Lower abundance of human gut virus species is associated with cancer cachexia

Axel Stang (a.stang@asklepios.com)  
Asklepios Hospital Barmbek

Hauke Weilert  
Asklepios Hospital Barmbek

Thomas Illig  
Institute of Epidemiology, Helmholtz Center Munich

Rudi Schmidt  
Asklepios Hospitals

Katharina Burbach  
CEMET GmbH

Raphael Gronauer  
Connexome GmbH

Martin Seifert  
Connexome GmbH

Article

Keywords: gut microbiome, gut viruses, gut bacteriophages, cancer cachexia, shotgun sequencing, metagenomics

Posted Date: July 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1756753/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Cancer cachexia has been linked to gut bacterial alterations, but alterations of gut viruses, mostly bacteriophages, have not yet been explored. We performed shotgun metagenomic sequencing of DNA from stool samples of 78 cachectic and 42 non-cachectic cancer patients. K-mer-based matching to reference databases revealed abundance variations of bacteria and viruses. Beyond bacterial alterations, cachectic patients exhibited significantly lower bacteriophage abundance, predominantly affecting Caudovirales and Siphoviridae species (double-stranded DNA) but also Inoviridae and Microviridae families (single-stranded DNA). Machine learning models exploiting the data for classification between cachectic and non-cachectic state yielded an AUC of 0.704. Caudovirales and Siphoviridae species were among the top-most important classifiers. AUC increased to 0.850 with solely antibiotic-exposed samples from 20 cachectic and 10 non-cachectic patients. This study is the first to suggest a link between cancer cachexia and intestinal bacteriophage richness. This could constitute a new basis for hypothesis-driven research in cancer cachexia diagnosis and treatment.

Introduction

Cancer cachexia is a common, poorly understood clinical syndrome with profound adverse effects on patients’ quality and lengths of survival\(^1\). It has recently emerged that altered gut bacteria may be implicated in the development of cancer cachexia\(^2\). Bacteria dominate research on the human gut microbiome to date, and 16S rRNA gene sequencing has been the mainstay for decades\(^3\). The availability of shotgun sequencing metagenomics of whole-community DNA within stool samples enables explorations of the entire microbial community\(^4\). However, there are critical factors to realize the full potential of shotgun metagenomics. Firstly, the success for assigning taxonomic labels to DNA fragments depends on effective tools for querying datasets against reference databases and the reference database choice as well\(^5\). Secondly, exploratory analyses aimed at detecting differentially abundant taxa in clinical metagenome samples should include statistical methods (direct species-level testing between sample groups) and machine learning (ML) classifiers (train models to label groups of samples)\(^6\). Consistent results obtained by both approaches enhance the confidence of findings, as they complement and validate each other\(^7\).

The ~10\(^{12}\) human gut viruses, mostly bacteriophages (phages), have not yet been explored in the context of cancer cachexia. Gut phages have been termed as “known unknown” and “dark matter” to describe our limited knowledge about their taxonomic composition\(^8\). Two key issues restrict research on human gut phage populations (phageome). Firstly, without a marker gene ubiquitously present in phages like 16S gene in bacteria, it is difficult to characterize the phageome diversity\(^5\)-\(^8\). Secondly, only ~10% of phage sequences recovered from the gut microbiome match to known sequences in existing viral databases\(^9,10\). To improve metagenomic studies on the human gut phageome, combinatory uses of nucleotide databases and advanced bioinformatics softwares utilizing k-mers (all possible sequences of
length k from reads of DNA sequencing) are worth exploring, as they theoretically improve the sequence reads in a sample and the fraction of reads assignable to a known virus.\textsuperscript{5-7,11,12}

In this study, we performed shotgun metagenomic sequencing of total DNA from stool samples of patients with and without cachexia. Matching the obtained sequences based on k-mers with classical microbiome and nucleotide reference databases, we aimed to assess statistically differential cachexia-related taxa abundances. Applying a ML strategy, we complementary aimed to define classifiers that separate cachexia states with good accuracy.

