7122 (Dec. 2006), pp. 1022–1023. ISSN: 1476-4687. DOI: 10.1038/4441022a. URL: https://doi.org/10.1038/4441022a.

[10] Yong Fan and Oluf Pedersen. “Gut microbiota in human metabolic health and disease.” In: *Nature Reviews Microbiology* 19.1 (Jan. 2021), pp. 55–71. ISSN: 1740-1534. DOI: 10.1038/s41579-020-0433-9. URL: https://doi.org/10.1038/s41579-020-0433-9.

[11] Tanya Yatsunenko et al. “Human gut microbiome viewed across age and geography.” In: *Nature* 486.7402 (June 2012), pp. 222–227. ISSN: 1476-4687. DOI: 10.1038/nature11053. URL: https://doi.org/10.1038/nature11053.

[12] Eugene Rosenberg. “Diversity of bacteria within the human gut and its contribution to the functional unity of holobionts.” In: *npj Biofilms and Microbiomes* 10.1 (Nov. 2024), p. 134. ISSN: 2055-5008. DOI: 10.1038/s41522-024-00580-y. URL: https://doi.org/10.1038/s41522-024-00580-y.

[13] Jonathan U. Peled et al. “Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell Transplantation.” In: *New England Journal of Medicine* 382.9 (2020), pp. 822–834. DOI: 10.1056/NEJMoa1900623. eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1900623. URL: https://www.nejm.org/doi/full/10.1056/NEJMoa1900623.

[14] Paolo Manghi et al. “MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice.” In: *Cell Reports* 42.5 (2023), p. 112464. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2023.112464. URL: https://www.sciencedirect.com/science/article/pii/S2211124723004758.

[15] Mantas Sereika et al. “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.” In: *Nature Methods* 19.7 (July 2022), pp. 823–826. ISSN: 1548-7105. URL: https://doi.org/10.1038/s41592-022-01539-7.

[16] Andreas Walker et al. “Characterization of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infection Clusters Based on Integrated Genomic Surveillance, Outbreak Analysis and Contact Tracing in an Urban Setting.” In: *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 74 (6 Mar. 2022), pp. 1039–1046. URL: https://doi.org/10.1093/cid/ciab588.

[17] Jim Shaw and Yun William Yu. “Rapid species-level metagenome profiling and containment estimation with sylph.” In: *Nature Biotechnology* (Oct. 2024). ISSN: 1546-1696. URL: https://doi.org/10.1038/s41587-024-02412-y.

[18] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences.” In: *Bioinformatics* 34.18 (May 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/18/3094/48919122/bioinformatics_34_18_3094.pdf. URL: https://doi.org/10.1093/bioinformatics/bty191.

[19] Derrick E. Wood, Jennifer Lu, and Ben Langmead. “Improved metagenomic analysis with Kraken 2.” In: *Genome Biology* 20.1 (Nov. 2019), p. 257. ISSN: 1474-760X. URL: https://doi.org/10.1186/s13059-019-1891-0.

[20] Pol Fernández et al. “A 160 Gbp fork fern genome shatters size record for eukaryotes.” In: *iScience* 27.6 (June 2024). ISSN: 2589-0042. DOI: 10.1016/j.isci.2024.109889. URL: https://doi.org/10.1016/j.isci.2024.109889.

[21] Paul I. Costea et al. “Towards standards for human fecal sample processing in metagenomic studies.” In: *Nature Biotechnology* 35.11 (Nov. 2017), pp. 1069–1076. ISSN: 1546-1696. URL: https://doi.org/10.1038/nbt.3960.

[22] Andrew Maltez Thomas and Nicola Segata. “Multiple levels of the unknown in microbiome research.” In: *BMC Biology* 17.1 (June 2019), p. 48. ISSN: 1741-7007. URL: https://doi.org/10.1186/s12915-019-0667-z.

[23] National Center for Biotechnology Information. *RefSeq Growth Statistics*. Available at: https://www.ncbi.nlm.nih.gov/refseq/statistics Accessed June 3, 2025.

[24] Daniel J. Nasko et al. “RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification.” In: *Genome Biology* 19.1 (Oct. 2018), p. 165. ISSN: 1474-760X. URL: https://doi.org/10.1186/s13059-018-1554-6.

[25] Marie-Madlen Pust and Burkhard Tümmler. “Bacterial low-abundant taxa are key determinants of a healthy airway metagenome in the early years of human life.” In: *Computational and Structural Biotechnology Journal* 20 (2022), pp. 175–186. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2021.12.008. URL: https://www.sciencedirect.com/science/article/pii/S200103702100516X.