Results

Ethics statement. The prospective exploratory cohort study operates under the name “MiBiTuKa-Study” (Trial registration number: ST-U069, Study Box German Cancer Society). The study protocol was approved by the ethics committees of the General Medical Councils (Aerztekammer) of the City Hamburg prior to data extraction (Ethics committee Aerztekammer Hamburg, Protocol number: V5649, Date: 23.10.2017). Data were collected, then patient identifiers were removed, and no patient identifiable data are reported in the analysis. All patients included into the database provided written informed consent for scientific evaluation of the data. All data processing and analyses presented in this study have been conducted in accordance with the Helsinki declaration.

Study population. Among a total of 120 cancer patients (63 [52.5\%] men, median age: 68 years), 78 (65\%) were recruited in the cachexia and 42 (35\%) in the non-cachexia group. 41 patients had colorectal adenocarcinoma, 32 pancreatic adenocarcinoma, 30 gastric adenocarcinoma, 12 hepatocellular carcinoma, and 5 peritoneal carcinosis from ovarian cancer. Except the Body-mass index (BMI, Fig. 1, \(P < 0.001\)), clinical characteristics (sex, age, dietary and alcohol intake, smoking, diabetes, medication, distribution of cancer types) were comparable between cachectic and non-cachectic patients (Supplementary Table 1, \(P > 0.05\), respectively). At the time of stool sampling, 20 cachectic (25.6\%) and 10 non-cachectic patients (23.8\%) had received antibiotic treatment within \(\leq 2\) weeks, whereas all other patients were not exposed to antibiotics for \(\geq 3\) months.

Alterations of gut bacteria in cancer cachexia. Regarding the overall variation in bacterial community composition at the species level, alpha-diversity analysis (Shannon index, ANOVA testing) showed no significant differences according to cachexia state (Fig. 2a, \(P = 0.066\)) but significant differences according to cancer type (Fig. 2b, \(P = 0.048\)) and antibiotic pre-treatment (Fig. 2c, \(P < 0.001\)). Beta-diversity analysis (Principal coordinates analysis based on Bray-Curtis dissimilarities, PERMANOVA testing) showed significant differences according to cachexia state (Fig. 3a, \(P = 0.035\)), cancer type (Fig. 3b, \(P = 0.003\)) and antibiotic pre-treatment (Fig. 3c, \(P < 0.001\)). Regardless to the reference database
approach applied, post-processing significance analyses (FDR-adjusted \( P < 0.05 \)) outlined largely identical statistically differentially abundant bacterial species according to cachectic state among taxa passing the threshold of \( \geq 0.001 \) mean basic abundance in the whole dataset. Specifically, \textit{Faecalibacterium prausnitzii}, \textit{Roseburia intestinalis}, \textit{Prevotella copri}, \textit{Streptococcus} species, and \textit{Lachnospiraceae} species showed significantly lower abundance in cachectic patients inferred from both the Standard-database (Fig. 4a, Supplementary Table 2a) and the NT-database (Fig. 4b, Supplementary Table 2b). Antibiotic-exposed microbiome composition (Fig. 4c) differed from the unexposed state (Fig. 4a, Fig. 4b). Specifically, among the initial statistically differential abundant species, only \textit{Faecalibacterium prausnitzii} remained significantly lower abundant in antibiotic-exposed cachectic patients, while other bacterial species, such as \textit{Clostridium scindens}, \textit{Coprococcus comes}, \textit{Anaerobutyricum hallii}, \textit{Lacticaseibacillus rhamnosus}, and \textit{Eubacterium rectale}, newly emerged as significant low-abundant marker of cachexia (Fig. 4c, Supplementary Table 2c).

**Alterations of gut bacteriophages in cancer cachexia.** Among viruses, the \textit{Caudovirales} order, which comprise double-stranded (ds) DNA bacteriophages, namely members of the \textit{Siphoviridae} families, dominated the gut phageome detected by the NT-based approach and were significantly (FDR-adjusted \( P < 0.05 \)) lower abundant in the cachexia cohort among taxa passing the threshold of \( \geq 0.001 \) mean basic abundance in the whole dataset (Fig. 4b, Supplementary Table 2b). Significant under-representation of dsDNA bacteriophages was also found in the antibiotic-exposed cachexia cohort microbiome (Fig. 4c). Notably, further ds-DNA bacteriophages, namely \textit{Myoviridae} phages, and single-stranded (ss) DNA bacteriophages comprising the \textit{Inoviridae} and \textit{Microviridae} families, showed also significantly reduced abundance in the cachexia cohort when referred to test for significant differences of taxa below the threshold of 0.001 mean basic abundance (Supplementary Table 2d; for significant differences in separate analysis of antibiotic-exposed microbiomes see Supplementary Table 2e).

**Machine learning dataset classification.** Random forest (RF) models exploiting combinatorial effects from microbiome-derived taxonomy and abundance data for classification between cachectic and non-cachectic state performed lowest with the Standard-database approach (AUC 0.642, Fig. 5a). RF-based classification performance improved with the NT-database approach, when taking bacteriophages as classifiers into account, namely members of the \textit{Caudovirales} order and \textit{Siphoviridae} families (AUC 0.704, Fig. 5b). Performance greatly increased with the NT-database approach when taken into account additionally antibiotic microbiome exposure (AUC 0.850, Fig. 5c). Detailed information on the confusion matrices and classification metrics (e.g. Precision, Recall, F-Measure) of RF-based models are listed in Supplementary Table 3.

**Discussion**
Cancer cachexia represents a poorly understood unmet clinical need\(^1\). Cancer cachexia has been linked to gut bacterial alterations\(^2\), but changes of gut viruses are unknown. We pursued this by shotgun metagenomic sequencing of total DNA from stool samples of cancer patients with and without cachexia. Matching the obtained sequences with reference databases, we found that, beyond bacterial alterations, cachectic patients exhibited significantly lower bacteriophage abundance. Applying a ML strategy, gut bacteriophages were top-ranked classifiers for separating cachexia states with good accuracy. To our best knowledge, this is the first study to suggest a link between gut bacteriophages and cancer cachexia. This link could constitute a new direction in gut microbiome research and a significant advance in the context of cachexia diagnosis.

Our work supports a relative high output for detection of bacteriophages in human gut by using shotgun metagenomes of stool samples without enrichment for viral-like particles (VLPs). This is important as shotgun metagenomes are much more readily available than VLP metagenomes\(^13,14\). Further, our work supports that computational approaches applying k-mer-based matching to nucleotide libraries may overcome challenges that typically affect assembly-/ and read mapping-based taxonomic profiling of viral genomes\(^5-7,11\). Viruses are poorly annotated by default pipelines, and a high fraction (~90%) of the ~5.8%-22% phage sequences contained in whole-community shotgun DNA in human feces cannot be mapped in viral databases\(^8-10,15,16\) or linked to a bacterial host\(^17\). Until comprehensive genome-/ or proteome-based taxonomic phage reference catalogues are available, k-mer-based nucleotide database-dependant matching may extent the scope of applying metagenomics to studying the human gut phageome, commonly referred to as “known unknown” or “viral dark matter” component of the gut microbiome\(^8,10,15,18\).

Consistent with the dominance observed in previous human gut phageomes\(^12-14\), dsDNA phages of the Caudovirales order (mainly Siphoviridae) dominated the gut phageome in cachectic and non-cachectic patients. The human gut phageome, however, is not yet fully characterized\(^19\). Further gut phage groups like CrAss-like phages continue to be discovered\(^20,21\), and current methods fail to detect RNA phages effectively\(^22,23\). In addition, the gut phageome composition is influenced by genetic, dietary, environmental, individual, and age-related factors\(^24,25\). Despite this, a study found a set of 23 gut phages shared across >50% of individuals\(^26\), leading to the concept of a healthy human core gut phageome\(^8,9,26\). Notably, obesity and metabolic disorders associate with both increased\(^27\) and decreased\(^28\) gut phage richness, suggesting also lower and upper limits for a healthy gut phage reservoir. Our analysis revealed that cachectic patients had a decreased gut bacteriophages richness, predominantly affecting dsDNA phages of the Caudovirales order (Siphoviridae\(^8\), Myoviridae\(^8\)), but also ssDNA phages (Inoviridae\(^29\), Microviridae\(^30\)). Whether this cachexia-related decrease is unique to dsDNA and ssDNA phages or if other gut phage groups behave similarly, yet remains to be determined.