[26] Ray A. Wickenheiser. “Forensic genealogy, bioethics and the Golden State Killer case.” In: *Forensic science international: Synergy* 1 (2019), pp. 114–125.

[27] Rebecca Skloot. “Cells that save lives are a mother’s legacy.” In: *The New York times on the Web* (2001), A15–A17.

[28] Jonathan J. M. Landry et al. “The genomic and transcriptomic landscape of a HeLa cell line.” en. In: *G3 (Bethesda)* 3.8 (Aug. 2013), pp. 1213–1224.

[29] Andrew J. Pakstis et al. “SNPs for a universal individual identification panel.” en. In: *Hum Genet* 127.3 (Mar. 2010), pp. 315–324.

[30] Nestor Maslej et al. “Artificial intelligence index report 2025.” In: *arXiv preprint arXiv:2504.07139* (2025).

[31] Donghee Shin. “The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI.” In: *International Journal of Human-Computer Studies* 146 (2021), p. 102551. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2020.102551. URL: https://www.sciencedirect.com/science/article/pii/S1071581920301531.

[32] similarweb. *Tob Websites Ranking*. Available at: https://www.similarweb.com/top-websites/ Accessed July 18, 2025.

[33] Vasileios Ntinopoulos et al. “Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation.” en. In: *BMJ Health Care Inform.* 32.1 (Jan. 2025), e101139.

[34] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system.” In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[35] Maximilian Christ et al. “Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh A Python package).” In: *Neurocomputing* 307 (May 2018). DOI: 10.1016/j.neucom.2018.03.067.

[36] Angus Dempster, François Petitjean, and Geoffrey I. Webb. “ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels.” In: *Data Mining and Knowledge Discovery* 34.5 (Sept. 2020), pp. 1454–1495. ISSN: 1573-756X. DOI: 10.1007/s10618-020-00701-z. URL: https://doi.org/10.1007/s10618-020-00701-z.

[37] Nicola Segata et al. “Metagenomic microbial community profiling using unique clade-specific marker genes.” en. In: *Nat. Methods* 9.8 (June 2012), pp. 811–814.

[38] Rebecca Fröhlich. “Entwicklung einer webbasierten Anwendung zur interaktiven Analyse klinischer Daten.” Bachelor’s Thesis. July 2021.

[39] Jonathan Bobak et al. “Dynamic Mortality Risk Prediction in Myelodysplastic Syndromes Using Longitudinal Clinical Data.” In: *medRxiv* (2025). DOI: 10.1101/2025.07.21.25331775. eprint: https://www.medrxiv.org/content/early/2025/07/21/2025.07.21.25331775.full.pdf. URL: https://www.medrxiv.org/content/early/2025/07/21/2025.07.21.25331775.

[40] Alexander Geissler et al. “A nationwide digital maturity assessment of hospitals Results from the German DigitalRadar.” In: *Health Policy and Technology* 13.4 (2024), p. 100904. ISSN: 2211-8837. DOI: https://doi.org/10.1016/j.hlpt.2024.100904. URL: https://www.sciencedirect.com/science/article/pii/S2211883724000674.

[41] Federico Toth. “Integration vs separation in the provision of health care: 24 OECD countries compared.” In: *Health Economics, Policy and Law* 15.2 (2020), pp. 160–172. DOI: 10.1017/S1744133118000476.

[42] Bundesministerium für Gesundheit. *Die elektronische Patientenakte (ePA) für alle*. Available at: https://www.bundesgesundheitsministerium.de/themen/digitalisierung/elektronische-patientenakte/epa-fuer-alle.html Accessed July 14, 2025.

[43] Volker Amelung et al. *Ergebnisse der zweiten nationalen Reifegradmessung deutscher Krankenhäuser 2024*. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Gesundheit/Berichte/DigitalRadar_Zwischenbericht_t2_barrierefrei.pdf.

[44] Pierre Marijon, Philipp Spohr, and Antoine Limasset. “Correcting Long-Reads with k-mers: A Dream Comes True.” In: Nov. 2020. URL: https://seqbim2020.sciencesconf.org/data/pages/seqBIM_2020_paper_5.pdf.

[45] Florian Serger. “SWGTS - Application extension and performance optimization.” Bachelor’s Thesis. Apr. 2025.