In addition to reduced gut bacteriophage richness, we found cachexia-related changes in the bacteriome composition. Finding similarity within (α-diversity) but dissimilarity between (β-diversity) samples across cohorts together with low-abundant Faecalibacterium prausnitzii\(^31\) in our cachectic
patients is well supported by a recent study reporting identical gut bacteriome alterations in 31 cachectic lung cancer patients. Notably, in antibiotic-exposed microbiomes, some bacterial species disappeared (e.g., Lachnospiraceae, Roseburia intestinalis, Prevotella copri, Streptococcus species) while new species emerged (e.g., Lacticaseibacillus rhamnosus, Eubacterium rectale) as markers of cachexia. Because prophage inducing antibiotic and non-antibiotic medications were equally distributed across cohorts, drug-induced phageome-mediated effects are unlikely to have a major impact on the gut bacteriome alterations in cachexia. Some studies reported cancer type-specific fecal bacterial signatures, making metagenomic data more complex. Our analysis is heterogeneous for cancer types, which associated with bacteriome diversity, but subsample sizes are too small for producing meaningful data in this respect. Further studies are called for to evaluate whether there are influences from the cancer type or whether generalized pan-cancer gut bacteriome and/or phageome alterations can be applied to cancer cachexia.

Bacteriophage production in gut of healthy adults is dominated by lysogenic replication cycles of temperate phages that coexist and replicate with their bacterial host, with a switch to lytic cycle in temperate phages from gut lumen towards mucosa surface. Balance and spatial distribution of lysogenic and lytic processes is crucial in fostering the amplification of gut phage populations and in controlling the density, diversity and network-interactions inside gut-associated bacterial communities. Caudovirales, the most abundant phage order in our study, target all main bacterial phyla in gut, namely Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria, and are mostly found integrated in bacterial genomes as prophages. With shotgun metagenomic data, it is impossible to discern DNA packaged in phage particles from prophage sequences in bacterial genomes. Whether reduced gut phage colonization combined with bacterial dysbiosis is driven by prophage and/or lytic cycle inactivation, disruption of spatial refuges of phage-bacteria interactions between phage-resistant and phage-susceptible and/or commensal and pathogenic bacteria is yet to be determined. Accumulated information on phage-bacteria dynamics specific to cancer cachexia may open the possibility of developing targeted phage-based treatments against cancer cachexia.

Complementary to single species-level statistical analyses, we applied a ML strategy to assess combinatorial effects of taxa pattern for separating cachexia states. Despite of small datasets and operations on subsamples through data splits into training and test data, the trained RF model separated cachexia states with good accuracy in a small validation set. Bacteriophages, namely Caudovirales and Siphoviridae species, were top-ranked classifiers, and model’s performance increased taken into account antibiotic treatment. Although antibiotics are described prophage inducers, they predominantly cause bactericidal effects and bacterial dysbiosis in the gut. This dysbiosis-inducing effect is visible in our study on lower abundant bacterial taxa and bacterial phyla dissimilarity within and between samples across antibiotic-treated and untreated patients. As antibiotic exposure adds complexity, linking antibiotics, gut microbial structure, gut phage load and cancer cachexia remains to be studied in more detail. However, from a clinical perspective, our analysis suggests that antibiotic medication in cancer patients would not negatively affect fecal microbiome-based signatures for cancer cachexia diagnosis.
Our study has limitations. Firstly, non-cachectic cancer patients served as controls, and we did not measure longitudinal data. Future trials should include a tumor-free healthy control cohort and collect samples from the same individual over time to exclude nonrepresentative outliers and to gain insight into dynamics of interest. Secondly, cancer cachexia was defined by weight loss. New scores may provide more accurate criteria to distinguish cachectic from malnourished patients. Thirdly, our microbiome datasets have many features but small sample size, making the detection of statistically significant differences between sets of samples difficult and findings prone to study-driven error and bias. However, finding largely identical results with statistical and ML-based analyses strengthen the reliability of our findings, as the two methods differently process the data and complement and validate each other. Despite this, our results are based on an exploratory single-centre analysis and warrant confirmatory testing in an independent multicentre validation, ideally combined with functional studies to investigate causality relations.

In conclusion, beyond bacterial alterations, this study is the first to show a link between reduced intestinal bacteriophage richness and cachexia in cancer patients. Indicating the biomedical relevance of studying the human gut phageome and filling a gap in our understanding of cancer cachexia, our results could serve as a new basis for hypothesis-driven research towards exploring gut phageome-bacteriome interactions in the onset, progression and treatment of cancer cachexia and beyond.

Methods

Study cohort. A prospective single-center cohort study design was applied. Adult patients presenting with newly diagnosed, histologic proven metastatic disease from gastric, colorectal, pancreatic, liver and ovarian cancer (peritoneal carcinosis) were identified between 2019 and 2021 at the Asklepios Hospital Barmbek. Patients were recruited prior to any anticancer treatment and included if they had no acute or chronic diarrhoea and/or acute gastrointestinal illness including ileus, inflammatory bowel disease, autoimmune diseases, immunosuppressive therapy including corticosteroids, acquired immunodeficiency syndrome, active infection, organ dysfunction or failure, or abdominal surgery. To address the impact of antibiotic administration, cut-offs were set to either ≤2 weeks exposure or ≥3 months non-exposure for antibiotic treatment to include or exclude patients. The criterion for cachexia was weight loss within ≤3 months before actual cancer diagnosis of ≥5% or ≥2% in patients with a BMI <20 kg/m². Clinical data, including age, sex, stage of diseases and laboratory values, were retrieved from hospital charts. Life style and medication data, including BMI, smoking, alcohol intake, and dietary nutrients, were collected during face-to-face interviews through a structured questionnaire. The ethics committees of the General Medical Councils (Aerztekammer) of the City Hamburg (V5649, 23.10.2017) approved the study, and all patients provided written informed consent.

Sample collection. Stool samples were collected using a stool catcher to avoid toilet contamination. Using a sterile spoon, two pieces of stool (cherry stone size, ~3-5 g of feces) were transferred into a sterile
plastic tube filled with 3.5 mL RNA stabilizing solution (Biosepar, Simbach, Austria), suspended with the integrated screw cap stirrer until obtaining a homogeneous fecal sample mixture, and refrigerated at 4°C in a portable cooler with ice packs. Samples were then stored within ≤12 hours at -80°C for an average of 1.2 ± 0.6 years until further processing.

**Shotgun metagenomic sequencing.** The wet lab part of the metagenomics sequencing including DNA extraction from stool samples and DNA sequencing was performed applying the metagenome sequencing service from Eurofins Genomics (Konstanz, Germany). In brief, stool samples were thawed on ice, aliquoted (1.000 µL), and then DNA was extracted with the DNeasy PowerSoil Pro Kit (Qiagen, Hilden, Germany) following the manufacturer’s protocol. For enzymatic DNA fragmentation, DNA was quantified for each sample and diluted to a concentration of 0.3 ng/µL. The sequencing library was built from 5µL of DNA at 0.3 ng/µL (1.5 ng in total) with the NEB-Kit NEBNext Ultra II FS DNA Library Prep Kit for Illumina. Samples were then sequenced in 2x150 bp paired-end mode on an Illumina NovaSeq platform (Illumina, San Diego, California, USA) in two runs to a target number of 20 million raw reads per sample. Data were de-multiplexed and quality checked using FastQC and MultiQC. The number of reads per sample range from a maximum of 34 million reads to a minimum of 19.6 million reads per sample. All samples passed the per sequence quality check. Adapter content was above 5% but below 10 % for five samples (only warming issued, no failure). The rest showed an adapter content below 5 % and passed without warning. K-mer bias for “GGGGGGG” was observed for most samples which is due and normal for the random prime based library.

**Taxonomy and abundance estimation.** The k-mer based sequence mapper kraken2 (version 2.1.2) was applied to assign metagenomic sequencing reads to taxonomic labels with two different mapping libraries with a confidence score of 0.6 for sequence mappings. Firstly, a default option of kraken2 was used to build a library on the standard-database (file-size: ~40 GB), which is based on taxonomic information and complete genomes in the NCBI Reference Sequence (RefSeq) for the bacterial, archaeal, and viral domains, along with the human genome and a collection of known vectors (UniVec_Core). The library was built with the command “kraken2-built-standard” on 16th of December 2021. Secondly, a kraken2 database was built using the NBCI nucleotide selection database (file-size: ~200 GB), which is also basis for the blast nucleotide search (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz). The data were downloaded with the command “kraken2-built—download-library nt” on 22nd of October 2021 to build a library on the NT-database with a k-mer size of 40 nucleotides. All reads were assigned to the taxonomy with the highest total hits of k-mers matched by pruning the general taxonomic trees affiliated with mapped sequences. Relative taxonomic abundance was calculated using the Bracken algorithm. The two kraken’s original taxonomic classification assignments based on the Standard-database and the NT-database underwent probabilistic reassignment by Bracken using the command “bracken-build” to establish the according libraries for the 150 bp read length.
**Machine learning classification.** All machine learning (ML) approaches used the Waikato Environment for Knowledge Analysis (Weka)\(^{54}\) as workbench and random forest (RF), as this algorithm has been shown to outperform, on average, other ML tools for microbiome data\(^{55}\). In addition, a meta classifier approach was used for making its base RF classifier cost-sensitive to balance false positives and false negatives. The RF-based approaches run on the input features inferred by the Standard-database and the NT-database approach. A 10-fold cross validation was applied, each fold contained a balanced proportion of cachectic and non-cachectic patients. For feature (taxa) selection, taxa with statistically differential abundance (\(p < 0.05\)) between the cachexia and non-cachexia cohort and a minimum mean basic abundance of \(\geq 0.001\) (\(\geq 0.1\%\)) in the whole dataset were used. After applying the trained RF model to classify the left-out test set, classification performance was estimated on the test sets using the receiver operating characteristic curves (ROC, pairs of observations with concordant ordering of classifications and true values)\(^{56}\). Single classifier (taxa) importance contributing to the RF model classification accuracy was determined based on mean impurity decrease and number of nodes using that classifier (Gini importance)\(^{57}\).

**Statistical analysis.** Categorical data were summarized with frequency and percentages and compared using the chi-squared test or Fisher´s exact test. Continuous data were summarized as means and interquartile range (if distributed normally) or medians and ranges (if not distributed normally) and compared using the two-tailed \(t\)-test or Wilcoxon´s rank sum test. Alpha and beta gut bacteriome diversity was calculated with R package vegan (version 2.5-7)\(^{58}\) and plotted with the ggplot2 package\(^{59}\). Kruskal-Wallis analysis of variance (ANOVA) test was used to determine statistically significant differences for alpha diversity based on Shannon index distribution between groups\(^{60}\). Pairwise permutational multivariate analysis of variance (PERMANOVA) with 999 random permutations was performed to assess statistically significant differences for beta diversity using principal coordinates analysis (PCoA) based on Bray-Curtis dissimilarity between groups\(^{61}\). Single species-level abundance analyses between patient cohorts were carried out using two-tailed unpaired \(t\)-tests for pairwise comparisons. The resulting \(p\)-values were adjusted for multiple testing using Benjamini-Hochberg´s adjustment\(^{62}\). An adjusted \(p\)-value (false discovery rate [FDR] \(>25\%\)) of \(<0.05\) was defined as cut-off for significant lower or higher abundant taxa between cohorts. After applying an FDR correction for the \(p\)-values for taxa below a minimum mean basic abundance of \(\geq 0.001\) (\(\geq 0.1\%\)) in the whole dataset, we did not get significant differences between cohorts below the 0.05 cut-off. To still attend to the tendencies among cohorts, we referred to the unadjusted \(p\)-values of \(<0.05\) in a separate correlation analysis.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.
Data Availability

All relevant data are provided within the manuscript and its supplementary information files. Raw data cannot be made publicly available due to ethical and legal restrictions and privacy concerns because shotgun metagenomics may contain human DNA. For analysis reproducibility, raw data are available from the corresponding author on reasonable request and in accordance with the EU-GDPR regulatory regime.

Code availability


References

35. Martinovic; A., Cocuzzi, R., Arioli, S. & Mora, D. Streptococcus thermophilus: to survive, or not to survive the gastrointestinal tract, that is the question! Nutrients 12, 2175 (2020).

**Declarations**

**Acknowledgements**

The study was supported by a research grant (#3465) awarded to A.S by Asklepios Proresearch, Asklepios Hospitals Hamburg GmbH, Hamburg, Germany, and supported by the Asklepios Campus Hamburg of the Semmelweis University, Budapest, Hungary. The authors thank Nadine Emmerich, Katrin Ohltmann and Sabine Drießelmann for excellent technical assistance.

**Author contributions**

A.S.; H.W., and R.S. designed the study. R.S. and K.B. performed the shotgun metagenomic sequencing. R.S., R.G., and M.S. carried out the quality control of the metagenomics data. A.S., H.W., T.I., R.S., K.B., R.G. and M.S. offered support during the analysis and interpretation of the metagenomics data. A.S., R.S., K.B., R.G. and M.S. conducted the statistical, computational and machine-learning procedure analysis of the data. A.S., H.W., T.I., and M.S. drafted the manuscript. All authors read, revised and approved the final manuscript.

**Competing Interests**

The authors declare no conflict of interest.

**Figures**
Body-mass index distribution between groups. Violin-plots illustrate BMI variation within groups. Two-tailed unpaired t-test was performed to determine statistical significance of BMI difference between groups ($P < 0.001$).
**Figure 2**

**Alpha diversity of gut bacteriome between groups.** (a) Shannon index variation according to cachexia state ($P = 0.066$). (b) Shannon index variation according to cancer type ($P = 0.048$). (c) Shannon index variation according to antibiotic exposure ($P < 0.001$). Pairwise ANOVA was performed to assess statistical significance. Boxes represent IQRs with the median as the horizontal line and the whiskers depicting the lowest and highest values within 1.5-fold IQR.

**Figure 3**

**Beta diversity of gut bacteriome between groups.** (a) Principal coordinate analysis (PCoA) of Bray-Curtis distances according to cachexia state ($P = 0.035$). (b) PCoA of Bray-Curtis distances according to cancer type ($P = 0.003$). (c) PCoA of Bray-Curtis distances according to antibiotic exposure ($P < 0.001$). Pairwise PERMANOVA with 999 permutations was performed to assess statistical significance.
Figure 4

**Significant single species-level abundance changes according to cachectic state.** (a) Abundance changes of bacteria inferred by the standard approach. (b) Abundance changes of bacteria and bacteriophages inferred by the NT approach. (c) Abundance changes of bacteria and bacteriophages inferred by the NT approach under antibiotic exposure. A threshold of mean basic abundance of \( \geq 0.001 \) (\( \geq 0.1\% \)) in the whole dataset was used to select taxa for abundance analysis. Two-tailed unpaired \( t \)-tests were performed to assess statistical significance for pairwise comparisons and further adjusted for multiple comparisons using the Benjamini-Hochberg correction (FDR-adjusted \( P < 0.05 \)).

![Figure 4](image)

Figure 5

**ROC curves from random forest classifiers for classifying between cachectic and non-cachectic state.** (a) ROC of microbiome input features inferred by the standard approach. (b) ROC of microbiome input features inferred by the NT approach. (c) ROC of microbiome input features inferred by the NT approach under antibiotic exposure. Boxes below ROC curves show rankings of classifier importance of individual taxa for RF model classification performance.

![Figure 5](image)

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.docx
- SupplementaryTable2a.docx
- SupplementaryTable2b.docx
- SupplementaryTable2c.docx
- SupplementaryTable2d.docx
- SupplementaryTable2e.docx
- SupplementaryTable3.doc